Model Compression

Presented by : Ashutosh Adhikari

Neural Networks Can be Too Huge !!

- NNs have been growing a lot more complex with time
- Objective : learn efficient NNs, prune redundant parameters, connections
- Helps in reducing the processing time
- Reduces the *run-time* memory requirement

Categories of Model Compression

- Parameter pruning and sharing
- Low-rank factorization
- transferred/compact convolutional filters
- Knowledge distillation

Parameter Pruning and Sharing

- One of the oldest techniques
- Optimal Brain Damage :
  - objective function to characterize importance of parameters
  - Delete the less important parameters
  - Done using second derivative and some other approximation

- Quantization and binarization

Model Compression & Computer Vision

- A lot of work on Model Compression for Computer Vision problems
- Many Convolutional Neural Network specific approaches developed
- Channel Pruning has been successful
  - CondenseNets - Group the features; prune the less important
    - Device a methodology to *learn* the groups
  - Network Slimming

Pretrained Models for NLP

- Paradigm of pre-trained models for NLP
  - Transformer based models (BERT, GPT)
  - BERT ⇒ Transformer based model with 300M parameters !!
- Pre-trained models are huge and cumbersome
- All the major works use knowledge distillation
- Future Work!!

Knowledge Distillation

- Model Agnostic approach
- Student-teacher system
- Teacher ⇒ Larger model, knows more
- Student ⇒ Smaller model, is limited
- Allow the student to learn “rich” representations from the teacher
- Using class probabilities produced by the teacher.
- Add a regression objective for “distilling knowledge”

Questions?

References

1. Y. LeCun, J. S. Denker, S. A. Solla, R. E. Howard, and L. D.Jackel. Optimal brain damage. In NIPS, volume 2, pages 598–605, 1989.

2. Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang.Learning efficient convolutional networks through networkslimming. In 2017 IEEE International Conference on Computer Vision (ICCV), pages 2755–2763, 2017.

3. G. Huang, S. Liu, L. van der Maaten, and K. Q. Weinberger. Condensenet: An efficient densenet using learned group convolutions.CVPR, 2018.

4. Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks.arXiv preprint arXiv:1710.09282, 2017.