

Рёма Сато

Национальный институт информатики, Чиода, Япония

Ускорение глубоких нейронных сетей



Москва, 2027

Справочная литература

Abouelenin, Abdelrahman, Ashfaq, Atabak, Atkinson, Adam, et al. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. arXiv, **abs/2503.01743**.

Ainsworth, Samuel K., Hayase, Jonathan, and Srinivasa, Siddhartha S. 2023. Git ReBasin: Merging Models modulo Permutation Symmetries. In: *Proceedings of the 11th International Conference on Learning Representations, ICLR, Kigali, Rwanda, May 1–5, 2023*. <https://openreview.net/forum?id=CQsmMYmlP5T>.

Akiba, Takuya, Shing, Makoto, Tang, Yujin, Sun, Qi, and Ha, David. 2025. Evolutionary Optimization of Model Merging Recipes. *Nat. Mach. Intell.*, **7**, 195–204. <https://doi.org/10.1038/s42256-024-00975-8>.

Allen-Zhu, Zeyuan, and Li, Yuanzhi. 2023. Towards Understanding Ensemble, Knowledge Distillation and Self-Distillation in Deep Learning. In: *Proceedings of the 11th International Conference on Learning Representations, ICLR, Kigali, Rwanda, May 1–5, 2023*. <https://openreview.net/forum?id=Uuf2q9TfXGA>.

Alquier, Pierre. 2024. User-Friendly Introduction to PAC-Bayes Bounds. *Found. Trends Mach. Learn.*, **17** (2), 174–303.

Aminabadi, Reza Yazdani, Rajbhandari, Samyam, Awan, Ammar Ahmad, et al. 2022. DeepSpeed-Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale. Pages 1–15 of: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC. IEEE.

Ao, Shuang, Li, Xiang, and Ling, Charles X. 2017. Fast Generalized Distillation for Semi-Supervised Domain Adaptation. Pages 1719–1725 of: *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. AAAI Press.

Arora, Sanjeev, Liang, Yingyu, and Ma, Tengyu. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In: *Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, France, April 24–26, 2017*. <https://openreview.net/forum?id=SyK00v5xx>.

Arora, Sanjeev, Cohen, Nadav, Hu, Wei, and Luo, Yuping. 2019. Implicit Regularization in Deep Matrix Factorization. Pages 7411–7422 of: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS*. Curran Associates.

Arora, Sanjeev, Li, Zhiyuan, and Panigrahi, Abhishek. 2022. Understanding Gradient Descent on the Edge of Stability in Deep Learning. Pages 948–1024 of: *Proceedings of the 39th International Conference on Machine Learning, ICML*. PMLR.

Ba, Jimmy, and Caruana, Rich. 2014. Do Deep Nets Really Need to Be Deep? Pages 2654–2662 of: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, NeurIPS*. Curran Associates.

Bai, Jinze, Bai, Shuai, Chu, Yunfei, et al. 2023. Qwen Technical Report. arXiv, **abs/2309.16609**.

Banner, Ron, Hubara, Itay, Hoffer, Elad, and Soudry, Daniel. 2018. Scalable Methods for 8-Bit Training of Neural Networks. Pages 5151–5159 of: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS*. Curran Associates.

Barrett, David G. T., and Dherin, Benoit. 2021. Implicit Gradient Regularization. In: *Proceedings of the 9th International Conference on Learning Representations, ICLR, Virtual Event, May 3–7, 2021*. <https://openreview.net/forum?id=3q5lqUrkcF>.

Bartoldson, Brian R., Morcos, Ari S., Barbu, Adrian, and Erlebacher, Gordon. 2020. The Generalization–Stability Tradeoff in Neural Network Pruning. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*. Curran Associates.

Bellec, Guillaume, Kappel, David, Maass, Wolfgang, and Legenstein, Robert. 2018. Deep Rewiring: Training Very Sparse Deep Networks. In: *Proceedings of the 6th International Conference on Learning Representations, ICLR, Vancouver, BC, Canada, April 30–May 3, 2018*. https://openreview.net/forum?id=BJ_wN01C-.

Bengio, Yoshua, Léonard, Nicholas, and Courville, Aaron C. 2013. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. arXiv, **abs/1308.3432**.

Blalock, Davis W., Ortiz, Jose Javier Gonzalez, Frankle, Jonathan, and Gutttag, John V. 2020. What Is the State of Neural Network Pruning? In: *Proceedings of Machine Learning and Systems 2020, MLSys Austin, TX, USA, March 2–4, 2020*. https://proceedings.mlsys.org/paper_files/paper/2020/hash/6c44dc73014d66ba49b28d483a8f8b0d-Abstract.html.

Brown, Tom B., Mann, Benjamin, Ryder, Nick, et al. 2020. Language Models Are FewShot Learners. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*. Curran Associates.

Bucila, Cristian, Caruana, Rich, and Niculescu-Mizil, Alexandru. 2006. Model Compression. Pages 535–541 of: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD*. ACM.

Cai, Han, Gan, Chuang, Wang, Tianzhe, Zhang, Zhekai, and Han, Song. 2020. Oncefor-All: Train One Network and Specialize It for Efficient Deployment. In: *Proceedings of the 8th International Conference on Learning Representations, ICLR, Virtual Event, April 26–30, 2020*. <https://openreview.net/forum?id=HylxE1HKwS>.

Cai, Tianle, Li, Yuhong, Geng, Zhengyang, et al. 2024. Medusa: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads. In: *Proceedings of the 41st International Conference on Machine Learning, ICML*. PMLR.

Cervenka, Eole. 2022. All You Need Is One GPU: Inference Benchmark for Stable Diffusion. Lambda Labs. <https://lambdalabs.com/blog/inference-benchmark-stable-diffusion>. Accessed December 19, 2023.

Charikar, Moses. 2002. Similarity Estimation Techniques from Rounding Algorithms. Pages 380–388 of: *Proceedings of 34th Annual ACM Symposium on Theory of Computing, STOC*. ACM.

Chaudhari, Pratik, Choromanska, Anna, Soatto, Stefano, et al. 2017. Entropy-SGD: Biasing Gradient Descent into Wide Valleys. In: *Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, France, April 24–26, 2017*. <https://openreview.net/forum?id=B1YfAfcgl>.

Chee, Jerry, Cai, Yaohui, Kuleshov, Volodymyr, and Sa, Christopher De. 2023. QuIP: 2-Bit Quantization of Large Language Models with Guarantees. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS*. Curran Associates.

Chen, Angelica, Schwartz-Ziv, Ravid, Cho, Kyunghyun, Leavitt, Matthew L., and Saphra, Naomi. 2024. Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLs. In: *Proceedings of the 12th International Conference on Learning Representations, ICLR, Vienna, Austria, May 7–11, 2024*. <https://openreview.net/forum?id=MO5PiKHELW>.

Chen, Charlie, Borgeaud, Sebastian, Irving, Geoffrey, et al. 2023. Accelerating Large Language Model Decoding with Speculative Sampling. arXiv, **abs/2302.01318**.

Chen, Hanting, Wang, Yunhe, Xu, Chang, et al. 2019. Data-Free Learning of Student Networks. Pages 3513–3521 of: *Proceedings of the 2019 IEEE International Conference on Computer Vision, ICCV*. IEEE.

Chen, Lingjiao, Zaharia, Matei, and Zou, James Y. 2020. FrugalML: How to Use ML Prediction APIs More Accurately and Cheaply. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*. Curran Associates.

Chen, Ting, Kornblith, Simon, Swersky, Kevin, Norouzi, Mohammad, and Hinton, Geoffrey E. 2020. Big Self-Supervised Models Are Strong Semi-Supervised Learners. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*. Curran Associates.

Chijiwa, Daiki. 2024. Transferring Learning Trajectories of Neural Networks. In: *Proceedings of the 12th International Conference on Learning Representations, ICLR, Vienna, Austria, May 7–11, 2024*. <https://openreview.net/forum?id=bWNJFD1l8M>.

Child, Rewon, Gray, Scott, Radford, Alec, and Sutskever, Ilya. 2019. Generating Long Sequences with Sparse Transformers. arXiv, **abs/1904.10509**.

Choi, Jungwook, Wang, Zhuo, Venkataramani, Swagath, et al. 2018. PACT: Parameterized Clipping Activation for Quantized Neural Networks. arXiv, **abs/1805.06085**.

Chollet, François. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. Pages 1800–1807 of: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

Choromanski, Krzysztof Marcin, Likhoshesterov, Valerii, Dohan, David, et al. 2021. Rethinking Attention with Performers. In: *Proceedings of the 9th International Conference on Learning Representations, ICLR, Virtual Event, May 3–7, 2021*. <https://openreview.net/forum?id=Ua6zukOWRH>.

Chu, Xiangxiang, Tian, Zhi, Zhang, Bo, Wang, Xinlong, and Shen, Chunhua. 2023. Conditional Positional Encodings for Vision Transformers. In: *Proceedings*

of the 11th International Conference on Learning Representations, ICLR, Kigali, Rwanda, May 1–5, 2023. <https://openreview.net/forum?id=3KWnuT-R1bh>.

Clark, Kevin, Khandelwal, Urvashi, Levy, Omer, and Manning, Christopher D. 2019. What Does BERT Look at? An Analysis of BERT’s Attention. Pages 276–286 of: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. ACL.

Cohen, Jeremy, Kaur, Simran, Li, Yuanzhi, Kolter, J. Zico, and Talwalkar, Ameet. 2021. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability. In: *Proceedings of the 9th International Conference on Learning Representations, ICLR, Virtual Event, May 3–7, 2021*. <https://openreview.net/forum?id=jh-rTtvkGeM>.

Cortes, Corinna, DeSalvo, Giulia, and Mohri, Mehryar. 2016. Learning with Rejection. Pages 67–82 of: *Proceedings of the 27th International Conference on Algorithmic Learning Theory, ALT*. Springer.

Courbariaux, Matthieu, Bengio, Yoshua, and David, Jean-Pierre. 2015. Binary-Connect: Training Deep Neural Networks with Binary Weights during Propagations. Pages 3123–3131 of: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, NeurIPS*. Curran Associates.

Croitoru, Ioana, Bogolin, Simion-Vlad, Leordeanu, Marius, et al. 2021. Teach-Text: CrossModal Generalized Distillation for Text-Video Retrieval. Pages 11563–11573 of: *Proceedings of the 2019 IEEE International Conference on Computer Vision, ICCV*. IEEE.

Dao, Tri. 2024. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. In: *Proceedings of the 12th International Conference on Learning Representations, ICLR, Vienna, Austria, May 7–11, 2024*. <https://openreview.net/forum?id=mZn2Xyh9Ec>.

Dao, Tri, Fu, Daniel Y., Ermon, Stefano, Rudra, Atri, and Ré, Christopher. 2022. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS*. Curran Associates.

Dehghani, Mostafa, Gouws, Stephan, Vinyals, Oriol, Uszkoreit, Jakob, and Kaiser, Lukasz. 2019. Universal Transformers. In: *Proceedings of the 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, May 6–9, 2019*. <https://openreview.net/forum?id=HyzdRiR9Y7>.

Dettmers, Tim, Lewis, Mike, Shleifer, Sam, and Zettlemoyer, Luke. 2022a. 8-Bit Optimizers via Block-wise Quantization. In: *Proceedings of the 10th International Conference on Learning Representations, ICLR, Virtual Event, April 25–29, 2022*. <https://openreview.net/forum?id=shpkpVXzo3h>.

Dettmers, Tim, Lewis, Mike, Belkada, Younes, and Zettlemoyer, Luke. 2022b. GPT3.int8(): 8-Bit Matrix Multiplication for Transformers at Scale. In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS*. Curran Associates.

Dettmers, Tim, Pagnoni, Artidoro, Holtzman, Ari, and Zettlemoyer, Luke. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. In: *Advances in Neural Informa-*

tion Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, *NeurIPS*. Curran Associates.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Pages 4171–4186 of: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. ACL.

Dinh, Laurent, Pascanu, Razvan, Bengio, Samy, and Bengio, Yoshua. 2017. Sharp Minima Can Generalize for Deep Nets. Pages 1019–1028 of: *Proceedings of the 34th International Conference on Machine Learning, ICML*. PLMR.

Dosovitskiy, Alexey, Beyer, Lucas, Kolesnikov, Alexander, et al. 2021. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: *Proceedings of the 9th International Conference on Learning Representations, ICLR, Virtual Event, May 3–7, 2021*. <https://openreview.net/forum?id=YicbFdNTTy>.

Draxler, Felix, Veschgini, Kambis, Salmhofer, Manfred, and Hamprecht, Fred A. 2018. Essentially No Barriers in Neural Network Energy Landscape. Pages 1308–1317 of: *Proceedings of the 35th International Conference on Machine Learning, ICML*. PLMR.

Elbayad, Maha, Gu, Jiatao, Grave, Edouard, and Auli, Michael. 2020. Depth-Adaptive Transformer. In: *Proceedings of the 8th International Conference on Learning Representations, ICLR, Virtual Event, April 26–30, 2020*. <https://openreview.net/forum?id=Slg7KhVKPH>.

Elfwing, Stefan, Uchibe, Eiji, and Doya, Kenji. 2018. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. *Neural Networks*, 107, 3–11.

Elhoushi, Mostafa, Shrivastava, Akshat, Liskovich, Diana, et al. 2024. LayerSkip: Enabling Early Exit Inference and Self-Speculative Decoding. Pages 12622–12642 of: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL.

Entezari, Rahim, Sedghi, Hanie, Saukh, Olga, and Neyshabur, Behnam. 2022. The Role of Permutation Invariance in Linear Mode Connectivity of Neural Networks. In: *Proceedings of the 10th International Conference on Learning Representations, ICLR, Virtual Event, April 25–29, 2022*. <https://openreview.net/forum?id=dNigytemkL>.

Evci, Utku, Gale, Trevor, Menick, Jacob, Castro, Pablo Samuel, and Elsen, Erich. 2020. Rigging the Lottery: Making All Tickets Winners. Pages 2943–2952 of: *Proceedings of the 37th International Conference on Machine Learning, ICML*. PMLR.

Fakoor, Rasool, Mueller, Jonas, Erickson, Nick, Chaudhari, Pratik, and Smola, Alexander J. 2020. Fast, Accurate, and Simple Models for Tabular Data via Augmented Distillation. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*. Curran Associates.

Fang, Zhiyuan, Wang, Jianfeng, Wang, Lijuan, Zhang, Lei, Yang, Yezhou, and Liu, Zicheng. 2021. SEED: Self-Supervised Distillation for Visual Representation. In: *Proceedings of the 9th International Conference on Learning Representations, ICLR, Virtual Event, May 3–7, 2021*. <https://openreview.net/forum?id=AHm3dbp7D1D>.

Fedyunin, Vitaly. 2020. *Channels Last Memory Format in PyTorch*. PyTorch. https://pytorch.org/tutorials/intermediate/memory_format_tutorial.html. Accessed June 4, 2024.

Fernandez, Jared, Kahn, Jacob, Na, Clara, Bisk, Yonatan, and Strubell, Emma. 2023. The Framework Tax: Disparities Between Inference Efficiency in NLP Research and Deployment. Pages 1588–1600 of: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*. ACL.

Foret, Pierre, Kleiner, Ariel, Mobahi, Hossein, and Neyshabur, Behnam. 2021. Sharpness-Aware Minimization for Efficiently Improving Generalization. In: *Proceedings of the 9th International Conference on Learning Representations, ICLR, Virtual Event, May 3–7, 2021*. <https://openreview.net/forum?id=6Tm1mposlRm>.

Fort, Stanislav, Hu, Huiyi, and Lakshminarayanan, Balaji. 2019. Deep Ensembles: A Loss Landscape Perspective. arXiv, **abs/1912.02757**.

Frankle, Jonathan, and Carbin, Michael. 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In: *Proceedings of the 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, May 6–9, 2019*. <https://openreview.net/forum?id=rJl-b3RcF7>.

Frankle, Jonathan, Dziugaite, Gintare Karolina, Roy, Daniel M., and Carbin, Michael. 2020. Linear Mode Connectivity and the Lottery Ticket Hypothesis. Pages 3259–3269 of: *Proceedings of the 37th International Conference on Machine Learning, ICML*. PMLR.

Frantar, Elias, Ashkboos, Saleh, Hoefler, Torsten, and Alistarh, Dan. 2023. OPTQ: Accurate Quantization for Generative Pre-trained Transformers. In: *Proceedings of the 11th International Conference on Learning Representations, ICLR, Kigali, Rwanda, May 1–5, 2023*. <https://openreview.net/pdf?id=tcbBPnfwxS>.

Fu, Daniel Y., Arora, Simran, Grogan, Jessica, et al. 2023. Monarch Mixer: A Simple Sub-Quadratic GEMM-Based Architecture. In: *NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates.

Fu, Yichao, Bailis, Peter, Stoica, Ion, and Zhang, Hao. 2024. Break the Sequential Dependency of LLM Inference Using Lookahead Decoding. In: *Proceedings of the 41st International Conference on Machine Learning, ICML*. PLMR.

Gal, Yarin, and Ghahramani, Zoubin. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Pages 1050–1059 of: *Proceedings of the 33rd International Conference on Machine Learning, ICML*. PMLR.

Gale, Trevor, Zaharia, Matei, Young, Cliff, and Elsen, Erich. 2020. Sparse GPU Kernels for Deep Learning. Page 17 of: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC*. IEEE/ACM.

Garipov, Timur, Izmailov, Pavel, Podoprikin, Dmitrii, Vetrov, Dmitry P., and Wilson, Andrew Gordon. 2018. Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. Pages 8803–8812 of: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS*. Curran Associates.

Ge, Suyu, Zhang, Yunan, Liu, Liyuan, et al. 2024. Model Tells You What to Discard: Adaptive KV Cache Compression for LLMs. In: *Proceedings of the 12th*

International Conference on Learning Representations, ICLR, Vienna, Austria, May 7–11, 2024. <https://openreview.net/forum?id=uNrFpDPMYo>.

Georgiadis, Georgios. 2019. Accelerating Convolutional Neural Networks via Activation Map Compression. Pages 7085–7095 of: *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

Goodfellow, Ian, Bengio, Yoshua, and Courville, Aaron. 2016. *Deep Learning*. MIT Press. www.deeplearningbook.org.

Google. 2022. *Model Optimization*. Google AI Edge. https://ai.google.dev/edge/litert/models/model_optimization. Accessed May 23, 2025.

Gou, Jianping, Yu, Baosheng, Maybank, Stephen J., and Tao, Dacheng. 2021. Knowledge Distillation: A Survey. *Int. J. Comput. Vis.*, 129 (6), 1789–1819.

Goyal, Priya, Dollár, Piotr, Girshick, Ross B., et al. 2017. Accurate, Large Mini-batch SGD: Training ImageNet in 1 Hour. arXiv, **abs/1706.02677**.

Graves, Alex. 2016. Adaptive Computation Time for Recurrent Neural Networks. arXiv, **abs/1603.08983**.

Grinsztajn, Léo, Oyallon, Edouard, and Varoquaux, Gaël. 2022. Why Do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data? In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS*. Curran Associates.

Gromov, Andrey, Tirumala, Kushal, Shapourian, Hassan, Glorioso, Paolo, and Roberts, Daniel A. 2025. The Unreasonable Ineffectiveness of the Deeper Layers. In: *The 13th International Conference on Learning Representations, ICLR, Singapore, April 24–28, 2025.* <https://openreview.net/forum?id=ngmEcEer8a>.

Gu, Albert, and Dao, Tri. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv, **abs/2312.00752**.

Gu, Albert, Johnson, Isys, Goel, Karan, et al. 2021. Combining Recurrent, Convolutional, and Continuous-Time Models with Linear State Space Layers. Pages 572–585 of: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS*. Curran Associates.

Gu, Albert, Goel, Karan, and Ré, Christopher. 2022. Efficiently Modeling Long Sequences with Structured State Spaces. In: *Proceedings of the 10th International Conference on Learning Representations, ICLR, Virtual Event, April 25–29, 2022.* <https://openreview.net/forum?id=uYLFoz1vIAC>.

Guedj, Benjamin. 2019. A Primer on PAC-Bayesian Learning. arXiv, **abs/1901.05353**.

Gupta, Suyog, Agrawal, Ankur, Gopalakrishnan, Kailash, and Narayanan, Pritish. 2015. Deep Learning with Limited Numerical Precision. Pages 1737–1746 of: *Proceedings of the 32nd International Conference on Machine Learning, ICML*. PMLR.

Han, Cheng, Wang, Qifan, Cui, Yiming, et al. 2024. Facing the Elephant in the Room: Visual Prompt Tuning or Full Finetuning? In: *Proceedings of the 12th International Conference on Learning Representations, ICLR, Vienna, Austria, May 7–11, 2024.* <https://openreview.net/forum?id=blx4iOI0xn>.

Han, Song, Pool, Jeff, Tran, John, and Dally, William J. 2015. Learning Both Weights and Connections for Efficient Neural Network. Pages 1135–1143 of: *Ad-*

vances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, NeurIPS. Curran Associates.

Han, Song, Mao, Huizi, and Dally, William J. 2016. Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding. In: *Proceedings of the 4th International Conference on Learning Representations, ICLR, San Juan, Puerto Rico, May 2–4, 2016*. <https://openreview.net/forum?id=n4Bp24yb5L>.

Hassibi, Babak, and Stork, David G. 1992. Second-Order Derivatives for Network Pruning: Optimal Brain Surgeon. Pages 164–171 of: *Advances in Neural Information Processing Systems 5, NeurIPS*. Morgan Kaufmann.

He, Jiaao, and Zhai, Jidong. 2024. FastDecode: High-Throughput GPU-Efficient LLM Serving Using Heterogeneous Pipelines. arXiv, **abs/2403.11421**.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. 2014. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. Pages 346–361 of: *Proceedings of the 13th European Conference on Computer Vision, ECCV*. Springer.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. 2016. Deep Residual Learning for Image Recognition. Pages 770–778 of: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

He, Yihui, Zhang, Xiangyu, and Sun, Jian. 2017. Channel Pruning for Accelerating Very Deep Neural Networks. Pages 1398–1406 of: *Proceedings of the 2017 IEEE International Conference on Computer Vision, ICCV*. IEEE.

Hendrycks, Dan, and Gimpel, Kevin. 2016. Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units. arXiv, **abs/1606.08415**.

Hinton, Geoffrey E., and van Camp, Drew. 1993. Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights. Pages 5–13 of: *Proceedings of the 6th Annual ACM Conference on Computational Learning Theory, COLT*. ACM.

Hinton, Geoffrey E., Vinyals, Oriol, and Dean, Jeffrey. 2014. *Dark Knowledge*. www.ttic.edu/dl/dark14.pdf.

Hinton, Geoffrey E., Vinyals, Oriol, and Dean, Jeffrey. 2015. Distilling the Knowledge in a Neural Network. arXiv, **abs/1503.02531**.

Hochreiter, Sepp, and Schmidhuber, Jürgen. 1997. Flat Minima. *Neural Comput.*, 9 (1), 1–42.

Hoeffler, Torsten, Alistarh, Dan, Ben-Nun, Tal, Dryden, Nikoli, and Peste, Alexandra. 2021. Sparsity in Deep Learning: Pruning and Growth for Efficient Inference and Training in Neural Networks. *J. Mach. Learn. Res.*, 22 (241), 1–124.

Hoffer, Elad, Hubara, Itay, and Soudry, Daniel. 2017. Train Longer, Generalize Better: Closing the Generalization Gap in Large Batch Training of Neural Networks. Pages 1731–1741 of: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, NeurIPS*. Curran Associates.

Horn, Roger A., and Johnson, Charles R. 2012. *Matrix Analysis*. Cambridge University Press.

Howard, Andrew, Pang, Ruoming, Adam, Hartwig, et al. 2019. Searching for MobileNetV3. Pages 1314–1324 of: *Proceedings of the 2019 IEEE International Conference on Computer Vision, ICCV*. IEEE.

Howard, Andrew G., Zhu, Menglong, Chen, Bo, et al. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv, **abs/1704.04861**.

Hu, Edward J., Shen, Yelong, Wallis, Phillip, et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In: *Proceedings of the 10th International Conference on Learning Representations, ICLR, Virtual Event, April 25–29, 2022*. <https://openreview.net/forum?id=nZeVKeeFYf9>.

Huang, Gao, Liu, Zhuang, van der Maaten, Laurens, and Weinberger, Kilian Q. 2017. Densely Connected Convolutional Networks. Pages 2261–2269 of: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

Huang, Gao, Chen, Danlu, Li, Tianhong, et al. 2018. Multi-Scale Dense Networks for Resource Efficient Image Classification. In: *Proceedings of the 6th International Conference on Learning Representations, ICLR, Vancouver, BC, Canada, April 30–May 3, 2018*. <https://openreview.net/forum?id=Hk2almxAb>.

Huang, W. Ronny, Emam, Zeyad, Goldblum, Micah, et al. 2020. Understanding Generalization through Visualizations. Pages 87–97 of: *“I Can’t Believe It’s Not Better!” at NeurIPS Workshops*. PLMR.

Hubara, Itay, Courbariaux, Matthieu, Soudry, Daniel, El-Yaniv, Ran, and Bengio, Yoshua. 2016. Binarized Neural Networks. Pages 4107–4115 of: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, NeurIPS*. Morgan Kaufmann.

IEEE Standards Board. 1985. IEEE Standard for Binary Floating-Point Arithmetic. ANSI/IEEE Std 754-1985, 1–20.

Ilharco, Gabriel, Ribeiro, Marco Túlio, Wortsman, Mitchell, et al. 2023. Editing Models with Task Arithmetic. In: *Proceedings of the 11th International Conference on Learning Representations, ICLR, Kigali, Rwanda, May 1–5, 2023*. <https://openreview.net/forum?id=6tOKwf8-jrj>.

Ilyas, Andrew, Santurkar, Shibani, Tsipras, Dimitris, et al. 2019. Adversarial Examples Are Not Bugs, They Are Features. Pages 125–136 of: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS*. Morgan Kaufmann.

Indyk, Piotr, and Motwani, Rajeev. 1998. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. Pages 604–613 of: *Proceedings of the 30th Annual ACM Symposium on the Theory of Computing, STOC*. ACM.

Intel. 2023. Intel Architecture Instruction Set Extensions and Future Features Programming Reference. Intel. www.intel.com/content/www/us/en/content-details/671368/intel-architecture-instruction-set-extensions-programming-reference.html.

Ioannou, Yani, Robertson, Duncan P., Shotton, Jamie, Cipolla, Roberto, and Criminisi, Antonio. 2016. Training CNNs with Low-Rank Filters for Efficient Im-

ge Classification. In: *Proceedings of the 4th International Conference on Learning Representations, ICLR, San Juan, Puerto Rico, May 2–4, 2016*. <https://openreview.net/forum?id=oD-rwshwqzh>.

Islam, Md. Amirul, Jia, Sen, and Bruce, Neil D. B. 2020. How Much Position Information Do Convolutional Neural Networks Encode? In: *Proceedings of the 8th International Conference on Learning Representations, ICLR, Virtual Event, April 26–30, 2020*. <https://openreview.net/forum?id=rJeB36NKvB>.

Izmailov, Pavel, Podoprikin, Dmitrii, Garipov, Timur, Vetrov, Dmitry P., and Wilson, Andrew Gordon. 2018. Averaging Weights Leads to Wider Optima and Better Generalization. Pages 876–885 of: *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence, UAI*. AUAI Press.

Jacob, Benoit, Kligys, Skirmantas, Chen, Bo, et al. 2018. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. Pages 2704–2713 of: *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

Jaderberg, Max, Vedaldi, Andrea, and Zisserman, Andrew. 2014. Speeding up Convolutional Neural Networks with Low-Rank Expansions. In: *Proceedings of the 25th British Machine Vision Conference*. British Machine Vision Association.

Jang, Eric, Gu, Shixiang, and Poole, Ben. 2017. Categorical Reparameterization with Gumbel-Softmax. In: *Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, France, April 24–26, 2017*. <https://openreview.net/forum?id=rkE3y85ee>.

Jiang, Albert Q., Sablayrolles, Alexandre, Mensch, Arthur, et al. 2023. Mistral 7B. arXiv, **abs/2310.06825**.

Jiang, Huiqiang, Wu, Qianhui, Lin, Chin-Yew, Yang, Yuqing, and Qiu, Lili. 2023. LLMingua: Compressing Prompts for Accelerated Inference of Large Language Models. Pages 13358–13376 of: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*. ACL.

Jiao, Xiaoqi, Yin, Yichun, Shang, Lifeng, et al. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. Pages 4163–4174 of: *Findings of the Association for Computational Linguistics: EMNLP*. ACL.

Jin, Tian, Carbin, Michael, Roy, Daniel M., Frankle, Jonathan, and Dziugaite, Gintare Karolina. 2022. Pruning’s Effect on Generalization Through the Lens of Training and Regularization. In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS*. Curran Associates.

Juneja, Jeevesh, Bansal, Rachit, Cho, Kyunghyun, Sedoc, João, and Saphra, Naomi. 2023. Linear Connectivity Reveals Generalization Strategies. In: *Proceedings of the 11th International Conference on Learning Representations, ICLR, Kigali, Rwanda, May 1–5, 2023*. <https://openreview.net/forum?id=hY6MOJHl3uL>.

Kalimeris, Dimitris, Kaplun, Gal, Nakkiran, Preetum, et al. 2019. SGD on Neural Networks Learns Functions of Increasing Complexity. Pages 3491–3501 of: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS*. Curran Associates.

Kang, SeongKu, Hwang, Junyoung, Kweon, Wonbin, and Yu, Hwanjo. 2021. Topology Distillation for Recommender System. Pages 829–839 of: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD*. ACM.

Katharopoulos, Angelos, Vyas, Apoorv, Pappas, Nikolaos, and Fleuret, François. 2020. Transformers Are RNNs: Fast Autoregressive Transformers with Linear Attention. Pages 5156–5165 of: *Proceedings of the 37th International Conference on Machine Learning, ICML*. PMLR.

Kendall, Alex, and Gal, Yarin. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? Pages 5574–5584 of: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, NeurIPS*. Curran Associates.

Keskar, Nitish Shirish, Mudigere, Dheevatsa, Nocedal, Jorge, Smelyanskiy, Mikhail, and Tang, Ping Tak Peter. 2017. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In: *Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, France, April 24–26, 2017*. <https://openreview.net/forum?id=H1oyRlYgg>.

Kitaev, Nikita, Kaiser, Lukasz, and Levskaya, Anselm. 2020. Reformer: The Efficient Transformer. In: *Proceedings of the 8th International Conference on Learning Representations, ICLR, Virtual Event, April 26–30, 2020*. <https://openreview.net/forum?id=rkgNKkHtvB>.

Krishnamoorthi, Raghuraman. 2018. Quantizing Deep Convolutional Networks for Efficient Inference: A Whitepaper. arXiv, **abs/1806.08342**.

Krizhevsky, Alex. 2010. *Convolutional deep belief networks on cifar-10*. <https://www.cs.toronto.edu/kriz/conv-cifar10-aug2010.pdf>.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. Pages 1106–1114 of: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, NeurIPS*. Curran Associates.

Kurtz, Mark, Kopinsky, Justin, Gelashvili, Rati, et al. 2020. Inducing and Exploiting Activation Sparsity for Fast Inference on Deep Neural Networks. Pages 5533–5543 of: *Proceedings of the 37th International Conference on Machine Learning, ICML*. PMLR.

Kusupati, Aditya, Bhatt, Gantavya, Rege, Aniket, et al. 2022. Matryoshka Representation Learning. In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS*. Curran Associates.

Lakshminarayanan, Balaji, Pritzel, Alexander, and Blundell, Charles. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. Pages 6402–6413 of: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, NeurIPS*. Curran Associates.

LeCun, Yann, Denker, John S., and Solla, Sara A. 1989. Optimal Brain Damage. In: *Advances in Neural Information Processing Systems 2, NeurIPS*. Morgan Kaufmann. Lee, Namhoon, Ajanthan, Thalaiyasingam, and Torr, Philip H. S. 2019.

SNIP: SingleShot Network Pruning Based on Connection Sensitivity. In: *Proceedings of the 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, May 6–9, 2019*. <https://openreview.net/forum?id=B1VZqjAcYX>.

Leviathan, Yaniv, Kalman, Matan, and Matias, Yossi. 2023. Fast Inference from Transformers via Speculative Decoding. In: *Proceedings of the 40th International Conference on Machine Learning, ICML*. PMLR.

Leviathan, Yaniv, Kalman, Matan, and Matias, Yossi. 2024. *Looking Back at Speculative Decoding*. Google Research. <https://research.google/blog/looking-back-at-speculative-decoding/>. Accessed July 7, 2025.

Li, Hao, Kadav, Asim, Durdanovic, Igor, Samet, Hanan, and Graf, Hans Peter. 2017. Pruning Filters for Efficient ConvNets. In: *Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, France, April 24–26, 2017*. <https://openreview.net/forum?id=rJqFGTslg>.

Li, Hao, Xu, Zheng, Taylor, Gavin, Studer, Christoph, and Goldstein, Tom. 2018. Visualizing the Loss Landscape of Neural Nets. Pages 6391–6401 of: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS*. Curran Associates.

Li, Yucheng. 2023. Unlocking Context Constraints of LLMs: Enhancing Context Efficiency of LLMs with Self-Information-Based Content Filtering. arXiv, **abs/2304.12102**.

Li, Yucheng, Dong, Bo, Guerin, Frank, and Lin, Chenghua. 2023. Compressing Context to Enhance Inference Efficiency of Large Language Models. Pages 6342–6353 of: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*. ACL.

Li, Yuhong, Huang, Yingbing, Yang, Bowen, et al. 2024. SnapKV: LLM Knows What You Are Looking for Before Generation. In: *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS*. Curran Associates.

Liu, Hanxiao, Simonyan, Karen, and Yang, Yiming. 2019. DARTS: Differentiable Architecture Search. In: *Proceedings of the 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, May 6–9, 2019*. <https://openreview.net/forum?id=S1eYHoC5FX>.

Liu, Jing, Cai, Jianfei, and Zhuang, Bohan. 2021. Sharpness-Aware Quantization for Deep Neural Networks. arXiv, **abs/2111.12273**.

Liu, Liyuan, Dong, Chengyu, Liu, Xiaodong, Yu, Bin, and Gao, Jianfeng. 2023. Bridging Discrete and Backpropagation: Straight-Through and Beyond. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS*. Curran Associates.

Liu, Xiaoxuan, Hu, Lanxiang, Bailis, Peter, et al. 2024. Online Speculative Decoding. In: *Proceedings of the 41st International Conference on Machine Learning, ICML*. PMLR.

Liu, Zechun, Zhao, Changsheng, Fedorov, Igor, et al. 2025. SpinQuant: LLM Quantization with Learned Rotations. In: *Proceedings of the 13th International Conference on Learning Representations, ICLR, Singapore, April 24–28, 2025*. <https://openreview.net/forum?id=ogO6DGE6FZ>.

Liu, Zhuang, Sun, Mingjie, Zhou, Tinghui, Huang, Gao, and Darrell, Trevor. 2019. Rethinking the Value of Network Pruning. In: *Proceedings of the 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, May 6–9, 2019*. <https://openreview.net/forum?id=rJlnB3C5Ym>.

Liu, Zichang, Wang, Jue, Dao, Tri, et al. 2023a. Deja Vu: Contextual Sparsity for Efficient LLMs at Inference Time. Pages 22137–22176 of: *Proceedings of the 40th International Conference on Machine Learning, ICML*. PMLR.

Liu, Zichang, Desai, Aditya, Liao, Fangshuo, et al. 2023b. Scissorhands: Exploiting the Persistence of Importance Hypothesis for LLM KV Cache Compression at Test Time. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS*. Curran Associates.

Liu, Ziming, Kitouni, Ouail, Nolte, Niklas, et al. 2022. Towards Understanding Grokking: An Effective Theory of Representation Learning. In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS*. Curran Associates.

Lopez-Paz, David, Bottou, Léon, Schölkopf, Bernhard, and Vapnik, Vladimir. 2016. Unifying Distillation and Privileged Information. In: *Proceedings of the 4th International Conference on Learning Representations, ICLR, San Juan, Puerto Rico, May 2–4, 2016*. <https://openreview.net/forum?id=lzlyCrlCzrg>.

Loshchilov, Ilya, and Hutter, Frank. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In: *Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, France, April 24–26, 2017*. <https://openreview.net/forum?id=Skq89Scxx>.

Loshchilov, Ilya, and Hutter, Frank. 2019. Decoupled Weight Decay Regularization. In: *Proceedings of the 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, May 6–9, 2019*. <https://openreview.net/forum?id=Bkg6RiCqY7>.

Luo, Zelun, Hsieh, Jun-Ting, Jiang, Lu, Niebles, Juan Carlos, and Fei-Fei, Li. 2018. Graph Distillation for Action Detection with Privileged Modalities. Pages 174–192 of: *Proceedings of the 15th European Conference on Computer Vision, ECCV*. Springer.

Lyu, Kaifeng, Li, Zhiyuan, and Arora, Sanjeev. 2022. Understanding the Generalization Benefit of Normalization Layers: Sharpness Reduction. In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS*. Curran Associates.

Ma, Shuming, Wang, Hongyu, Ma, Lingxiao, et al. 2024. The Era of 1-Bit LLMs: All Large Language Models Are in 1.58 Bits. arXiv, **abs/2402.17764**.

Maddox, Wesley J., Izmailov, Pavel, Garipov, Timur, Vetrov, Dmitry P., and Wilson, Andrew Gordon. 2019. A Simple Baseline for Bayesian Uncertainty in Deep Learning. Pages 13132–13143 of: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS*. Curran Associates.

Malach, Eran, Yehudai, Gilad, Shalev-Shwartz, Shai, and Shamir, Ohad. 2020. Proving the Lottery Ticket Hypothesis: Pruning Is All You Need. Pages 6682–6691

of: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*. PMLR.

Manku, Gurmeet Singh, Jain, Arvind, and Sarma, Anish Das. 2007. Detecting NearDuplicates for Web Crawling. Pages 141–150 of: *Proceedings of the 16th International Conference on World Wide Web*. ACM.

McAllester, David A. 2003. PAC-Bayesian Stochastic Model Selection. *Mach. Learn.*, **51**(1), 5–21.

McCandlish, Sam, Kaplan, Jared, Amodei, Dario, and Team, OpenAI Dota. 2018. An Empirical Model of Large-Batch Training. arXiv, **abs/1812.06162**.

McKinstry, Jeffrey L., Esser, Steven K., Appuswamy, Rathinakumar, et al. 2018. Discovering Low-Precision Networks Close to Full-Precision Networks for Efficient Embedded Inference. arXiv, **abs/1809.04191**.

Miao, Xupeng, Oliaro, Gabriele, Zhang, Zhihao, et al. 2024. SpecInfer: Accelerating Large Language Model Serving with Tree-Based Speculative Inference and Verification. Pages 932–949 of: *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS*. ACM.

Micaelli, Paul, and Storkey, Amos J. 2019. Zero-Shot Knowledge Transfer via Adversarial Belief Matching. Pages 9547–9557 of: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS*. Curran Associates.

Michel, Paul, Levy, Omer, and Neubig, Graham. 2019. Are Sixteen Heads Really Better than One? Pages 14014–14024 of: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS*. Curran Associates.

Micikevicius, Paulius, Stosic, Dusan, Burgess, Neil, et al. 2022. FP8 Formats for Deep Learning. arXiv, **abs/2209.05433**.

Migacz, Szymon. 2017. *8-bit Inference with TensorRT*. www.cse.iitd.ac.in/rijurekha/course/tensorrt.pdf. Accessed 20 December 2023.

Mirzadeh, Iman, Alizadeh, Keivan, Mehta, Sachin, et al. 2024. ReLU Strikes Back: Exploiting Activation Sparsity in Large Language Models. In: *Proceedings of the 12th International Conference on Learning Representations, ICLR, Vienna, Austria, May 7–11, 2024*. <https://openreview.net/forum?id=osoWxY8q2E>.

Mirzadeh, Seyed-Iman, Farajtabar, Mehrdad, Li, Ang, et al. 2020. Improved Knowledge Distillation via Teacher Assistant. Pages 5191–5198 of: *Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI*. AAAI Press.

Mishra, Asit K., and Marr, Debbie. 2018. Apprentice: Using Knowledge Distillation Techniques to Improve Low-Precision Network Accuracy. In: *Proceedings of the 6th International Conference on Learning Representations, ICLR, Vancouver, BC, Canada, April 30–May 3, 2018*. <https://openreview.net/forum?id=B1ae1lZRb>.

Mishra, Asit K., Latorre, Jorge Albericio, Pool, Jeff, et al. 2021. Accelerating Sparse Deep Neural Networks. arXiv, **abs/2104.08378**.

Morcos, Ari S., Barrett, David G. T., Rabinowitz, Neil C., and Botvinick, Matthew M. 2018. On the Importance of Single Directions for Generalization. In: *Pro-*

ceedings of the 6th International Conference on Learning Representations, ICLR, Vancouver, BC, Canada, April 30–May 3, 2018. <https://openreview.net/forum?id=r1iuQjxCZ>.

Morcos, Ari S., Yu, Haonan, Paganini, Michela, and Tian, Yuandong. 2019. One Ticket to Win Them All: Generalizing Lottery Ticket Initializations across Datasets and Optimizers. Pages 4933–4943 of: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS*. Curran Associates.

Mozannar, Hussein, and Sontag, David A. 2020. Consistent Estimators for Learning to Defer to an Expert. Pages 7076–7087 of: *Proceedings of the 37th International Conference on Machine Learning, ICML*. PMLR.

Narasimhan, Harikrishna, Jitkrittum, Wittawat, Rawat, Ankit Singh, et al. 2025. Faster Cascades via Speculative Decoding. In: *The 13th International Conference on Learning Representations, ICLR, Singapore, April 24–28, 2025.* <https://openreview.net/forum?id=vo9t20wsmd>.

Neyshabur, Behnam, Tomioka, Ryota, and Srebro, Nathan. 2015. In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. In: *Proceedings of the 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, May 7–9, 2015.* <https://openreview.net/forum?id=SEgVxhoY8lc>.

Neyshabur, Behnam, Sedghi, Hanie, and Zhang, Chiyuan. 2020. What Is Being Transferred in Transfer Learning? In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*. Curran Associates.

NVIDIA. 2020. *NVIDIA A100 Tensor Core GPU Architecture*. NVIDIA. <https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidiaampere-architecture-whitepaper.pdf>.

Ortiz-Jiménez, Guillermo, Favero, Alessandro, and Frossard, Pascal. 2023. Task Arithmetic in the Tangent Space: Improved Editing of Pre-Trained Models. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS*. Curran Associates.

Pan, Zhuoshi, Wu, Qianhui, Jiang, Huiqiang, et al. 2024. LLMingua-2: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression. Pages 963–981 of: *Findings of the Association for Computational Linguistics*. ACL.

Patry, Nicolas. 2022. *Removing Autocast for 35-25% Speedup (Autocast Considered Harmful)*. GitHub. <https://github.com/huggingface/diffusers/pull/511>. Accessed December 19, 2023.

Patterson, David, and Hennessy, John. 2013. *Computer Organization and Design: The Hardware/Software Interface*. Morgan Kaufmann.

Phuong, Mary, and Lampert, Christoph. 2019. Distillation-Based Training for MultiExit Architectures. Pages 1355–1364 of: *Proceedings of the 2019 IEEE International Conference on Computer Vision, ICCV*. IEEE.

Pool, Jeff, Sawarkar, Abhishek, and Rodge, Jay. 2021. *Accelerating Inference with Sparsity Using the NVIDIA Ampere Architecture and NVIDIA TensorRT*. NVIDIA Developer. <https://developer.nvidia.com/blog/accelerating-inference-with-sparsity-using-ampere-and-tensorrt/>. Accessed March 29, 2024.

Power, Alethea, Burda, Yuri, Edwards, Harrison, Babuschkin, Igor, and Misra, Vedant. 2022. Grokking: Generalization beyond Overfitting on Small Algorithmic Datasets. arXiv, **abs/2201.02177**.

Rahimi, Ali, and Recht, Benjamin. 2007. Random Features for Large-Scale Kernel Machines. Pages 1177–1184 of: *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, NeurIPS*. Curran Associates.

Ramanujan, Vivek, Wortsman, Mitchell, Kembhavi, Aniruddha, Farhadi, Ali, and Rastegari, Mohammad. 2020. What's Hidden in a Randomly Weighted Neural Network? Pages 11890–11899 of: *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

Rao, Yongming, Zhao, Wenliang, Zhu, Zheng, Lu, Jiwen, and Zhou, Jie. 2021. Global Filter Networks for Image Classification. Pages 980–993 of: *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS*. Curran Associates.

Rastegari, Mohammad, Ordonez, Vicente, Redmon, Joseph, and Farhadi, Ali. 2016. XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks. Pages 525–542 of: *Proceedings of the 14th European Conference of Computer Vision, ECCV*. Springer.

Reed, Russell. 1993. Pruning Algorithms: A Survey. *IEEE Trans. Neural Networks*, 4 (5), 740–747.

Ren, Jie, Rajbhandari, Samyam, Aminabadi, Reza Yazdani, et al. 2021. ZeRO-Offload: Democratizing Billion-Scale Model Training. Pages 551–564 of: *Proceedings of the 2021 USENIX Annual Technical Conference, USENIX ATC*. USENIX Association.

Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick, and Ommer, Björn. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. Pages 10674–10685 of: *Proceedings of the 2022 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

Romero, Adriana, Ballas, Nicolas, Kahou, Samira Ebrahimi, et al. 2015. FitNets: Hints for Thin Deep Nets. In: *Proceedings of the 3rd International Conference on Learning Representations, ICLR, San Diego, CA, USA, May 7–9, 2015*. <https://open-review.net/forum?id=SWN-lzy85l9>.

Sakamoto, Keitaro, and Sato, Issei. 2022. Analyzing Lottery Ticket Hypothesis from PAC-Bayesian Theory Perspective. In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS*. Curran Associates.

Samaria, Ferdinand, and Harter, Andy. 1994. Parameterisation of a Stochastic Model for Human Face Identification. Pages 138–142 of: *Proceedings of 2nd IEEE Workshop on Applications of Computer Vision, WACV*. IEEE.

Sandler, Mark, Howard, Andrew G., Zhu, Menglong, Zhmoginov, Andrey, and Chen, Liang-Chieh. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. Pages 4510–4520 of: *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

Sato, Ryoma, Yamada, Makoto, and Kashima, Hisashi. 2022. Re-evaluating Word Mover's Distance. Pages 19231–19249 of: *Proceedings of the 39th International Conference on Machine Learning, ICML*. PMLR.

Schölkopf, Bernhard, and Smola, Alexander J. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Schuster, Tal, Fisch, Adam, Gupta, Jai, et al. 2022. Confident Adaptive Language Modeling. In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS*. Curran Associates.

Shah, Jay, Bikshandi, Ganesh, Zhang, Ying, et al. 2024. FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-Precision. In: *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS*. Curran Associates.

Shawe-Taylor, John, and Cristianini, Nello. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press.

Shazeer, Noam, and Stern, Mitchell. 2018. Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. Pages 4603–4611 of: *Proceedings of the 35th International Conference on Machine Learning, ICML*. PMLR.

Shen, Dinghan, Wang, Guoyin, Wang, Wenlin, et al. 2018. Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms. Pages 440–450 of: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. ACL.

Sheng, Ying, Zheng, Lianmin, Yuan, Binhang, et al. 2023. FlexGen: High-Throughput Generative Inference of Large Language Models with a Single GPU. Pages 31094–31116 of: *Proceedings of the 40th International Conference on Machine Learning, ICML*. PMLR.

Shing, Makoto, Misaki, Kou, Bao, Han, Yokoi, Sho, and Akiba, Takuya. 2025. TAID: Temporally Adaptive Interpolated Distillation for Efficient Knowledge Transfer in Language Models. In: *Proceedings of the 13th International Conference on Learning Representations, ICLR, Singapore, April 24–28, 2025*. <https://openreview.net/forum?id=cqsw28DuMW>.

Singh, Sidak Pal, and Jaggi, Martin. 2020. Model Fusion via Optimal Transport. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*. Curran Associates.

Smith, Jimmy T. H., Warrington, Andrew, and Linderman, Scott W. 2023. Simplified State Space Layers for Sequence Modeling. In: *Proceedings of the 11th International Conference on Learning Representations, ICLR, Kigali, Rwanda, May 1–5, 2023*. <https://openreview.net/forum?id=Ai8Hw3AXqks>.

Somepalli, Gowthami, Fowl, Liam, Bansal, Arpit, et al. 2022. Can Neural Nets Learn the Same Model Twice? Investigating Reproducibility and Double Descent from the Decision Boundary Perspective. Pages 13689–13698 of: *Proceedings of the 2022 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

Soudry, Daniel, Hoffer, Elad, Nacson, Mor Shpigel, Gunasekar, Suriya, and Srebro, Nathan. 2018. The Implicit Bias of Gradient Descent on Separable Data. *J. Mach. Learn. Res.*, 19 (70), 1–57.

Spector, Benjamin, and Ré, Christopher. 2023. Accelerating LLM Inference with Staged Speculative Decoding. arXiv, **abs/2308.04623**.

Sun, Mingjie, Chen, Xinlei, Kolter, J. Zico, and Liu, Zhuang. 2024a. Massive Activations in Large Language Models. arXiv, **abs/2402.17762**.

Sun, Mingjie, Liu, Zhuang, Bair, Anna, and Kolter, J. Zico. 2024b. A Simple and Effective Pruning Approach for Large Language Models. In: *The 12th International Conference on Learning Representations, ICLR, Vienna, Austria, May 7–11, 2024*. <https://openreview.net/forum?id=PxoFut3dWW>.

Sun, Xiao, Wang, Naigang, Chen, Chia-Yu, et al. 2020. Ultra-Low Precision 4-Bit Training of Deep Neural Networks. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*. Curran Associates.

Sun, Zhiqing, Yu, Hongkun, Song, Xiaodan, et al. 2020. MobileBERT: A Compact Task-Agnostic BERT for Resource-Limited Devices. Pages 2158–2170 of: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL.

Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jonathon, and Wojna, Zbigniew. 2016. Rethinking the Inception Architecture for Computer Vision. Pages 2818–2826 of: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

Tan, Mingxing, Chen, Bo, Pang, Ruoming, et al. 2019. MnasNet: Platform-Aware Neural Architecture Search for Mobile. Pages 2820–2828 of: *Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

Tang, Hanlin, Lin, Yang, Lin, Jing, et al. 2025. RazorAttention: Efficient KV Cache Compression Through Retrieval Heads. In: *The 13th International Conference on Learning Representations, ICLR, Singapore, April 24–28, 2025*. <https://openreview.net/forum?id=tkiZQL04w>.

Touvron, Hugo, Cord, Matthieu, Douze, Matthijs, et al. 2021. Training Data-Efficient Image Transformers and Distillation Through Attention. Pages 10347–10357 of: *Proceedings of the 38th International Conference on Machine Learning, ICML*. PMLR.

Touvron, Hugo, Martin, Louis, Stone, Kevin, et al. 2023a. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv, **abs/2307.09288**.

Touvron, Hugo, Lavril, Thibaut, Izacard, Gautier, et al. 2023b. LLaMA: Open and Efficient Foundation Language Models. arXiv, **abs/2302.13971**.

Tseng, Albert, Chee, Jerry, Sun, Qingyao, Kuleshov, Volodymyr, and Sa, Christopher De. 2024. QuIP#: Even Better LLM Quantization with Hadamard Incoherence and Lattice Codebooks. In: *Proceedings of the 41st International Conference on Machine Learning, ICML*. PMLR.

Tsipras, Dimitris, Santurkar, Shibani, Engstrom, Logan, Turner, Alexander, and Madry, Aleksander. 2019. Robustness May Be at Odds with Accuracy. In: *Proceedings of the 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, May 6–9, 2019*. <https://openreview.net/forum?id=SyxAb30cY7>.

Ulyanov, Dmitry, Vedaldi, Andrea, and Lempitsky, Victor S. 2018. Deep Image Prior. Pages 9446–9454 of: *Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

van den Oord, Aaron, Vinyals, Oriol, and Kavukcuoglu, Koray. 2017. Neural Discrete Representation Learning. Pages 6306–6315 of: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, NeurIPS*. Curran Associates.

Vanhoucke, Vincent, Senior, Andrew, and Mao, Mark Z. 2011. Improving the Speed of Neural Networks on CPUs. In: *Deep Learning and Unsupervised Feature Learning Workshop, NeurIPS*. <https://research.google/pubs/improving-the-speed-of-neural-networks-on-cpus/>.

Vapnik, Vladimir, and Izmailov, Rauf. 2015. Learning Using Privileged Information: Similarity Control and Knowledge Transfer. *J. Mach. Learn. Res.*, 16, 2023–2049.

Vapnik, Vladimir, and Vashist, Akshay. 2009. A New Learning Paradigm: Learning Using Privileged Information. *Neural Networks*, 22 (5–6), 544–557.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, et al. 2017. Attention Is All You Need. Pages 5998–6008 of: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, NeurIPS*. Curran Associates.

Veit, Andreas, Wilber, Michael J., and Belongie, Serge J. 2016. Residual Networks Behave Like Ensembles of Relatively Shallow Networks. Pages 550–558 of: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, NeurIPS*. Curran Associates.

Viola, Paul A., and Jones, Michael J. 2001. Rapid Object Detection Using a Boosted Cascade of Simple Features. Pages 511–518 of: *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*. IEEE.

Voita, Elena, Talbot, David, Moiseev, Fedor, Sennrich, Rico, and Titov, Ivan. 2019. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. Pages 5797–5808 of: *Proceedings of the 57th Conference of the Association for Computational Linguistics*. ACL.

Wang, Hongyu, Ma, Shuming, Dong, Li, et al. 2023. BitNet: Scaling 1-Bit Transformers for Large Language Models. arXiv, **abs/2310.11453**.

Wang, Mengjiao, Liu, Rujie, Nada, Hajime, et al. 2019. Improved Knowledge Distillation for Training Fast Low Resolution Face Recognition Model. Pages 2655–2661 of: *Proceedings of the 2019 IEEE International Conference on Computer Vision Workshops, ICCV Workshops*. IEEE.

Wang, Shibo, and Kanwar, Pankaj. 2019. *BFLOAT16: The Secret to High Performance on Cloud TPUs*. Google Cloud. <https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus>. Accessed December 19, 2023.

Wang, Wenhui, Wei, Furu, Dong, Li, et al. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*. Curran Associates.

Wang, Xin, Zheng, Yu, Wan, Zhongwei, and Zhang, Mi. 2025. SVD-LLM: Truncation-Aware Singular Value Decomposition for Large Language Model Compression.

sion. In: *The 13th International Conference on Learning Representations, ICLR, Singapore, April 24–28, 2025*. <https://openreview.net/forum?id=LNyIUouhdt>.

Wang, Zheng, Li, Juncheng B., Qu, Shuhui, Metze, Florian, and Strubell, Emma. 2022. SQuAT: Sharpness- and Quantization-Aware Training for BERT. arXiv, **abs/2210.07171**.

Wen, Wei, Wu, Chunpeng, Wang, Yandan, Chen, Yiran, and Li, Hai. 2016. Learning Structured Sparsity in Deep Neural Networks. Pages 2074–2082 of: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, NeurIPS*. Curran Associates.

Wortsman, Mitchell, Ilharco, Gabriel, Gadre, Samir Yitzhak, et al. 2022. Model Soups: Averaging Weights of Multiple Fine-Tuned Models Improves Accuracy Without Increasing Inference Time. Pages 23965–23998 of: *Proceedings of the 39th International Conference on Machine Learning, ICML*. PMLR.

Wu, Hao, Judd, Patrick, Zhang, Xiaojie, Isaev, Mikhail, and Micikevicius, Paulius. 2020. Integer Quantization for Deep Learning Inference: Principles and Empirical Evaluation. arXiv, **abs/2004.09602**.

Wu, Wenhao, Wang, Yizhong, Xiao, Guangxuan, Peng, Hao, and Fu, Yao. 2025. Retrieval Head Mechanistically Explains Long-Context Factuality. In: *Proceedings of the 13th International Conference on Learning Representations, ICLR, Singapore, April 24–28, 2025*. <https://openreview.net/forum?id=EytBpUGB1Z>.

Xiao, Guangxuan, Tian, Yuandong, Chen, Beidi, Han, Song, and Lewis, Mike. 2024. Efficient Streaming Language Models with Attention Sinks. In: *Proceedings of the 12th International Conference on Learning Representations, ICLR, Vienna, Austria, May 7–11, 2024*. <https://openreview.net/forum?id=NG7sS51zVF>.

Xiao, Guangxuan, Tang, Jiaming, Zuo, Jingwei, et al. 2025. DuoAttention: Efficient Long-Context LLM Inference with Retrieval and Streaming Heads. In: *The 13th International Conference on Learning Representations, ICLR, Singapore, April 24–28, 2025*. <https://openreview.net/forum?id=cFu7ze7xUm>.

Xiong, Yunyang, Zeng, Zhanpeng, Chakraborty, Rudrasis, et al. 2021. Nyström-former: A Nyström-based Algorithm for Approximating Self-Attention. Pages 14138–14148 of: *Proceedings of the 35th AAAI Conference on Artificial Intelligence*. AAAI Press.

Yang, Tien-Ju, Howard, Andrew G., Chen, Bo, et al. 2018. NetAdapt: Platform-Aware Neural Network Adaptation for Mobile Applications. Pages 289–304 of: *Proceedings of the 15th European Conference of Computer Vision, ECCV*. Springer.

Young, Alex, Chen, Bei, Li, Chao, et al. 2024. Yi: Open Foundation Models by 01.AI. arXiv, **abs/2403.04652**.

Yu, Jiahui, and Huang, Thomas S. 2019. Universally Slimmable Networks and Improved Training Techniques. Pages 1803–1811 of: *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV*. IEEE.

Yu, Jiahui, Yang, Linjie, Xu, Ning, Yang, Jianchao, and Huang, Thomas S. 2019. Slimmable Neural Networks. In: *Proceedings of the 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, May 6–9, 2019*. <https://openreview.net/forum?id=H1gMCsAqY7>.

Yu, Le, Yu, Bowen, Yu, Haiyang, Huang, Fei, and Li, Yongbin. 2023. Language Models Are Super Mario: Absorbing Abilities from Homologous Models as a Free Lunch. arXiv, **abs/2311.03099**.

Yu, Xiyu, Liu, Tongliang, Wang, Xinchao, and Tao, Dacheng. 2017. On Compressing Deep Models by Low-Rank and Sparse Decomposition. Pages 67–76 of: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

Yuan, Li, Tay, Francis E. H., Li, Guilin, Wang, Tao, and Feng, Jiashi. 2020. Revisiting Knowledge Distillation via Label Smoothing Regularization. Pages 3902–3910 of: *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE.

Yuan, Zhihang, Shang, Yuzhang, Song, Yue, et al. 2023. ASVD: Activation-Aware Singular Value Decomposition for Compressing Large Language Models. arXiv, **abs/2312.05821**.

Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, and Vinyals, Oriol. 2017. Understanding Deep Learning Requires Rethinking Generalization. In: *Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, France, April 24–26, 2017*. <https://openreview.net/forum?id=Sy8gdB9xx>.

Zhang, Jun, Wang, Jue, Li, Huan, et al. 2024. Draft and Verify: Lossless Large Language Model Acceleration via Self-Speculative Decoding. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL.

Zhang, Linfeng, Song, Jiebo, Gao, Anni, et al. 2019. Be Your Own Teacher: Improve the Performance of Convolutional Neural Networks via Self Distillation. Pages 3712–3721 of: *Proceedings of the 2019 IEEE International Conference on Computer Vision, ICCV*. IEEE.

Zhang, Susan, Roller, Stephen, Goyal, Naman, et al. 2022. OPT: Open Pre-trained Transformer Language Models. arXiv, **abs/2205.01068**.

Zhang, Xiangyu, Zou, Jianhua, He, Kaiming, and Sun, Jian. 2016. Accelerating Very Deep Convolutional Networks for Classification and Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38 (10), 1943–1955.

Zhang, Yushun, Chen, Congliang, Li, Ziniu, et al. 2025. Adam-Mini: Use Fewer Learning Rates to Gain More. In: *Proceedings of the 13th International Conference on Learning Representations, ICLR, Singapore, April 24–28, 2025*. <https://openreview.net/forum?id=iBExhaU3Lc>.

Zhang, Zhenyu, Sheng, Ying, Zhou, Tianyi, et al. 2023. H2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models. In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS*. Curran Associates.

Zhang, Zhilu, and Sabuncu, Mert R. 2020. Self-Distillation as Instance-Specific Label Smoothing. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS*. Curran Associates.

Zhao, Ruizhe, Vogel, Brian, Ahmed, Tanvir, and Luk, Wayne. 2020. Reducing Underflow in Mixed Precision Training by Gradient Scaling. Pages 2922–2928

of: *Proceedings of the 29th International Joint Conference on Artificial Intelligence, IJCAI*. International Joint Conferences on Artificial Intelligence Organization.

Zhou, Hattie, Lan, Janice, Liu, Rosanne, and Yosinski, Jason. 2019. Deconstructing Lottery Tickets: Zeros, Signs, and the Supermask. Pages 3592–3602 of: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS*. Curran Associates.

Zhou, Helong, Song, Liangchen, Chen, Jiajie, et al. 2021. Rethinking Soft Labels for Knowledge Distillation: A Bias–Variance Tradeoff Perspective. In: *Proceedings of the 9th International Conference on Learning Representations, ICLR, Virtual Event, May 3–7, 2021*. <https://openreview.net/forum?id=gIHd-5X324>.

Zhou, Yongchao, Lyu, Kaifeng, Rawat, Ankit Singh, et al. 2024. DistillSpec: Improving Speculative Decoding via Knowledge Distillation. In: *The 12th International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. <https://openreview.net/forum?id=rsY6J3ZaTF>.