Bibliography

T. Akiba, M. Shing, Y. Tang, et al. Evolutionary Optimization of Model Merging Recipes. *CoRR*, abs/2403.13187, 2024. [pdf]. 157

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer Normalization. *CoRR*, abs/1607.06450, 2016. [pdf]. 83

R. Balestriero, M. Ibrahim, V. Sobal, et al. A Cookbook of Self-Supervised Learning. *CoRR*, abs/2304.12210, 2023. [pdf]. 162

A. Baydin, B. Pearlmutter, A. Radul, and J. Siskind. Automatic differentiation in machine learning: a survey. *CoRR*, abs/1502.05767, 2015. [pdf]. 42

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine learning and the bias-variance trade-off. *CoRR*, abs/1812.11118, 2018. [pdf]. 50

I. Beltagy, M. Peters, and A. Cohan. Longformer: The Long-Document Transformer. *CoRR*, abs/2004.05150, 2020. [pdf]. 91

R. Bommasani, D. Hudson, E. Adeli, et al. On the Opportunities and Risks of Foundation Models. *CoRR*, abs/2108.07258, 2021. [pdf]. 140

J. Bradbury, S. Merity, C. Xiong, and R. Socher. Quasi-Recurrent Neural Networks. *CoRR*, abs/1611.01576, 2016. [pdf]. 159

T. Brown, B. Mann, N. Ryder, et al. Language Models are Few-Shot Learners. *CoRR*, abs/2005.14165, 2020. [pdf]. 54, 113, 139

S. Bubeck, V. Chandrasekaran, R. Eldan, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712, 2023. [pdf]. 140

T. Chen, B. Xu, C. Zhang, and C. Guestrin. Training Deep Nets with Sublinear Memory Cost. *CoRR*, abs/1604.06174, 2016. [pdf]. 43

K. Cho, B. van Merrienboer, Ç. Gülçehre, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *CoRR*, abs/1406.1078, 2014. [pdf]. 158

A. Chowdhery, S. Narang, J. Devlin, et al. PaLM: Scaling Language Modeling with Pathways. *CoRR*, abs/2204.02311, 2022. [pdf]. 9, 54, 140

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, December 1989. [pdf]. 99

J. Deng, W. Dong, R. Socher, et al. ImageNet: A Large-Scale Hierarchical Image Database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. [pdf]. 51

T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. *CoRR*, abs/2305.14314, 2023. [pdf]. 155

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805, 2018. [pdf]. 54, 115, 162

A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*, abs/2010.11929, 2020. [pdf]. 113, 114

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980. [pdf]. 2

Y. Gal and Z. Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *CoRR*, abs/1506.02142, 2015. [pdf]. 78

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010. [pdf]. 44, 62

X. Glorot, A. Bordes, and Y. Bengio. Deep Sparse Rectifier Neural Networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011. [pdf]. 71

A. Gomez, M. Ren, R. Urtasun, and R. Grosse. The Reversible Residual Network: Backpropagation Without Storing Activations. *CoRR*, abs/1707.04585, 2017. [pdf]. 43

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. Generative Adversarial Networks. *CoRR*, abs/1406.2661, 2014. [pdf]. 160

A. Gu and T. Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *CoRR*, abs/2312.00752, 2023. [pdf]. 159

A. Gu, K. Goel, and C. Ré. Efficiently Modeling Long Sequences with Structured State Spaces. *CoRR*, abs/2111.00396, 2021. [pdf]. 159

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. *CoRR*, abs/1512.03385, 2015. [pdf]. 52, 84, 85, 103, 105

D. Hendrycks and K. Gimpel. Gaussian Error Linear Units (GELUs). *CoRR*, abs/1606.08415, 2016. [pdf]. 73

D. Hendrycks, K. Zhao, S. Basart, et al. Natural Adversarial Examples. *CoRR*, abs/1907.07174, 2019. [pdf]. 132

J. Ho, A. Jain, and P. Abbeel. Denoising Diffusion Probabilistic Models. *CoRR*, abs/2006.11239, 2020. [pdf]. 142, 143, 144

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. [pdf]. 158

N. Houlsby, A. Giurgiu, S. Jastrzebski, et al. Parameter-Efficient Transfer Learning for NLP. *CoRR*, abs/1902.00751, 2019. [pdf]. 153

E. Hu, Y. Shen, P. Wallis, et al. LoRA: Low-Rank Adaptation of Large Language Models. *CoRR*, abs/2106.09685, 2021. [pdf]. 153

G. Ilharco, M. Ribeiro, M. Wortsman, et al. Editing Models with Task Arithmetic. *CoRR*, abs/2212.04089, 2022. [pdf]. 156

S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning (ICML)*, 2015. [pdf]. 80

A. Jiang, A. Sablayrolles, A. Mensch, et al. Mistral 7B. *CoRR*, abs/2310.06825, 2023. [pdf]. 157

J. Kaplan, S. McCandlish, T. Henighan, et al. Scaling Laws for Neural Language Models. *CoRR*, abs/2001.08361, 2020. [pdf]. 52, 53

A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5294–5303, 2020. [pdf]. 91

D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6980, 2014. [pdf]. 39

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *CoRR*, abs/1312.6114, 2013. [pdf]. 160

T. Kojima, S. Gu, M. Reid, et al. Large Language Models are Zero-Shot Reasoners. *CoRR*, abs/2205.11916, 2022. [pdf]. 149

A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Neural Information Processing Systems (NIPS)*, 2012. [pdf]. 8, 101

Y. LeCun, B. Boser, J. S. Denker, et al. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. [pdf]. 8

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324, 1998. [pdf]. 101, 102

P. Lewis, E. Perez, A. Piktus, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *CoRR*, abs/2005.11401, 2020. [pdf]. 149

W. Liu, D. Anguelov, D. Erhan, et al. SSD: Single Shot MultiBox Detector. *CoRR*, abs/1512.02325, 2015. [pdf]. 121, 123

Llama.cpp. Llama.cpp git repository, June 2023. [web]. 150, 151

J. Long, E. Shelhamer, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *CoRR*, abs/1411.4038, 2014. [pdf]. 84, 85, 127

S. Ma, H. Wang, L. Ma, et al. The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. *CoRR*, abs/2402.17764, 2024. [pdf]. 152

A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *proceedings of the ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013. [pdf]. 72

V. Mnih, K. Kavukcuoglu, D. Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. [pdf]. 135, 136

A. Nichol, P. Dhariwal, A. Ramesh, et al. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *CoRR*, abs/2112.10741, 2021. [pdf]. 145

L. Ouyang, J. Wu, X. Jiang, et al. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155, 2022. [pdf]. 141

R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning (ICML)*, 2013. [pdf]. 44

A. Radford, J. Kim, C. Hallacy, et al. Learning Transferable Visual Models From Natural Language Supervision. *CoRR*, abs/2103.00020, 2021. [pdf]. 131, 133

A. Radford, J. Kim, T. Xu, et al. Robust Speech Recognition via Large-Scale Weak Supervision. *CoRR*, abs/2212.04356, 2022. [pdf]. 129

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving Language Understanding by Generative Pre-Training, 2018. [pdf]. 109, 112, 139

A. Radford, J. Wu, R. Child, et al. Language Models are Unsupervised Multitask Learners, 2019. [pdf]. 112, 162

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015. [pdf]. 84, 85, 127

P. Sahoo, A. Singh, S. Saha, et al. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. *CoRR*, abs/2402.07927, 2024. [pdf]. 147

F. Scarselli, M. Gori, A. C. Tsoi, et al. The Graph Neural Network Model. *IEEE Transactions on Neural Networks (TNN)*, 20(1):61–80, 2009. [pdf]. 161

R. Sennrich, B. Haddow, and A. Birch. Neural Machine Translation of Rare Words with Subword Units. *CoRR*, abs/1508.07909, 2015. [pdf]. 34

J. Sevilla, L. Heim, A. Ho, et al. Compute Trends Across Three Eras of Machine Learning. *CoRR*, abs/2202.05924, 2022. [pdf]. 8, 52, 54

J. Sevilla, P. Villalobos, J. F. Cerón, et al. Parameter, Compute and Data Trends in Machine Learning, May 2023. [web]. 55

K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*, abs/1409.1556, 2014. [pdf]. 101

N. Srivastava, G. Hinton, A. Krizhevsky, et al. Dropout: A Simple Way to Prevent Neural

Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15:1929–1958, 2014. [pdf]. 77

M. Telgarsky. Benefits of depth in neural networks. *CoRR*, abs/1602.04485, 2016. [pdf]. 47

H. Touvron, T. Lavril, G. Izacard, et al. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971, 2023. [pdf]. 151

A. Vaswani, N. Shazeer, N. Parmar, et al. Attention Is All You Need. *CoRR*, abs/1706.03762, 2017. [pdf]. 84, 87, 97, 108, 109, 110

J. Wei, X. Wang, D. Schuurmans, et al. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *CoRR*, abs/2201.11903, 2022. [pdf]. 149

B. Xu, A. Yang, J. Lin, et al. ExpertPrompting: Instructing Large Language Models to be Distinguished Experts. *CoRR*, abs/2305.14688, 2023. [pdf]. 147

P. Yadav, D. Tam, L. Choshen, et al. TIES-Merging: Resolving Interference When Merging Models. *CoRR*, abs/2306.01708, 2023. [pdf]. 157

L. Yu, B. Yu, H. Yu, et al. Language Models are Super Mario: Absorbing Abilities from

Homologous Models as a Free Lunch. *CoRR*, abs/2311.03099, 2023. [pdf]. 157

J. Zbontar, L. Jing, I. Misra, et al. Barlow Twins: Self-Supervised Learning via Redundancy Reduction. *CoRR*, abs/2103.03230, 2021. [pdf]. 162

M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision (ECCV)*, 2014. [pdf]. 69

H. Zhao, J. Shi, X. Qi, et al. Pyramid Scene Parsing Network. *CoRR*, abs/1612.01105, 2016. [pdf]. 127, 128

J. Zhou, C. Wei, H. Wang, et al. iBOT: Image BERT Pre-Training with Online Tokenizer. *CoRR*, abs/2111.07832, 2021. [pdf]. 163