

参考文献

- (-1). *JMLR*. 618, 649
- (-1a). Icm1'08. In *ICML'08*. ACM. 649, 674
- (-1b). Icm1'11. In *ICML'11*. 628, 634
- (-1c). Icm1'13. In *ICML'13*. 635, 660
- (-1). International conference on learning representations 2014. In *ICLR'2014*. 661, 675
- (-1). Nips'13. In *NIPS26*. NIPS Foundation. 629, 635
- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. 24, 183, 380
- Ackley, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, **9**, 147–169. 486, 559
- Alain, G. and Bengio, Y. (2013). What regularized auto-encoders learn from the data generating distribution. In *ICLR'2013*, *arXiv:1211.4246*. 433, 439, 445
- Alain, G., Bengio, Y., Yao, L., Éric Thibodeau-Laufer, Yosinski, J., and Vincent, P. (2015). GSNs: Generative stochastic networks. *arXiv:1503.05571*. 436, 607
- Anderson, E. (1935). The Irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, **59**, 2–5. 18

- Ba, J., Mnih, V., and Kavukcuoglu, K. (2014). Multiple object recognition with visual attention. *arXiv:1412.7755*. 591
- Bachman, P. and Precup, D. (2015). Variational generative stochastic networks with collaborative shaping. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1964–1972. 611
- Bacon, P.-L., Bengio, E., Pineau, J., and Precup, D. (2015). Conditional computation in neural networks using a decision-theoretic approach. In *2nd Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM 2015)*. 383
- Bagnell, J. A. and Bradley, D. M. (2009). Differentiable sparse coding. In *NIPS'2009*, pages 113–120. 425
- Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *ICLR'2015, arXiv:1409.0473*. 23, 89, 339, 356, 358, 395, 404, 405
- Bahl, L. R., Brown, P., de Souza, P. V., and Mercer, R. L. (1987). Speech recognition with continuous-parameter hidden Markov models. *Computer, Speech and Language*, **2**, 219–234. 390
- Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, **2**, 53–58. 245
- Baldi, P., Brunak, S., Frasconi, P., Soda, G., and Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**(11), 937–946. 337
- Baldi, P., Sadowski, P., and Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, **5**. 24
- Ballard, D. H., Hinton, G. E., and Sejnowski, T. J. (1983). Parallel vision computation. *Nature*. 385
- Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, **1**, 295–311. 128
- Barron, A. E. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. on Information Theory*, **39**, 930–945. 172
- Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. Oxford University Press. 418
- Basilevsky, A. (1994). *Statistical Factor Analysis and Related Methods: Theory and Applications*. Wiley. 418

- Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A., Bouchard, N., Warde-Farley, D., and Bengio, Y. (2012a). Theano: new features and speed improvements. Submitted to the Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, <http://www.iro.umontreal.ca/lisa/publications2/index.php/publications/show/551>. 23, 73, 380
- Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I. J., Bergeron, A., Bouchard, N., and Bengio, Y. (2012b). Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop. 182, 191
- Basu, S. and Christensen, J. (2013). Teaching classification boundaries to humans. In *AAAI'2013*. 280
- Baxter, J. (1995). Learning internal representations. In *Proceedings of the 8th International Conference on Computational Learning Theory (COLT'95)*, pages 311–320, Santa Cruz, California. ACM Press. 211
- Bayer, J. and Osendorfer, C. (2014). Learning stochastic recurrent networks. *ArXiv e-prints*. 228
- Becker, S. and Hinton, G. (1992). A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, **355**, 161–163. 462
- Behnke, S. (2001). Learning iterative image reconstruction in the neural abstraction pyramid. *Int. J. Computational Intelligence and Applications*, **1**(4), 427–438. 440
- Beiu, V., Quintana, J. M., and Avedillo, M. J. (2003). VLSI implementations of threshold logic-a comprehensive survey. *Neural Networks, IEEE Transactions on*, **14**(5), 1217–1243. 384
- Belkin, M. and Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS'01)*, Cambridge, MA. MIT Press. 210
- Belkin, M. and Niyogi, P. (2003a). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, **15**(6), 1373–1396. 443
- Belkin, M. and Niyogi, P. (2003b). Using manifold structure for partially labeled classification. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15 (NIPS'02)*, Cambridge, MA. MIT Press. 141
- Bengio, E., Bacon, P.-L., Pineau, J., and Precup, D. (2015a). Conditional computation in neural networks for faster models. arXiv:1511.06297. 383

- Bengio, S. and Bengio, Y. (2000a). Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Transactions on Neural Networks, special issue on Data Mining and Knowledge Discovery*, **11**(3), 550–557. 603
- Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. (2015b). Scheduled sampling for sequence prediction with recurrent neural networks. Technical report, arXiv:1506.03099. 327
- Bengio, Y. (1991). *Artificial Neural Networks and their Application to Sequence Recognition*. Ph.D. thesis, McGill University, (Computer Science), Montreal, Canada. 347
- Bengio, Y. (2000). Gradient-based optimization of hyperparameters. *Neural Computation*, **12**(8), 1889–1900. 370
- Bengio, Y. (2002). New distributed probabilistic language models. Technical Report 1215, Dept. IRO, Université de Montréal. 397
- Bengio, Y. (2009). *Learning deep architectures for AI*. Now Publishers. 174, 531
- Bengio, Y. (2013). Deep learning of representations: looking forward. In *Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 1–37. Springer, also in arXiv at <http://arxiv.org/abs/1305.0445>. 382
- Bengio, Y. (2015). Early inference in energy-based models approximates back-propagation. Technical Report arXiv:1510.02777, Université de Montréal. 560
- Bengio, Y. and Bengio, S. (2000b). Modeling high-dimensional discrete data with multi-layer neural networks. In *NIPS 12*, pages 400–406. MIT Press. 602, 603, 604, 606
- Bengio, Y. and Delalleau, O. (2009). Justifying and generalizing contrastive divergence. *Neural Computation*, **21**(6), 1601–1621. 438, 520
- Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. In *JML (1)*, pages 1089–1105. 107
- Bengio, Y. and LeCun, Y. (2007a). Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. 17
- Bengio, Y. and LeCun, Y. (2007b). Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large Scale Kernel Machines*. MIT Press. 17
- Bengio, Y. and Monperrus, M. (2005). Non-local manifold tangent learning. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17 (NIPS'04)*, pages 129–136. MIT Press. 138, 444

- Bengio, Y. and Sénécal, J.-S. (2003). Quick training of probabilistic neural nets by importance sampling. In *Proceedings of AISTATS 2003*. 400
- Bengio, Y. and Sénécal, J.-S. (2008). Adaptive importance sampling to accelerate training of a neural probabilistic language model. *IEEE Trans. Neural Networks*, **19**(4), 713–722. 400
- Bengio, Y., De Mori, R., Flammia, G., and Kompe, R. (1991). Phonetically motivated acoustic parameters for continuous speech recognition using artificial neural networks. In *Proceedings of EuroSpeech'91*. 21, 390
- Bengio, Y., De Mori, R., Flammia, G., and Kompe, R. (1992). Neural network-Gaussian mixture hybrid for speech recognition or density estimation. In *NIPS 4*, pages 175–182. Morgan Kaufmann. 390
- Bengio, Y., Frasconi, P., and Simard, P. (1993). The problem of learning long-term dependencies in recurrent networks. In *IEEE International Conference on Neural Networks*, pages 1183–1195, San Francisco. IEEE Press. (invited paper). 344
- Bengio, Y., Simard, P., and Frasconi, P. (1994a). Learning long-term dependencies with gradient descent is difficult. *IEEE Tr. Neural Nets*. 16
- Bengio, Y., Simard, P., and Frasconi, P. (1994b). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, **5**(2), 157–166. 343, 344, 345
- Bengio, Y., Simard, P., and Frasconi, P. (1994c). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, **5**(2), 157–166. 351
- Bengio, Y., Latendresse, S., and Dugas, C. (1999). Gradient-based learning of hyper-parameters. In *Learning Conference*. 370
- Bengio, Y., Ducharme, R., and Vincent, P. (2001a). A neural probabilistic language model. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13 (NIPS'00)*, pages 933–938. MIT Press. 16
- Bengio, Y., Ducharme, R., and Vincent, P. (2001b). A neural probabilistic language model. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *NIPS'2000*, pages 932–938. MIT Press. 380, 394, 396, 402, 406, 410
- Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A neural probabilistic language model. *JMLR*, **3**, 1137–1155. 396, 402
- Bengio, Y., Delalleau, O., and Le Roux, N. (2006a). The curse of highly variable functions for local kernel machines. In *NIPS'2005*. 137

- Bengio, Y., Larochelle, H., and Vincent, P. (2006b). Non-local manifold Parzen windows. In *NIPS'2005*. MIT Press. 138, 444
- Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007a). Greedy layer-wise training of deep networks. In *NIPS'2006*. 13, 276
- Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007b). Greedy layer-wise training of deep networks. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pages 153–160. MIT Press. 173
- Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007c). Greedy layer-wise training of deep networks. In *Adv. Neural Inf. Proc. Sys. 19*, pages 153–160. 275
- Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007d). Greedy layer-wise training of deep networks. In *NIPS 19*, pages 153–160. MIT Press. 276, 451, 452
- Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *ICML'09*. ACM. 279
- Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S. (2013a). Better mixing via deep representations. In *ICML'2013*. 514
- Bengio, Y., Léonard, N., and Courville, A. (2013b). Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv:1308.3432. 382, 383, 588, 590
- Bengio, Y., Yao, L., Alain, G., and Vincent, P. (2013c). Generalized denoising auto-encoders as generative models. In *NIPS'2013*. 433, 607, 608
- Bengio, Y., Courville, A., and Vincent, P. (2013d). Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **35**(8), 1798–1828. 473
- Bengio, Y., Thibodeau-Laufer, E., Alain, G., and Yosinski, J. (2014). Deep generative stochastic networks trainable by backprop. In *ICML'2014*. 607, 608, 609, 610
- Bennett, C. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, **22**(2), 245–268. 536
- Bennett, J. and Lanning, S. (2007). The Netflix prize. 408
- Berger, A. L., Della Pietra, V. J., and Della Pietra, S. A. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**, 39–71. 403
- Berglund, M. and Raiko, T. (2013). Stochastic gradient estimate variance in contrastive divergence and persistent contrastive divergence. *CoRR*, **abs/1312.6002**. 523

- Bergstra, J. (2011). *Incorporating Complex Cells into Neural Networks for Pattern Classification*. Ph.D. thesis, Université de Montréal. 219
- Bergstra, J. and Bengio, Y. (2009). Slow, decorrelated features for pretraining complex cell-like networks. In *NIPS 22*, pages 99–107. MIT Press. 421
- Bergstra, J. and Bengio, Y. (2011). Random search for hyper-parameter optimization. *The Learning Workshop*, Fort Lauderdale, Florida. 369
- Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Machine Learning Res.*, **13**, 281–305. 369, 370
- Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010a). Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*. Oral Presentation. 23, 73
- Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010b). Theano: a CPU and GPU math expression compiler. In *Proc. SciPy*. 182, 191
- Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010c). Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*. 380
- Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *NIPS'2011*. 371
- Berkes, P. and Wiskott, L. (2005). Slow feature analysis yields a rich repertoire of complex cell properties. *Journal of Vision*, **5**(6), 579–602. 423
- Bertsekas, D. P. and Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific. 93
- Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, **24**(3), 179–195. 525
- Bishop, C. M. (1994). Mixture density networks. 163
- Bishop, C. M. (1995a). Regularization and complexity control in feed-forward networks. In *Proceedings International Conference on Artificial Neural Networks ICANN'95*, volume 1, page 141–148. 208, 215
- Bishop, C. M. (1995b). Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, **7**(1), 108–116. 208

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. 87, 126
- Blum, A. L. and Rivest, R. L. (1992). Training a 3-node neural network is NP-complete. 250
- Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, **36**(4), 929--865. 100
- Bonnet, G. (1964). Transformations des signaux aléatoires à travers les systèmes non linéaires sans mémoire. *Annales des Télécommunications*, **19**(9-10), 203-220. 588
- Bordes, A., Weston, J., Collobert, R., and Bengio, Y. (2011). Learning structured embeddings of knowledge bases. In *AAAI 2011*. 411, 412
- Bordes, A., Glorot, X., Weston, J., and Bengio, Y. (2012). Joint learning of words and meaning representations for open-text semantic parsing. *AISTATS'2012*. 343, 411, 412
- Bordes, A., Glorot, X., Weston, J., and Bengio, Y. (2013a). A semantic matching energy function for learning with multi-relational data. *Machine Learning: Special Issue on Learning Semantics*. 411
- Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013b). Translating embeddings for modeling multi-relational data. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2787-2795. Curran Associates, Inc. 411
- Bornschein, J. and Bengio, Y. (2015). Reweighted wake-sleep. In *ICLR'2015, arXiv:1406.2751*. 592
- Bornschein, J., Shabaniyan, S., Fischer, A., and Bengio, Y. (2015). Training bidirectional Helmholtz machines. Technical report, arXiv:1506.03877. 592
- Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, pages 144-152, New York, NY, USA. ACM. 16, 123
- Bottou, L. (1998). Online algorithms and stochastic approximations. In D. Saad, editor, *Online Learning in Neural Networks*. Cambridge University Press, Cambridge, UK. 253
- Bottou, L. (2011). From machine learning to machine reasoning. Technical report, arXiv.1102.1808. 341, 342
- Bottou, L. (2015). Multilayer neural networks. Deep Learning Summer School. 374

- Bottou, L. and Bousquet, O. (2008a). The tradeoffs of large scale learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS'07)*, volume 20. MIT Press, Cambridge, MA. 241
- Bottou, L. and Bousquet, O. (2008b). The tradeoffs of large scale learning. In *NIPS'2008*. 252
- Boulanger-Lewandowski, N., Bengio, Y., and Vincent, P. (2012). Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In *ICML'12*. 585
- Boureau, Y., Ponce, J., and LeCun, Y. (2010). A theoretical analysis of feature pooling in vision algorithms. In *Proc. International Conference on Machine learning (ICML'10)*. 292
- Boureau, Y., Le Roux, N., Bach, F., Ponce, J., and LeCun, Y. (2011). Ask the locals: multi-way local pooling for image recognition. In *Proc. International Conference on Computer Vision (ICCV'11)*. IEEE. 293
- Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, **59**, 291–294. 429
- Bourlard, H. and Wellekens, C. (1989). Speech pattern discrimination and multi-layered perceptrons. *Computer Speech and Language*, **3**, 1–19. 390
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA. 82
- Brady, M. L., Raghavan, R., and Slawny, J. (1989). Back-propagation fails to separate where perceptrons succeed. *IEEE Transactions on Circuits and Systems*, **36**(5), 665–674. 243
- Brakel, P., Stroobandt, D., and Schrauwen, B. (2013). Training energy-based models for time-series imputation. *Journal of Machine Learning Research*, **14**, 2771–2797. 576, 596
- Brand, M. (2003a). Charting a manifold. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15 (NIPS'02)*, pages 961–968. MIT Press. 141
- Brand, M. (2003b). Charting a manifold. In *NIPS'2002*, pages 961–968. MIT Press. 443
- Breiman, L. (1994). Bagging predictors. *Machine Learning*, **24**(2), 123–140. 220
- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA. 125
- Bridle, J. S. (1990). Alphanets: a recurrent ‘neural’ network architecture with a hidden Markov model interpretation. *Speech Communication*, **9**(1), 83–92. 160

- Briggman, K., Denk, W., Seung, S., Helmstaedter, M. N., and Turaga, S. C. (2009). Maximin affinity learning of image segmentation. In *NIPS'2009*, pages 1865–1873. 306
- Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational linguistics*, **16**(2), 79–85. 18
- Brown, P. F., Pietra, V. J. D., DeSouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-based n -gram models of natural language. *Computational Linguistics*, **18**, 467–479. 394
- Bryson, A. and Ho, Y. (1969). *Applied optimal control: optimization, estimation, and control*. Blaisdell Pub. Co. 194
- Bryson, Jr., A. E. and Denham, W. F. (1961). A steepest-ascent method for solving optimum programming problems. Technical Report BR-1303, Raytheon Company, Missile and Space Division. 194
- Buciluă, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541. ACM. 381
- Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*. 596
- Cai, M., Shi, Y., and Liu, J. (2013). Deep maxout neural networks for speech recognition. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 291–296. IEEE. 167
- Carreira-Perpiñan, M. A. and Hinton, G. E. (2005). On contrastive divergence learning. In *AISTATS'2005*, pages 33–40. 520
- Caruana, R. (1993). Multitask connectionist learning. In *Proceedings of the 1993 Connectionist Models Summer School*, pages 372–379. 210
- Cauchy, A. (1847). Méthode générale pour la résolution de systèmes d'équations simultanées. In *Compte rendu des séances de l'académie des sciences*, pages 536–538. 74, 194
- Cayton, L. (2005). Algorithms for manifold learning. Technical Report CS2008-0923, UCSD. 141
- Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, **41**(3), 15. 90

- Chapelle, O., Weston, J., and Schölkopf, B. (2003). Cluster kernels for semi-supervised learning. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15 (NIPS'02)*, pages 585–592, Cambridge, MA. MIT Press. 210
- Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-Supervised Learning*. MIT Press, Cambridge, MA. 210, 462
- Chellapilla, K., Puri, S., and Simard, P. (2006). High Performance Convolutional Neural Networks for Document Processing. In Guy Lorette, editor, *Tenth International Workshop on Frontiers in Handwriting Recognition*, La Baule (France). Université de Rennes 1, Suvisoft. <http://www.suvisoft.com>. 20, 21, 379
- Chen, B., Ting, J.-A., Marlin, B. M., and de Freitas, N. (2010). Deep learning of invariant spatio-temporal features from video. NIPS*2010 Deep Learning and Unsupervised Feature Learning Workshop. 307
- Chen, S. F. and Goodman, J. T. (1999). An empirical study of smoothing techniques for language modeling. *Computer, Speech and Language*, **13**(4), 359–393. 393, 394, 402
- Chen, T., Du, Z., Sun, N., Wang, J., Wu, C., Chen, Y., and Temam, O. (2014a). DianNao: A small-footprint high-throughput accelerator for ubiquitous machine-learning. In *Proceedings of the 19th international conference on Architectural support for programming languages and operating systems*, pages 269–284. ACM. 384
- Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., and Zhang, Z. (2015). MXNet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274*. 23
- Chen, Y., Luo, T., Liu, S., Zhang, S., He, L., Wang, J., Li, L., Chen, T., Xu, Z., Sun, N., *et al.* (2014b). DaDianNao: A machine-learning supercomputer. In *Microarchitecture (MICRO), 2014 47th Annual IEEE/ACM International Symposium on*, pages 609–622. IEEE. 384
- Chilimbi, T., Suzue, Y., Apacible, J., and Kalyanaraman, K. (2014). Project Adam: Building an efficient and scalable deep learning training system. In *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI'14)*. 381
- Cho, K., Raiko, T., and Ilin, A. (2010a). Parallel tempering is efficient for learning restricted Boltzmann machines. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2010)*, Barcelona, Spain. 514
- Cho, K., Raiko, T., and Ilin, A. (2010b). Parallel tempering is efficient for learning restricted Boltzmann machines. In *IJCNN'2010*. 524

- Cho, K., Raiko, T., and Ilin, A. (2011). Enhanced gradient and adaptive learning rate for training restricted Boltzmann machines. In *ICML'2011*, pages 105–112. 575
- Cho, K., Van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014a). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics. 338
- Cho, K., van Merriënboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)*. 403
- Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014c). On the properties of neural machine translation: Encoder-decoder approaches. *ArXiv e-prints*, **abs/1409.1259**. 351
- Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2014). The loss surface of multilayer networks. 244, 245
- Chorowski, J., Bahdanau, D., Cho, K., and Bengio, Y. (2014). End-to-end continuous speech recognition using attention-based recurrent NN: First results. arXiv:1412.1602. 392
- Christianson, B. (1992). Automatic Hessians by reverse accumulation. *IMA Journal of Numerical Analysis*, **12**(2), 135–150. 193
- Chrupala, G., Kadar, A., and Alishahi, A. (2015). Learning language through pictures. arXiv 1506.03694. 351
- Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS'2014 Deep Learning workshop, arXiv 1412.3555. 351, 392
- Chung, J., Gülçehre, Ç., Cho, K., and Bengio, Y. (2015a). Gated feedback recurrent neural networks. In *ICML'15*. 351
- Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A., and Bengio, Y. (2015b). A recurrent latent variable model for sequential data. In *NIPS'2015*. 596
- Ciresan, D., Meier, U., Masci, J., and Schmidhuber, J. (2012). Multi-column deep neural network for traffic sign classification. *Neural Networks*, **32**, 333–338. 22, 174

- Ciresan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. (2010). Deep big simple neural nets for handwritten digit recognition. *Neural Computation*, **22**, 1–14. 20, 21, 379
- Coates, A. and Ng, A. Y. (2011). The importance of encoding versus training with sparse coding and vector quantization. In *ICML'2011*. 21, 220, 425
- Coates, A., Lee, H., and Ng, A. Y. (2011). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*. 310, 387
- Coates, A., Huval, B., Wang, T., Wu, D., Catanzaro, B., and Andrew, N. (2013). Deep learning with COTS HPC systems. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28 (3), pages 1337–1345. JMLR Workshop and Conference Proceedings. 20, 21, 310, 381
- Cohen, N., Sharir, O., and Shashua, A. (2015). On the expressive power of deep learning: A tensor analysis. arXiv:1509.05009. 472
- Collobert, R. (2004). *Large Scale Machine Learning*. Ph.D. thesis, Université de Paris VI, LIP6. 170
- Collobert, R. (2011). Deep learning for efficient discriminative parsing. In *AISTATS'2011*. 89, 406
- Collobert, R. and Weston, J. (2008a). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML'2008*. 401, 406
- Collobert, R. and Weston, J. (2008b). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML'2008*. 455
- Collobert, R., Bengio, S., and Bengio, Y. (2001). A parallel mixture of SVMs for very large scale problems. Technical Report 12, IDIAP. 383
- Collobert, R., Bengio, S., and Bengio, Y. (2002). Parallel mixture of SVMs for very large scale problem. *Neural Computation*. 383
- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011a). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, **12**, 2493–2537. 279, 406, 455, 456
- Collobert, R., Kavukcuoglu, K., and Farabet, C. (2011b). Torch7: A Matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*. 23, 182, 380

- Comon, P. (1994). Independent component analysis - a new concept? *Signal Processing*, **36**, 287–314. 419
- Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, **20**, 273–297. 16, 123
- Couprie, C., Farabet, C., Najman, L., and LeCun, Y. (2013). Indoor semantic segmentation using depth information. In *International Conference on Learning Representations (ICLR2013)*. 22, 174
- Courbariaux, M., Bengio, Y., and David, J.-P. (2015). Low precision arithmetic for deep learning. In *Arxiv:1412.7024, ICLR'2015 Workshop*. 384
- Courville, A., Bergstra, J., and Bengio, Y. (2011a). Unsupervised models of images by spike-and-slab RBMs. In *ICML'2011*. 477
- Courville, A., Bergstra, J., and Bengio, Y. (2011b). Unsupervised models of images by spike-and-slab RBMs. In *ICM (1b)*. 581
- Courville, A., Desjardins, G., Bergstra, J., and Bengio, Y. (2014). The spike-and-slab RBM and extensions to discrete and sparse data distributions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **36**(9), 1874–1887. 583
- Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory, 2nd Edition*. Wiley-Interscience. 66
- Cox, D. and Pinto, N. (2011). Beyond simple features: A large-scale feature search approach to unconstrained face recognition. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 8–15. IEEE. 310
- Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press. 118, 252
- Crick, F. H. C. and Mitchison, G. (1983). The function of dream sleep. *Nature*, **304**, 111–114. 518
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, **2**, 303–314. 171
- Dahl, G. E., Ranzato, M., Mohamed, A., and Hinton, G. E. (2010). Phone recognition with the mean-covariance restricted Boltzmann machine. In *Advances in Neural Information Processing Systems (NIPS)*. 22
- Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, **20**(1), 33–42. 391

- Dahl, G. E., Sainath, T. N., and Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout. In *ICASSP'2013*. 391
- Dahl, G. E., Jaitly, N., and Salakhutdinov, R. (2014). Multi-task neural networks for QSAR predictions. arXiv:1406.1231. 24
- Dauphin, Y. and Bengio, Y. (2013). Stochastic ratio matching of RBMs for sparse high-dimensional inputs. In *NIP (1)*. 528
- Dauphin, Y., Glorot, X., and Bengio, Y. (2011). Large-scale learning of embeddings with reconstruction sampling. In *ICML'2011*. 401
- Dauphin, Y., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *NIPS'2014*. 244, 245
- Davis, A., Rubinstein, M., Wadhwa, N., Mysore, G., Durand, F., and Freeman, W. T. (2014). The visual microphone: Passive recovery of sound from video. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, **33**(4), 79:1–79:10. 385
- Dayan, P. (1990). Reinforcement comparison. In *Connectionist Models: Proceedings of the 1990 Connectionist Summer School*, San Mateo, CA. 590
- Dayan, P. and Hinton, G. E. (1996). Varieties of Helmholtz machine. *Neural Networks*, **9**(8), 1385–1403. 592
- Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The Helmholtz machine. *Neural computation*, **7**(5), 889–904. 592
- Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., and Ng, A. Y. (2012). Large scale distributed deep networks. In *NIPS'2012*. 23, 381
- Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, **5**(3), 142–150. 566
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**(6), 391–407. 406, 410
- Delalleau, O. and Bengio, Y. (2011). Shallow vs. deep sum-product networks. In *NIPS*. 17, 472
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*. 18

- Deng, J., Berg, A. C., Li, K., and Fei-Fei, L. (2010a). What does classifying more than 10,000 image categories tell us? In *Proceedings of the 11th European Conference on Computer Vision: Part V, ECCV'10*, pages 71–84, Berlin, Heidelberg. Springer-Verlag. 18
- Deng, L. and Yu, D. (2014). Deep learning – methods and applications. *Foundations and Trends in Signal Processing*. 391
- Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., and Hinton, G. (2010b). Binary coding of speech spectrograms using a deep auto-encoder. In *Interspeech 2010*, Makuhari, Chiba, Japan. 22
- Denil, M., Bazzani, L., Larochelle, H., and de Freitas, N. (2012). Learning where to attend with deep architectures for image tracking. *Neural Computation*, **24**(8), 2151–2184. 313
- Denton, E., Chintala, S., Szlam, A., and Fergus, R. (2015). Deep generative image models using a Laplacian pyramid of adversarial networks. *NIPS*. 599, 612
- Desjardins, G. and Bengio, Y. (2008). Empirical evaluation of convolutional RBMs for vision. Technical Report 1327, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal. 583
- Desjardins, G., Courville, A. C., Bengio, Y., Vincent, P., and Delalleau, O. (2010). Tempered Markov chain Monte Carlo for training of restricted Boltzmann machines. In *International Conference on Artificial Intelligence and Statistics*, pages 145–152. 514, 524
- Desjardins, G., Courville, A., and Bengio, Y. (2011). On tracking the partition function. In *NIPS'2011*. 537
- Devlin, J., Zbib, R., Huang, Z., Lamar, T., Schwartz, R., and Makhoul, J. (2014). Fast and robust neural network joint models for statistical machine translation. In *Proc. ACL'2014*. 403
- Devroye, L. (2013). *Non-Uniform Random Variate Generation*. SpringerLink : Bücher. Springer New York. 593
- DiCarlo, J. J. (2013). Mechanisms underlying visual object recognition: Humans vs. neurons vs. machines. *NIPS Tutorial*. 24, 312
- Dinh, L., Krueger, D., and Bengio, Y. (2014). NICE: Non-linear independent components estimation. arXiv:1410.8516. 421
- Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2014). Long-term recurrent convolutional networks for visual recognition and description. arXiv:1411.4389. 90

- Donoho, D. L. and Grimes, C. (2003). Hessian eigenmaps: new locally linear embedding techniques for high-dimensional data. Technical Report 2003-08, Dept. Statistics, Stanford University. 141, 443
- Dosovitskiy, A., Springenberg, J. T., and Brox, T. (2015). Learning to generate chairs with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1538–1546. 594, 601
- Doya, K. (1993). Bifurcations of recurrent neural networks in gradient descent learning. *IEEE Transactions on Neural Networks*, **1**, 75–80. 343, 345
- Dreyfus, S. E. (1962). The numerical solution of variational problems. *Journal of Mathematical Analysis and Applications*, **5(1)**, 30–45. 194
- Dreyfus, S. E. (1973). The computational solution of optimal control problems with time lag. *IEEE Transactions on Automatic Control*, **18(4)**, 383–385. 194
- Drucker, H. and LeCun, Y. (1992). Improving generalisation performance using double back-propagation. *IEEE Transactions on Neural Networks*, **3(6)**, 991–997. 233
- Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*. 261
- Dudik, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning, ICML '11*. 410
- Dugas, C., Bengio, Y., Bélisle, F., and Nadeau, C. (2001). Incorporating second-order functional knowledge for better option pricing. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13 (NIPS'00)*, pages 472–478. MIT Press. 61, 170
- Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*. 600
- El Hahi, S. and Bengio, Y. (1996). Hierarchical recurrent neural networks for long-term dependencies. In *NIPS 8*. MIT Press. 340, 348
- Elkahky, A. M., Song, Y., and He, X. (2015). A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*, pages 278–288. 408
- Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, **48**, 781–799. 279

- Erhan, D., Manzagol, P.-A., Bengio, Y., Bengio, S., and Vincent, P. (2009). The difficulty of training deep architectures and the effect of unsupervised pre-training. In *AISTATS'2009*, pages 153–160. 174
- Erhan, D., Bengio, Y., Courville, A., Manzagol, P., Vincent, P., and Bengio, S. (2010). Why does unsupervised pre-training help deep learning? *J. Machine Learning Res.* 452, 454, 455, 456
- Fahlman, S. E., Hinton, G. E., and Sejnowski, T. J. (1983). Massively parallel architectures for AI: NETL, thistle, and Boltzmann machines. In *Proceedings of the National Conference on Artificial Intelligence AAAI-83*. 486, 559
- Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J. C., Zitnick, C. L., and Zweig, G. (2015). From captions to visual concepts and back. arXiv:1411.4952. 90
- Farabet, C., LeCun, Y., Kavukcuoglu, K., Culurciello, E., Martini, B., Akselrod, P., and Talay, S. (2011). Large-scale FPGA-based convolutional networks. In R. Bekkerman, M. Bilenko, and J. Langford, editors, *Scaling up Machine Learning: Parallel and Distributed Approaches*. Cambridge University Press. 447
- Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(8), 1915–1929. 22, 174, 306
- Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(4), 594–611. 459
- Finn, C., Tan, X. Y., Duan, Y., Darrell, T., Levine, S., and Abbeel, P. (2015). Learning visual feature spaces for robotic manipulation with deep spatial autoencoders. *arXiv preprint arXiv:1509.06113*. 23
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188. 18, 92
- Földiák, P. (1989). Adaptive network for optimal linear feature extraction. In *International Joint Conference on Neural Networks (IJCNN)*, volume 1, pages 401–405, Washington 1989. IEEE, New York. 421
- Franzius, M., Sprekeler, H., and Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. 423

- Franzius, M., Wilbert, N., and Wiskott, L. (2008). Invariant object recognition with slow feature analysis. In *Proceedings of the 18th international conference on Artificial Neural Networks, Part I*, ICANN '08, pages 961–970, Berlin, Heidelberg. Springer-Verlag. 423
- Frasconi, P., Gori, M., and Sperduti, A. (1997). On the efficient classification of data structures by neural networks. In *Proc. Int. Joint Conf. on Artificial Intelligence*. 341, 342
- Frasconi, P., Gori, M., and Sperduti, A. (1998). A general framework for adaptive processing of data structures. *IEEE Transactions on Neural Networks*, **9**(5), 768–786. 341, 342
- Freund, Y. and Schapire, R. E. (1996a). Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of Thirteenth International Conference*, pages 148–156, USA. ACM. 222
- Freund, Y. and Schapire, R. E. (1996b). Game theory, on-line prediction and boosting. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 325–332. 222
- Frey, B. J. (1998). *Graphical models for machine learning and digital communication*. MIT Press. 602
- Frey, B. J., Hinton, G. E., and Dayan, P. (1996). Does the wake-sleep algorithm learn good density estimators? In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems 8 (NIPS'95)*, pages 661–670. MIT Press, Cambridge, MA. 557
- Frobenius, G. (1908). Über matrizen aus positiven elementen, s. *B. Preuss. Akad. Wiss. Berlin, Germany*. 508
- Fukushima, K. (1975). Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, **20**, 121–136. 14, 195, 451
- Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, **36**, 193–202. 14, 20, 21, 195, 313
- Gal, Y. and Ghahramani, Z. (2015). Bayesian convolutional neural networks with Bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158*. 227
- Gallinari, P., LeCun, Y., Thiria, S., and Fogelman-Soulie, F. (1987). Memoires associatives distribuees. In *Proceedings of COGNITIVA 87*, Paris, La Villette. 440

- Garcia-Duran, A., Bordes, A., Usunier, N., and Grandvalet, Y. (2015). Combining two and three-way embeddings models for link prediction in knowledge bases. *arXiv preprint arXiv:1506.00999*. 412
- Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1993). Darpa timit acoustic-phonetic continuous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon Technical Report N*, **93**, 27403. 390
- Garson, J. (1900). The metric system of identification of criminals, as used in Great Britain and Ireland. *The Journal of the Anthropological Institute of Great Britain and Ireland*, (2), 177–227. 18
- Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural computation*, **12**(10), 2451–2471. 349, 352
- Ghahramani, Z. and Hinton, G. E. (1996). The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, Dpt. of Comp. Sci., Univ. of Toronto. 417
- Gillick, D., Brunk, C., Vinyals, O., and Subramanya, A. (2015). Multilingual language processing from bytes. *arXiv preprint arXiv:1512.00103*. 406
- Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. 363
- Giudice, M. D., Manera, V., and Keysers, C. (2009). Programmed to learn? The ontogeny of mirror neurons. *Dev. Sci.*, **12**(2), 350--363. 560
- Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *AISTATS'2010*. 258
- Glorot, X., Bordes, A., and Bengio, Y. (2011a). Deep sparse rectifier neural networks. In *AISTATS'2011*. 15, 150, 170, 195
- Glorot, X., Bordes, A., and Bengio, Y. (2011b). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML'2011*. 433
- Glorot, X., Bordes, A., and Bengio, Y. (2011c). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICM (1b)*, pages 97–110. 457
- Goldberger, J., Roweis, S., Hinton, G. E., and Salakhutdinov, R. (2005). Neighbourhood components analysis. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17 (NIPS'04)*. MIT Press. 101

- Gong, S., McKenna, S., and Psarrou, A. (2000). *Dynamic Vision: From Images to Face Recognition*. Imperial College Press. 142, 443
- Goodfellow, I., Le, Q., Saxe, A., and Ng, A. (2009). Measuring invariances in deep networks. In Y. Bengio, D. Schuurmans, C. Williams, J. Lafferty, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22 (NIPS'09)*, pages 646–654. 219
- Goodfellow, I., Koenig, N., Muja, M., Pantofaru, C., Sorokin, A., and Takayama, L. (2010). Help me help you: Interfaces for personal robots. In *Proc. of Human Robot Interaction (HRI)*, Osaka, Japan. ACM Press, ACM Press. 88
- Goodfellow, I., Mirza, M., Xiao, D., Courville, A., and Bengio, Y. (2014a). An empirical investigation of catastrophic forgetting in gradient-based neural networks. In *ICLR'14*. 168
- Goodfellow, I. J. (2010). Technical report: Multidimensional, downsampled convolution for autoencoders. Technical report, Université de Montréal. 302
- Goodfellow, I. J. (2014). On distinguishability criteria for estimating generative models. In *International Conference on Learning Representations, Workshops Track*. 531, 598
- Goodfellow, I. J., Courville, A., and Bengio, Y. (2011). Spike-and-slab sparse coding for unsupervised feature discovery. In *NIPS Workshop on Challenges in Learning Hierarchical Models*. 454, 458
- Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013a). Maxout networks. In *ICML'2013*. 167
- Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013b). Maxout networks. In *ICM (1c)*, pages 1319–1327. 227, 292, 312
- Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013c). Maxout networks. Technical Report arXiv:1302.4389, Université de Montréal. 387
- Goodfellow, I. J., Mirza, M., Courville, A., and Bengio, Y. (2013d). Multi-prediction deep Boltzmann machines. In *NIP (1)*. 89, 526, 572, 574, 575, 576, 577, 596
- Goodfellow, I. J., Warde-Farley, D., Lamblin, P., Dumoulin, V., Mirza, M., Pascanu, R., Bergstra, J., Bastien, F., and Bengio, Y. (2013e). Pylearn2: a machine learning research library. *arXiv preprint arXiv:1308.4214*. 23, 380
- Goodfellow, I. J., Courville, A., and Bengio, Y. (2013f). Scaling up spike-and-slab models for unsupervised feature learning. *IEEE T. PAMI*, pages 1902–1914. 425, 426, 555

- Goodfellow, I. J., Courville, A., and Bengio, Y. (2013g). Scaling up spike-and-slab models for unsupervised feature learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(8), 1902–1914. 583
- Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014b). Explaining and harnessing adversarial examples. *CoRR*, **abs/1412.6572**. 230, 231, 233, 473, 474
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014c). Generative adversarial networks. In *NIPS'2014*. 464, 588, 597, 598, 601
- Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., and Shet, V. (2014d). Multi-digit number recognition from Street View imagery using deep convolutional neural networks. In *International Conference on Learning Representations*. 22, 89, 174, 175, 334, 359, 382
- Goodfellow, I. J., Vinyals, O., and Saxe, A. M. (2015). Qualitatively characterizing neural network optimization problems. In *International Conference on Learning Representations*. 244, 245, 246, 248
- Goodman, J. (2001). Classes for fast maximum entropy training. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Utah. 397
- Gori, M. and Tesi, A. (1992). On the problem of local minima in backpropagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-14**(1), 76–86. 243
- Gosset, W. S. (1908). The probable error of a mean. *Biometrika*, **6**(1), 1–25. Originally published under the pseudonym “Student”. 18
- Gouws, S., Bengio, Y., and Corrado, G. (2014). BilBOWA: Fast bilingual distributed representations without word alignments. Technical report, arXiv:1410.2455. 406, 459
- Graf, H. P. and Jackel, L. D. (1989). Analog electronic neural network circuits. *Circuits and Devices Magazine, IEEE*, **5**(4), 44–49. 384
- Graves, A. (2011). Practical variational inference for neural networks. In *NIPS'2011*. 208
- Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in Computational Intelligence. Springer. 320, 336, 351, 392
- Graves, A. (2013). Generating sequences with recurrent neural networks. Technical report, arXiv:1308.0850. 164, 349, 351, 354, 358
- Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *ICML'2014*. 349

- Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, **18**(5), 602–610. 337
- Graves, A. and Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *NIPS'2008*, pages 545–552. 337
- Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML'2006*, pages 369–376, Pittsburgh, USA. 392
- Graves, A., Liwicki, M., Bunke, H., Schmidhuber, J., and Fernández, S. (2008). Unconstrained on-line handwriting recognition with recurrent neural networks. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *NIPS'2007*, pages 577–584. 337
- Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., and Schmidhuber, J. (2009). A novel connectionist system for unconstrained handwriting recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **31**(5), 855–868. 349
- Graves, A., Mohamed, A., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *ICASSP'2013*, pages 6645–6649. 337, 340, 349, 392
- Graves, A., Wayne, G., and Danihelka, I. (2014). Neural Turing machines. arXiv:1410.5401. 23, 356
- Grefenstette, E., Hermann, K. M., Suleyman, M., and Blunsom, P. (2015). Learning to transduce with unbounded memory. In *NIPS'2015*. 356
- Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2015). LSTM: a search space odyssey. *arXiv preprint arXiv:1503.04069*. 352
- Gregor, K. and LeCun, Y. (2010a). Emergence of complex-like cells in a temporal product network with local receptive fields. Technical report, arXiv:1006.0448. 300
- Gregor, K. and LeCun, Y. (2010b). Learning fast approximations of sparse coding. In L. Bottou and M. Littman, editors, *Proceedings of the Twenty-seventh International Conference on Machine Learning (ICML-10)*. ACM. 558
- Gregor, K., Danihelka, I., Mnih, A., Blundell, C., and Wierstra, D. (2014). Deep autoregressive networks. In *International Conference on Machine Learning (ICML'2014)*. 592
- Gregor, K., Danihelka, I., Graves, A., and Wierstra, D. (2015). DRAW: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*. 596

- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, **13**(1), 723–773. 601
- Guillaume Desjardins, Karen Simonyan, R. P. K. K. (2015). Natural neural networks. Technical report, arXiv:1507.00210. 273
- Gulcehre, C. and Bengio, Y. (2013). Knowledge matters: Importance of prior information for optimization. Technical Report arXiv:1301.4083, Universite de Montreal. 22
- Guo, H. and Gelfand, S. B. (1992). Classification trees with neural network feature extraction. *Neural Networks, IEEE Transactions on*, **3**(6), 923–933. 383
- Gupta, S., Agrawal, A., Gopalakrishnan, K., and Narayanan, P. (2015). Deep learning with limited numerical precision. *CoRR*, **abs/1502.02551**. 384
- Gutmann, M. and Hyvarinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS'10)*. 529
- Hadsell, R., Sermanet, P., Ben, J., Erkan, A., Han, J., Muller, U., and LeCun, Y. (2007). Online learning for offroad robots: Spatial label propagation to learn long-range traversability. In *Proceedings of Robotics: Science and Systems*, Atlanta, GA, USA. 386
- Hajnal, A., Maass, W., Pudlak, P., Szegedy, M., and Turan, G. (1993). Threshold circuits of bounded depth. *J. Comput. System. Sci.*, **46**, 129–154. 172
- Håstad, J. (1986). Almost optimal lower bounds for small depth circuits. In *Proceedings of the 18th annual ACM Symposium on Theory of Computing*, pages 6–20, Berkeley, California. ACM Press. 172
- Håstad, J. and Goldmann, M. (1991). On the power of small-depth threshold circuits. *Computational Complexity*, **1**, 113–129. 172
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning: data mining, inference and prediction*. Springer Series in Statistics. Springer Verlag. 126
- He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *arXiv preprint arXiv:1502.01852*. 23, 167
- Hebb, D. O. (1949). *The Organization of Behavior*. Wiley, New York. 13, 15, 560
- Henaff, M., Jarrett, K., Kavukcuoglu, K., and LeCun, Y. (2011). Unsupervised learning of sparse features for scalable audio classification. In *ISMIR'11*. 447

- Henderson, J. (2003). Inducing history representations for broad coverage statistical parsing. In *HLT-NAACL*, pages 103–110. 406
- Henderson, J. (2004). Discriminative training of a neural network statistical parser. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 95. 406
- Henniges, M., Puertas, G., Bornschein, J., Eggert, J., and Lücke, J. (2010). Binary sparse coding. In *Latent Variable Analysis and Signal Separation*, pages 450–457. Springer. 546
- Herault, J. and Ans, B. (1984). Circuits neuronaux à synapses modifiables: Décodage de messages composites par apprentissage non supervisé. *Comptes Rendus de l' Académie des Sciences*, **299(III-13)**, 525--528. 419
- Hinton, G., Deng, L., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012a). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, **29(6)**, 82–97. 22, 391
- Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*. 381
- Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, **40**, 185–234. 421
- Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, **46(1)**, 47–75. 356
- Hinton, G. E. (1999). Products of experts. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN)*, volume 1, pages 1–6, Edinburgh, Scotland. IEE. 486
- Hinton, G. E. (2000). Training products of experts by minimizing contrastive divergence. Technical Report GCNU TR 2000-004, Gatsby Unit, University College London. 519, 578
- Hinton, G. E. (2006). To recognize shapes, first learn to generate images. Technical Report UTML TR 2006-003, University of Toronto. 451
- Hinton, G. E. (2007a). How to do backpropagation in a brain. Invited talk at the NIPS'2007 Deep Learning Workshop. 560
- Hinton, G. E. (2007b). Learning multiple layers of representation. *Trends in cognitive sciences*, **11(10)**, 428–434. 564
- Hinton, G. E. (2010). A practical guide to training restricted Boltzmann machines. Technical Report UTML TR 2010-003, Comp. Sc., University of Toronto. 519

- Hinton, G. E. (2012). Tutorial on deep learning. IPAM Graduate Summer School: Deep Learning, Feature Learning. 262
- Hinton, G. E. and Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society of London*. 128
- Hinton, G. E. and McClelland, J. L. (1988). Learning representations by recirculation. In *NIPS'1987*, pages 358–366. 429
- Hinton, G. E. and Roweis, S. (2003). Stochastic neighbor embedding. In *NIPS'2002*. 443
- Hinton, G. E. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, **313**(5786), 504–507. 435, 448, 451, 452, 454
- Hinton, G. E. and Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, chapter 7, pages 282–317. MIT Press, Cambridge. 486, 559
- Hinton, G. E. and Sejnowski, T. J. (1999). *Unsupervised learning: foundations of neural computation*. MIT press. 462
- Hinton, G. E. and Shallice, T. (1991). Lesioning an attractor network: investigations of acquired dyslexia. *Psychological review*, **98**(1), 74. 12
- Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length, and Helmholtz free energy. In *NIPS'1993*. 429
- Hinton, G. E., Sejnowski, T. J., and Ackley, D. H. (1984a). Boltzmann machines: Constraint satisfaction networks that learn. Technical Report TR-CMU-CS-84-119, Carnegie-Mellon University, Dept. of Computer Science. 486
- Hinton, G. E., Sejnowski, T. J., and Ackley, D. H. (1984b). Boltzmann machines: Constraint satisfaction networks that learn. Technical Report TR-CMU-CS-84-119, Carnegie-Mellon University, Dept. of Computer Science. 559
- Hinton, G. E., McClelland, J., and Rumelhart, D. (1986). Distributed representations. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 77–109. MIT Press, Cambridge. 16, 194, 449
- Hinton, G. E., Revow, M., and Dayan, P. (1995a). Recognizing handwritten digits using mixtures of linear models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7 (NIPS'94)*, pages 1015–1022. MIT Press, Cambridge, MA. 417

- Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995b). The wake-sleep algorithm for unsupervised neural networks. *Science*, **268**, 1558–1161. 431, 557
- Hinton, G. E., Dayan, P., and Revow, M. (1997). Modelling the manifolds of images of hand-written digits. *IEEE Transactions on Neural Networks*, **8**, 65–74. 426
- Hinton, G. E., Welling, M., Teh, Y. W., and Osindero, S. (2001). A new view of ICA. In *Proceedings of 3rd International Conference on Independent Component Analysis and Blind Signal Separation (ICA'01)*, pages 746–751, San Diego, CA. 419
- Hinton, G. E., Osindero, S., and Teh, Y. (2006a). A fast learning algorithm for deep belief nets. *Neural Computation*, **18**, 1527–1554. 13, 17, 21, 506, 564, 565
- Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006b). A fast learning algorithm for deep belief nets. *Neural Computation*, **18**, 1527–1554. 125, 451, 452
- Hinton, G. E., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012b). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, **29**(6), 82–97. 89
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012c). Improving neural networks by preventing co-adaptation of feature detectors. Technical report, arXiv:1207.0580. 205, 226
- Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012d). Improving neural networks by preventing co-adaptation of feature detectors. Technical report, arXiv:1207.0580. 229
- Hinton, G. E., Vinyals, O., and Dean, J. (2014). Dark knowledge. Invited talk at the BayLearn Bay Area Machine Learning Symposium. 381
- Hochreiter, S. (1991a). Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, T.U. München. 343, 344
- Hochreiter, S. (1991b). Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, Institut für Informatik, Lehrstuhl Prof. Brauer, Technische Universität München. 16
- Hochreiter, S. and Schmidhuber, J. (1995). Simplifying neural nets by discovering flat minima. In *Advances in Neural Information Processing Systems 7*, pages 529–536. MIT Press. 209
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**(8), 1735–1780. 16, 349, 351

- Hochreiter, S., Bengio, Y., and Frasconi, P. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In J. Kolen and S. Kremer, editors, *Field Guide to Dynamical Recurrent Networks*. IEEE Press. 351
- Holi, J. L. and Hwang, J.-N. (1993). Finite precision error analysis of neural network hardware implementations. *Computers, IEEE Transactions on*, **42**(3), 281–290. 384
- Holt, J. L. and Baker, T. E. (1991). Back propagation simulations using limited precision calculations. In *Neural Networks, 1991., IJCNN-91-Seattle International Joint Conference on*, volume 2, pages 121–126. IEEE. 384
- Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359–366. 171
- Hornik, K., Stinchcombe, M., and White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural networks*, **3**(5), 551–560. 171
- Hsu, F.-H. (2002). *Behind Deep Blue: Building the Computer That Defeated the World Chess Champion*. Princeton University Press, Princeton, NJ, USA. 2
- Huang, F. and Ogata, Y. (2002). Generalized pseudo-likelihood estimates for Markov random fields on lattice. *Annals of the Institute of Statistical Mathematics*, **54**(1), 1–18. 525
- Huang, P.-S., He, X., Gao, J., Deng, L., Acero, A., and Heck, L. (2013). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 2333–2338. ACM. 408
- Hubel, D. and Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology (London)*, **195**, 215–243. 311
- Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurons in the cat’s striate cortex. *Journal of Physiology*, **148**, 574–591. 311
- Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat’s visual cortex. *Journal of Physiology (London)*, **160**, 106–154. 311
- Huszar, F. (2015). How (not) to train your generative model: schedule sampling, likelihood, adversary? *arXiv:1511.05101*. 596
- Hutter, F., Hoos, H., and Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In *LION-5*. Extended version as UBC Tech report TR-2010-10. 371

- Hyotyniemi, H. (1996). Turing machines are recurrent neural networks. In *STeP'96*, pages 13–24. 325
- Hyvärinen, A. (1999). Survey on independent component analysis. *Neural Computing Surveys*, **2**, 94–128. 419
- Hyvärinen, A. (2005a). Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, **6**, 695–709. 437
- Hyvärinen, A. (2005b). Estimation of non-normalized statistical models using score matching. *J. Machine Learning Res.*, **6**. 526
- Hyvärinen, A. (2007a). Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, **18**, 1529–1531. 527
- Hyvärinen, A. (2007b). Some extensions of score matching. *Computational Statistics and Data Analysis*, **51**, 2499–2512. 527
- Hyvärinen, A. and Hoyer, P. O. (1999). Emergence of topography and complex cell properties from natural images using extensions of ica. In *NIPS*, pages 827–833. 421
- Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, **12**(3), 429–439. 420
- Hyvärinen, A., Karhunen, J., and Oja, E. (2001a). *Independent Component Analysis*. Wiley-Interscience. 419
- Hyvärinen, A., Hoyer, P. O., and Inki, M. O. (2001b). Topographic independent component analysis. *Neural Computation*, **13**(7), 1527–1558. 421
- Hyvärinen, A., Hurri, J., and Hoyer, P. O. (2009). *Natural Image Statistics: A probabilistic approach to early computational vision*. Springer-Verlag. 316
- Iba, Y. (2001). Extended ensemble Monte Carlo. *International Journal of Modern Physics*, **C12**, 623–656. 514
- Inayoshi, H. and Kurita, T. (2005). Improved generalization by adding both auto-association and hidden-layer noise to neural-network-based-classifiers. *IEEE Workshop on Machine Learning for Signal Processing*, pages 141–146. 440
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. 88, 271, 273

- Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural networks*, **1**(4), 295–307. 261
- Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, **3**, 79–87. 163, 383
- Jaeger, H. (2003). Adaptive nonlinear system identification with echo state networks. In *Advances in Neural Information Processing Systems 15*. 345
- Jaeger, H. (2007a). Discovering multiscale dynamical features with hierarchical echo state networks. Technical report, Jacobs University. 340
- Jaeger, H. (2007b). Echo state network. *Scholarpedia*, **2**(9), 2330. 345
- Jaeger, H. (2012). Long short-term memory in echo state networks: Details of a simulation study. Technical report, Technical report, Jacobs University Bremen. 346
- Jaeger, H. and Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, **304**(5667), 78–80. 21, 345
- Jaeger, H., Lukosevicius, M., Popovici, D., and Siewert, U. (2007). Optimization and applications of echo state networks with leaky- integrator neurons. *Neural Networks*, **20**(3), 335–352. 348
- Jain, V., Murray, J. F., Roth, F., Turaga, S., Zhigulin, V., Briggman, K. L., Helmstaedter, M. N., Denk, W., and Seung, H. S. (2007). Supervised learning of image restoration with convolutional networks. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE. 306
- Jaitly, N. and Hinton, G. (2011). Learning a better representation of speech soundwaves using restricted Boltzmann machines. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5884–5887. IEEE. 390
- Jaitly, N. and Hinton, G. E. (2013). Vocal tract length perturbation (VTLP) improves speech recognition. In *ICML'2013*. 207
- Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009a). What is the best multi-stage architecture for object recognition? In *Proc. International Conference on Computer Vision (ICCV'09)*, pages 2146–2153. IEEE. 15, 167
- Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009b). What is the best multi-stage architecture for object recognition? In *ICCV'09*. 20, 21, 150, 195, 310, 447

- Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, **78**, 2690–2693. 533, 536
- Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press. 47
- Jean, S., Cho, K., Memisevic, R., and Bengio, Y. (2014). On using very large target vocabulary for neural machine translation. arXiv:1412.2007. 403
- Jelinek, F. and Mercer, R. L. (1980). Interpolated estimation of Markov source parameters from sparse data. In E. S. Gelsema and L. N. Kanal, editors, *Pattern Recognition in Practice*. North-Holland, Amsterdam. 393, 402
- Jia, Y. (2013). Caffe: An open source convolutional architecture for fast feature embedding. <http://caffe.berkeleyvision.org/>. 23, 182
- Jia, Y., Huang, C., and Darrell, T. (2012). Beyond spatial pyramids: Receptive field learning for pooled image features. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3370–3377. IEEE. 293
- Jim, K.-C., Giles, C. L., and Horne, B. G. (1996). An analysis of noise in recurrent neural networks: convergence and generalization. *IEEE Transactions on Neural Networks*, **7**(6), 1424–1438. 208
- Jordan, M. I. (1998). *Learning in Graphical Models*. Kluwer, Dordrecht, Netherlands. 16
- Joulin, A. and Mikolov, T. (2015). Inferring algorithmic patterns with stack-augmented recurrent nets. *arXiv preprint arXiv:1503.01007*. 356
- Jozefowicz, R., Zaremba, W., and Sutskever, I. (2015). An empirical evaluation of recurrent network architectures. In *ICML'2015*. 260, 351, 352
- Judd, J. S. (1989). *Neural Network Design and the Complexity of Learning*. MIT press. 250
- Jutten, C. and Herault, J. (1991). Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, **24**, 1–10. 419
- Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, c., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R. C., Mirza, M., Jean, S., Carrier, P. L., Dauphin, Y., Boulanger-Lewandowski, N., Aggarwal, A., Zumer, J., Lamblin, P., Raymond, J.-P., Desjardins, G., Pascanu, R., Warde-Farley, D., Torabi, A., Sharma, A., Bengio, E., Côté, M., Konda, K. R., and Wu, Z. (2013). Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*. 174

- Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *EMNLP'2013*. 403
- Kalchbrenner, N., Danihelka, I., and Graves, A. (2015). Grid long short-term memory. *arXiv preprint arXiv:1507.01526*. 338
- Kamyshanska, H. and Memisevic, R. (2015). The potential energy of an autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 439
- Karpathy, A. and Li, F.-F. (2015). Deep visual-semantic alignments for generating image descriptions. In *CVPR'2015*. arXiv:1412.2306. 90
- Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*. 18
- Karush, W. (1939). *Minima of Functions of Several Variables with Inequalities as Side Constraints*. Master's thesis, Dept. of Mathematics, Univ. of Chicago. 85
- Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **ASSP-35**(3), 400–401. 393, 402
- Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2008). Fast inference in sparse coding algorithms with applications to object recognition. Technical report, Computational and Biological Learning Lab, Courant Institute, NYU. Tech Report CBLL-TR-2008-12-01. 447
- Kavukcuoglu, K., Ranzato, M.-A., Fergus, R., and LeCun, Y. (2009). Learning invariant features through topographic filter maps. In *CVPR'2009*. 447
- Kavukcuoglu, K., Sermanet, P., Boureau, Y.-L., Gregor, K., Mathieu, M., and LeCun, Y. (2010). Learning convolutional feature hierarchies for visual recognition. In *NIPS'2010*. 310, 447
- Kelley, H. J. (1960). Gradient theory of optimal flight paths. *ARS Journal*, **30**(10), 947–954. 194
- Khan, F., Zhu, X., and Mutlu, B. (2011). How do humans teach: On curriculum learning and teaching dimension. In *Advances in Neural Information Processing Systems 24 (NIPS'11)*, pages 1449–1457. 280
- Kim, S. K., McAfee, L. C., McMahan, P. L., and Olukotun, K. (2009). A highly scalable restricted Boltzmann machine FPGA implementation. In *Field Programmable Logic and Applications, 2009. FPL 2009. International Conference on*, pages 367–372. IEEE. 384

- Kindermann, R. (1980). *Markov Random Fields and Their Applications (Contemporary Mathematics ; V. 1)*. American Mathematical Society. 482
- Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 262
- Kingma, D. and LeCun, Y. (2010a). Regularized estimation of image statistics by score matching. In *NIPS'2010*. 438
- Kingma, D. and LeCun, Y. (2010b). Regularized estimation of image statistics by score matching. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1126–1134. 528
- Kingma, D., Rezende, D., Mohamed, S., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *NIPS'2014*. 363
- Kingma, D. P. (2013). Fast gradient-based inference with continuous latent variable models in auxiliary form. Technical report, arxiv:1306.0733. 558, 588, 594
- Kingma, D. P. and Welling, M. (2014a). Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*. 588, 597
- Kingma, D. P. and Welling, M. (2014b). Efficient gradient-based inference through transformations between bayes nets and neural nets. Technical report, arxiv:1402.0480. 588
- Kirkpatrick, S., Jr., C. D. G., , and Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, **220**, 671–680. 279
- Kiros, R., Salakhutdinov, R., and Zemel, R. (2014a). Multimodal neural language models. In *ICML'2014*. 90
- Kiros, R., Salakhutdinov, R., and Zemel, R. (2014b). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv:1411.2539 [cs.LG]*. 90, 349
- Klementiev, A., Titov, I., and Bhattarai, B. (2012). Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*. 406, 459
- Knowles-Barley, S., Jones, T. R., Morgan, J., Lee, D., Kasthuri, N., Lichtman, J. W., and Pfister, H. (2014). Deep learning for the connectome. *GPU Technology Conference*. 24
- Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press. 496, 506, 551

- Konig, Y., Bourlard, H., and Morgan, N. (1996). REMAP: Recursive estimation and maximization of a posteriori probabilities – application to transition-based connectionist speech recognition. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems 8 (NIPS'95)*. MIT Press, Cambridge, MA. 390
- Koren, Y. (2009). The BellKor solution to the Netflix grand prize. 222, 408
- Kotzias, D., Denil, M., de Freitas, N., and Smyth, P. (2015). From group to individual labels using deep features. In *ACM SIGKDD*. 93
- Koutnik, J., Greff, K., Gomez, F., and Schmidhuber, J. (2014). A clockwork RNN. In *ICML'2014*. 348
- Kočiský, T., Hermann, K. M., and Blunsom, P. (2014). Learning Bilingual Word Representations by Marginalizing Alignments. In *Proceedings of ACL*. 404
- Krause, O., Fischer, A., Glasmachers, T., and Igel, C. (2013). Approximation properties of DBNs with binary hidden units and real-valued visible units. In *ICML'2013*. 472
- Krizhevsky, A. (2010). Convolutional deep belief networks on CIFAR-10. Technical report, University of Toronto. Unpublished Manuscript: <http://www.cs.utoronto.ca/~kriz/conv-cifar10-aug2010.pdf>. 380
- Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto. 18, 477
- Krizhevsky, A. and Hinton, G. E. (2011). Using very deep autoencoders for content-based image retrieval. In *ESANN*. 448
- Krizhevsky, A., Sutskever, I., and Hinton, G. (2012a). ImageNet classification with deep convolutional neural networks. In *NIPS'2012*. 20, 21, 88, 174, 317
- Krizhevsky, A., Sutskever, I., and Hinton, G. (2012b). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS'2012)*. 22, 386, 389
- Krueger, K. A. and Dayan, P. (2009). Flexible shaping: how learning in small steps helps. *Cognition*, **110**, 380–394. 279
- Kuhn, H. W. and Tucker, A. W. (1951). Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, Berkeley, Calif. University of California Press. 85

- Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., and Socher, R. (2015a). Ask me anything: Dynamic memory networks for natural language processing. Technical report, arXiv:1506.07285. 356
- Kumar, A., Irsoy, O., Su, J., Bradbury, J., English, R., Pierce, B., Ondruska, P., Iyyer, M., Gulrajani, I., and Socher, R. (2015b). Ask me anything: Dynamic memory networks for natural language processing. *arXiv:1506.07285*. 412
- Kumar, M. P., Packer, B., and Koller, D. (2010). Self-paced learning for latent variable models. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1189–1197. 279
- Lang, K. J. and Hinton, G. E. (1988). The development of the time-delay neural network architecture for speech recognition. Technical Report CMU-CS-88-152, Carnegie-Mellon University. 313, 319, 347
- Lang, K. J., Waibel, A. H., and Hinton, G. E. (1990). A time-delay neural network architecture for isolated word recognition. *Neural networks*, **3**(1), 23–43. 319
- Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for contextual multi-armed bandits. In *NIPS'2008*, pages 1096--1103. 409
- Lappalainen, H., Giannakopoulos, X., Honkela, A., and Karhunen, J. (2000). Nonlinear independent component analysis using ensemble learning: Experiments and discussion. In *Proc. ICA*. Citeseer. 420
- Larochelle, H. and Bengio, Y. (2008a). Classification using discriminative restricted Boltzmann machines. In *ICML'2008*. 210, 586, 610
- Larochelle, H. and Bengio, Y. (2008b). Classification using discriminative restricted Boltzmann machines. In *ICM (1a)*, pages 536–543. 219, 453
- Larochelle, H. and Hinton, G. E. (2010). Learning to combine foveal glimpses with a third-order Boltzmann machine. In *Advances in Neural Information Processing Systems 23*, pages 1243–1251. 313
- Larochelle, H. and Murray, I. (2011). The Neural Autoregressive Distribution Estimator. In *AISTATS'2011*. 602, 604, 605
- Larochelle, H., Erhan, D., and Bengio, Y. (2008). Zero-data learning of new tasks. In *AAAI Conference on Artificial Intelligence*. 459
- Larochelle, H., Bengio, Y., Louradour, J., and Lamblin, P. (2009). Exploring strategies for training deep neural networks. In *JML (1)*, pages 1–40. 455

- Lasserre, J. A., Bishop, C. M., and Minka, T. P. (2006). Principled hybrids of generative and discriminative models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'06)*, pages 87–94, Washington, DC, USA. IEEE Computer Society. 210, 218
- Le, Q., Ngiam, J., Chen, Z., hao Chia, D. J., Koh, P. W., and Ng, A. (2010). Tiled convolutional neural networks. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23 (NIPS'10)*, pages 1279–1287. 300
- Le, Q., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B., and Ng, A. (2011). On optimization methods for deep learning. In *Proc. ICML'2011*. ACM. 270
- Le, Q., Ranzato, M., Monga, R., Devin, M., Corrado, G., Chen, K., Dean, J., and Ng, A. (2012). Building high-level features using large scale unsupervised learning. In *ICML'2012*. 20, 21
- Le Roux, N. and Bengio, Y. (2008). Representational power of restricted Boltzmann machines and deep belief networks. *Neural Computation*, **20**(6), 1631–1649. 472, 560
- Le Roux, N. and Bengio, Y. (2010). Deep belief networks are compact universal approximators. *Neural Computation*, **22**(8), 2192–2207. 472
- LeCun, Y. (1985). Une procédure d'apprentissage pour Réseau à seuil assymétrique. In *Cognitive 85: A la Frontière de l'Intelligence Artificielle, des Sciences de la Connaissance et des Neurosciences*, pages 599–604, Paris 1985. CESTA, Paris. 194
- LeCun, Y. (1986). Learning processes in an asymmetric threshold network. In E. Bienenstock, F. Fogelman-Soulié, and G. Weisbuch, editors, *Disordered Systems and Biological Organization*, pages 233–240. Springer-Verlag, Berlin, Les Houches 1985. 298
- LeCun, Y. (1987). *Modèles connexionistes de l'apprentissage*. Ph.D. thesis, Université de Paris VI. 16, 429, 440
- LeCun, Y. (1989). Generalization and network design strategies. Technical Report CRG-TR-89-4, University of Toronto. 281, 298
- LeCun, Y., Jackel, L. D., Boser, B., Denker, J. S., Graf, H. P., Guyon, I., Henderson, D., Howard, R. E., and Hubbard, W. (1989). Handwritten digit recognition: Applications of neural network chips and automatic learning. *IEEE Communications Magazine*, **27**(11), 41–46. 314
- LeCun, Y., Bottou, L., Orr, G. B., and Müller, K.-R. (1998a). Efficient backprop. In *Neural Networks, Tricks of the Trade*, Lecture Notes in Computer Science LNCS 1524. Springer Verlag. 265

- LeCun, Y., Bottou, L., Orr, G. B., and Müller, K. (1998b). Efficient backprop. In *Neural Networks, Tricks of the Trade*. 365
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998c). Gradient based learning applied to document recognition. *Proc. IEEE*. 14, 16, 18, 21, 317, 390, 392
- LeCun, Y., Kavukcuoglu, K., and Farabet, C. (2010). Convolutional networks and applications in vision. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, pages 253–256. IEEE. 317
- L’Ecuyer, P. (1994). Efficiency improvement and variance reduction. In *Proceedings of the 1994 Winter Simulation Conference*, pages 122--132. 589
- Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., and Tu, Z. (2014). Deeply-supervised nets. *arXiv preprint arXiv:1409.5185*. 278
- Lee, H., Battle, A., Raina, R., and Ng, A. (2007). Efficient sparse coding algorithms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19 (NIPS’06)*, pages 801–808. MIT Press. 544
- Lee, H., Ekanadham, C., and Ng, A. (2008). Sparse deep belief net model for visual area V2. In *NIPS’07*. 219
- Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In L. Bottou and M. Littman, editors, *Proceedings of the Twenty-sixth International Conference on Machine Learning (ICML’09)*. ACM, Montreal, Canada. 310, 583, 584
- Lee, Y. J. and Grauman, K. (2011). Learning the easy things first: self-paced visual category discovery. In *CVPR’2011*. 279
- Leibniz, G. W. (1676). Memoir using the chain rule. (Cited in TMME 7:2&3 p 321-332, 2010). 194
- Lenat, D. B. and Guha, R. V. (1989). *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc. 2
- Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, **6**, 861--867. 171, 172
- Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly Journal of Applied Mathematics*, **II(2)**, 164–168. 266

- L'Hôpital, G. F. A. (1696). *Analyse des infiniment petits, pour l'intelligence des lignes courbes*. Paris: L'Imprimerie Royale. 194
- Li, Y., Swersky, K., and Zemel, R. S. (2015). Generative moment matching networks. *CoRR*, **abs/1502.02761**. 600
- Lin, T., Horne, B. G., Tino, P., and Giles, C. L. (1996). Learning long-term dependencies is not as difficult with NARX recurrent neural networks. *IEEE Transactions on Neural Networks*, **7**(6), 1329–1338. 347
- Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. In *Proc. AAAI'15*. 412
- Linde, N. (1992). The machine that changed the world, episode 3. Documentary miniseries. 2
- Lindsey, C. and Lindblad, T. (1994). Review of hardware neural networks: a user's perspective. In *Proc. Third Workshop on Neural Networks: From Biology to High Energy Physics*, pages 195--202, Isola d'Elba, Italy. 384
- Linnainmaa, S. (1976). Taylor expansion of the accumulated rounding error. *BIT Numerical Mathematics*, **16**(2), 146–160. 194
- LISA (2008). Deep learning tutorials: Restricted Boltzmann machines. Technical report, LISA Lab, Université de Montréal. 501
- Long, P. M. and Servedio, R. A. (2010). Restricted Boltzmann machines are hard to approximately evaluate or simulate. In *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*. 561
- Lotter, W., Kreiman, G., and Cox, D. (2015). Unsupervised learning of visual structure using predictive generative networks. *arXiv preprint arXiv:1511.06380*. 464, 465
- Lovelace, A. (1842). Notes upon L. F. Menabrea's "Sketch of the Analytical Engine invented by Charles Babbage". 1
- Lu, L., Zhang, X., Cho, K., and Renals, S. (2015). A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition. In *Proc. Interspeech*. 392
- Lu, T., Pál, D., and Pál, M. (2010). Contextual multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 485–492. 409
- Luenberger, D. G. (1984). *Linear and Nonlinear Programming*. Addison Wesley. 270
- Lukoševičius, M. and Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, **3**(3), 127–149. 345

- Luo, H., Shen, R., Niu, C., and Ullrich, C. (2011). Learning class-relevant features and class-irrelevant features via a hybrid third-order RBM. In *International Conference on Artificial Intelligence and Statistics*, pages 470–478. 586
- Luo, H., Carrier, P. L., Courville, A., and Bengio, Y. (2013). Texture modeling with convolutional spike-and-slab RBMs and deep extensions. In *AISTATS'2013*. 90
- Lyu, S. (2009). Interpretation and generalization of score matching. In *Proceedings of the Twenty-fifth Conference in Uncertainty in Artificial Intelligence (UAI'09)*. 527
- Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., and Svetnik, V. (2015). Deep neural nets as a method for quantitative structure - activity relationships. *J. Chemical information and modeling*. 452
- Maas, A. L., Hannun, A. Y., and Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing*. 167
- Maass, W. (1992). Bounds for the computational power and learning complexity of analog neural nets (extended abstract). In *Proc. of the 25th ACM Symp. Theory of Computing*, pages 335–344. 172
- Maass, W., Schnitger, G., and Sontag, E. D. (1994). A comparison of the computational power of sigmoid and Boolean threshold circuits. *Theoretical Advances in Neural Computation and Learning*, pages 127–151. 172
- Maass, W., Natschlaeger, T., and Markram, H. (2002). Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, **14**(11), 2531–2560. 345
- MacKay, D. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press. 66
- Maclaurin, D., Duvenaud, D., and Adams, R. P. (2015). Gradient-based hyperparameter optimization through reversible learning. *arXiv preprint arXiv:1502.03492*. 370
- Mao, J., Xu, W., Yang, Y., Wang, J., and Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv:1412.6632 [cs.CV]*. 90
- Marcotte, P. and Savard, G. (1992). Novel approaches to the discrimination problem. *Zeitschrift für Operations Research (Theory)*, **36**, 517–545. 237
- Marlin, B. and de Freitas, N. (2011). Asymptotic efficiency of deterministic estimators for discrete energy-based models: Ratio matching and pseudolikelihood. In *UAI'2011*. 526, 528

- Marlin, B., Swersky, K., Chen, B., and de Freitas, N. (2010). Inductive principles for restricted Boltzmann machine learning. In *AISTATS'2010*, pages 509–516. 522, 527
- Marquardt, D. W. (1963). An algorithm for least-squares estimation of non-linear parameters. *Journal of the Society of Industrial and Applied Mathematics*, **11**(2), 431–441. 266
- Marr, D. and Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, **194**. 313
- Martens, J. (2010). Deep learning via Hessian-free optimization. In *ICML'2010*, pages 735–742. 259
- Martens, J. and Medabalimi, V. (2014). On the expressive efficiency of sum product networks. *arXiv:1411.7717*. 472
- Martens, J. and Sutskever, I. (2011). Learning recurrent neural networks with Hessian-free optimization. In *Proc. ICML'2011*. ACM. 352, 353
- Mase, S. (1995). Consistency of the maximum pseudo-likelihood estimator of continuous state space Gibbsian processes. *The Annals of Applied Probability*, **5**(3), pp. 603–612. 525
- McClelland, J., Rumelhart, D., and Hinton, G. (1995). The appeal of parallel distributed processing. In *Computation & intelligence*, pages 305–341. American Association for Artificial Intelligence. 15
- McCulloch, W. S. and Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, 115–133. 13
- Mead, C. and Ismail, M. (2012). *Analog VLSI implementation of neural systems*, volume 80. Springer Science & Business Media. 384
- Melchior, J., Fischer, A., and Wiskott, L. (2013). How to center binary deep Boltzmann machines. *arXiv preprint arXiv:1311.1354*. 575
- Memisevic, R. and Hinton, G. E. (2007). Unsupervised learning of image transformations. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'07)*. 586
- Memisevic, R. and Hinton, G. E. (2010). Learning to represent spatial transformations with factored higher-order Boltzmann machines. *Neural Computation*, **22**(6), 1473–1492. 586
- Mesnil, G., Dauphin, Y., Glorot, X., Rifai, S., Bengio, Y., Goodfellow, I., Lavoie, E., Muller, X., Desjardins, G., Warde-Farley, D., Vincent, P., Courville, A., and Bergstra, J. (2011). Unsupervised and transfer learning challenge: a deep learning approach. In *JMLR W&CP: Proc. Unsupervised and Transfer Learning*, volume 7. 174, 454, 458

- Mesnil, G., Rifai, S., Dauphin, Y., Bengio, Y., and Vincent, P. (2012). Surfing on the manifold. Learning Workshop, Snowbird. 607
- Miikkulainen, R. and Dyer, M. G. (1991). Natural language processing with modular PDP networks and distributed lexicon. *Cognitive Science*, **15**, 343–399. 406
- Mikolov, T. (2012). *Statistical Language Models based on Neural Networks*. Ph.D. thesis, Brno University of Technology. 353
- Mikolov, T., Deoras, A., Kombrink, S., Burget, L., and Cernocky, J. (2011a). Empirical evaluation and combination of advanced language modeling techniques. In *Proc. 12th annual conference of the international speech communication association (INTERSPEECH 2011)*. 402
- Mikolov, T., Deoras, A., Povey, D., Burget, L., and Cernocky, J. (2011b). Strategies for training large scale neural network language models. In *Proc. ASRU'2011*. 279, 402
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In *International Conference on Learning Representations: Workshops Track*. 456
- Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. Technical report, arXiv:1309.4168. 459
- Minka, T. (2005). Divergence measures and message passing. *Microsoft Research Cambridge UK Tech Rep MSRTR2005173*, **72**(TR-2005-173). 533
- Minsky, M. L. and Papert, S. A. (1969). *Perceptrons*. MIT Press, Cambridge. 14
- Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*. 599
- Mishkin, D. and Matas, J. (2015). All you need is a good init. *arXiv preprint arXiv:1511.06422*. 259
- Misra, J. and Saha, I. (2010). Artificial neural networks in hardware: A survey of two decades of progress. *Neurocomputing*, **74**(1), 239–255. 384
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York. 87
- Miyato, T., Maeda, S., Koyama, M., Nakae, K., and Ishii, S. (2015). Distributional smoothing with virtual adversarial training. In *ICLR*. Preprint: arXiv:1507.00677. 231
- Mnih, A. and Gregor, K. (2014). Neural variational inference and learning in belief networks. In *ICML'2014*. 590, 591, 592

- Mnih, A. and Hinton, G. E. (2007). Three new graphical models for statistical language modelling. In Z. Ghahramani, editor, *Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML'07)*, pages 641–648. ACM. 396
- Mnih, A. and Hinton, G. E. (2009). A scalable hierarchical distributed language model. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21 (NIPS'08)*, pages 1081–1088. 397
- Mnih, A. and Kavukcuoglu, K. (2013). Learning word embeddings efficiently with noise-contrastive estimation. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2265–2273. Curran Associates, Inc. 401, 530
- Mnih, A. and Teh, Y. W. (2012). A fast and simple algorithm for training neural probabilistic language models. In *ICML'2012*, pages 1751–1758. 401
- Mnih, V. and Hinton, G. (2010). Learning to detect roads in high-resolution aerial images. In *Proceedings of the 11th European Conference on Computer Vision (ECCV)*. 90
- Mnih, V., Larochelle, H., and Hinton, G. (2011). Conditional restricted Boltzmann machines for structure output prediction. In *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*. 585
- Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., and Wierstra, D. (2013). Playing Atari with deep reinforcement learning. Technical report, arXiv:1312.5602. 93
- Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. (2014). Recurrent models of visual attention. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *NIPS'2014*, pages 2204–2212. 591
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidgeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, **518**, 529–533. 23
- Mobahi, H. and Fisher, III, J. W. (2015). A theoretical analysis of optimization by Gaussian continuation. In *AAAI'2015*. 279
- Mobahi, H., Collobert, R., and Weston, J. (2009). Deep learning from temporal coherence in video. In L. Bottou and M. Littman, editors, *Proceedings of the 26th International Conference on Machine Learning*, pages 737–744, Montreal. Omnipress. 421
- Mohamed, A., Dahl, G., and Hinton, G. (2009). Deep belief networks for phone recognition. 391

- Mohamed, A., Sainath, T. N., Dahl, G., Ramabhadran, B., Hinton, G. E., and Picheny, M. A. (2011). Deep belief networks using discriminative features for phone recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5060–5063. IEEE. 391
- Mohamed, A., Dahl, G., and Hinton, G. (2012a). Acoustic modeling using deep belief networks. *IEEE Trans. on Audio, Speech and Language Processing*, **20**(1), 14–22. 391
- Mohamed, A., Hinton, G., and Penn, G. (2012b). Understanding how deep belief networks perform acoustic modelling. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4273–4276. IEEE. 391
- Moller, M. (1993). *Efficient Training of Feed-Forward Neural Networks*. Ph.D. thesis, Aarhus University, Aarhus, Denmark. 270
- Montavon, G. and Muller, K.-R. (2012). Deep Boltzmann machines and the centering trick. In G. Montavon, G. Orr, and K.-R. Müller, editors, *Neural Networks: Tricks of the Trade*, volume 7700 of *Lecture Notes in Computer Science*, pages 621–637. Preprint: <http://arxiv.org/abs/1203.3783>. 575
- Montúfar, G. (2014). Universal approximation depth and errors of narrow belief networks with discrete units. *Neural Computation*, **26**. 472
- Montúfar, G. and Ay, N. (2011). Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. *Neural Computation*, **23**(5), 1306–1319. 472
- Montufar, G. F., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *NIPS'2014*. 17, 172, 173
- Mor-Yosef, S., Samueloff, A., Modan, B., Navot, D., and Schenker, J. G. (1990). Ranking the risk factors for cesarean: logistic regression analysis of a nationwide study. *Obstet Gynecol*, **75**(6), 944–7. 2
- Morin, F. and Bengio, Y. (2005). Hierarchical probabilistic neural network language model. In *AISTATS'2005*. 397, 399
- Mozer, M. C. (1992). The induction of multiscale temporal structure. In J. M. S. Hanson and R. Lippmann, editors, *Advances in Neural Information Processing Systems 4 (NIPS'91)*, pages 275–282, San Mateo, CA. Morgan Kaufmann. 348
- Murphy, K. P. (2012). *Machine Learning: a Probabilistic Perspective*. MIT Press, Cambridge, MA, USA. 56, 87, 126

- Murray, B. U. I. and Larochelle, H. (2014). A deep and tractable density estimator. In *ICML'2014*. 164, 606
- Nair, V. and Hinton, G. (2010a). Rectified linear units improve restricted Boltzmann machines. In *ICML'2010*. 150, 170
- Nair, V. and Hinton, G. E. (2009). 3d object recognition with deep belief nets. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1339–1347. Curran Associates, Inc. 586
- Nair, V. and Hinton, G. E. (2010b). Rectified linear units improve restricted Boltzmann machines. In L. Bottou and M. Littman, editors, *Proceedings of the Twenty-seventh International Conference on Machine Learning (ICML-10)*, pages 807–814. ACM. 14
- Narayanan, H. and Mitter, S. (2010). Sample complexity of testing the manifold hypothesis. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1786–1794. 141
- Naumann, U. (2008). Optimal Jacobian accumulation is NP-complete. *Mathematical Programming*, **112**(2), 427–441. 191
- Navigli, R. and Velardi, P. (2005). Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **27**(7), 1075--1086. 412
- Neal, R. and Hinton, G. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge, MA. 541
- Neal, R. M. (1990). Learning stochastic feedforward networks. Technical report. 591
- Neal, R. M. (1993). Probabilistic inference using Markov chain Monte-Carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto. 581
- Neal, R. M. (1994). Sampling from multimodal distributions using tempered transitions. Technical Report 9421, Dept. of Statistics, University of Toronto. 514
- Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. Springer. 228
- Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, **11**(2), 125–139. 533, 535, 536

- Neal, R. M. (2005). Estimating ratios of normalizing constants using linked importance sampling. 536, 537
- Nesterov, Y. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, **27**, 372–376. 256
- Nesterov, Y. (2004). *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London. 256
- Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. (2011). Reading digits in natural images with unsupervised feature learning. Deep Learning and Unsupervised Feature Learning Workshop, NIPS. 18
- Ney, H. and Kneser, R. (1993). Improved clustering techniques for class-based statistical language modelling. In *European Conference on Speech Communication and Technology (Eurospeech)*, pages 973–976, Berlin. 394
- Ng, A. (2015). Advice for applying machine learning. <https://see.stanford.edu/materials/aimlcs229/ML-advice.pdf>. 359
- Niesler, T. R., Whittaker, E. W. D., and Woodland, P. C. (1998). Comparison of part-of-speech and automatically derived category-based language models for speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 177–180. 394
- Ning, F., Delhomme, D., LeCun, Y., Piano, F., Bottou, L., and Barbano, P. E. (2005). Toward automatic phenotyping of developing embryos from videos. *Image Processing, IEEE Transactions on*, **14**(9), 1360–1371. 306
- Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer. 82, 85
- Norouzi, M. and Fleet, D. J. (2011). Minimal loss hashing for compact binary codes. In *ICML'2011*. 448
- Nowlan, S. J. (1990). Competing experts: An experimental investigation of associative mixture models. Technical Report CRG-TR-90-5, University of Toronto. 383
- Nowlan, S. J. and Hinton, G. E. (1992). Adaptive soft weight tying using Gaussian mixtures. In J. M. S. Hanson and R. Lippmann, editors, *Advances in Neural Information Processing Systems 4 (NIPS'91)*, pages 993–1000, San Mateo, CA. Morgan Kaufmann. 122
- Olshausen, B. and Field, D. J. (2005). How close are we to understanding V1? *Neural Computation*, **17**, 1665–1699. 14

- Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**, 607–609. 128, 219, 316, 423
- Olshausen, B. A., Anderson, C. H., and Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.*, **13**(11), 4700–4719. 383
- Opper, M. and Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural computation*, **21**(3), 786–792. 588
- Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1717–1724. IEEE. 456
- Osindero, S. and Hinton, G. E. (2008). Modeling image patches with a directed hierarchy of Markov random fields. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS'07)*, pages 1121–1128, Cambridge, MA. MIT Press. 539
- Ovid and Martin, C. (2004). *Metamorphoses*. W.W. Norton. 1
- Paccanaro, A. and Hinton, G. E. (2000). Extracting distributed representations of concepts and relations from positive and negative propositions. In *International Joint Conference on Neural Networks (IJCNN)*, Como, Italy. IEEE, New York. 411, 412
- Paine, T. L., Khorrami, P., Han, W., and Huang, T. S. (2014). An analysis of unsupervised pre-training in light of recent advances. *arXiv preprint arXiv:1412.6597*. 454
- Palatucci, M., Pomerleau, D., Hinton, G. E., and Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1410–1418. Curran Associates, Inc. 459
- Parker, D. B. (1985). Learning-logic. Technical Report TR-47, Center for Comp. Research in Economics and Management Sci., MIT. 194
- Pascanu, R., Mikolov, T., and Bengio, Y. (2013a). On the difficulty of training recurrent neural networks. In *ICML'2013*. 247, 343, 348, 353, 354, 355
- Pascanu, R., Mikolov, T., and Bengio, Y. (2013b). On the difficulty of training recurrent neural networks. In *ICM (1c)*. 345
- Pascanu, R., Gulcehre, C., Cho, K., and Bengio, Y. (2014a). How to construct deep recurrent neural networks. In *ICLR*. 17, 228, 340, 341, 349, 392

- Pascanu, R., Montufar, G., and Bengio, Y. (2014b). On the number of inference regions of deep feed forward networks with piece-wise linear activations. In ICL (1). 469
- Pati, Y., Rezaifar, R., and Krishnaprasad, P. (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers*, pages 40–44. 220
- Pearl, J. (1985). Bayesian networks: A model of self-activated memory for evidential reasoning. In *Proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine*, pages 329–334. 480
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. 48
- Perron, O. (1907). Zur theorie der matrices. *Mathematische Annalen*, **64**(2), 248–263. 508
- Petersen, K. B. and Pedersen, M. S. (2006). The matrix cookbook. Version 20051003. 27
- Peterson, G. B. (2004). A day of great illumination: B. F. Skinner’s discovery of shaping. *Journal of the Experimental Analysis of Behavior*, **82**(3), 317–328. 279
- Pham, D.-T., Garat, P., and Jutten, C. (1992). Separation of a mixture of independent sources through a maximum likelihood approach. In *EUSIPCO*, pages 771–774. 419
- Pham, P.-H., Jelaca, D., Farabet, C., Martini, B., LeCun, Y., and Culurciello, E. (2012). NeFlow: dataflow vision processing system-on-a-chip. In *Circuits and Systems (MWSCAS), 2012 IEEE 55th International Midwest Symposium on*, pages 1044–1047. IEEE. 384
- Pinheiro, P. H. O. and Collobert, R. (2014). Recurrent convolutional neural networks for scene labeling. In *ICML’2014*. 306
- Pinheiro, P. H. O. and Collobert, R. (2015). From image-level to pixel-level labeling with convolutional networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 306
- Pinto, N., Cox, D. D., and DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS Comput Biol*, **4**. 388
- Pinto, N., Stone, Z., Zickler, T., and Cox, D. (2011). Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 35–42. IEEE. 310

- Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, **46**(1), 77–105. 341
- Polyak, B. and Juditsky, A. (1992). Acceleration of stochastic approximation by averaging. *SIAM J. Control and Optimization*, **30**(4), 838–855. 274
- Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, **4**(5), 1–17. 253
- Poole, B., Sohl-Dickstein, J., and Ganguli, S. (2014). Analyzing noise in autoencoders and deep networks. *CoRR*, **abs/1406.1831**. 207
- Poon, H. and Domingos, P. (2011). Sum-product networks for deep learning. In *Learning Workshop*, Fort Lauderdale, FL. 472
- Presley, R. K. and Haggard, R. L. (1994). A fixed point implementation of the backpropagation learning algorithm. In *Southeastcon'94. Creative Technology Transfer-A Global Affair., Proceedings of the 1994 IEEE*, pages 136–138. IEEE. 384
- Price, R. (1958). A useful theorem for nonlinear devices having Gaussian inputs. *IEEE Transactions on Information Theory*, **4**(2), 69–72. 588
- Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, **435**(7045), 1102–1107. 312
- Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*. 470, 471, 599
- Raiko, T., Yao, L., Cho, K., and Bengio, Y. (2014). Iterative neural autoregressive distribution estimator (NADE-k). Technical report, arXiv:1406.1485. 576, 605
- Raina, R., Madhavan, A., and Ng, A. Y. (2009a). Large-scale deep unsupervised learning using graphics processors. In L. Bottou and M. Littman, editors, *Proceedings of the Twenty-sixth International Conference on Machine Learning (ICML'09)*, pages 873–880, New York, NY, USA. ACM. 21
- Raina, R., Madhavan, A., and Ng, A. Y. (2009b). Large-scale deep unsupervised learning using graphics processors. In *ICML'2009*. 379
- Ramsey, F. P. (1926). Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, chapter 7, pages 156–198. McMaster University Archive for the History of Economic Thought. 49

- Ranzato, M. and Hinton, G. H. (2010). Modeling pixel means and covariances using factorized third-order Boltzmann machines. In *CVPR'2010*, pages 2551–2558. 581
- Ranzato, M., Poultney, C., Chopra, S., and LeCun, Y. (2007a). Efficient learning of sparse representations with an energy-based model. In *NIPS'2006*. 13, 433, 451, 452
- Ranzato, M., Poultney, C., Chopra, S., and LeCun, Y. (2007b). Efficient learning of sparse representations with an energy-based model. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pages 1137–1144. MIT Press. 17
- Ranzato, M., Huang, F., Boureau, Y., and LeCun, Y. (2007c). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *CVPR'07*. 310
- Ranzato, M., Boureau, Y., and LeCun, Y. (2008). Sparse feature learning for deep belief networks. In *NIPS'2007*. 433
- Ranzato, M., Krizhevsky, A., and Hinton, G. E. (2010a). Factored 3-way restricted Boltzmann machines for modeling natural images. In *Proceedings of AISTATS 2010*. 579, 580
- Ranzato, M., Mnih, V., and Hinton, G. (2010b). Generating more realistic images using gated MRFs. In *NIPS'2010*. 581
- Rao, C. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, **37**, 81–89. 118, 252
- Rasmus, A., Valpola, H., Honkala, M., Berglund, M., and Raiko, T. (2015). Semi-supervised learning with ladder network. *arXiv preprint arXiv:1507.02672*. 363, 453
- Recht, B., Re, C., Wright, S., and Niu, F. (2011). Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *NIPS'2011*. 380
- Reichert, D. P., Seriès, P., and Storkey, A. J. (2011). Neuronal adaptation for sampling-based probabilistic inference in perceptual bistability. In *Advances in Neural Information Processing Systems*, pages 2357–2365. 569
- Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *ICML'2014*. Preprint: arXiv:1401.4082. 558, 588, 594
- Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. (2011a). Contractive auto-encoders: Explicit invariance during feature extraction. In *ICML'2011*. 445, 446, 447

- Rifai, S., Mesnil, G., Vincent, P., Muller, X., Bengio, Y., Dauphin, Y., and Glorot, X. (2011b). Higher order contractive auto-encoder. In *ECML PKDD*. 445, 446
- Rifai, S., Dauphin, Y., Vincent, P., Bengio, Y., and Muller, X. (2011c). The manifold tangent classifier. In *NIPS'2011*. 233, 446
- Rifai, S., Dauphin, Y., Vincent, P., Bengio, Y., and Muller, X. (2011d). The manifold tangent classifier. In *NIPS'2011*. Student paper award. 233
- Rifai, S., Bengio, Y., Dauphin, Y., and Vincent, P. (2012). A generative process for sampling contractive auto-encoders. In *ICML'2012*. 607
- Ringach, D. and Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, **28**(2), 147–166. 314
- Roberts, S. and Everson, R. (2001). *Independent component analysis: principles and practice*. Cambridge University Press. 420
- Robinson, A. J. and Fallside, F. (1991). A recurrent error propagation network speech recognition system. *Computer Speech and Language*, **5**(3), 259–274. 21, 390
- Rockafellar, R. T. (1997). *Convex analysis*. princeton landmarks in mathematics. 82
- Romero, A., Ballas, N., Ebrahimi Kahou, S., Chassang, A., Gatta, C., and Bengio, Y. (2015). Fitnets: Hints for thin deep nets. In *ICLR'2015*, *arXiv:1412.6550*. 277
- Rosen, J. B. (1960). The gradient projection method for nonlinear programming. part i. linear constraints. *Journal of the Society for Industrial and Applied Mathematics*, **8**(1), pp. 181–217. 83
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386–408. 13, 21
- Rosenblatt, F. (1962). *Principles of Neurodynamics*. Spartan, New York. 21
- Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, **27**(3), 832–837. 13
- Roweis, S. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**(5500). 141, 443
- Roweis, S., Saul, L., and Hinton, G. (2002). Global coordination of local linear models. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS'01)*, Cambridge, MA. MIT Press. 417

- Rubin, D. B. *et al.* (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, **12**(4), 1151–1172. 611
- Rumelhart, D., Hinton, G., and Williams, R. (1986a). Learning representations by back-propagating errors. *Nature*, **323**, 533–536. 13, 194, 406, 410
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986b). Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, chapter 8, pages 318–362. MIT Press, Cambridge. 18, 21, 194
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986c). Learning representations by back-propagating errors. *Nature*, **323**, 533–536. 16, 175, 319
- Rumelhart, D. E., McClelland, J. L., and the PDP Research Group (1986d). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press, Cambridge. 15, 22
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2014a). ImageNet Large Scale Visual Recognition Challenge. 18
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., *et al.* (2014b). Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*. 23
- Russel, S. J. and Norvig, P. (2003). *Artificial Intelligence: a Modern Approach*. Prentice Hall. 77
- Rust, N., Schwartz, O., Movshon, J. A., and Simoncelli, E. (2005). Spatiotemporal elements of macaque V1 receptive fields. *Neuron*, **46**(6), 945–956. 313
- Sainath, T., Mohamed, A., Kingsbury, B., and Ramabhadran, B. (2013). Deep convolutional neural networks for LVCSR. In *ICASSP 2013*. 391
- Salakhutdinov, R. (2010). Learning in Markov random fields using tempered transitions. In Y. Bengio, D. Schuurmans, C. Williams, J. Lafferty, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22 (NIPS'09)*. 514
- Salakhutdinov, R. and Hinton, G. (2009a). Deep Boltzmann machines. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 5, pages 448–455. 20, 21, 452, 566, 569, 572, 574
- Salakhutdinov, R. and Hinton, G. (2009b). Semantic hashing. In *International Journal of Approximate Reasoning*. 448

- Salakhutdinov, R. and Hinton, G. E. (2007a). Learning a nonlinear embedding by preserving class neighbourhood structure. In *Proceedings of AISTATS-2007*. 450
- Salakhutdinov, R. and Hinton, G. E. (2007b). Semantic hashing. In *SIGIR'2007*. 448
- Salakhutdinov, R. and Hinton, G. E. (2008). Using deep belief nets to learn covariance kernels for Gaussian processes. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS'07)*, pages 1249–1256, Cambridge, MA. MIT Press. 210
- Salakhutdinov, R. and Larochelle, H. (2010). Efficient learning of deep Boltzmann machines. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*, *JMLR W&CP*, volume 9, pages 693–700. 557
- Salakhutdinov, R. and Mnih, A. (2008). Probabilistic matrix factorization. In *NIPS'2008*. 408
- Salakhutdinov, R. and Murray, I. (2008). On the quantitative analysis of deep belief networks. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML'08)*, volume 25, pages 872–879. ACM. 536, 566
- Salakhutdinov, R., Mnih, A., and Hinton, G. (2007). Restricted Boltzmann machines for collaborative filtering. In *ICML*. 408
- Sanger, T. D. (1994). Neural network learning control of robot manipulators using gradually increasing task difficulty. *IEEE Transactions on Robotics and Automation*, **10**(3). 279
- Saul, L. K. and Jordan, M. I. (1996). Exploiting tractable substructures in intractable networks. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems 8 (NIPS'95)*. MIT Press, Cambridge, MA. 544
- Saul, L. K., Jaakkola, T., and Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, **4**, 61–76. 21, 592
- Savich, A. W., Moussa, M., and Areibi, S. (2007). The impact of arithmetic representation on implementing mlp-bp on fpgas: A study. *Neural Networks, IEEE Transactions on*, **18**(1), 240–252. 384
- Saxe, A. M., Koh, P. W., Chen, Z., Bhand, M., Suresh, B., and Ng, A. (2011). On random weights and unsupervised feature learning. In *Proc. ICML'2011*. ACM. 310
- Saxe, A. M., McClelland, J. L., and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR*. 244, 245, 258

- Schaul, T., Antonoglou, I., and Silver, D. (2014). Unit tests for stochastic optimization. In *International Conference on Learning Representations*. 263
- Schmidhuber, J. (1992). Learning complex, extended sequences using the principle of history compression. *Neural Computation*, **4**(2), 234–242. 340
- Schmidhuber, J. (1996). Sequential neural text compression. *IEEE Transactions on Neural Networks*, **7**(1), 142–146. 406
- Schmidhuber, J. (2012). Self-delimiting neural networks. *arXiv preprint arXiv:1210.0118*. 333
- Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press. 601
- Schölkopf, B., Burges, C. J. C., and Smola, A. J. (1998a). *Advances in kernel methods: support vector learning*. MIT Press, Cambridge, MA. 141
- Schölkopf, B., Smola, A., and Müller, K.-R. (1998b). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**, 1299–1319. 443
- Schölkopf, B., Burges, C. J. C., and Smola, A. J. (1999). *Advances in Kernel Methods — Support Vector Learning*. MIT Press, Cambridge, MA. 16, 124
- Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012). On causal and anticausal learning. In *ICML'2012*, pages 1255–1262. 465
- Schuster, M. (1999). On supervised learning from sequential data with applications for speech recognition. 164
- Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, **45**(11), 2673–2681. 336
- Schwenk, H. (2007). Continuous space language models. *Computer speech and language*, **21**, 492–518. 396
- Schwenk, H. (2010). Continuous space language models for statistical machine translation. *The Prague Bulletin of Mathematical Linguistics*, **93**, 137–146. 402
- Schwenk, H. (2014). Cleaned subset of WMT '14 dataset. 18
- Schwenk, H. and Bengio, Y. (1998). Training methods for adaptive boosting of neural networks. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Processing Systems 10 (NIPS'97)*, pages 647–653. MIT Press. 222

- Schwenk, H. and Gauvain, J.-L. (2002). Connectionist language modeling for large vocabulary continuous speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 765–768, Orlando, Florida. 396
- Schwenk, H., Costa-jussà, M. R., and Fonollosa, J. A. R. (2006). Continuous space language models for the IWSLT 2006 task. In *International Workshop on Spoken Language Translation*, pages 166–173. 402
- Seide, F., Li, G., and Yu, D. (2011). Conversational speech transcription using context-dependent deep neural networks. In *Interspeech 2011*, pages 437–440. 22
- Sejnowski, T. (1987). Higher-order Boltzmann machines. In *AIP Conference Proceedings 151 on Neural Networks for Computing*, pages 398–403. American Institute of Physics Inc. 586
- Series, P., Reichert, D. P., and Storkey, A. J. (2010). Hallucinations in Charles Bonnet syndrome induced by homeostasis: a deep Boltzmann machine model. In *Advances in Neural Information Processing Systems*, pages 2020–2028. 569
- Sermanet, P., Chintala, S., and LeCun, Y. (2012). Convolutional neural networks applied to house numbers digit classification. In *International Conference on Pattern Recognition (ICPR 2012)*. 388
- Sermanet, P., Kavukcuoglu, K., Chintala, S., and LeCun, Y. (2013). Pedestrian detection with unsupervised multi-stage feature learning. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR'13)*. IEEE. 22, 174
- Shilov, G. (1977). *Linear Algebra*. Dover Books on Mathematics Series. Dover Publications. 27
- Siegelmann, H. (1995). Computation beyond the Turing limit. *Science*, **268**(5210), 545–548. 324
- Siegelmann, H. and Sontag, E. (1991). Turing computability with neural nets. *Applied Mathematics Letters*, **4**(6), 77–80. 324
- Siegelmann, H. T. and Sontag, E. D. (1995). On the computational power of neural nets. *Journal of Computer and Systems Sciences*, **50**(1), 132–150. 324, 325, 345
- Sietsma, J. and Dow, R. (1991). Creating artificial neural networks that generalize. *Neural Networks*, **4**(1), 67–79. 207
- Simard, D., Steinkraus, P. Y., and Platt, J. C. (2003). Best practices for convolutional neural networks. In *ICDAR'2003*. 317

- Simard, P. and Graf, H. P. (1994). Backpropagation without multiplication. In *Advances in Neural Information Processing Systems*, pages 232–239. 384
- Simard, P., Victorri, B., LeCun, Y., and Denker, J. (1992). Tangent prop - A formalism for specifying selected invariances in an adaptive network. In *NIPS'1991*. 232, 233, 301
- Simard, P. Y., LeCun, Y., and Denker, J. (1993). Efficient pattern recognition using a new transformation distance. In *NIPS'92*. 232
- Simard, P. Y., LeCun, Y. A., Denker, J. S., and Victorri, B. (1998). Transformation invariance in pattern recognition — tangent distance and tangent propagation. *Lecture Notes in Computer Science*, **1524**. 232
- Simons, D. J. and Levin, D. T. (1998). Failure to detect changes to people during a real-world interaction. *Psychonomic Bulletin & Review*, **5**(4), 644–649. 463
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*. 275
- Sjöberg, J. and Ljung, L. (1995). Overtraining, regularization and searching for a minimum, with application to neural networks. *International Journal of Control*, **62**(6), 1391–1407. 215
- Skinner, B. F. (1958). Reinforcement today. *American Psychologist*, **13**, 94–99. 279
- Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, chapter 6, pages 194–281. MIT Press, Cambridge. 486, 499, 561
- Snoek, J., Larochelle, H., and Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *NIPS'2012*. 371
- Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011a). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *NIPS'2011*. 341, 342
- Socher, R., Manning, C., and Ng, A. Y. (2011b). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the Twenty-Eighth International Conference on Machine Learning (ICML'2011)*. 341
- Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011c). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *EMNLP'2011*. 341, 342

- Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013a). Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP'2013*. 341, 342
- Socher, R., Ganjoo, M., Manning, C. D., and Ng, A. Y. (2013b). Zero-shot learning through cross-modal transfer. In *27th Annual Conference on Neural Information Processing Systems (NIPS 2013)*. 459
- Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. 610, 611
- Sohn, K., Zhou, G., and Lee, H. (2013). Learning and selecting features jointly with point-wise gated Boltzmann machines. In *ICML'2013*. 586
- Solomonoff, R. J. (1989). A system for incremental learning based on algorithmic probability. 279
- Sontag, E. D. (1998). VC dimension of neural networks. *NATO ASI Series F Computer and Systems Sciences*, **168**, 69–96. 467, 470
- Sontag, E. D. and Sussman, H. J. (1989). Backpropagation can give rise to spurious local minima even for networks without hidden layers. *Complex Systems*, **3**, 91–106. 243
- Sparkes, B. (1996). *The Red and the Black: Studies in Greek Pottery*. Routledge. 1
- Spitkovsky, V. I., Alshawi, H., and Jurafsky, D. (2010). From baby steps to leapfrog: how “less is more” in unsupervised dependency parsing. In *HLT'10*. 279
- Squire, W. and Trapp, G. (1998). Using complex variables to estimate derivatives of real functions. *SIAM Rev.*, **40**(1), 110–112. 373
- Srebro, N. and Shraibman, A. (2005). Rank, trace-norm and max-norm. In *Proceedings of the 18th Annual Conference on Learning Theory*, pages 545–560. Springer-Verlag. 206
- Srivastava, N. (2013). *Improving Neural Networks With Dropout*. Master’s thesis, U. Toronto. 456
- Srivastava, N. and Salakhutdinov, R. (2012). Multimodal learning with deep Boltzmann machines. In *NIPS'2012*. 460
- Srivastava, N., Salakhutdinov, R. R., and Hinton, G. E. (2013). Modeling documents with deep Boltzmann machines. *arXiv preprint arXiv:1309.6865*. 566

- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, **15**, 1929–1958. 222, 227, 228, 229, 574
- Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Highway networks. *arXiv:1505.00387*. 278
- Steinkrau, D., Simard, P. Y., and Buck, I. (2005). Using GPUs for machine learning algorithms. *2013 12th International Conference on Document Analysis and Recognition*, **0**, 1115–1119. 379
- Stoyanov, V., Ropson, A., and Eisner, J. (2011). Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 15 of *JMLR Workshop and Conference Proceedings*, pages 725–733, Fort Lauderdale. Supplementary material (4 pages) also available. 576, 596
- Sukhbaatar, S., Szlam, A., Weston, J., and Fergus, R. (2015). Weakly supervised memory networks. *arXiv preprint arXiv:1503.08895*. 356
- Supancic, J. and Ramanan, D. (2013). Self-paced learning for long-term tracking. In *CVPR'2013*. 280
- Sussillo, D. (2014). Random walks: Training very deep nonlinear feed-forward networks with smart initialization. *CoRR*, **abs/1412.6558**. 248, 259, 260, 344
- Sutskever, I. (2012). *Training Recurrent Neural Networks*. Ph.D. thesis, Department of computer science, University of Toronto. 347, 353
- Sutskever, I. and Hinton, G. E. (2008). Deep narrow sigmoid belief networks are universal approximators. *Neural Computation*, **20**(11), 2629–2636. 592
- Sutskever, I. and Tieleman, T. (2010). On the Convergence Properties of Contrastive Divergence. In *AISTATS'2010*. 521
- Sutskever, I., Hinton, G., and Taylor, G. (2009). The recurrent temporal restricted Boltzmann machine. In *NIPS'2008*. 585
- Sutskever, I., Martens, J., and Hinton, G. E. (2011). Generating text with recurrent neural networks. In *ICML'2011*, pages 1017–1024. 406
- Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *ICML*. 256, 347, 353

- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS'2014, arXiv:1409.3215*. 23, 89, 338, 349, 351, 403
- Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction*. MIT Press. 93
- Sutton, R. S., Mcallester, D., Singh, S., and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *NIPS'1999*, pages 1057--1063. MIT Press. 590
- Swersky, K., Ranzato, M., Buchman, D., Marlin, B., and de Freitas, N. (2011). On autoencoders and score matching for energy based models. In *ICML'2011*. ACM. 438
- Swersky, K., Snoek, J., and Adams, R. P. (2014). Freeze-thaw Bayesian optimization. *arXiv preprint arXiv:1406.3896*. 371
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014a). Going deeper with convolutions. Technical report, arXiv:1409.4842. 20, 21, 174, 222, 231, 278, 295
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. (2014b). Intriguing properties of neural networks. *ICLR*, **abs/1312.6199**. 230, 233
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the Inception Architecture for Computer Vision. *ArXiv e-prints*. 209, 275
- Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. In *CVPR'2014*. 88
- Tandy, D. W. (1997). *Works and Days: A Translation and Commentary for the Social Sciences*. University of California Press. 1
- Tang, Y. and Elasmith, C. (2010). Deep networks for robust visual recognition. In *Proceedings of the 27th International Conference on Machine Learning, June 21-24, 2010, Haifa, Israel*. 207
- Tang, Y., Salakhutdinov, R., and Hinton, G. (2012). Deep mixtures of factor analysers. *arXiv preprint arXiv:1206.4635*. 417
- Taylor, G. and Hinton, G. (2009). Factored conditional restricted Boltzmann machines for modeling motion style. In L. Bottou and M. Littman, editors, *Proceedings of the Twenty-sixth International Conference on Machine Learning (ICML'09)*, pages 1025–1032, Montreal, Quebec, Canada. ACM. 585

- Taylor, G., Hinton, G. E., and Roweis, S. (2007). Modeling human motion using binary latent variables. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pages 1345–1352. MIT Press, Cambridge, MA. 585
- Teh, Y., Welling, M., Osindero, S., and Hinton, G. E. (2003). Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, **4**, 1235–1260. 419
- Tenenbaum, J., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**(5500), 2319–2323. 141, 443, 456
- Theis, L., van den Oord, A., and Bethge, M. (2015). A note on the evaluation of generative models. arXiv:1511.01844. 596, 613
- Thompson, J., Jain, A., LeCun, Y., and Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS'2014*. 306
- Thrun, S. (1995). Learning to play the game of chess. In *NIPS'1994*. 232
- Tibshirani, R. J. (1995). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, **58**, 267–288. 204
- Tieleman, T. (2008). Training restricted Boltzmann machines using approximations to the likelihood gradient. In *ICML'2008*, pages 1064–1071. 521
- Tieleman, T. and Hinton, G. (2009). Using fast weights to improve persistent contrastive divergence. In *ICML'2009*. 524
- Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal components analysis. *Journal of the Royal Statistical Society B*, **61**(3), 611–622. 419
- Torralba, A., Fergus, R., and Weiss, Y. (2008). Small codes and large databases for recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'08)*, pages 1–8. 448
- Touretzky, D. S. and Minton, G. E. (1985). Symbols among the neurons: Details of a connectionist inference architecture. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'85*, pages 238–243, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. 15
- Tu, K. and Honavar, V. (2011). On the utility of curricula in unsupervised learning of probabilistic grammars. In *IJCAI'2011*. 279
- Turaga, S. C., Murray, J. F., Jain, V., Roth, F., Helmstaedter, M., Briggman, K., Denk, W., and Seung, H. S. (2010). Convolutional networks can learn to generate affinity graphs for image segmentation. *Neural Computation*, **22**, 511–538. 306

- Turian, J., Ratinov, L., and Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Proc. ACL'2010*, pages 384–394. 455
- Töscher, A., Jahrer, M., and Bell, R. M. (2009). The BigChaos solution to the Netflix grand prize. 408
- Uria, B., Murray, I., and Larochelle, H. (2013). Rnade: The real-valued neural autoregressive density-estimator. In *NIPS'2013*. 605, 606
- van den Oörd, A., Dieleman, S., and Schrauwen, B. (2013). Deep content-based music recommendation. In *NIPS'2013*. 408
- van der Maaten, L. and Hinton, G. E. (2008). Visualizing data using t-SNE. *J. Machine Learning Res.*, **9**. 406, 443
- Vanhoucke, V., Senior, A., and Mao, M. Z. (2011). Improving the speed of neural networks on CPUs. In *Proc. Deep Learning and Unsupervised Feature Learning NIPS Workshop*. 378, 384
- Vapnik, V. N. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin. 100
- Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer, New York. 100
- Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, **16**, 264–280. 100
- Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural Computation*, **23**(7). 438, 439, 440, 608
- Vincent, P. and Bengio, Y. (2003). Manifold Parzen windows. In *NIPS'2002*. MIT Press. 444
- Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008a). Extracting and composing robust features with denoising autoencoders. In *ICM (1a)*, pages 1096–1103. 207
- Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008b). Extracting and composing robust features with denoising autoencoders. In *ICML 2008*. 440
- Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Machine Learning Res.*, **11**. 440
- Vincent, P., de Brébisson, A., and Bouthillier, X. (2015). Efficient exact gradient update for training deep networks with very large sparse targets. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1108–1116. Curran Associates, Inc. 396

- Vinyals, O., Kaiser, L., Koo, T., Petrov, S., Sutskever, I., and Hinton, G. (2014a). Grammar as a foreign language. *arXiv preprint arXiv:1412.7449*. 349
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2014b). Show and tell: a neural image caption generator. *arXiv 1411.4555*. 349
- Vinyals, O., Fortunato, M., and Jaitly, N. (2015a). Pointer networks. *arXiv preprint arXiv:1506.03134*. 356
- Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015b). Show and tell: a neural image caption generator. In *CVPR'2015*. *arXiv:1411.4555*. 90
- Viola, P. and Jones, M. (2001). Robust real-time object detection. In *International Journal of Computer Vision*. 382
- Visin, F., Kastner, K., Cho, K., Matteucci, M., Courville, A., and Bengio, Y. (2015). ReNet: A recurrent neural network based alternative to convolutional networks. *arXiv preprint arXiv:1505.00393*. 338
- Von Melchner, L., Pallas, S. L., and Sur, M. (2000). Visual behaviour mediated by retinal projections directed to the auditory pathway. *Nature*, **404**(6780), 871–876. 14
- Wager, S., Wang, S., and Liang, P. (2013). Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems 26*, pages 351–359. 228
- Waibel, A., Hanazawa, T., Hinton, G. E., Shikano, K., and Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, 328–339. 319, 386, 390
- Wan, L., Zeiler, M., Zhang, S., LeCun, Y., and Fergus, R. (2013). Regularization of neural networks using dropconnect. In *ICML'2013*. 229
- Wang, S. and Manning, C. (2013). Fast dropout training. In *ICML'2013*. 228
- Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014a). Knowledge graph and text jointly embedding. In *Proc. EMNLP'2014*. 411
- Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014b). Knowledge graph embedding by translating on hyperplanes. In *Proc. AAAI'2014*. 412
- Warde-Farley, D., Goodfellow, I. J., Courville, A., and Bengio, Y. (2014). An empirical analysis of dropout in piecewise linear networks. In *ICL (1)*. 225, 228, 229
- Wawrzynek, J., Asanovic, K., Kingsbury, B., Johnson, D., Beck, J., and Morgan, N. (1996). Spert-II: A vector microprocessor system. *Computer*, **29**(3), 79–86. 384

- Weaver, L. and Tao, N. (2001). The optimal reward baseline for gradient-based reinforcement learning. In *Proc. UAI'2001*, pages 538–545. 590
- Weinberger, K. Q. and Saul, L. K. (2004a). Unsupervised learning of image manifolds by semidefinite programming. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR'04)*, volume 2, pages 988–995, Washington D.C. 141
- Weinberger, K. Q. and Saul, L. K. (2004b). Unsupervised learning of image manifolds by semidefinite programming. In *CVPR'2004*, pages 988–995. 443
- Weiss, Y., Torralba, A., and Fergus, R. (2008). Spectral hashing. In *NIPS*, pages 1753–1760. 448
- Welling, M., Zemel, R. S., and Hinton, G. E. (2002). Self supervised boosting. In *Advances in Neural Information Processing Systems*, pages 665–672. 600
- Welling, M., Hinton, G. E., and Osindero, S. (2003a). Learning sparse topographic representations with products of Student-t distributions. In *NIPS'2002*. 581
- Welling, M., Zemel, R., and Hinton, G. E. (2003b). Self-supervised boosting. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15 (NIPS'02)*, pages 665–672. MIT Press. 531
- Welling, M., Rosen-Zvi, M., and Hinton, G. E. (2005). Exponential family harmoniums with an application to information retrieval. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17 (NIPS'04)*, volume 17, Cambridge, MA. MIT Press. 578
- Verbos, P. J. (1981). Applications of advances in nonlinear sensitivity analysis. In *Proceedings of the 10th IFIP Conference, 31.8 - 4.9, NYC*, pages 762–770. 194
- Weston, J., Bengio, S., and Usunier, N. (2010). Large scale image annotation: learning to rank with joint word-image embeddings. *Machine Learning*, **81**(1), 21–35. 343
- Weston, J., Chopra, S., and Bordes, A. (2014). Memory networks. *arXiv preprint arXiv:1410.3916*. 356, 412
- Widrow, B. and Hoff, M. E. (1960). Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, volume 4, pages 96–104. IRE, New York. 13, 18, 20, 21
- Wikipedia (2015). List of animals by number of neurons — Wikipedia, the free encyclopedia. [Online; accessed 4-March-2015]. 20, 21

- Williams, C. K. I. and Agakov, F. V. (2002). Products of Gaussians and Probabilistic Minor Component Analysis. *Neural Computation*, **14**(5), 1169–1182. 582
- Williams, C. K. I. and Rasmussen, C. E. (1996). Gaussian processes for regression. In D. Touretzky, M. Mozer, and M. Hasselmo, editors, *Advances in Neural Information Processing Systems 8 (NIPS'95)*, pages 514–520. MIT Press, Cambridge, MA. 124
- Williams, R. J. (1992). Simple statistical gradient-following algorithms connectionist reinforcement learning. *Machine Learning*, **8**, 229–256. 588, 589
- Williams, R. J. and Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, **1**, 270–280. 192
- Wilson, D. R. and Martinez, T. R. (2003). The general inefficiency of batch training for gradient descent learning. *Neural Networks*, **16**(10), 1429–1451. 239
- Wilson, J. R. (1984). Variance reduction techniques for digital simulation. *American Journal of Mathematical and Management Sciences*, **4**(3), 277--312. 589
- Wiskott, L. and Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, **14**(4), 715–770. 421, 422
- Wolpert, D. and MacReady, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, **1**, 67–82. 250
- Wolpert, D. H. (1996). The lack of a priori distinction between learning algorithms. *Neural Computation*, **8**(7), 1341–1390. 102
- Wu, R., Yan, S., Shan, Y., Dang, Q., and Sun, G. (2015). Deep image: Scaling up image recognition. arXiv:1501.02876. 381
- Wu, Z. (1997). Global continuation for distance geometry problems. *SIAM Journal of Optimization*, **7**, 814–836. 279
- Xiong, H. Y., Barash, Y., and Frey, B. J. (2011). Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics*, **27**(18), 2554–2562. 228
- Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *ICML'2015, arXiv:1502.03044*. 90, 349, 591
- Yildiz, I. B., Jaeger, H., and Kiebel, S. J. (2012). Re-visiting the echo state property. *Neural networks*, **35**, 1–9. 346

- Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *NIPS '14*, pages 3320–3328. Curran Associates, Inc. 277, 456
- Younes, L. (1998). On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates. In *Stochastics and Stochastics Models*, pages 177–228. 521
- Yu, D., Wang, S., and Deng, L. (2010). Sequential labeling using deep-structured conditional random fields. *IEEE Journal of Selected Topics in Signal Processing*. 276
- Zaremba, W. and Sutskever, I. (2014). Learning to execute. arXiv 1410.4615. 280
- Zaremba, W. and Sutskever, I. (2015). Reinforcement learning neural Turing machines. *arXiv:1505.00521*. 358
- Zaslavsky, T. (1975). *Facing Up to Arrangements: Face-Count Formulas for Partitions of Space by Hyperplanes*. Number no. 154 in *Memoirs of the American Mathematical Society*. American Mathematical Society. 469
- Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *ECCV'14*. 5
- Zeiler, M. D., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q., Nguyen, P., Senior, A., Vanhoucke, V., Dean, J., and Hinton, G. E. (2013). On rectified linear units for speech processing. In *ICASSP 2013*. 391
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2015). Object detectors emerge in deep scene CNNs. *ICLR'2015*, arXiv:1412.6856. 470
- Zhou, J. and Troyanskaya, O. G. (2014). Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. In *ICML'2014*. 610
- Zhou, Y. and Chellappa, R. (1988). Computation of optical flow using a neural network. In *Neural Networks, 1988., IEEE International Conference on*, pages 71–78. IEEE. 290
- Zöhrer, M. and Pernkopf, F. (2014). General stochastic networks for classification. In *NIPS'2014*. 610

术语

- 绝对值整流 absolute value rectification 167, 172, 173
- 准确率 accuracy 91, 360, 372–375
- 声学 acoustic 392
- 激活函数 activation function 147, 245, 257, 258, 260, 271–273, 277, 278
- AdaGrad** AdaGrad 261, 262
- 对抗 adversarial 464
- 对抗样本 adversarial example 230, 231
- 对抗训练 adversarial training 230, 231, 233, 474
- 几乎处处 almost everywhere 64
- 几乎必然 almost sure 114
- 几乎必然收敛 almost sure convergence 114
- 选择性剪接数据集 alternative splicing dataset 456
- 原始采样 Ancestral Sampling 494, 495, 507, 513, 557, 565, 569, 591, 606, 610
- 退火重要采样 annealed importance sampling 533–537, 566, 571, 612
- 专用集成电路 application-specific integrated circuit 384
- 近似贝叶斯计算 approximate Bayesian computation 611
- 近似推断 approximate inference 490, 497, 499, 539–542, 556–558
- 架构 architecture 170
- 人工智能 artificial intelligence 1–4, 6–10, 16, 17, 21, 47, 49, 136, 138, 141, 279, 362, 377, 385, 411, 416, 444, 462, 471, 472, 476, 614
- 人工神经网络 artificial neural network 12, 13, 20, 21, 377
- 渐近无偏 asymptotically unbiased 109

异步随机梯度下降 Asynchronous Stochastic Gradient Descent 380

异步 asynchronous 240

注意力机制 attention mechanism 313, 339, 358, 382, 383, 404, 405, 596

属性 attribute 411

自编码器 autoencoder xv, 4, 20, 21, 169, 233, 234, 245, 260, 293, 301–303, 322, 373, 420, 425, 426, 428–442, 445–448, 450, 452, 453, 464, 474, 514, 521, 528, 551, 558, 595, 596, 601, 606–608

自动微分 automatic differentiation 191

自动语音识别 Automatic Speech Recognition 390, 391

自回归网络 auto-regressive network 592, 602, 603, 605, 606

反向传播 back propagate 425

反向传播 back propagation 147, 175, 406, 429, 530, 560, 575–577, 585–588, 592, 594–596, 610

回退 back-off 478

反向传播 backprop 153, 176, 181, 182, 185, 187, 188, 384, 385

通过时间反向传播 back-propagation through time 326–328, 586

反向传播 backward propagation 257–259, 271, 326, 328, 329, 345–347, 354, 355, 358

词袋 bag of words 401

Bagging bootstrap aggregating 220–222, 224, 225, 229

bandit bandit 409

批量 batch viii, 237–239, 251–253, 256, 261, 273

批标准化 batch normalization 230, 271–273, 362, 455, 456

贝叶斯误差 Bayes error 102, 103, 360

贝叶斯规则 Bayes' rule 63, 64, 119, 463, 465, 535

贝叶斯推断 Bayesian inference 87, 121, 122, 450

贝叶斯网络 Bayesian network 480, 483, 496, 566

贝叶斯概率 Bayesian probability 49

贝叶斯统计 Bayesian statistics 118

基准 bechmark 106, 360

信念网络 belief network 21, 480, 592, 603

Bernoulli 分布 Bernoulli distribution 56, 61, 157–159, 369, 435, 548, 570, 573, 593

基准 baseline 362, 363, 375

BFGS BFGS 270

偏置 bias in affine function 96, 199, 202, 243, 257, 260, 326, 334, 350–354, 371, 396, 408, 546, 547, 559, 564, 565, 567, 571, 572, 575, 579, 580, 582, 585

偏差 bias in statistics 197, 198, 265, 400

有偏 biased 240, 248

有偏重要采样 biased importance sampling 400, 505

偏差 bias 114

二元语法 bigram 393, 400

二元关系 binary relation 410

二值稀疏编码 binary sparse coding 546–551

比特 bit 66

块坐标下降 block coordinate descent 274

块吉布斯采样 block Gibbs Sampling 500, 510, 563

玻尔兹曼分布 Boltzmann distribution 485

玻尔兹曼机 Boltzmann Machine 248, 260, 293, 485, 486, 500, 513, 520, 559–561, 567, 571, 575, 576, 578, 579, 584–586, 596, 607

Boosting Boosting 222, 229

桥式采样 bridge sampling 533, 536, 537

广播 broadcasting 29

磨合 Burning-in 509, 518–520, 522, 523, 571, 573

变分法 calculus of variations 155, 156, 544, 545, 551, 554, 555

容量 capacity 98, 99, 101, 104, 106, 114, 215, 222, 237, 359, 364–367, 381, 382, 394, 401, 402, 430, 431, 433, 434, 436, 440, 441, 470, 523, 528

级联 cascade 382, 384

灾难遗忘 catastrophic forgetting 168

范畴分布 categorical distribution 56, 369

因果因子 causal factor 466, 470, 472, 473

因果模型 causal modeling 53

中心差分 centered difference 373

- 中心极限定理 central limit theorem 58, 504
- 链式法则 chain rule 53, 76, 525
- 混沌 chaos 258
- 弦 chord 492, 493
- 弦图 chordal graph 493
- 梯度截断 clip gradient 164
- 截断梯度 clipping the gradient 353
- 团 clique 482–486, 491–494, 496, 497, 539, 543, 565
- 团势能 clique potential 482, 484, 485
- 闭式解 closed form solution 206, 420, 422
- 级联 coalesced 379, 383
- 编码 code 429–431, 433–435, 445, 447, 448
- 协同过滤 collaborative filtering 407, 408
- 列 column 28
- 列空间 column space 33
- 共因 common cause 489
- 完全图 complete graph 491
- 复杂细胞 complex cell 312
- 计算图 computational graph 176, 247, 320–322, 328–330, 341, 355, 498, 576, 596, 603
- 计算机视觉 Computer Vision 218, 363, 377, 384–386, 389, 421, 470
- 概念漂移 concept drift 457, 458
- 条件计算 conditional computation 382
- 条件概率 conditional probability 52, 53, 64, 69, 524
- 条件独立的 conditionally independent 53, 418, 481, 487, 488, 492
- 共轭 conjugate 268
- 共轭方向 conjugate directions 267
- 共轭梯度 conjugate gradient 267–270
- 联结主义 connectionism 12, 13, 15, 16, 19, 377, 559
- 一致性 consistency 114
- 约束优化 constrained optimization 82, 83, 85, 220, 485

- 特定环境下的独立 context-specific independences 488
- contextual bandit** contextual bandit 409, 410
- 延拓法 continuation method 278, 279
- 收缩 contractive 346, 445–447
- 收缩自编码器 contractive autoencoder 434, 438, 440, 442, 445, 446, 606, 607
- 对比散度 contrastive divergence 248, 438, 519–523, 527, 529, 564, 565, 571, 574, 575, 580–582, 586, 608
- 凸优化 Convex optimization 82, 241–243, 261, 274
- 卷积 convolution 281, 282, 450, 499
- 卷积玻尔兹曼机 Convolutional Boltzmann Machine 293
- 卷积玻尔兹曼机 convolutional Boltzmann machine 584
- 卷积网络 convolutional net 472
- 卷积网络 convolutional network 20, 21, 144, 175, 242, 246, 281, 282, 285, 287, 288, 290, 293–297, 299, 301, 302, 304, 306–314, 317–319, 337, 338, 360, 362, 363, 375, 379, 391, 395, 402, 403, 408, 456, 469, 470, 583, 584, 594, 601, 604
- 卷积神经网络 convolutional neural network 145, 218, 229, 281, 284, 285, 290, 295, 306
- 坐标上升 coordinate ascent 541, 543, 572
- 坐标下降 coordinate descent 274
- 共父 coparent 539, 547
- 相关系数 correlation 55
- 代价 cost 119, 134, 243–246, 248, 252, 257, 360, 361, 365, 370, 455, 506
- 代价函数 cost function 26, 74, 76, 78, 87, 104, 115, 116, 132–134, 152, 201, 203, 204, 208, 209, 214, 215, 231, 235–237, 242–249, 251, 252, 255, 269, 271, 272, 274, 275, 278, 279, 353, 360, 365, 375, 413, 421, 423, 433, 437, 438, 453, 465, 506, 524, 575, 588, 601
- 协方差 covariance 55, 60, 202, 220, 427
- 协方差矩阵 covariance matrix 55, 58, 60, 418, 427
- 协方差 **RBM** covariance RBM 580, 581
- 覆盖 coverage 361, 375
- 准则 criterion 74, 210, 251, 254, 256, 262–265, 267, 269, 322, 327, 345, 401, 435, 437–439, 446, 447, 575, 586, 594, 596, 608, 610
- 临界点 critical point 74–77, 79–82, 242–245, 249, 250, 266, 453, 551, 553

- 临界温度 critical temperatures 514
- 互相关函数 cross-correlation 283
- 交叉熵 cross-entropy 68, 116, 153–156, 189–191, 194, 330, 333, 396, 397
- 累积函数 cumulative function 504
- 课程学习 curriculum learning 279, 280, 327
- 维数灾难 curse of dimensionality 135, 136, 138, 394, 395, 468, 473, 603
- 曲率 curvature 78–81, 99, 201, 242, 253, 266
- 控制论 cybernetics 12, 13
- 衰减 damping 551
- 数据生成分布 data generating distribution 97, 236, 240, 241, 251
- 数据生成过程 data generating process 97, 449
- 数据并行 data parallelism 380
- 数据点 data point 92
- 数据集 dataset 87, 92–95, 97, 98, 101, 104, 106, 107, 113–115, 118, 119, 125, 128, 131, 133, 134, 141
- 数据集增强 dataset augmentation 386, 389
- 决策树 decision tree 125, 127, 382–384, 466
- 解码器 decoder 4, 338, 339, 402–404, 417, 420, 421, 423–427, 429–431, 434–436, 439–441, 447, 469, 595
- 分解 decompose 38
- 深度信念网络 deep belief network 17, 21, 310, 452, 472, 520, 536, 538, 562, 564–566, 568, 569, 572, 584, 591, 609
- 深度玻尔兹曼机 Deep Boltzmann Machine xiv, 20, 21, 452, 513, 520, 523, 526, 527, 538, 539, 551, 557, 562, 564, 566–577, 584, 609
- 深度回路 deep circuit 472
- 深度前馈网络 deep feedforward network 145, 147, 391, 417, 428
- 深度生成模型 deep generative model 452
- 深度学习 deep learning 1, 4–7, 10–15, 17, 18, 22–24, 26, 73, 74, 76, 79, 82, 87–89, 92, 93, 100, 105, 125, 128, 132, 133, 135–138, 141, 144, 197, 198, 210–212, 230, 235, 237, 239, 248, 251, 256, 261, 262, 266, 269, 270, 275, 345, 358, 362, 364, 371, 374, 377, 379, 381, 383–386, 390–392, 407, 408, 410, 412, 415, 416, 444, 448, 456, 458, 462, 466, 472, 474–476, 484, 496–499, 501, 506, 507, 510, 516, 518, 521, 526, 538, 539, 542, 543, 555

- 深度模型 deep model 93, 235, 236, 241, 243, 245, 257, 263, 277, 452, 522, 526
- 深度网络 deep network 144, 211, 258, 272, 278, 471
- 信任度 degree of belief 49
- 去噪 denoising 90, 92, 433, 437, 438, 440, 445, 476, 528
- 去噪自编码器 denoising autoencoder xv, 207, 433, 434, 436–440, 442, 445, 454, 457, 588, 606–611
- 去噪得分匹配 denoising score matching 438, 528
- 依赖 dependency 474, 476, 488, 492, 496
- 深度 depth 145
- 导数 derivative 74, 76, 77, 81, 86
- 描述 description 70
- 设计矩阵 design matrix 93–95, 129
- 细致平衡 detailed balance 608
- 探测级 detector stage 290
- 确定性 deterministic 238
- 对角矩阵 diagonal matrix 36
- 微分熵 differential entropy 67, 552
- 微分方程 differential equation 255
- 降维 dimensionality reduction 406, 429, 448
- Dirac delta 函数 Dirac delta function 59
- Dirac 分布 dirac distribution 59, 60, 528, 542, 543, 553, 554
- 有向 directed 69
- 有向图模型 directed graphical model 331, 334, 418, 462, 480–482, 491, 494, 495, 591, 603
- 有向模型 Directed Model 481, 482, 485, 488, 490–492, 495, 507, 538, 557, 565, 566, 594
- 方向导数 directional derivative 76, 77
- 判别 RBM discriminative RBM 453
- 判别器网络 discriminator network 597
- 分布式表示 distributed representation 16, 138, 228, 394–396, 404, 406–408, 410–412, 444, 449, 459, 466–471, 473, 498, 499
- 深度神经网络 DNN 247, 261, 262, 265, 271, 273, 381, 384, 391, 450–453, 471, 566
- 领域自适应 domain adaption 457

- 点积 dot product 30, 35, 123, 124
- 双反向传播 double backprop 233, 474
- 双重分块循环矩阵 doubly block circulant matrix 284, 307
- 降采样 downsampling 293, 298
- Dropout** Dropout 208, 222–230, 252, 257, 362, 364, 366, 367, 381, 383, 391, 455, 456, 574, 576, 588, 600
- Dropout Boosting** Dropout Boosting 229
- d-分离** d-separation 488, 490
- 动态规划 dynamic programming 188
- 动态结构 dynamic structure 382, 383
- 提前终止 early stopping 212–217, 237, 258, 362, 454, 455
- 回声状态网络 echo state network 21, 345–348
- 有效容量 effective capacity 100
- 特征分解 eigendecomposition 37–39
- 特征值 eigenvalue 37
- 特征向量 eigenvector 37
- 基本单位向量 elementary basis vectors 485
- 元素对应乘积 element-wise product 30
- 嵌入 embedding 442, 443
- 经验分布 empirical distribution 59, 60, 236, 238, 528
- 经验频率 empirical frequency 59
- 经验风险 empirical risk 236
- 经验风险最小化 empirical risk minimization 236, 237
- 编码器 encoder 4, 338, 339, 402–404, 421, 424–427, 429–432, 434–440, 442, 443, 445, 447, 451, 558, 595, 596
- 端到端的 end-to-end 359, 362, 363, 374, 392, 496
- 能量函数 energy function 485, 486, 499, 500, 511, 518, 559–561, 566, 567, 575, 578–583, 586
- 基于能量的模型 Energy-based model 485–487, 499, 506, 507, 510, 511, 513, 514, 559, 561, 566, 583
- 集成 ensemble 197, 220–223, 225–227, 229, 381, 402, 450

- 集成学习 ensemble learning 420
- 轮 epoch 242, 374
- 轮数 epochs 213
- 等式约束 equality constraint 83, 84
- 均衡分布 Equilibrium Distribution 508, 509
- 等变 equivariance 286
- 等变表示 equivariant representations 285
- 误差条 error bar 103
- 误差函数 error function 74
- 误差度量 error metric 359, 360
- 错误率 error rate 91, 360, 361, 366
- 估计量 estimator 108–115, 197, 456, 468, 520, 523
- 欧几里得范数 Euclidean norm 34
- 欧拉-拉格朗日方程 Euler-Lagrange Equation 552
- 证据下界 evidence lower bound 539, 540, 543, 544, 548, 565
- 样本 example 13, 23, 88, 90–95, 97, 99, 100, 102, 106, 107, 109, 110, 112–119, 123–125, 128, 129, 131–133, 135–138, 141, 210
- 额外误差 excess error 252, 256
- 期望 expectation 54, 56
- 期望最大化 expectation maximization 419, 541–544, 595
- E 步 expectation step 541
- 期望值 expected value 54
- 经验 experience, E 87, 88, 92, 94, 95
- 专家网络 expert network 383
- 相消解释 explaining away 538, 550, 565
- 相消解释作用 explaining away effect 489
- 解释因子 explanatory factort 463, 471, 473, 474
- 梯度爆炸 exploding gradient 248
- 利用 exploitation 409, 410
- 探索 exploration 409, 410

- 指数分布 exponential distribution 58
- 因子 factor 482–484, 486, 493, 494, 559, 585
- 因子分析 factor analysis 418, 420, 426
- 因子图 factor graph 493, 494
- 因子 factorial 417, 425, 426, 501, 544, 551, 562, 563, 568, 569
- 分解 factorization 69, 70
- 分解的 factorized 474
- 变差因素 factors of variation 4, 6, 173, 470, 472, 473
- 快速 Dropout fast dropout 228
- 快速持续性对比散度 fast persistent contrastive divergence 524
- 可行 feasible 83, 84, 86
- 特征 feature 88, 92–96, 98, 99, 104, 123–125, 128–131
- 特征提取器 feature extractor 422, 425, 453, 469, 543
- 特征映射 feature map 282, 389
- 特征选择 feature selection 204
- 反馈 feedback 145
- 前向 feedforward 145
- 前馈分类器 feedforward classifier 464
- 前馈网络 feedforward network 145–150, 156, 169, 171, 172, 174, 193–196, 245, 247, 248, 259, 276, 319, 321, 330, 334, 337, 344, 347, 361, 362, 429, 432, 434, 435, 437, 449, 450, 464, 465, 472, 474, 593
- 前馈神经网络 feedforward neural network 145–148, 151, 153, 165, 171, 175, 246, 434
- 现场可编程门阵列 field programmable gated array 384
- 精调 fine-tune 451, 452, 455, 520
- 精调 fine-tuning 275, 276, 425, 565
- 有限差分 finite difference 373
- 第一层 first layer 145
- 不动点方程 fixed point equation 545, 549, 550, 554, 556, 557, 569, 572
- 定点运算 fixed-point arithmetic 378
- 翻转 flip 283

- 浮点运算 float-point arithmetic 378
- 遗忘门 forget gate 350–352
- 前向模式累加 forward mode accumulation 192
- 前向传播 forward propagation 175, 182, 183, 257–259, 285, 301, 302, 309, 325, 326, 338, 346, 349
- 傅立叶变换 Fourier transform 308, 309
- 中央凹 fovea 313
- 自由能 free energy 487
- 频率派概率 frequentist probability 49
- 频率派统计 frequentist statistics 118
- Frobenius 范数** Frobenius norm 35, 41, 44, 45
- F 分数** F-score 361
- 全 full 297
- 泛函 functional 155, 551–555
- 泛函导数 functional derivative 551–554
- Gabor 函数** Gabor function 314–317
- Gamma 分布** Gamma distribution 581
- 门控 gated 349–352, 355
- 门控循环网络 gated recurrent net 362
- 门控循环单元 gated recurrent unit 349, 351, 362
- 门控 RNN gated RNN 349, 351
- 选通器 gater 383
- 高斯分布 Gaussian distribution xxvi, 57, 58, 60, 68, 154, 156, 162, 163, 165, 295, 418, 426, 554, 555, 578, 580, 581, 587, 588, 595, 600, 602
- 高斯核 Gaussian kernel 124, 466
- 高斯混合模型 Gaussian Mixture Model 60, 61, 390, 391, 496
- 高斯混合体 Gaussian mixtures 466
- 高斯输出分布 Gaussian output distribution 155
- 高斯 RBM Gaussian RBM 579–581
- Gaussian-Bernoulli RBM** Gaussian-Bernoulli RBM xiv, 578–580

- 通用 GPU general purpose GPU 379
- 泛化 generalization 97, 99, 136, 137, 146–149, 151, 172, 174, 194, 197, 198, 257, 277, 364, 381, 386, 389, 425, 457–459, 465, 468, 469, 472
- 泛化误差 generalization error 97, 100–102, 114, 236, 239–241, 250, 252, 257, 259, 261, 362, 364–367, 425
- 泛化 generalize 257, 457–459, 468–470, 473, 592, 603, 606
- 广义函数 generalized function 59
- 广义 Lagrange 函数 generalized Lagrange function 83, 84, 204
- 广义 Lagrangian generalized Lagrangian 83, 85
- 广义伪似然 generalized pseudolikelihood 525, 526, 576
- 广义伪似然估计 generalized pseudolikelihood estimator 525
- 广义得分匹配 generalized score matching 527, 528
- 生成式对抗框架 generative adversarial framework 465
- 生成式对抗网络 generative adversarial network 464, 465, 513, 531, 592, 597–601
- 生成模型 generative model 385, 417, 419, 420, 422, 425, 426, 428, 431–433, 440, 453, 464, 465, 470, 471, 498, 513, 515, 531, 537, 557–559, 564–566, 585, 587, 591, 592, 594, 596, 600, 611–614
- 生成式建模 generative modeling 594, 595, 597, 602, 610–613
- 生成矩匹配网络 generative moment matching network 600, 601
- 生成随机网络 generative stochastic network xv, 431, 607–611
- 生成器网络 generator network 592–595, 597, 599–601
- 吉布斯分布 Gibbs distribution 484
- Gibbs 采样 Gibbs Sampling 495, 499–501, 510–513, 515, 522, 527, 565, 568, 570, 573, 576, 581, 582
- 吉布斯步数 Gibbs steps 519, 521, 523, 573
- 全局对比度归一化 Global contrast normalization 387–389
- 全局极小值 global minima 245, 246
- 全局最小点 global minimum 75, 76, 82, 85, 243, 244, 249, 279
- 梯度 gradient 76–78, 82, 83, 85, 86, 199–201, 203, 205, 214, 215, 323, 326–330, 343, 344, 346, 347, 349, 352–358, 438, 439
- 梯度上升 gradient ascent 548

- 梯度截断 gradient clipping 246, 248, 258, 354, 355
- 梯度下降 gradient descent 74, 75, 77–83, 85, 123, 132–134, 205, 206, 215, 237, 238, 242, 245–247, 249, 251–255, 258, 259, 266, 272–274, 354, 365, 371, 380, 381, 405, 421, 429, 437, 447, 453, 470, 511, 541, 545, 549, 577, 589, 594
- 图模型 graphical model 69, 331–334, 396, 475, 476, 479, 481, 487, 488, 491, 494–499, 501, 538, 543, 544, 550, 551, 554, 559, 561, 562, 566, 567, 591, 603
- 图形处理器 Graphics Processing Unit 239, 378–380, 383, 384
- 贪心 greedy 451, 452
- 贪心算法 greedy algorithm 275, 451
- 贪心逐层预训练 greedy layer-wise pretraining 310, 572, 575, 576
- 贪心逐层训练 greedy layer-wise training 572
- 贪心逐层无监督预训练 greedy layer-wise unsupervised pretraining 450–452
- 贪心监督预训练 greedy supervised pretraining 275, 276
- 贪心无监督预训练 greedy unsupervised pretraining 452, 574
- 网格搜索 grid search 368–370
- Hadamard 乘积** Hadamard product xxv, 30
- 汉明距离 Hamming distance 528
- 硬专家混合体 hard mixture of experts 383
- 硬双曲正切函数 hard tanh 170
- 簧风琴 harmonium 499, 561
- 哈里斯链 Harris Chain 509
- Helmholtz 机** Helmholtz machine 431, 592
- Hessian** Hessian xxv, 78–82, 200, 201, 203, 204, 215, 239, 242, 244, 246, 248, 253, 266–268, 270, 271, 279, 352, 453, 454, 575
- 异方差 heteroscedastic 162
- 隐藏层 hidden layer 5, 13, 146–148, 150, 165, 171–173, 184, 188, 190, 195, 224, 271, 272, 275–278, 301, 324, 429, 434, 440, 446, 448, 449, 471, 472, 527, 538, 561–571, 573, 580, 591, 604, 606, 610
- 隐马尔可夫模型 Hidden Markov Model 390–392

- 隐藏单元 hidden unit vi, 5, 15, 16, 20, 21, 148, 154, 156, 165, 166, 168–172, 175, 190, 195, 206–208, 211, 215, 218, 220, 222–224, 226, 229, 230, 243, 257, 260, 273, 295, 300, 321, 323–327, 329, 332, 334, 335, 339, 345, 348–350, 352, 363, 365–368, 374, 382, 383, 387, 405, 421, 434, 437, 445, 446, 466, 469–472, 492, 496, 499, 500, 510, 517, 520, 522, 527, 539, 541, 543, 546, 547, 550, 560–562, 565–571, 575, 578–582, 584–586, 592, 594, 602, 604–606, 613
- 隐藏变量 hidden variable 526, 538
- 爬山 hill climbing 77
- 超参数 hyperparameter 253, 254, 259, 261–264, 359, 363–371, 375, 455
- 超参数优化 hyperparameter optimization 368
- 假设空间 hypothesis space 98
- 同分布的 identically distributed 97
- 可辨认的 identifiable 243
- 单位矩阵 identity matrix xxiii, 31
- 独立同分布假设 i.i.d. assumption 97
- 病态 ill conditioning 242
- 不道德 immorality 491, 492
- 重要采样 Importance Sampling 400, 401, 504–506, 532–536, 592, 596
- 相互独立的 independent 53, 97
- 独立成分分析 independent component analysis 418–422
- 独立同分布 independent identically distributed 503, 531
- 独立子空间分析 independent subspace analysis 421
- 索引 index of matrix 27, 28
- 指示函数 indicator function 58
- 不等式约束 inequality constraint 83–85
- 推断 inference xiv, 2, 208, 225, 227–229, 393, 394, 415, 431, 432, 497, 542, 559, 560, 565, 567–571, 573–577, 582, 586, 592–596, 598, 600, 605, 606, 613
- 无限 infinite 456
- 信息检索 information retrieval 448
- 内积 inner product 123
- 输入 input 282, 453

- 输入分布 input distribution 453, 454, 457
- 干预查询 intervention query 53
- 不变 invariant 291
- 求逆 invert 579
- Isomap** Isomap 456
- 各向同性 isotropic 58, 61
- Jacobian** Jacobian xxv, 77, 78, 176, 178, 180, 185–187, 233, 278, 329, 343, 345, 346, 373, 421, 445, 446
- Jacobian 矩阵** Jacobian matrix 65, 178, 192
- 联合概率分布 joint probability distribution 50, 52, 53, 69, 559–561, 566
- Karush-Kuhn-Tucker** Karush-Kuhn-Tucker 83–85, 204, 206
- 核函数 kernel function 123, 282
- 核机器 kernel machine 124, 125, 146, 210, 345, 466, 564
- 核方法 kernel method 124
- 核技巧 kernel trick 123, 124, 133, 146
- KL 散度** KL divergence 116, 219, 539, 545
- 知识库 knowledge base 2, 411, 412
- 知识图谱 knowledge graph 412
- Krylov 方法** Krylov method 193
- KL 散度** Kullback-Leibler (KL) divergence xxvi, 67, 68
- 标签 label 92, 94, 124, 136, 453, 459, 470, 472
- 标注 labeled 363, 364, 375, 450, 454, 456, 458, 459, 461, 462
- 拉格朗日乘子 Lagrange multiplier 552, 553
- 语言模型 language model 355, 392–394, 402, 403, 406, 410, 506
- Laplace 分布** Laplace distribution 58
- 大学习步骤 large learning step 544
- 潜在 latent 163, 418, 419, 426, 431, 451, 463, 496, 522, 560, 561, 597, 599, 602, 609
- 潜层 latent layer 561

- 潜变量** latent variable xiii, 60, 163, 243, 396, 417–419, 429, 431–433, 435, 452, 462, 466, 472, 486, 487, 496–499, 501, 512, 514, 517, 521, 527, 538, 539, 541, 542, 544, 545, 548, 554, 560–562, 564–567, 576, 592, 594–596, 607, 609
- 大数定理** Law of large number 503
- 逐层的** layer-wise 451
- L-BFGS** L-BFGS 270, 271
- 渗漏整流线性单元** Leaky ReLU 167, 362
- 渗漏单元** leaky unit 347–349
- 学成** learned 450, 454, 458, 459, 465, 467, 470, 473, 474, 557, 558, 592
- 学习近似推断** learned approximate inference 447
- 学习器** learner 106, 138, 240, 457, 459, 463, 469, 472, 473
- 学习率** learning rate 77, 79, 133, 235, 239, 242, 251, 252, 254, 256, 261–264, 266, 268, 271, 362, 363, 365–368, 372, 523, 524, 573, 589, 591
- 勒贝格可积** Lebesgue-integrable 517
- 左特征向量** left eigenvector 37
- 左奇异向量** left singular vector 40
- 莱布尼兹法则** Leibniz's rule 517
- 似然** likelihood 49
- 线搜索** line search 77, 83, 269
- 线性自回归网络** linear auto-regressive network 602
- 线性分类器** linear classifier 237, 428, 449, 453, 458, 467, 470
- 线性组合** linear combination 33
- 线性相关** linear dependence 33
- 线性因子模型** linear factor model 417, 418, 420, 421, 423, 425, 426, 428, 501, 543, 579
- 线性模型** linear model 14, 198, 203, 204, 206, 215, 228, 231, 560, 602
- 线性回归** linear regression 87, 94, 96–98, 100, 101, 104, 108, 117–119, 121–123, 133, 134, 198, 200–203, 205, 206, 219, 228, 260, 345, 428, 544, 602
- 线性阈值单元** linear threshold units 469, 470
- 线性无关** linearly independent 33
- 链接预测** link prediction 412
- 链接重要采样** linked importance sampling 537

- Lipschitz** Lipschitz 82
- Lipschitz 常数** Lipschitz constant 82
- Lipschitz 连续** Lipschitz continuous 82
- 流体状态机** liquid state machine 345
- 局部条件概率分布** local conditional probability distribution 480
- 局部不变性先验** local constancy prior 136
- 局部对比度归一化** local contrast normalization 388, 389
- 局部下降** local descent 250
- 局部核** local kernel 137, 466
- 局部极大值** local maxima 127, 245
- 局部极大点** local maximum 74, 75, 79, 80, 244, 549
- 局部极小值** local minima 243–245, 249, 279, 453
- 局部极小点** local minimum 74–76, 79, 80, 82, 213, 214, 237, 243, 244, 249, 255, 453
- 对数尺度** logarithmic scale 368, 369
- 逻辑回归** logistic regression 2, 3, 6, 123, 146, 153, 155, 177, 198, 206, 231, 310, 362, 367, 397, 530, 560, 600, 602
- logistic sigmoid** logistic sigmoid vi, 61, 62, 122, 157, 159, 168, 171
- 分对数** logit 63, 158
- 对数线性模型** log-linear model 486
- 长短期记忆** long short-term memory ix, 16, 22, 260, 278, 349–353, 355, 356, 358, 362, 392
- 长期依赖** long-term dependency 247, 341, 343–345, 347, 348, 351, 355
- 环** loop 492, 493
- 环状信念传播** loopy belief propagation 498, 499
- 损失** loss 91, 116, 132, 528, 576
- 损失函数** loss function 74, 107, 134, 219, 236, 237, 245, 248, 249, 253, 278, 325–327, 355, 365, 396, 401, 422, 430, 431, 433, 435, 447, 585, 587, 602
- 机器学习** machine learning 2–4, 7, 10, 12–18, 20, 24, 26, 72, 86–95, 97–100, 102, 104, 105, 108, 112, 113, 118, 119, 123, 126, 132, 134, 135, 138, 139, 141, 197, 204, 206–208, 220, 222, 232, 234–237, 240, 241, 251, 252, 260, 279, 319, 353, 359–364, 371, 372, 374, 377, 378, 380–382, 401, 407, 408, 410, 411, 429, 440, 443, 449, 453, 458, 473, 474, 476, 486, 490, 496, 498, 502, 506, 518, 519, 542, 551, 552, 557

- 机器学习模型 machine learning model 452
- 机器翻译 machine translation 362, 459
- 主对角线 main diagonal 29
- 流形 manifold 139, 141, 142, 233, 426, 427, 438–446, 473, 474, 496, 511, 513, 597
- 流形假设 manifold hypothesis 140
- 流形学习 manifold learning 139, 434, 442–444, 597
- 边缘概率分布 marginal probability distribution 52
- 马尔可夫链 Markov Chain xv, 506–514, 518–524, 527, 534, 566, 569, 571, 573, 607–610
- 马尔可夫链蒙特卡罗 Markov Chain Monte Carlo 415, 504, 506, 507, 509, 511, 513, 518–520, 524, 528, 534, 563, 569, 575, 606, 609–611
- 马尔可夫网络 Markov network 482, 486, 496, 500
- 马尔可夫随机场 Markov random field 482, 486
- 掩码 mask 222–225, 228, 229
- 矩阵 matrix 28
- 矩阵逆 matrix inversion 31, 32
- 矩阵乘积 matrix product 29
- 最大范数 max norm 35
- 池 pool 291, 293, 294
- 最大池化 max pooling 290–293, 301, 469, 602
- 极大值 maxima 244, 245
- M 步 maximization step 541, 542
- 最大后验 Maximum A Posteriori v, 121, 122, 204, 392, 432, 542–544, 558, 582
- 最大似然 maximum likelihood 420, 424, 516, 545, 546
- 最大似然估计 maximum likelihood estimation 115–119, 121, 122, 134, 238, 393, 520, 525, 529, 543, 545
- 最大平均偏差 maximum mean discrepancy 601
- maxout** maxout 213, 243, 259, 278, 292, 317, 362
- maxout** 单元 maxout unit 167, 168, 172, 317, 365
- 平均绝对误差 mean absolute error 156
- 均值和协方差 **RBM** mean and covariance RBM 580–583
- 学生 t 分布均值乘积 mean product of Student t -distribution 580–583

- 均方误差 mean squared error 95, 96, 103, 104, 113, 116–118, 120, 129, 148, 154–156, 158, 194, 195, 345, 422, 430, 435, 437, 464, 465, 590, 595, 608
- 均值-协方差 **RBM** mean-covariance restricted Boltzmann machine 486
- 均匀场 meanfield 21, 568–570, 572–574, 576, 577, 584, 592, 596, 605
- 均值场 mean-field 544–551, 554, 557, 558
- 测度论 measure theory 64
- 零测度 measure zero 64
- 记忆网络 memory network 356, 358, 412
- 信息传输 message passing 551
- 小批量 minibatch viii, 132, 183, 189, 190, 221–223, 237–241, 248, 251–254, 256, 259, 261–265, 270, 272, 320, 353, 354, 374, 380, 383, 429, 436, 453, 502, 509, 519, 521, 523, 541, 573, 577
- 小批量随机 minibatch stochastic 239
- 极小值 minima 245, 249
- 极小点 minimum 250, 251, 553
- 混合 Mixing 511–515, 521–524
- 混合时间 Mixing Time 509, 510
- 混合密度网络 mixture density network 163
- 混合分布 mixture distribution 59
- 专家混合体 mixture of experts 383, 466
- 模态 modality 460
- 峰值 mode xiii, 511–515, 520, 522–524, 551
- 模型 model 452
- 模型平均 model averaging 220–222
- 模型压缩 model compression 381
- 模型可辨识性 model identifiability 243
- 模型并行 model parallelism 380
- 矩 moment 600, 601, 611
- 矩匹配 moment matching 600, 611
- 动量 momentum 253–256, 261, 263, 264, 277, 362
- 蒙特卡罗 Monte Carlo 227, 400, 502–504, 506, 515, 518, 524, 532, 557, 581, 589, 595
- Moore-Penrose 伪逆** Moore-Penrose pseudoinverse xxv, 41, 99, 105

- 道德化 moralization 491, 492
- 道德图 moralized graph 491, 492
- 多层感知机 multilayer perceptron 5, 20, 21, 145, 188, 189, 194, 275, 276, 298, 340, 341, 403, 440, 471, 560, 565, 566, 568, 569, 574, 575, 586
- 多峰值 multimodal 533, 550, 611
- 多模态学习 multimodal learning 460
- 多项式分布 multinomial distribution 56
- Multinoulli** 分布 multinoulli distribution 56, 59, 60, 73, 159, 163
- 多预测深度玻尔兹曼机 multi-prediction deep Boltzmann machine 575–577, 596, 607
- 多任务学习 multitask learning 210, 211, 457, 458
- 多维正态分布 multivariate normal distribution 58, 418, 512
- 朴素贝叶斯 naive Bayes 2
- 奈特 nats 66
- 自然语言处理 Natural Language Processing 246, 363, 377, 392, 395, 396, 406, 407, 410, 455
- 最近邻 nearest neighbor 137, 450, 466–468
- 最近邻图 nearest neighbor graph 443
- 最近邻回归 nearest neighbor regression 101, 125
- 负定 negative definite 38
- 负部函数 negative part function 63
- 负相 negative phase 517–520, 522–524, 526, 527, 557, 561, 571, 572
- 半负定 negative semidefinite 38
- Nesterov** 动量 Nesterov momentum 256
- 网络 network 145
- 神经自回归密度估计器 neural auto-regressive density estimator xiv, 602, 604–606
- 神经自回归网络 neural auto-regressive network 603–606
- 神经语言模型 Neural Language Model 394, 396, 397, 399, 401, 402, 406, 411
- 神经机器翻译 Neural Machine Translation 395
- 神经网络 neural network 12–17, 19–23, 197–199, 205–207, 215, 218, 221, 222, 225, 229–232, 234, 235, 241–250, 257, 258, 261, 262, 266, 267, 269, 270, 273–275, 277–280, 319, 341, 349, 356, 358, 377–379, 384, 387, 390–392, 395, 396, 401, 402, 405, 406, 408, 411, 429, 444, 447, 452–455, 466, 470, 506, 556, 587

- 神经网络图灵机 neural Turing machine 356, 357
- 牛顿法 Newton's method 81, 82, 85, 242, 243, 245, 250, 266–268, 270, 274
- n -gram n -gram 393, 394, 396, 397, 401–403, 467, 478
- 没有免费午餐定理 no free lunch theorem 102, 105, 472
- 噪声 noise 101, 140, 239, 248, 253, 279, 362, 363, 453, 528–531
- 噪声分布 noise distribution 529–531
- 噪声对比估计 noise-contrastive estimation 529–531
- 非凸 nonconvex 241, 243–246, 262, 266, 275, 279
- 非分布式 nondistributed 467–469
- 非分布式表示 nondistributed representation 466–468
- 非线性共轭梯度 nonlinear conjugate gradients 269, 270
- 非线性独立成分估计 nonlinear independent components estimation 420, 421
- 非参数 non-parametric 100, 394, 442–444
- 范数 norm 34
- 正态分布 normal distribution 57, 58, 61, 504, 553
- 正规方程 normal equation 96, 98, 99, 133, 148
- 归一化的 normalized 51
- 标准初始化 normalized initialization 258
- 数值 numeric value 182
- 数值优化 numerical optimization 235, 242, 246
- 对象识别 object recognition 246, 362, 364, 385, 389, 390, 423, 425, 459, 612
- 目标 objective 455
- 目标函数 objective function 74, 77, 84, 197–202, 204, 205, 213, 214, 217, 221, 236–238, 241, 246–248, 250, 252, 253, 265–267, 269, 274, 278, 279, 353, 359, 368, 374, 450, 470, 525, 527, 564, 572
- 奥卡姆剃刀 Occam's razor 100
- one-hot** one-hot 125, 131, 161, 193, 394, 395, 454, 455, 459, 466, 468, 586, 603
- 一次学习 one-shot learning 459
- 在线 online 238
- 在线学习 online learning 240

- 操作 operation 176
- 最佳容量 optimal capacity 101, 103, 114
- 原点 origin 33
- 正交 orthogonal 36
- 正交矩阵 orthogonal matrix 37
- 标准正交 orthonormal 36, 39
- 输出 output 453
- 输出层 output layer 145
- 过完备 overcomplete 431, 434, 582, 583
- 过估计 overestimation 506
- 过拟合 overfitting 98, 99, 105, 114, 197, 198, 215, 237, 241, 252, 258, 359, 363, 365, 366, 372, 375, 381, 450, 454, 455, 478, 613
- 过拟合机制 overfitting regime 101
- 上溢 overflow 72, 73, 535
- 并行分布式处理 Parallel Distributed Processing 194
- 并行回火 parallel tempering 514, 524, 537
- 参数 parameter 94
- 参数服务器 parameter server 381
- 参数共享 parameter sharing 218, 225, 229, 285, 286, 288, 300, 313, 319, 320, 322, 323, 332, 333, 402, 601, 602, 604
- 有参情况 parametric case 118
- 参数化整流线性单元 parametric ReLU 167, 362
- 偏导数 partial derivative 76, 77, 445, 551
- 配分函数 Partition Function 415, 484, 486, 502, 506, 515, 516, 518, 519, 524, 525, 527–529, 531–537, 557, 559–561, 564, 565, 571, 578, 583, 584, 598
- 性能度量 performance measures 87, 88, 91, 95, 361, 362
- 性能度量 performance metrics 359, 360, 362, 370, 372, 374, 375
- 置换不变性 permutation invariant 296
- 持续性对比散度 persistent contrastive divergence 521, 523, 564, 572, 575, 581, 582
- 音素 phoneme 390–392, 457

- 语音 phonetic 392
- 分段 piecewise 362
- 点估计 point estimator 108
- 策略 policy 409, 410
- 策略梯度 policy gradient 383
- 池化 pooling 207, 229, 281, 287, 290–295, 299, 306, 309, 310, 312, 313, 386, 421
- 池化函数 pooling function 290
- 病态条件 poor conditioning 74, 81, 239, 242, 246, 248, 250, 253, 454
- 正定 positive definite 38
- 正部函数 positive part function 63
- 正相 positive phase 517–520, 523, 524, 557, 560, 571
- 半正定 positive semidefinite 38
- 后验概率 posterior probability 60
- 幂方法 power method 248
- PR** 曲线 PR curve 361
- 精度 precision 57, 361, 373, 612
- 精度矩阵 precision matrix 58
- 预测稀疏分解 predictive sparse decomposition 447
- 预训练 pretraining 275–278, 391, 425, 451–456, 498, 521, 527
- 初级视觉皮层 primary visual cortex 311
- 主成分分析 principal components analysis xi, 42–44, 128–130, 134, 210, 235, 302, 388, 418–420, 422, 424, 426, 427, 430, 441, 446, 448
- 先验概率 prior probability 60
- 先验概率分布 prior probability distribution 118, 295
- 概率 **PCA** probabilistic PCA 418–420, 426, 538, 539
- 概率密度函数 probability density function 51, 52, 57–59, 64, 503, 551–553, 598
- 概率分布 probability distribution 47, 50–56, 58–61, 66, 67, 69, 70, 360, 472, 516, 529, 531
- 概率质量函数 probability mass function 50, 51, 90, 560, 571
- 专家之积 product of expert 486
- 乘法法则 product rule 53

- 成比例 proportional 70
- 提议分布 proposal distribution 400, 532, 534–536
- 伪似然 pseudolikelihood 524–530, 571
- 象限对 quadrature pair 316
- 量子力学 quantum mechanics 48
- 径向基函数 radial basis function 124, 146, 170, 471
- 随机搜索 random search 369–371
- 随机变量 random variable 49–56, 58–60, 64, 65, 67, 69, 70, 472, 525, 530, 534
- 值域 range 33
- 比率匹配 ratio matching 527, 528, 564
- 召回率 recall 361, 382, 612
- 接受域 receptive field 287, 295
- 再循环 recirculation 429
- 推荐系统 recommender system 407–409
- 重构 reconstruction 429, 430, 436–439, 441, 442, 445–447, 608, 609
- 重构误差 reconstruction error 419, 422, 426, 427, 431, 433, 437, 438, 440, 445, 446, 448, 454, 514, 608
- 整流线性 rectified linear 151, 167, 230, 243, 273, 290
- 整流线性变换 rectified linear transformation 152
- 整流线性单元 rectified linear unit 14, 15, 150, 151, 165–168, 170–172, 177, 195, 233, 278, 362, 375, 391, 433, 455
- 整流网络 rectifier network 172, 173, 195
- 循环 recurrence 450
- 循环卷积网络 recurrent convolutional network 307
- 循环网络 recurrent network 145, 246–248, 307, 319–324, 326, 330, 333, 338, 341, 343–347, 349, 350, 353, 354, 357, 412, 417, 440, 474, 550, 576, 577
- 循环神经网络 recurrent neural network ix, 21, 22, 144, 145, 208, 228, 247, 306, 318–325, 328, 330–341, 343–345, 348, 349, 352, 355, 358, 392, 403, 550, 585, 586, 596
- 回归 regression 103
- 正则化 regularization 104, 105, 118, 122, 197–206, 208, 209, 212–220, 222, 227–236, 258, 355, 359, 362, 364–366, 387–389, 422, 431, 432, 434, 438, 440, 446, 453, 455, 472

- 正则化 regularize 239, 365, 421, 422, 455, 456, 514, 528, 575, 584, 588
- 正则化项 regularizer 104, 122, 126, 134, 362, 452, 454, 455, 467
- 强化学习 reinforcement learning 23, 93, 232, 383, 409, 410, 458, 557, 588, 590
- 关系 relation 410–412
- 关系型数据库 relational database 411
- 重参数化 reparametrization 575, 588
- 重参数化技巧 reparametrization trick 588, 594, 610
- 表示 representation 2–7, 16, 210, 219, 220, 297, 357, 367, 394, 395, 403, 404, 411, 430, 431, 433, 440–442, 448
- 表示学习 representation learning 4, 403, 417, 419, 448–450, 452, 457, 458, 461–463, 466, 472–474, 501, 514
- 表示容量 representational capacity 100, 104
- 储层计算 reservoir computing 345
- 受限玻尔兹曼机 Restricted Boltzmann Machine 228, 301, 391, 408, 437, 438, 448, 450, 472, 490, 499–501, 510, 514, 515, 517, 519–523, 533, 536–538, 561–568, 571, 572, 574, 575, 578, 579, 581, 583, 585, 586, 591, 600, 605, 609, 610
- 反向相关 reverse correlation 314
- 反向模式累加 reverse mode accumulation 191
- 岭回归 ridge regression 199
- 右特征向量 right eigenvector 37
- 右奇异向量 right singular vector 40
- 风险 risk 236
- 行 row 28
- 扫视 saccade 313
- 鞍点 saddle point 75, 76, 79, 80, 82, 244–246, 248, 249, 266, 267
- 无鞍牛顿法 saddle-free Newton method 245
- 相同 same 297, 298
- 样本均值 sample mean 110
- 样本方差 sample variance 110, 111
- 饱和 saturate 61
- 标量 scalar 27

- 得分 score 437–440, 526, 527
- 得分匹配 score matching 437, 438, 445, 526–530, 606
- 二阶导数 second derivative 77–80
- 二阶导数测试 second derivative test 80
- 第二层 second layer 145
- 二阶方法 second-order method 245
- 自对比估计 self-contrastive estimation 531
- 自信息 self-information 66
- 语义哈希 semantic hashing 448
- 半受限玻尔兹曼机 semi-restricted Boltzmann Machine 539
- 半监督 semi-supervised 363, 415
- 半监督学习 semi-supervised learning 209, 210, 231, 450, 452, 454, 462, 463, 473
- 可分离的 separable 309, 449, 453
- 分离的 separate 473
- 分离 separation 487, 488, 495
- 情景 setting 458, 459, 469, 471
- 浅度回路 shadow circuit 472
- 香农熵 Shannon entropy xxvi, 66, 67
- 香农 shannons 66
- 塑造 shaping 279, 560, 611
- 短列表 shortlist 396, 397
- sigmoid** sigmoid 157–162, 168, 169, 195, 278, 362, 425, 511
- sigmoid 信念网络** sigmoid Belief Network 591, 592
- 简单细胞 simple cell 311
- 奇异的 singular 34
- 奇异值 singular value 39, 40
- 奇异值分解 singular value decomposition 39–41, 130, 408
- 奇异向量 singular vector 39
- 跳跃连接 skip connection 340, 341, 347, 348
- 慢特征分析 slow feature analysis 421–423, 474

- 慢性原则 slowness principle 421–423
- 平滑 smoothing 394
- 平滑先验 smoothness prior 136
- softmax** softmax 449
- softmax 函数** softmax function 72, 73, 209, 226, 227, 325, 328, 372, 375, 383
- softmax 单元** softmax unit 375
- softplus** softplus 170
- softplus 函数** softplus function 61–63, 158, 170
- 生成子空间 span 33
- 稀疏 sparse 203, 204, 218–220, 227, 431–434, 440
- 稀疏激活 sparse activation 195
- 稀疏编码 sparse coding 274, 423–426, 432, 440, 447, 451, 490, 492, 496, 501, 527, 538, 543, 544, 551, 558, 582, 583, 591
- 稀疏连接 sparse connectivity 285–287
- 稀疏初始化 sparse initialization 259
- 稀疏交互 sparse interactions 285
- 稀疏权重 sparse weights 285
- 谱半径 spectral radius 345–347
- 语音识别 Speech Recognition 362, 377, 381, 390–392, 457
- sphering** sphering 388
- 尖峰和平板 spike and slab 317, 425, 426
- 尖峰和平板 **RBM** spike and slab RBM 580–583
- 虚假模态 spurious modes 520, 522
- 方阵 square 34
- 标准差 standard deviation 54, 112, 238, 272, 273, 386–389
- 标准差 standard error 57, 111, 112, 238
- 标准正态分布 standard normal distribution 57
- 声明 statement 47, 48
- 平稳的 stationary 333
- 平稳分布 Stationary Distribution 508–510, 512

- 驻点 stationary point 74, 84
- 统计效率 statistic efficiency 118
- 统计学习理论 statistical learning theory 97
- 统计量 statistics 108
- 最陡下降 steepest descent 247
- 随机 stochastic 238, 239
- 随机课程 stochastic curriculum 280
- 随机梯度上升 Stochastic Gradient Ascent 541
- 随机梯度下降 stochastic gradient descent 14, 87, 132, 133, 205, 206, 216, 222, 228, 238–242, 246, 251–254, 256, 258, 270, 277, 344, 353, 354, 356, 362, 380, 437, 506, 518, 574, 575, 589, 606
- 随机矩阵 Stochastic Matrix 508
- 随机最大似然 stochastic maximum likelihood 521–524, 526, 528, 529, 564, 565, 568, 571–574, 576
- 流 stream 240
- 步幅 stride 287, 291, 293, 294, 297, 298, 301, 302, 306
- 结构学习 structure learning 496, 498
- 结构化概率模型 structured probabilistic model 47, 69, 70, 472, 475, 477, 479–482, 495, 498, 559
- 结构化变分推断 structured variational inference 544
- 亚原子 subatomic 48
- 子采样 subsample 502
- 求和法则 sum rule 52
- 和-积网络 sum-product network 472
- 监督 supervised 92, 210, 211, 218, 231, 236, 310, 311, 317, 379, 425, 440, 449–453, 455, 557, 584
- 监督学习 supervised learning xxvii, 87, 92–94, 101, 107, 116, 122, 123, 125, 126, 134, 140, 144, 210, 232, 236, 342, 362, 397, 407, 409, 410, 415, 432, 449, 450, 452, 453, 455–458, 462, 463, 472, 529, 594
- 监督学习算法 supervised learning algorithm 92
- 监督模型 supervised model 453
- 监督预训练 supervised pretraining 456
- 支持向量 support vector 124, 466

- 代理损失函数 surrogate loss function 237, 248
- 符号 symbol 182
- 符号表示 symbolic representation 182, 466, 468
- 对称 symmetric 36
- 切面距离 tangent distance 232
- 切平面 tangent plane 440, 443, 446
- 正切传播 tangent prop 232–234
- 目标 target 92–95, 101, 102, 105, 108, 116, 122, 128, 134, 135, 137, 138, 141
- 泰勒 taylor 79, 81, 203, 215, 242
- 导师驱动过程 teacher forcing 327, 328
- 温度 temperature 514
- 回火转移 tempered transition 514
- 回火 tempering 514
- 张量 tensor 28
- 测试误差 test error 97, 98, 101, 103, 241, 363, 365, 366, 371, 372, 375, 452, 454, 455
- 测试集 test set 91, 95, 97, 98, 106, 107, 112, 235, 237, 252, 277, 363, 364, 366, 372, 375, 454
- 碰撞情况 the collider case 489
- 绑定的权重 tied weights 285
- Tikhonov 正则** Tikhonov regularization 199
- 平铺卷积 tiled convolution 300, 301, 303, 305
- 时延神经网络 time delay neural network 314, 319, 391
- 时间步 time step 168, 247, 248, 265, 319–335, 339–341, 343, 346, 348–350, 352–354, 357, 392, 404, 405, 423, 577, 585, 605, 609, 610
- Toeplitz 矩阵** Toeplitz matrix 284
- 标记 token 392, 393, 411
- 容差 tolerance 85, 549
- 地质 **ICA** topographic ICA 421
- 训练误差 training error 97, 98, 100–103, 236, 241, 364–366, 372, 375, 454
- 训练集 training set 97, 98, 235–241, 243, 249, 251, 252, 254, 256, 260, 262–265, 267, 269, 274, 277, 280, 360, 362–364, 366, 367, 372, 373, 375, 462, 464, 468

- 转录 transcribe 89, 91, 94
- 转录系统 transcription system 359, 361, 372, 374, 375
- 迁移学习 transfer learning 454, 456–461, 604
- 转移 transition 322
- 转置 transpose 29
- 三角不等式 triangle inequality 34
- 三角形化 triangulate 493
- 三角形化图 triangulated graph 493
- 三元语法 trigram 393
- 无偏 unbiased 109, 240, 241, 251, 503–505, 528
- 无偏样本方差 unbiased sample variance 111
- 欠完备 undercomplete 430, 431
- 欠定的 underdetermined 552
- 欠估计 underestimation 506
- 欠拟合 underfitting 98, 99, 105, 114, 197–199, 241, 295, 359, 365, 366, 372, 373, 375, 598, 613
- 欠拟合机制 underfitting regime 101
- 下溢 underflow 72, 73
- 潜在 underlying 236, 237, 462–466, 470–474
- 潜在成因 underlying cause 461, 463, 473
- 无向 undirected 69
- 无向模型 undirected Model 482–488, 490–493, 495, 500, 502, 510, 515–518, 538, 557, 564, 565, 591
- 展开图 unfolded graph 322, 323, 326, 392
- 展开 unfolding 320–322, 340, 392
- 均匀分布 uniform distribution 51, 52, 55, 67, 165, 456
- 一元语法 unigram 393, 400
- 单峰值 unimodal 514, 556
- 单元 unit 146
- 单位范数 unit norm 36, 43
- 单位向量 unit vector 36

- 万能近似定理 universal approximation theorem 171, 172, 434
- 万能近似器 universal approximator 60, 471, 472, 560
- 万能函数近似器 universal function approximator 151
- 未标注 unlabeled 450, 454, 455, 459, 461, 463, 472
- 未归一化概率函数 unnormalized probability function 483, 484, 486, 493
- 非共享卷积 unshared convolution 299
- 无监督 unsupervised 20, 21, 92, 210, 218, 228, 363, 391, 415, 423, 425, 440, 447, 449–453, 455, 458, 459, 462, 463
- 无监督学习 unsupervised learning 87, 92–94, 107, 128, 134, 207, 210, 211, 234, 236, 363, 391, 415, 432, 443, 450–455, 457, 458, 462–464, 529, 610
- 无监督学习算法 unsupervised learning algorithm 92
- 无监督预训练 unsupervised pretraining 450, 452–457
- 有效 valid 284, 297, 298
- 验证集 validation set 106, 237, 242, 259, 368–370, 455
- 梯度消失与爆炸问题 vanishing and exploding gradient problem 247, 248, 259
- 梯度消失 vanishing gradient 248
- Vapnik-Chervonenkis 维度** Vapnik-Chervonenkis dimension 100, 467, 470
- 变量消去 variable elimination 547
- 方差 variance 54, 56, 57, 111, 197–199, 202, 206, 220
- 方差减小 variance reduction 589, 590
- 变分自编码器 variational auto-encoder 195, 431, 506, 558, 592, 594–597, 600, 606
- 变分导数 variational derivative 551
- 变分自由能 variational free energy 539
- 变分推断 variational inference 497, 499, 526
- 去噪 denoise 128, 386
- 向量 vector 27
- 虚拟对抗样本 virtual adversarial example 231
- 虚拟对抗训练 virtual adversarial training 452
- 可见层 visible layer 5
- V-结构** V-structure 489, 539

醒眠 wake sleep 557, 565, 592

warp warp 380, 383

支持向量机 support vector machine 123–125, 153, 310, 367, 522

无向图模型 undirected graphical model 516, 531

权重 weight 94

权重衰减 weight decay 104–106, 199–202, 205, 206, 209, 213, 215, 217, 218, 227, 228, 243, 258, 274, 364–367, 432, 454, 524, 544

权重比例推断规则 weight scaling inference rule 226–229

权重空间对称性 weight space symmetry 243

条件概率分布 conditional probability distribution 534

白化 whitening 388

宽度 width 146

赢者通吃 winner-take-all 161

正切传播 tangent propagation 474

流形正切分类器 manifold tangent classifier 474

词嵌入 word embedding 363, 395, 404, 406, 408, 454, 459

词义消歧 word-sense disambiguation 412

零数据学习 zero-data learning 459, 461

零次学习 zero-shot learning 459–461