Paper Review Assignment 2
AI Model Pruning

Chia-Chi Tsai (蔡家齊)
cctsai@gs.ncku.edu.tw
AI System Lab
Department of Electrical Engineering
National Cheng Kung University

Paper Readings and Review

- Paper related to AI Models **Pruning**
  - To understand model compression flow
  - To learn the AI model **pruning** methodology
  - To learn the AI model architecture search technique
- **Due**
  - **11/19 23:59**
- Requirement
  - Choose **at least one or more** papers
    - From recommended paper list
    - **Or any other paper as long as it related to the topics**
  - Summarize and write paper review in word/latex format
    - **LaTeX format is highly recommended**
  - Hand in **compiled pdf files** on moodle

Paper Readings and Review

- Reading reviews are free of format
- But the following review questions guide you through the paper reading process.
  - What are the **motivations** for this work?
  - What is the **proposed solution**?
  - What is the work's **evaluation** of the proposed solution?
  - What is your **analysis** of the identified problem, idea, and evaluation?
  - What are **future directions** for this research?
  - What **questions** are you left with?

Recommended Paper List

- Pruning Granularity
  - Han, S., Mao, H., & Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149.
  - Mao, H., Han, S., Pool, J., Li, W., Liu, X., Wang, Y., & Dally, W. J. (2017). Exploring the granularity of sparsity in convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 13-20).
- Pruning
  - He, Y., Zhang, X., & Sun, J. (2017). Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 1389-1397).
  - He, Y., Kang, G., Dong, X., Fu, Y., & Yang, Y. (2018). Soft filter pruning for accelerating deep convolutional neural networks. *arXiv preprint arXiv:1808.06866*.
  - Luo, J. H., Wu, J., & Lin, W. (2017). Thinet: A filter level pruning method for deep neural network compression. In *Proceedings of the IEEE international conference on computer vision* (pp. 5058-5066).
  - Liu, Z., Sun, M., Zhou, T., Huang, G., & Darrell, T. (2018). Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*.
  - Liu, Z., Mu, H., Zhang, X., Guo, Z., Yang, X., Cheng, K. T., & Sun, J. (2019). Metapruning: Meta learning for automatic neural network channel pruning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3296-3305).
  - Li, B., Wu, B., Su, J., & Wang, G. (2020, August). Eagleeye: Fast sub-net evaluation for efficient neural network pruning. In *European conference on computer vision* (pp. 639-654). Springer, Cham.
  - Wang, Y., Zhang, X., Xie, L., Zhou, J., Su, H., Zhang, B., & Hu, X. (2020, April). Pruning from scratch. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 12273-12280).
  - Sehwag, V., Wang, S., Mittal, P., & Jana, S. (2020). Hydra: Pruning adversarially robust neural networks. *Advances in Neural Information Processing Systems*, *33*, 19655-19666.
  - Lin, T., Stich, S. U., Barba, L., Dmitriev, D., & Jaggi, M. (2020). Dynamic model pruning with feedback. *arXiv preprint arXiv:2006.07253*.

Recommended Paper List

- Network Architecture Search
  - Zoph, B., & Le, Q. V. (2016). Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.
  - Liu, H., Simonyan, K., & Yang, Y. (2018). Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*.
  - Bender, G., Kindermans, P. J., Zoph, B., Vasudevan, V., & Le, Q. (2018, July). Understanding and simplifying one-shot architecture search. In *International conference on machine learning* (pp. 550-559). PMLR.
  - Cai, H., Gan, C., Wang, T., Zhang, Z., & Han, S. (2019). Once-for-all: Train one network and specialize it for efficient deployment. *arXiv preprint arXiv:1908.09791*.
  - Cai, H., Zhu, L., & Han, S. (2018). Proxylessnas: Direct neural architecture search on target task and hardware. *arXiv preprint arXiv:1812.00332*.
  - Li, Y., Zhao, P., Yuan, G., Lin, X., Wang, Y., & Chen, X. (2022). Pruning-as-Search: Efficient Neural Architecture Search via Channel Pruning and Structural Reparameterization. *arXiv preprint arXiv:2206.01198*. Li, X., Zhou, Y., Pan, Z., & Feng, J. (2019). Partial order pruning: for best speed/accuracy trade-off in neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9145-9153).
  - Wang, T., Wang, K., Cai, H., Lin, J., Liu, Z., Wang, H., ... & Han, S. (2020). Apq: Joint search for network architecture, pruning and quantization policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2078-2087).
  - Lin, M., Ji, R., Zhang, Y., Zhang, B., Wu, Y., & Tian, Y. (2020). Channel pruning via automatic structure search. *arXiv preprint arXiv:2001.08565*.
  - Dai, X., Chen, D., Liu, M., Chen, Y., & Yuan, L. (2020, August). Da-nas: Data adapted pruning for efficient neural architecture search. In *European Conference on Computer Vision* (pp. 584-600). Springer, Cham.