Mathematical Engineering of Deep Learning

Book Draft

Benoit Liquet, Sarat Moka and Yoni Nazarathy

February 28, 2024

Contents

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467*, 2016.

[2] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for boltzmann machines. *Cognitive Science*, 1985.

[3] J. Adler and S. Lunz. Banach Wasserstein GAN. *Advances in Neural Information Processing Systems*, 2018.

[4] C. C. Aggarwal. Neural networks and deep learning. *Springer*, 2018.

[5] A. Agresti. *Categorical data analysis*. John Wiley & Sons, 2003.

[6] A. Agresti. *Analysis of ordinal categorical data*. John Wiley & Sons, 2010.

[7] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974.

[8] K. Akuzawa, Y. Iwasawa, and Y. Matsuo. Expressive speech synthesis via modeling expressions with variational autoencoder. *arXiv:1804.02135*, 2018.

[9] P. Albertos and I. Mareels. *Feedback and control for everyone*. Springer, 2010.

[10] J. J. Allaire. *Deep Learning with R*. Simon and Schuster, 2018.

[11] A. Alotaibi. Deep generative adversarial networks for image-to-image translation: A review. *Symmetry*, 2020.

[12] S. Amari. A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers*, 1967.

[13] S. Amari. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers*, 1972.

[14] P. J. Antsaklis and A. N. Michel. *Linear systems*. Springer, 1997.

[15] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017.

[16] L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific Journal of mathematics*, 1966.

[17] S. Arora, Z. Li, and K. Lyu. Theoretical analysis of auto rate-tuning by batch normalization. *arXiv:1812.03981*, 2018.

[18] J. Atwood and D. Towsley. Diffusion-convolutional neural networks. *Advances in neural information processing systems*, 2016.

[19] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv:1607.06450*, 2016.

[20] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv:1409.0473*, 2014.

[21] P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 1989.

[22] W. Bao, J. Yue, and Y. Rao. A deep learning framework for financial time series using stacked autoencoders and long-short term memory. *PLoS ONE*, 2017.

[23] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[24] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 2018.

[25] L. M. Beda, L. N. Korolev, N. V. Sukkikh, and T. S. Frolova. Programs for automatic differentiation for the machine BESM (in Russian). *Technical report, Institute for Precise Mechanics and Computation Techniques, Academy of Science, Moscow, USSR*, 1959.

[26] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 2019.

[27] A. Ben-Tal and A. Nemirovski. *Lectures on modern convex optimization*. SIAM, Philadelphia, PA; MPS, Philadelphia, PA, 2001.

[28] Y. Bengio. Learning deep architectures for ai. *Foundations and Trends® in Machine Learning*, 2009.

[29] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*. 2012.

[30] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. *Advances in Neural Information Processing Systems*, 2006.

[31] J. O. Berger and R. L. Wolpert. The Likelihood Principle: A Review, Generalizations, and Statistical Implications. *Lecture Notes—Monograph Series*, 1988.

[32] D. Bertsekas, A. Nedić, and A. E. Ozdaglar. *Convex analysis and optimization*. Athena Scientific, 2003.

[33] D. P. Bertsekas. *Dynamic Programming and Optimal Control, Volume. II.* Athena Scientific, 3rd edition, 2007.

[34] D. P. Bertsekas. *Dynamic programming and optimal control: Volume I.* Athena Scientific, 2012.

[35] D. P. Bertsekas. *Nonlinear programming.* Athena Scientific, Third edition, 2016.

[36] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming.* Athena Scientific, 1996.

[37] D. Bertsimas and J. N. Tsitsiklis. *Introduction to linear optimization.* Athena Scientific, 1997.

[38] C. Biever. ChatGPT broke the Turing test-the race is on for new ways to assess AI. *Nature*, 2023.

[39] C. M. Bishop. *Pattern Recognition and Machine learning.* Springer, 2006.

[40] A. Bjerhammar. *Application of calculus of matrices to method of least squares: with special reference to geodetic calculations.* Elander, 1951.

[41] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.*, 2017.

[42] C. I. Bliss. The method of probits. *Science*, 1934.

[43] S. Bock and Weiß. A Proof of Local Convergence for the Adam Optimizer. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019.

[44] D. Böhning. Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 1992.

[45] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee. YOLACT: Real-Time Instance Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[46] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[47] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical optimization: Theoretical and practical aspects.* Springer Science & Business Media, 2006.

[48] L. Bottou. Online algorithms and stochastic approximations. *Online learning in neural networks*, 1998.

[49] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational Statistics*, 2010.

[50] L. Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade*. 2012.

[51] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 2018.

[52] L. Bottou and Y. LeCun. Large scale online learning. In *Advances in Neural Information Processing Systems*, 2003.

[53] H. Bourlard and Y. Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 1988.

[54] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. In *20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016*, 2016.

[55] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[56] S. Boyd and L. Vandenberghe. *Introduction to applied linear algebra: Vectors, matrices, and least squares*. Cambridge university press, 2018.

[57] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: Composable Transformations of Python+NumPy Programs. *http://github.com/google/jax*, 2018.

[58] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 1997.

[59] L. Breiman. Bagging predictors. *Machine Learning*, 1996.

[60] L. Breiman. Random forests. *Machine Learning*, 2001.

[61] L. Breiman. Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). *Statistical Science*, 2001.

[62] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and regression trees. Wadsworth & Brooks. *Cole Statistics/Probability Series*, 1984.

[63] J. S. Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*. 1990.

[64] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer Science & Business Media, 1991.

[65] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a "siamese" time delay neural network. *Advances in neural information processing systems*, 1993.

[66] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.

[67] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral networks and locally connected networks on graphs. *arXiv:1312.6203*, 2013.

[68] A. Burkov. *The Hundred-Page Machine Learning Book*. Andriy Burkov Quebec City, QC, Canada, 2019.

[69] O. Calin. *Deep Learning Architectures*. Springer, 2020.

[70] A. Canziani, A. Paszke, and E. Culurciello. An Analysis of Deep Neural Network Models for Practical Applications. *arXiv:1605.07678*, 2016.

[71] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P. A. Heng, and S. Z. Li. A survey on generative diffusion model. *arXiv:2209.02646*, 2022.

[72] A. Cauchy. Méthode générale pour la résolution des systèmes d'équations simultanées. *Comp. Rend. Sci. Paris*, 1847.

[73] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, et al. A survey on evaluation of large language models. *arXiv:2307.03109*, 2023.

[74] D. Charte, F. Charte, S. García, M. J. del Jesus, and F. Herrera. A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. *Information Fusion*, 2018.

[75] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 2002.

[76] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao. Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 2021.

[77] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[78] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*, 2016.

[79] P. S. Chib and P. Singh. Recent advancements in end-to-end autonomous driving using deep learning: A survey. *IEEE Transactions on Intelligent Vehicles*, 2023.

[80] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv:1409.1259*, 2014.

[81] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv:1406.1078*, 2014.

[82] D. Choi, C. J. Shallue, Z. Nado, J. Lee, C. J. Maddison, and G. E. Dahl. On empirical comparisons of optimizers for deep learning. *arXiv:1910.05446*, 2019.

[83] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[84] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, 2005.

[85] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv:1412.3555*, 2014.

[86] D. C. Cireșan, U. Meier, L. M. Gambardella, and J. Schmidhuber. Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation*, 2010.

[87] G. Claeskens and N. L. Hjort. Model selection and model averaging. *Cambridge Books*, 2008.

[88] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 1979.

[89] A. K. Cline and I. S. Dhillon. Computation of the singular value decomposition. In *Handbook of linear algebra*. 2006.

[90] N. Cohen, O. Sharir, and A. Shashua. On the Expressive Power of Deep Learning: A Tensor Analysis, 2016.

[91] D. Commenges and H. Jacqmin-Gadda. *Dynamical biostatistical models*. CRC Press, 2015.

[92] P. Congdon. *Bayesian Statistical Modelling*. John Wiley & Sons, 2007.

[93] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT press, 2022.

[94] D. R. Cox and D. V. Hinkley. *Theoretical statistics*. CRC Press, 1979.

[95] J. S. Cramer. The origins of logistic regression. *Tinbergen Institute Working Paper No. 2002-119/4*, 2002.

[96] F. A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[97] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 1989.

[98] Y. H. Dai. Convergence properties of the BFGS algorithm. *SIAM Journal on Optimization*, 2002.

[99] W. C. Davidon. Variable metric method for minimization. Technical report, Argonne National Lab., Lemont, Ill., 1959.

[100] W. C. Davidon. Variable metric method for minimization. *SIAM Journal on Optimization*, 1991.

[101] M. P. Deisenroth, A. A. Faisal, and C. S. Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020.

[102] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[103] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

[104] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 2021.

[105] L. DO Q. Numerically efficient methods for solving least squares problems. *Pennsylvania State University*, 2012.

[106] A. J. Dobson and A. G. Barnett. *An introduction to generalized linear models*. Chapman and Hall/CRC, 2018.

[107] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929*, 2020.

[108] J. C. Douma and J. T. Weedon. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. *Methods in Ecology and Evolution*, 2019.

[109] T. Dozat. Incorporating Nesterov momentum into Adam. *International Conference on Learning Representations (ICLR) Workshop*, 2016.

[110] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, 2022.

[111] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing*, 2022.

[112] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011.

[113] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. *arXiv:1610.07629*, 2016.

[114] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1936.

[115] B. Efron and T. Hastie. *Computer age statistical inference*. Cambridge University Press, 2016.

[116] R. Eldan and O. Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, 2016.

[117] J. L. Elman. Finding structure in time. *Cognitive science*, 1990.

[118] F. Emmert-Streib, S. Moutari, and M. Dehmer. A comprehensive survey of error measures for evaluating binary decision making in data science. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2019.

[119] D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 2009.

[120] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 2022.

[121] J. J. Faraway. *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman and Hall/CRC, 2016.

[122] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional Two-Stream Network Fusion for Video Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[123] R. Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

[124] R. Fletcher and M. J. D. Powell. A rapidly convergent descent method for minimization. *The Computer Journal*, 1963.

[125] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau. An introduction to deep reinforcement learning. *Foundations and Trends in Machine Learning*, 2018.

[126] K. Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 1969.

[127] K. Fukushima. c: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 1980.

[128] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning*, 2016.

[129] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez. A Survey on Deep Learning Techniques for Image and Video Semantic Segmentation. *Applied Soft Computing*, 2018.

[130] S. Garg and G. Ramakrishnan. Advances in Quantum Deep Learning: An Overview. *arXiv:2005.04316*, 2020.

[131] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv:1508.06576*, 2015.

[132] Gauss, C. F. Theoria Motus Corporum Coelestium. Perthes, Hamburg. *Translation reprinted as Theory of the Motions of the Heavenly Bodies Moving about the Sun in Conic Sections. Dover, New York, 1963*, 1809.

[133] A. Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media, 2019.

[134] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, 2017.

[135] R. Girshick. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[136] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.

[137] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010.

[138] Y. Goldberg. *Neural Network Methods for Natural Language Processing.* Springer Nature, 2022.

[139] G. H. Golub. Least squares, singular values and matrix approximations. *Aplikace matematiky*, 1968.

[140] G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. In *Linear algebra.* 1971.

[141] G. H. Golub and C. F. Van Loan. *Matrix computations.* JHU press, 2013.

[142] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning.* MIT Press, 2016.

[143] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2014.

[144] M. Gori, G. Monfardini, and F. Scarselli. A new model for learning in graph domains. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, 2005.

[145] A. Graves. Generating sequences with recurrent neural networks. *arXiv:1308.0850*, 2013.

[146] A. Graves, A. R. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, 2013.

[147] A. Griewank and A. Walther. *Evaluating derivatives: Principles and techniques of algorithmic differentiation*. SIAM, 2008.

[148] A. Grover and J. Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016.

[149] R. Gueorguieva, R. Rosenheck, and D. Zelterman. Dirichlet component regression and its applications to psychiatric data. *Computational Statistics & Data Analysis*, 2008.

[150] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[151] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, 2017.

[152] I. Gulrajani, K. Kumar, F. Ahmed, A. A. Taiga, F. Visin, D. Vazquez, and A. Courville. PixelVAE: A Latent Variable Model for Natural Images. In *International Conference on Learning Representations*, 2016.

[153] H. Guo, Y. Li, J. Shang, M. Gu, Y. Huang, and B. Gong. Learning from class-imbalanced data: Review of methods and applications. *Expert systems with applications*, 2017.

[154] I. Guyon, P. Albrecht, Y. LeCun, J. Denker, and W. Hubbard. Design of a neural network character recognizer for a touch terminal. *Pattern Recognition*, 1991.

[155] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, S. Mirjalili, et al. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 2023.

[156] W. Hamilton, Z. Ying, and J. Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 2017.

[157] W. L. Hamilton. *Graph Representation Learning*. Morgan & Claypool Publishers, 2020.

[158] D. J. Hand. Assessing the performance of classification methods. *International Statistical Review*, 2012.

[159] K. Hara, D. Saitoh, and H. Shouno. Analysis of dropout learning regarded as ensemble learning. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*, 2016.

[160] M. A. Hardy. *Regression with dummy variables*. Sage, 1993.

[161] D. Harrison Jr and D. L. Rubinfeld. Hedonic Housing Prices and the Demand for Clean Air. *Journal of Environmental Economics and Management*, 1978.

[162] G. M. Harshvardhan, M. K. Gourisaria, M. Pandey, and S. S. Rautaray. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 2020.

[163] W. Harvey, S. Naderiparizi, V. Masrani, C. Weilbach, and F. Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 2022.

[164] D. Hassabis. Artificial Intelligence: Chess Match of the Century. *Nature*, 2017.

[165] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick. Neuroscience-Inspired Artificial Intelligence. *Neuron*, 2017.

[166] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: Data mining, inference, and prediction*. Springer, 2009.

[167] T. Hastie, R. Tibshirani, and M. Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 2015.

[168] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Routledge, 2017.

[169] S. Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall, 1998.

[170] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[171] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 2015.

[172] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[173] R. Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural Networks for Perception*. 1992.

[174] M. Henaff, J. Bruna, and Y. LeCun. Deep convolutional networks on graph-structured data. *arXiv:1506.05163*, 2015.

[175] S. Herculano-Houzel. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in human neuroscience*, 2009.

[176] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv:2208.01626*, 2022.

[177] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving Linear Systems. *Journal of research of the National Bureau of Standards*, 1952.

[178] R. H. Hijazi and R. W. Jernigan. Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability & Statistics*, 2009.

[179] J. M. Hilbe. *Logistic regression models*. Chapman and Hall/CRC, 2009.

[180] G. E. Hinton, S. Osindero, and Y. W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006.

[181] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv:1207.0580*, 2012.

[182] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022.

[183] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[184] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[185] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 1982.

[186] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 1991.

[187] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[188] M. Z. Hossain, F. Sohel, M. Shiratuddin, and H. Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 2019.

[189] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 1933.

[190] J. Howard and S. Gugger. *Deep Learning for Coders with fastai and PyTorch*. O'Reilly Media, 2020.

[191] C. C. Hsu, H. T. Hwang, Y. C. Wu, Y. Tsao, and H. M. Wang. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *Interspeech 2017*, 2017.

[192] G. Hu, Y. Yang, D. Yi, J. Kittler, W. Christmas, S. Z. Li, and T. Hospedales. When Face Recognition Meets with Deep Learning: An Evaluation of Convolutional Neural Networks for Face Recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015.

[193] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. Toward controlled generation of text. In *International Conference on Machine Learning*, 2017.

[194] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao. Normalization techniques in training dnns: Methodology, analysis and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[195] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 1959.

[196] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 1962.

[197] P. J. Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *The Annals of Statistics*, 1973.

[198] R. J. Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, 3rd edition, 2021.

[199] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and$<$ 0.5 mb model size. *arXiv:1602.07360*, 2016.

[200] G. Iglesias, E. Talavera, and A. Díaz-Álvarez. A survey on GANs for computer vision: Recent research, analysis and taxonomy. *Computer Science Review*, 2023.

[201] M. Innes. Flux: Elegant Machine Learning with Julia. *Journal of Open Source Software*, 2018.

[202] S. Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. *Advances in Neural Information Processing Systems*, 2017.

[203] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 2015.

[204] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[205] A. G. Ivakhnenko. Polynomial theory of complex systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 1971.

[206] A. G. Ivakhnenko and V. G. Lapa. Cybernetic predicting devices. *Purdue Univ Lafayette Ind School of Electrical Engineering, appearing in The Defense Technical Information Center*, 1966.

[207] B. Jähne. *Digital image processing*. Springer Science & Business Media, 2005.

[208] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 1999.

[209] G. Jin, Y. Liang, Y. Fang, Z. Shao, J. Huang, J. Zhang, and Y. Zheng. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[210] W. Jin, R. Barzilay, and T. Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, 2018.

[211] J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 2019.

[212] I. T. Jolliffe. *Principal component analysis for special types of data*. Springer, 2002.

[213] L. V. Jospin, W. Buntine, F. Boussaid, H. Laga, and M. Bennamoun. Hands-on Bayesian Neural Networks–A Tutorial for Deep Learning Users. *arXiv:2007.06823*, 2020.

[214] R. Jozefowicz, W. Zaremba, and I. Sutskever. An empirical exploration of recurrent network architectures. In *International conference on machine learning*, 2015.

[215] Y. Jung. Multiple predicting K-fold cross-validation for model selection. *Journal of Nonparametric Statistics*, 2018.

[216] D. Jurafsky and J. H. Martin. *Speech and Language Processing*. Pearson, 2000.

[217] L. P. Kaelbling, M. L. Littman, and A. W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, 1996.

[218] N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013.

[219] M. Kang, J. Y. Zhu, R. Zhang, J. Park, E. Shechtman, S. Paris, and T. Park. Scaling up GANs for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[220] Z. Karevan and J. A. K. Suykens. Transductive lstm for time-series prediction: An application to weather forecasting. *Neural Networks*, 2020.

[221] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[222] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv:1710.10196*, 2017.

[223] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 2020.

[224] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019.

[225] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[226] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[227] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, 2022.

[228] D. Khurana, A. Koli, K. Khatter, and S. Singh. Natural language processing: State of the art, current trends and challenges. *Multimedia Tools and Applications*, 2023.

[229] J. Kiefer. Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 1953.

[230] J. Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1959.

[231] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 1952.

[232] J. H. Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 2009.

[233] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[234] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv:1312.6114*, 2013.

[235] D. P. Kingma and M. Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 2019.

[236] T. N. Kipf and M. Welling. Variational graph auto-encoders. *arXiv:1611.07308*, 2016.

[237] S. Kiranyaz, T. Ince, and M. Gabbouj. Optimization techniques: An overview. *Multidimensional Particle Swarm Optimization for Machine Learning and Pattern Recognition*, 2014.

[238] M. J. Kochenderfer and T. A. Wheeler. *Algorithms for optimization*. MIT Press, 2019.

[239] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012.

[240] D. P. Kroese, Z. Botev, T. Taimre, and R. Vaisman. *Data science and machine learning: Mathematical and statistical methods*. CRC Press, 2019.

[241] A. Krogh and J. Hertz. A Simple Weight Decay Can Improve Generalization. In *Advances in Neural Information Processing Systems*, 1991.

[242] J. Kuehn, S. Abadie, B. Liquet, and V. Roeber. A deep learning super-resolution model to speed up computations of coastal sea states. *Applied Ocean Research*, 2023.

[243] J. Kukačka, V. Golkov, and D. Cremers. Regularization for deep learning: A taxonomy. *arXiv:1710.10686*, 2017.

[244] H. Kwakernaak and R. Sivan. *Modern signal and systems*. Prentice Hall, 1991.

[245] P. Lafaye de Micheaux, R. Drouilhet, and B. Liquet. *The R software: Fundamentals of programming and statistical analysis*. Springer, 2013.

[246] H. Lai, S. Xiao, Y. Pan, Z. Cui, J. Feng, C. Xu, J. Yin, and S. Yan. Deep recurrent regression for facial landmark detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[247] K. J. Lang. A time-delay neural network architecture for speech recognition. *Technical Report, Carnegie-Mellon University*, 1988.

[248] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin. Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*, 2009.

[249] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 2015.

[250] Y. LeCun, B. Boser, J. Denker, D. Henderson, W. Hubbard, and L. Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 1989.

[251] 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*, 1989.

[252] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[253] C. Lemaréchal. Cauchy and the gradient method. *Doc Math Extra*, 2012.

[254] K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 1944.

[255] H. Levitt. Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, 1971.

[256] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen. Medical Image Classification with Convolutional Neural Network. In *2014 13th International Conference on Control Automation Robotics & Vision (ICARCV)*, 2014.

[257] Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv:1707.01926*, 2017.

[258] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[259] H. W. Lin, M. Tegmark, and D. Rolnick. Why does deep and cheap learning work so well? *Journal of Statistical Physics*, 2017.

[260] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv:1312.4400*, 2013.

[261] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi. Don't Use Large Mini-batches, Use Local SGD. In *International Conference on Learning Representations*, 2020.

[262] T. Lin, Y. Wang, X. Liu, and X. Qiu. A survey of transformers. *AI Open*, 2022.

[263] A. Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön. *Machine Learning - A First Course for Engineers and Scientists*. Cambridge University Press, 2022.

[264] H. Ling, K. Kreis, D. Li, S. W. Kim, A. Torralba, and S. Fidler. Editgan: High-precision semantic image editing. *Advances in Neural Information Processing Systems*, 2021.

[265] S. Linnainmaa. The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. *Master's Thesis (in Finnish), Univ. Helsinki*, 1970.

[266] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large-scale optimization. *Mathematical Programming*, 1989.

[267] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path Aggregation Network for Instance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[268] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv:1907.11692*, 2019.

[269] Y. Liu, Y. Zhang, Y. Wang, F. Hou, J. Yuan, J. Tian, Y. Zhang, Z. Shi, J. Fan, and Z. He. A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[270] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017.

[271] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are GANs created equal? A large-scale study. *Advances in Neural Information Processing Systems*, 2018.

[272] D. G. Luenberger and Y. Ye. *Linear and nonlinear programming*. Springer, Cham, Fifth edition, 2021.

[273] Y. Ma and J. Tang. *Deep Learning on Graphs*. Cambridge University Press, 2021.

[274] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[275] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[276] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, 2013.

[277] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, 1967.

[278] A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[279] K. M. Malan. A survey of advances in landscape analysis for optimization. *Algorithms*, 2021.

[280] C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT press, 1999.

[281] E. R. Mansfield and B. P. Helms. Detecting multicollinearity. *The American Statistician*, 1982.

[282] N. Mazyavkina, S. Sviridov, S. Ivanov, and E. Burnaev. Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research*, 2021.

[283] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 1943.

[284] S. Mei, Y. Bai, and A. Montanari. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 2018.

[285] G. Menghani. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys*, 2023.

[286] L. Mero, D. Yi, M. Dianati, and A. Mouzakitis. A survey on imitation learning techniques for end-to-end autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[287] A. Micheli. Neural network for graphs: A contextual constructive approach. *IEEE Transactions on Neural Networks*, 2009.

[288] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv:1301.3781*, 2013.

[289] G. A. Miller. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[290] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image Segmentation Using Deep Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[291] M. Minsky and S. A. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

[292] L. Mirsky. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 1960.

[293] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, 2014.

[294] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *arXiv:1312.5602*, 2013.

[295] A. Moghar and M. Hamiche. Stock market prediction using lstm recurrent neural network. *Procedia Computer Science*, 2020.

[296] D. C. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, 2017.

[297] E. H. Moore. On the reciprocal of the general algebraic matrix. *Bull. Am. Math. Soc.*, 1920.

[298] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

[299] K. P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

[300] F. Murtagh and P. Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2012.

[301] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *International conference on machine learning*, 2010.

[302] S. C. Narula and J. F. Wellington. The minimum sum of absolute errors regression: A state of the art survey. *International Statistical Review/Revue Internationale de Statistique*, 1982.

[303] Y. Nazarathy and H. Klok. *Statistics with Julia*. Springer, 2021.

[304] J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A*, 1972.

[305] Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 1983.

[306] A. Ng. Machine Learning Yearning. https://info.deeplearning.ai/machine-learning-yearning-book, 2017.

[307] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv:2112.10741*, 2021.

[308] M. A. Nielsen. *Neural networks and deep learning*. Determination press San Francisco, CA, 2015.

[309] M. Niepert, M. Ahmed, and K. Kutzkov. Learning convolutional neural networks for graphs. In *International conference on machine learning*, 2016.

[310] J. Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 1980.

[311] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 1999.

[312] H. Noh, S. Hong, and B. Han. Learning Deconvolution Network for Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[313] J. F. Nolan. *Analytical differentiation on a digital computer*. PhD thesis, Massachusetts Institute of Technology, 1953.

[314] A. B. Novikoff. On convergence proofs for perceptrons. Technical report, Stanford Research Inst. Menlo Park CA, 1963.

[315] N. A. Obuchowski and J. A. Bullen. Receiver Operating Characteristic (ROC) Curves: Review of Methods with Applications in Diagnostic Medicine. *Physics in Medicine & Biology*, 2018.

[316] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Machine Learning*, 2017.

[317] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu. Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016.

[318] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

[319] A. S. Pandya and R. B. Macy. *Pattern Recognition with Neural Networks in C++*. CRC Press, 1995.

[320] J. Papa. *PyTorch Pocket Reference: Building and Deploying Deep Learning Model*. O'Reilly Media, 2021.

[321] J. M. Papakonstantinou and R. A. Tapia. Origin and evolution of the secant method in one dimension. *The American Mathematical Monthly*, 2013.

[322] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 2021.

[323] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, 2013.

[324] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. *31st Conference on Neural Information Processing Systems (NIPS2017)*, 2017.

[325] J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect.* Basic Books, 2018.

[326] K. Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 1901.

[327] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014.

[328] R. Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge Philosophical Society*, 1955.

[329] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

[330] K. B. Petersen and M. S. Pedersen. The matrix cookbook. *Technical University of Denmark*, 2012.

[331] M. Phuong and M. Hutter. Formal algorithms for transformers. *arXiv:2207.09238*, 2022.

[332] E. Plaut. From principal subspaces to principal components with linear autoencoders. *arXiv:1804.10253*, 2018.

[333] T. Poggio, A. Banburski, and Q. Liao. Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences*, 2020.

[334] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao. Why and When Can Deep – but Not Shallow – Networks Avoid the Curse of Dimensionality: a Review, 2017.

[335] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 1964.

[336] S. J. D. Prince. *Understanding Deep Learning.* MIT Press, 2023.

[337] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

[338] C. Rackauckas, Y. Ma, J. Martensen, C. Warner, K. Zubov, R. Supekar, D. Skinner, A. Ramadhan, and A. Edelman. Universal differential equations for scientific machine learning. *arXiv:2001.04385*, 2020.

[339] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434*, 2015.

[340] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[341] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv:2112.11446*, 2021.

[342] L. Ramalho. *Fluent Python*. O'Reilly Media, Incorporated, 2021.

[343] R. Ranganath, D. Tran, and D. Blei. Hierarchical variational models. In *International conference on machine learning*. PMLR, 2016.

[344] W. Rawat and Z. Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 2017.

[345] S. J. Reddi, S. Kale, and S. Kumar. On the Convergence of Adam and Beyond. *arXiv:1904.09237*, 2019.

[346] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[347] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*. PMLR, 2015.

[348] S. Rezvani and X. Wang. A broad review on class imbalance learning techniques. *Applied Soft Computing*, 2023.

[349] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

[350] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, 1951.

[351] J. Rodriguez-Perez, C. Leigh, B. Liquet, C. Kermorvant, E. Peterson, D. Sous, and K. Mengersen. Detecting technical anomalies in high-frequency water-quality data using artificial neural networks. *Environmental Science & Technology*, 2020.

[352] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 2015.

[353] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 1958.

[354] H. H. Rosenbrock. An automatic method for finding the greatest or least value of a function. *The computer journal*, 1960.

[355] S. M. Ross. *A first course in probability*. Pearson, 2014.

[356] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv:1609.04747*, 2016.

[357] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 1986.

[358] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022.

[359] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. G. Lopes, B. K. Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022.

[360] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[361] R. Salakhutdinov and G. Hinton. Deep boltzmann machines. In *Artificial Intelligence and Statistics*, 2009.

[362] S. Sarao Mannelli and P. Urbani. Analytical study of momentum-based acceleration methods in paradigmatic high-dimensional non-convex problems. *Advances in Neural Information Processing Systems*, 2021.

[363] M. Sarigül and M. Avci. Performance comparison of different momentum techniques on deep reinforcement learning. *Journal of Information and Telecommunication*, 2018.

[364] N. Savage. How AI and Neuroscience Drive Each Other Forwards. *Nature*, 2019.

[365] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 2008.

[366] R. E. Schapire and Y. Freund. *Boosting: Foundations and Algorithms*. The MIT Press, 2012.

[367] J. Schmidhuber. Annotated history of modern AI and Deep learning. *arXiv:2212.11279*, 2022.

[368] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[369] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 1997.

[370] T. J. Sejnowski. *The Deep Learning Revolution*. MIT Press, 2018.

[371] Y. Seo, M. Defferrard, P. Vandergheynst, and X. Bresson. Structured sequence modeling with graph convolutional recurrent networks. In *Neural Information Processing: 25th International Conference, ICONIP, 2018, Proceedings, Part I 25*, 2018.

[372] P. Sermanet and Y. LeCun. Traffic Sign Recognition with Multi-Scale Convolutional Networks. In *The 2011 International Joint Conference on Neural Networks*, 2011.

[373] V. Sharma, M. Gupta, A. Kumar, and D. Mishra. Video Processing Using Deep Learning Techniques: A Systematic Literature Review. *IEEE Access*, 2021.

[374] A. Shashua and A. Levin. Ranking with large margin principle: Two approaches. *Advances in neural information processing systems*, 2002.

[375] Z. Shen, W. Bao, and D. S. Huang. Recurrent neural network for predicting transcription factor binding sites. *Scientific reports*, 2018.

[376] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv:1703.00810*, 2017.

[377] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016.

[378] J. S. Simonoff. *Smoothing Methods in Statistics*. Springer Science & Business Media, 2012.

[379] K. Simonyan and A. Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in Neural Information Processing Systems*, 2014.

[380] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014.

[381] S. A. Sisson, Y. Fan, and M. A. Beaumont, editors. *Handbook of Approximate Bayesian Computation*. CRC Press, Boca Raton, FL, 2019.

[382] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv:2201.11990*, 2022.

[383] I. Sobel. History and definition of the sobel operator. *Retrieved from the World Wide Web*, 2014.

[384] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

[385] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 2016.

[386] B. Speelpenning. *Compiling fast partial derivatives of functions given by algorithms*. University of Illinois at Urbana-Champaign, 1980.

[387] A. Sperduti and A. Starita. Supervised neural networks for the classification of structures. *IEEE Transactions on Neural Networks*, 1997.

[388] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 2014.

[389] S. M. Stigler. Gauss and the invention of least squares. *The Annals of Statistics*, 1981.

[390] P. Stoica and Y. Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 2004.

[391] G. Strang. *Linear algebra and learning from data*. Wellesley-Cambridge Press Cambridge, 2019.

[392] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 2014.

[393] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[394] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[395] M. Tan and Q. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning*, 2019.

[396] M. Tan and Q. V. Le. EfficientNetV2: Smaller Models and Faster Training. *International Conference on Machine Learning, PMLR*, 2021.

[397] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, 2015.

[398] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida. Deep Learning in Spiking Neural Networks. *Neural Networks*, 2019.

[399] M. Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, 2016.

[400] R. Thoppilan, D. D. Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. Lamda: Language models for dialog applications. *arXiv:2201.08239*, 2022.

[401] A. N. Tikhonov. On the stability of inverse problems. In *Comptes Rendus de l'Académie des Sciences de l'URSS*, 1943.

[402] A. C. Tsoi. Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, 1997.

[403] L. Tunstall, L. V. Werra, and T. Wolf. *Natural Language Processing with Transformers*. O'Reilly Media, Inc., 2022.

[404] A. M. Turing and J. Haugeland. *Computing Machinery and Intelligence*. MIT Press Cambridge, MA, 1950.

[405] I. Ulku and E. Akagündüz. A Survey on Deep Learning-Based Architectures for Semantic Segmentation on 2D Images. *Applied Artificial Intelligence*, 2022.

[406] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv:1607.08022*, 2016.

[407] R. van de Schoot, S. Depaoli, R. King, B. Kramer, K. Märtens, M. G. Tadesse, M. Vannucci, A. Gelman, D. Veen, J. Willemsen, and C. Yau. Bayesian Statistics and Modelling. *Nature Reviews Methods Primers*, 2021.

[408] C. Van Rijsbergen. Information Retrieval (Book 2nd ed), 1979.

[409] V. N. Vapnick. *Statistical Learning Theory*. Wiley, New York, 1998.

[410] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, 2017.

[411] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph Attention Networks. *International Conference on Learning Representations (ICLR)*, 2018.

[412] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang. Phoneme recognition using time-delay neural networks. In *Backpropagation*. 2013.

[413] L. Waikhom and R. Patgiri. A survey of graph neural networks in various learning paradigms: methods, applications, and challenges. *Artificial Intelligence Review*, 2023.

[414] J. Walker, C. Doersch, A. Gupta, and M. Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *Computer Vision–ECCV 2016: 14th European Conference, 2016, Proceedings, Part VII 14*, 2016.

[415] C. Y. Wang, A. Bochkovskiy, and H. Y. M. Liao. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[416] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *European Conference on Computer Vision*, 2016.

[417] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[418] Z. Wang, L. Zhao, and W. Xing. Styledifusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[419] C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 1992.

[420] R. E. Wengert. A simple automatic derivative evaluation program. *Communications of the ACM*, 1964.

[421] P. J. Werbos. Applications of advances in nonlinear sensitivity analysis. In *System Modeling and Optimization: Proceedings of the 10th IFIP Conference, 1981*, 2005.

[422] P. Wolfe. Convergence conditions for ascent methods. *SIAM Review*, 1969.

[423] P. Wolfe. Convergence conditions for ascent methods. II: Some corrections. *SIAM Review*, 1971.

[424] T. Wong and P. Yeh. Reliable accuracy estimates from k-fold cross-validation. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[425] H. Wu, Z. Xu, J. Zhang, W. Yan, and X. Ma. Face Recognition Based on Convolution Siamese Networks. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2017.

[426] Y. Wu and K. He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[427] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 2020.

[428] P. Xu, X. Zhu, and D. A. Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[429] A. Yang, B. Xiao, B. Wang, B. Zhang, C. Bian, C. Yin, C. Lv, D. Pan, D. Wang, D. Yan, et al. Baichuan 2: Open large-scale language models. *arXiv:2309.10305*, 2023.

[430] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2023.

[431] T. Yang, Q. Lin, and Z. Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv:1604.03257*, 2016.

[432] Y. Yang. Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*, 2005.

[433] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. XLNet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 2019.

[434] G. Yao, T. Lei, and J. Zhong. A Review of Convolutional-Neural-Network-Based Action Recognition. *Pattern Recognition Letters*, 2019.

[435] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122*, 2015.

[436] F. Yu, V. Koltun, and T. Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017.

[437] L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, 2017.

[438] Y. Yu, X. Si, C. Hu, and J. Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation*, 2019.

[439] Y. Yuan. Recent advances in trust region algorithms. *Mathematical Programming*, 2015.

[440] M. D. Zeiler. Adadelta: An adaptive learning rate method. *arXiv:1212.5701*, 2012.

[441] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 2014.

[442] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *2011 International Conference on Computer Vision*, 2011.

[443] X. Zeng and T. R. Martinez. Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence*, 2000.

[444] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt. Advances in variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[445] M. Zhang and Y. Chen. Link prediction based on graph neural networks. *Advances in neural information processing systems*, 2018.

[446] N. Zhang, S. Shen, A. Zhou, and Y. Jin. Application of lstm approach for modelling stress–strain behaviour of soil. *Applied Soft Computing*, 2021.

[447] Q. Zhang and S. Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 2018.

[448] W. Zhao, J. P. Queralta, and T. Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*, 2020.

[449] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Y. Nie, and J. R. Wen. A survey of large language models. *arXiv:2303.18223*, 2023.

[450] Y. Zhao, X. Li, W. Zhang, S. Zhao, M. Makkie, M. Zhang, Q. Li, and T. Liu. Modeling 4D fMRI Data via Spatio-Temporal Convolutional Neural Networks (ST-CNN). In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, 2018, Proceedings, Part III 11*, 2018.

[451] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun. Graph neural networks: A review of methods and applications. *AI open*, 2020.

[452] Q. Zhou, W. Chen, S. Song, J. Gardner, K. Weinberger, and Y. Chen. A reduction of the elastic net to support vector machines with an application to GPU computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

[453] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2005.

[454] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye. Object Detection in 20 Years: A Survey. *Proceedings of the IEEE*, 2023.

Index