Knowledge Distillation

First week of machine learning workshop

Outline

- Model compression
  - Quantization
  - Pruning
  - Knowledge distillation
- Distilling the Knowledge in a Neural Network
  - Softmax temperature
  - Dark knowledge
- Improved Knowledge Distillation via Teacher Assistant
  - Why distillation does not work when there is a huge gap between capacity of teacher and student
  - How to introduce teacher assistant
  - What is the optimum number of assistants?
- Distill Bert
  - Bert
  - How to compress Bert
- Quantized Neural Networks: Training Neural Networks with Low Precision Weights and Activations
  - What is quantization
  - Quantization during training or post training
  - 1-bit quantization

References

1. Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).
2. Mirzadeh, Seyed-Iman, et al. "Improved knowledge distillation via teacher assistant: Bridging the gap between student and teacher." arXiv preprint arXiv:1902.03393 (2019).
3. Lopez-Paz, David, et al. "Unifying distillation and privileged information." arXiv preprint arXiv:1511.03643 (2015).
4. Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).
5. Hubara, Itay, et al. "Quantized neural networks: Training neural networks with low precision weights and activations." The Journal of Machine Learning Research 18.1 (2017): 6869-6898.
6. Polino, Antonio, Razvan Pascanu, and Dan Alistarh. "Model compression via distillation and quantization." arXiv preprint arXiv:1802.05668 (2018).