References

1. Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, and Zhenyao Zhu, *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin*, arXiv:1512.02595 [cs] (2015) (en), arXiv: 1512.02595.
2. Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber, *Common Voice: A Massively-Multilingual Speech Corpus*, arXiv:1912.06670 [cs] (2020) (en), arXiv: 1912.06670.
3. Martin Arjovsky, Amar Shah, and Yoshua Bengio, *Unitary Evolution Recurrent Neural Networks*, arXiv:1511.06464 [cs, stat] (2016) (en), arXiv: 1511.06464.
4. David Arthur and Sergei Vassilvitskii, *K-means++: the advantages of careful seeding*, In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, 2007.
5. Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton, *Layer Normalization*, arXiv:1607.06450 [cs, stat] (2016) (en), arXiv: 1607.06450.
6. Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio, *Neural machine translation by jointly learning to align and translate*, January 2015, 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015 (English (US)).
7. Eric Battenberg, Jitong Chen, Rewon Child, Adam Coates, Yashesh Gaur, Yi Li, Hairong Liu, Sanjeev Satheesh, David Seetapun, Anuroop Sriram, and Zhenyao Zhu, *Exploring Neural Transducers for End-to-End Speech Recognition*, arXiv:1707.07413 [cs] (2017) (en), arXiv: 1707.07413.
8. Yoshua Bengio, *Practical recommendations for gradient-based training of deep architectures*, arXiv:1206.5533 [cs] (2012) (en), arXiv: 1206.5533.
9. Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle, *Greedy Layer-Wise Training of Deep Networks*, 2006, p. 8 (en).
10. Nils Bjorck, Carla P Gomes, Bart Selman, and Kilian Q Weinberger, *Understanding Batch Normalization*, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), 2018, p. 12 (en).
11. Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis, *Soft-NMS – Improving Object Detection With One Line of Code*, arXiv:1704.04503 [cs] (2017) (en), arXiv: 1704.04503.
12. D.S. Broomhead and D. Lowe, *Multivariable Functional Interpolation and Adaptive Networks*, Complex Systems **2** (1988), 321–355.
13. William Chan, Navdeep Jaitly, Quoc V. Le, and Oriol Vinyals, *Listen, Attend and Spell*, arXiv:1508.01211 [cs, stat] (2015) (en), arXiv: 1508.01211.
14. Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik, *Vicinal Risk Minimization*, Advances in Neural Information Processing Systems 13, 2000, p. 7 (en).
15. Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud, *Neural Ordinary Differential Equations*, NIPS (2018), 13 (en).
16. Chung-Cheng Chiu, Tara N. Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Ekaterina Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani, *State-of-the-art Speech Recognition With Sequence-to-Sequence Models*, arXiv:1712.01769 [cs, eess, stat] (2018) (en), arXiv: 1712.01769.
17. Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*, arXiv:1406.1078 [cs, stat] (2014) (en), arXiv: 1406.1078.
18. Anna Choromanska, Mikael Henaff, Michael Mathieu, Gerard Ben Arous, and Yann LeCun, *The Loss Surfaces of Multilayer Networks*, roceedings of the 18thInternational Con-ference on Artificial Intelligence and Statistics, 2015, p. 13 (en).

19. D. Ciresan, U. Meier, and J. Schmidhuber, *Multi-column deep neural networks for image classification*, 2012 IEEE Conference on Computer Vision and Pattern Recognition (Providence, RI), IEEE, June 2012, pp. 3642–3649 (en).
20. Dan Ciresan, Alessandro Giusti, Luca M Gambardella, and Jürgen Schmidhuber, *Deep Neural Networks Segment Neuronal Membranes in Electron Microscopy Images*, (2012), 9 (en).
21. Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter, *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)*, arXiv:1511.07289 [cs] (2016) (en), arXiv: 1511.07289.
22. Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve, *Wav2Letter: an End-to-End ConvNet-based Speech Recognition System*, arXiv:1609.03193 [cs] (2016) (en), arXiv: 1609.03193.
23. Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David, *BinaryConnect: Training Deep Neural Networks with binary weights during propagations*, Advances in Neural Information Processing Systems 28, 2015, p. 9 (en).
24. G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of Control, Signals and Systems **2** (1989), no. 4, 303–314 (en).
25. Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier, *Language modeling with gated convolutional networks*, 2017.
26. Yann N. Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio, *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*, Advances in Neural Information Processing Systems **27** (2014), 2933–2941 (en).
27. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805 [cs] (2019) (en), arXiv: 1810.04805.
28. John Duchi, Elad Hazan, and Yoram Singer, *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization*, **12** (2011), 2121–2159 (en).
29. Jeffrey L. Elman, *Finding Structure in Time*, Cognitive Science **14** (1990), no. 2, 179–211 (en), _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog1402_1.
30. Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent, *Visualizing higher-layer features of a deep network*, Tech. Report 1341, University of Montreal, June 2009, Also presented at the ICML 2009 Workshop on Learning Feature Hierarchies, Montréal, Canada.
31. Bernd Fritzke, *A growing neural gas network learns topologies*, Proceedings of the 7th International Conference on Neural Information Processing Systems (Cambridge, MA, USA), NIPS'94, MIT Press, January 1994, pp. 625–632.
32. Kunihiko Fukushima, *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*, Biological Cybernetics **36** (1980), no. 4, 193–202 (en).
33. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin, *Convolutional Sequence to Sequence Learning*, arXiv:1705.03122 [cs] (2017) (en), arXiv: 1705.03122.
34. Felix A. Gers, Jürgen A. Schmidhuber, and Fred A. Cummins, *Learning to forget: Continual prediction with lstm*, Neural Comput. **12** (2000), no. 10, 2451–2471.
35. R. Girshick, *Fast R-CNN*, 2015 IEEE International Conference on Computer Vision (ICCV), December 2015, ISSN: 2380-7504, pp. 1440–1448.
36. Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, *Rich feature hierarchies for accurate object detection and semantic segmentation*, arXiv:1311.2524 [cs] (2014) (en), arXiv: 1311.2524.
37. Xavier Glorot and Yoshua Bengio, *Understanding the difficulty of training deep feedforward neural networks*, roceedings of the13thInternational Conferenceon Artificial Intelligence and Statistics (AISTATS) 2010, 2010, p. 8 (en).
38. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning*, MIT Press, 2016, `http://www.deeplearningbook.org`.
39. A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, *A Novel Connectionist System for Unconstrained Handwriting Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence **31** (2009), no. 5, 855–868 (en).
40. Alex Graves, *Sequence Transduction with Recurrent Neural Networks*, arXiv:1211.3711 [cs, stat] (2012) (en), arXiv: 1211.3711.

41. Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, *Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks*, Proceedings of the 23rd International Conference on Machine Learning (New York, NY, USA), ICML '06, Association for Computing Machinery, 2006, pp. 369–376.
42. Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, *Speech Recognition with Deep Recurrent Neural Networks*, arXiv:1303.5778 [cs] (2013) (en), arXiv: 1303.5778.
43. Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis, *Hybrid computing using a neural network with dynamic external memory*, Nature **538** (2016), no. 7626, 471–476 (en), Number: 7626 Publisher: Nature Publishing Group.
44. K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, *Lstm: A search space odyssey*, IEEE Transactions on Neural Networks and Learning Systems **28** (2017), no. 10, 2222–2232.
45. Andreas Griewank, *Who Invented the Reverse Mode of Differentiation?*, Documenta Mathematica (2012), 12 (en).
46. Andreas Griewank and Andrea Walther, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, 2 édition ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, November 2008 (Anglais).
47. Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger, *On Calibration of Modern Neural Networks*, arXiv:1706.04599 [cs] (2017) (en), arXiv: 1706.04599.
48. Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng, *Deep Speech: Scaling up end-to-end speech recognition*, arXiv:1412.5567 [cs] (2014) (en), arXiv: 1412.5567.
49. J Hastad, *Almost optimal lower bounds for small depth circuits*, Proceedings of the eighteenth annual ACM symposium on Theory of computing (New York, NY, USA), STOC '86, Association for Computing Machinery, November 1986, pp. 6–20.
50. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, *Mask R-CNN*, arXiv:1703.06870 [cs] (2018) (en), arXiv: 1703.06870.
51. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, arXiv:1502.01852 [cs] (2015) (en), arXiv: 1502.01852.
52. ———, *Deep Residual Learning for Image Recognition*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Las Vegas, NV, USA), IEEE, June 2016, pp. 770–778 (en).
53. ———, *Identity Mappings in Deep Residual Networks*, arXiv:1603.05027 [cs] (2016) (en), arXiv: 1603.05027.
54. Yanzhang He, Tara N. Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, Qiao Liang, Deepti Bhatia, Yuan Shangguan, Bo Li, Golan Pundak, Khe Chai Sim, Tom Bagby, Shuo-yiin Chang, Kanishka Rao, and Alexander Gruenstein, *Streaming End-to-end Speech Recognition For Mobile Devices*, arXiv:1811.06621 [cs] (2018) (en), arXiv: 1811.06621.
55. Mikael Henaff, Arthur Szlam, and Yann LeCun, *Recurrent Orthogonal Networks and Long-Memory Tasks*, (2016), 9 (en).
56. G. E. Hinton, *Reducing the Dimensionality of Data with Neural Networks*, Science **313** (2006), no. 5786, 504–507 (en).
57. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, *Distilling the Knowledge in a Neural Network*, arXiv:1503.02531 [cs, stat] (2015) (en), arXiv: 1503.02531.
58. Sepp Hochreiter and Jürgen Schmidhuber, *Long short-term memory*, Neural Computation **9** (1997), no. 8, 1735–1780.
59. A. L. Hodgkin and A. F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve*, The Journal of Physiology **117** (1952), no. 4, 500–544.

60. Elad Hoffer, Itay Hubara, and Daniel Soudry, *Train longer, generalize better: closing the generalization gap in large batch training of neural networks*, arXiv:1705.08741 [cs, stat] (2017), arXiv: 1705.08741.
61. Kurt Hornik, *Approximation capabilities of multilayer feedforward networks*, Neural Networks **4** (1991), no. 2, 251–257 (en).
62. Andrew G Howard, *Some Improvements on Deep Convolutional Neural Network Based Image Classification*, (2013), 6 (en).
63. Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*, arXiv:1704.04861 [cs] (2017) (en), arXiv: 1704.04861.
64. Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger, *Snapshot ensembles: train 1, get m for free*, 2017, p. 14 (en).
65. Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger, *Densely Connected Convolutional Networks*, arXiv:1608.06993 [cs] (2018) (en), arXiv: 1608.06993.
66. Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio, *Binarized Neural Networks*, 30th Conference on Neural Information Processing Systems (NIPOS 2016), 2016, p. 9 (en).
67. Sergey Ioffe and Christian Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, International Conference on Machine Learning, PMLR, June 2015, ISSN: 1938-7228, pp. 448–456 (en).
68. Max Jaderberg, Karen Simonyan, and Andrew Zisserman, *Spatial Transformer Networks*, Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015, p. 9 (en).
69. Josef Hochreiter, *Untersuchungen zu dynamischen neuronalen Netzen*, Ph.D. thesis, 1991.
70. Kam-Chuen Jim, C. L. Giles, and B. G. Horne, *An analysis of noise in recurrent neural networks: convergence and generalization*, IEEE Transactions on Neural Networks **7** (1996), no. 6, 1424–1438.
71. Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang, *On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima*, arXiv:1609.04836 [cs, math] (2017) (en), arXiv: 1609.04836.
72. Diederik P. Kingma and Jimmy Ba, *Adam: A Method for Stochastic Optimization*, arXiv:1412.6980 [cs], 2015, arXiv: 1412.6980.
73. Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár, *Panoptic Feature Pyramid Networks*, arXiv:1901.02446 [cs] (2019) (en), arXiv: 1901.02446.
74. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, *ImageNet classification with deep convolutional neural networks*, Communications of the ACM **60** (2012), no. 6, 84–90.
75. David Krueger, Tegan Maharaj, János Kramár, Mohammad Pezeshki, Nicolas Ballas, Nan Rosemary Ke, Anirudh Goyal, Yoshua Bengio, Aaron Courville, and Christopher Pal, *ZONEOUT: REGULARIZING RNNS BY RANDOMLY PRESERVING HIDDEN ACTIVATIONS*, (2017), 11 (en).
76. Quoc V. Le, Navdeep Jaitly, and Geoffrey E. Hinton, *A Simple Way to Initialize Recurrent Networks of Rectified Linear Units*, arXiv:1504.00941 [cs] (2015) (en), arXiv: 1504.00941.
77. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, *Backpropagation Applied to Handwritten Zip Code Recognition*, Neural Computation **1** (1989), no. 4, 541–551.
78. Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller, *Efficient BackProp*, Neural Networks: Tricks of the Trade: Second Edition (Grégoire Montavon, Geneviève B. Orr, and Klaus-Robert Müller, eds.), Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 1998, pp. 9–48 (en).
79. Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M. Cohen, Huyen Nguyen, and Ravi Teja Gadde, *Jasper: An End-to-End Convolutional Neural Acoustic Model*, arXiv:1904.03288 [cs, eess] (2019) (en), arXiv: 1904.03288.
80. Yuanzhi Li, Colin Wei, and Tengyu Ma, *Towards Explaining the Regularization Effect of Initial Large Learning Rate in Training Neural Networks*, NIPS 2019, 2019, p. 12 (en).
81. Min Lin, Qiang Chen, and Shuicheng Yan, *Network In Network*, arXiv:1312.4400 [cs] (2014) (en), arXiv: 1312.4400.

82. Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, *Feature Pyramid Networks for Object Detection*, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Honolulu, HI), IEEE, July 2017, pp. 936–944 (en).
83. Jonathan Long, Evan Shelhamer, and Trevor Darrell, *Fully Convolutional Networks for Semantic Segmentation*, arXiv:1411.4038 [cs] (2015) (en), arXiv: 1411.4038.
84. Ilya Loshchilov and Frank Hutter, *SGDR: Stochastic Gradient Descent with Warm Restarts*, arXiv:1608.03983 [cs, math], 2017, arXiv: 1608.03983 (en).
85. Minh-Thang Luong, Hieu Pham, and Christopher D. Manning, *Effective Approaches to Attention-based Neural Machine Translation*, arXiv:1508.04025 [cs] (2015) (en), arXiv: 1508.04025.
86. Andrew L Maas, Awni Y Hannun, and Andrew Y Ng, *Rectifier Nonlinearities Improve Neural Network Acoustic Models*, Proceedings of the 30th International Conference on Machine Learning (ICML13), 2013, p. 6 (en).
87. Richard Maclin and Jude W. Shavlik, *Combining the predictions of multiple classifiers: using competitive learning to initialize neural networks*, Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1 (San Francisco, CA, USA), IJCAI'95, Morgan Kaufmann Publishers Inc., August 1995, pp. 524–530.
88. Warren S. McCulloch and Walter Pitts, *A logical calculus of the ideas immanent in nervous activity*, The bulletin of mathematical biophysics **5** (1943), no. 4, 115–133 (en).
89. Marvin Minksy and Seymour Papert, *Perceptrons: An Introduction to Computational Geometry*, 1969.
90. Guido F. Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio, *On the Number of Linear Regions of Deep Neural Networks*, Advances in Neural Information Processing Systems, vol. 27, 2014, pp. 2924–2932 (en).
91. T. Moon, H. Choi, H. Lee, and I. Song, *Rnndrop: A novel dropout for rnns in asr*, 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015, pp. 65–70.
92. Rafael Müller, Simon Kornblith, and Geoffrey Hinton, *When Does Label Smoothing Help?*, arXiv:1906.02629 [cs, stat] (2020) (en), arXiv: 1906.02629.
93. Vinod Nair and Geoffrey E. Hinton, *Rectified linear units improve restricted boltzmann machines*, Proceedings of the 27th International Conference on International Conference on Machine Learning (Madison, WI, USA), ICML'10, Omnipress, June 2010, pp. 807–814.
94. Michael A. Nielsen, *Neural Networks and Deep Learning*, 2015 (en), Publisher: Determination Press.
95. Augustus Odena, Vincent Dumoulin, and Chris Olah, *Deconvolution and Checkerboard Artifacts*, Distill **1** (2016), no. 10, e3 (en).
96. Chris Olah, *Calculus on computational graphs: backpropagation*, 2015, `http://colah.github.io/posts/2015-08-Backprop/`.
97. Chris Olah, Alexander Mordvintsev, and Ludwig Schubert, *Feature visualization*, Distill (2017), https://distill.pub/2017/feature-visualization.
98. J. Park and I. W. Sandberg, *Universal Approximation Using Radial-Basis-Function Networks*, Neural Computation **3** (1991), no. 2, 246–257.
99. Razvan Pascanu, Yann N. Dauphin, Surya Ganguli, and Yoshua Bengio, *On the saddle point problem for non-convex optimization*, arXiv:1405.4604 [cs] (2014) (en), arXiv: 1405.4604.
100. Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio, *On the difficulty of training recurrent neural networks*, Proceedings of the 30thInternational Conference on Machine Learning, 2013, p. 9 (en).
101. Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, *Automatic differentiation in PyTorch*, Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, p. 4 (en).
102. Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, *You Only Look Once: Unified, Real-Time Object Detection*, arXiv:1506.02640 [cs] (2016) (en), arXiv: 1506.02640.
103. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*, arXiv:1506.01497 [cs] (2016) (en), arXiv: 1506.01497.
104. Olaf Ronneberger, Philipp Fischer, and Thomas Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, arXiv:1505.04597 [cs] (2015) (en), arXiv: 1505.04597.

105. F. Rosenblatt, *The perceptron: A probabilistic model for information storage and organization in the brain.*, Psychological Review **65** (1958), no. 6, 386–408 (en).
106. David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, *Learning representations by back-propagating errors*, Nature **323** (1986), no. 6088, 533–536 (en).
107. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, *ImageNet Large Scale Visual Recognition Challenge*, International Journal of Computer Vision **115** (2015), no. 3, 211–252 (en).
108. Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, *MobileNetV2: Inverted Residuals and Linear Bottlenecks*, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (Salt Lake City, UT), IEEE, June 2018, pp. 4510–4520 (en).
109. Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Ma, *How Does Batch Normalization Help Optimization?*, 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), 2018, p. 11 (en).
110. J. Schmidhuber, *Learning Complex, Extended Sequences Using the Principle of History Compression*, Neural Computation **4** (1992), no. 2, 234–242, Conference Name: Neural Computation.
111. Jürgen Schmidhuber, *Deep learning in neural networks: An overview*, Neural Networks **61** (2015), 85–117.
112. M. Schuster and K.K. Paliwal, *Bidirectional recurrent neural networks*, IEEE Transactions on Signal Processing **45** (1997), no. 11, 2673–2681 (en).
113. Friedhelm Schwenker, Hans A. Kestler, and Günther Palm, *Three learning phases for radial-basis-function networks.*, Neural Networks **14** (2001), no. 4-5, 439–458.
114. Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*, International Journal of Computer Vision **128** (2020), no. 2, 336–359 (en), arXiv: 1610.02391.
115. Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, *Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network*, arXiv:1609.05158 [cs, stat] (2016) (en), arXiv: 1609.05158.
116. Wenzhe Shi, Jose Caballero, Lucas Theis, Ferenc Huszar, Andrew Aitken, Alykhan Tejani, Johannes Totz, Christian Ledig, and Zehan Wang, *Is the deconvolution layer the same as a convolutional layer?*, (2016), 7 (en).
117. Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*, arXiv:1312.6034 [cs] (2014) (en), arXiv: 1312.6034.
118. Karen Simonyan and Andrew Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, arXiv:1409.1556 [cs], April 2015, arXiv: 1409.1556 (en).
119. Leslie N. Smith, *A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay*, arXiv:1803.09820 [cs, stat] (2018), arXiv: 1803.09820.
120. Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller, *Striving for Simplicity: The All Convolutional Net*, arXiv:1412.6806 [cs] (2015) (en), arXiv: 1412.6806.
121. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, Journal of Machine Learning Research **15** (2014), no. 56, 1929–1958.
122. Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber, *Training Very Deep Networks*, (2015), 9 (en).
123. Ilya Sutskever, *Training recurrent neural networks*, Ph.D. thesis, CAN, 2013.
124. Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton, *On the importance of initialization and momentum in deep learning*, ICML, 2013, p. 14 (en).
125. Ilya Sutskever, James Martens, and Geoffrey Hinton, *Generating Text with Recurrent Neural Networks*, (2011), 8 (en).
126. Ilya Sutskever, Oriol Vinyals, and Quoc V. Le, *Sequence to Sequence Learning with Neural Networks*, arXiv:1409.3215 [cs] (2014) (en), arXiv: 1409.3215.

127. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, *Going Deeper with Convolutions*, arXiv:1409.4842 [cs] (2014) (en), arXiv: 1409.4842.
128. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna, *Rethinking the Inception Architecture for Computer Vision*, arXiv:1512.00567 [cs] (2015) (en), arXiv: 1512.00567.
129. Mingxing Tan and Quoc V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, arXiv:1905.11946 [cs, stat] (2020) (en), arXiv: 1905.11946.
130. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, *Attention is All you Need*, (2017), 11 (en).
131. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, *Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge*, IEEE Transactions on Pattern Analysis and Machine Intelligence **39** (2015), no. 4, 652–663 (en), arXiv: 1609.06647.
132. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, *Phoneme recognition using time-delay neural networks*, IEEE Transactions on Acoustics, Speech, and Signal Processing **37** (1989), no. 3, 328–339.
133. Paul J Werbos, *Backpropagation through time: what it does and how to do it*, Proceedings of the IEEE **78** (1990), no. 10, 1550–1560.
134. P.J. Werbos, *Application of advances in nonlinear sensitivity analysis*, Proc. of the 10th IFIP conference, 1981, pp. 762–770.
135. Bernard Widrow and Marcian E. Hoff, *Associative Storage and Retrieval of Digital Information in Networks of Adaptive "Neurons"*, Biological Prototypes and Synthetic Systems: Volume 1 Proceedings of the Second Annual Bionics Symposium sponsored by Cornell University and the General Electric Company, Advanced Electronics Center, held at Cornell University, August 30–September 1, 1961 (Eugene E. Bernard and Morley R. Kare, eds.), Springer US, Boston, MA, 1962, pp. 160–160 (en).
136. Ronald J. Williams and Jing Peng, *An efficient gradient-based algorithm for on-line training of recurrent network trajectories*, Neural Computation **2** (1990), no. 4, 490–501.
137. Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean, *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*, arXiv:1609.08144 [cs] (2016) (en), arXiv: 1609.08144.
138. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio, *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*, arXiv:1502.03044 [cs] (2016) (en), arXiv: 1502.03044.
139. Fisher Yu and Vladlen Koltun, *Multi-Scale Context Aggregation by Dilated Convolutions*, arXiv:1511.07122 [cs] (2016) (en), arXiv: 1511.07122.
140. Heiga Ze, Andrew Senior, and Mike Schuster, *Statistical parametric speech synthesis using deep neural networks*, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (Vancouver, BC, Canada), IEEE, May 2013, pp. 7962–7966 (en).
141. Neil Zeghidour, Nicolas Usunier, Gabriel Synnaeve, Ronan Collobert, and Emmanuel Dupoux, *End-to-End Speech Recognition from the Raw Waveform*, Interspeech 2018, ISCA, September 2018, pp. 781–785 (en).
142. Matthew D. Zeiler, *ADADELTA: An Adaptive Learning Rate Method*, arXiv:1212.5701 [cs] (2012), arXiv: 1212.5701.
143. Matthew D. Zeiler and Rob Fergus, *Visualizing and Understanding Convolutional Networks*, Computer Vision – ECCV 2014 (Cham) (David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, eds.), Lecture Notes in Computer Science, Springer International Publishing, 2014, pp. 818–833 (en).
144. Jian Zhang and Ioannis Mitliagkas, *YellowFin and the Art of Momentum Tuning*, arXiv:1706.03471 [cs, stat] (2018), arXiv: 1706.03471.