L46 Additional Reading List v1.7

Lecture 1

The following book chapters will be of interest to certain topics raised during these lectures. You can find access to these books via the Cambridge digital library:

- For more on the fundamentals of deep learning in chapters 10 and 11 of Geron's 2019 *"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow"*, 2nd Edition
- More can be found regarding the fundamentals of computer systems consult chapters 1 to 3 of Bryant and O'Hallaron's *"Computer Systems: A Programmer's Perspective"*, Global Edition
- Details and deeper discussion about deep learning acceleration (especially regarding hardware mapping) can be found in chapters 5, 6, 7, 8, and 9 of Vivienne Sze's 2020 textbook: "*Efficient Processing of Deep Neural Networks"*

The following informal blogs and class related materials are a good place to brush up on some of basic deep learning concepts and terminology covered in the lectures, if you feel the above books are too much to review:

- Cheat sheet for Stanford's CS 230 class:
  - https://github.com/afshinea/stanford-cs-230-deep-learning
- Gradient Descent Optimisations
  - http://ruder.io/optimizing-gradient-descent/
- Loss Function Comparison
  - http://rohanvarma.me/Loss-Functions/

During the lecture there were references to the following systems, blogs or papers:

- Nvidia Triton: https://developer.nvidia.com/blog/simplifying-ai-model-deployment-at-the-edge-with-triton-inference-server/
- Pipe Transformer: https://arxiv.org/abs/2102.03161
- "AI and compute": https://openai.com/research/ai-and-compute

Lecture 2

The following papers are touched upon during this lecture:

- MobileNets are discussed in Howard et al. (2017) https://arxiv.org/pdf/1704.04861.pdf
- ShuffleNets (only tangentially mentioned) are presented in Zhang et al. (2017) https://arxiv.org/pdf/1707.01083.pdf
- A good place to start regarding the pruning pipeline is to consult Wang et al. (2019) https://arxiv.org/pdf/1909.12579.pdf
- One of the first papers to study deep learning with limited numerical precision: https://proceedings.mlr.press/v37/gupta15.html
- MAGNet is described in more detail by Venkatesan et al. (2019) https://people.eecs.berkeley.edu/~ysshao/assets/papers/magnet2019-iccad.pdf
- The original distillation paper by Hinton et al. (2015) https://storage.googleapis.com/pub-tools-public-publication-data/pdf/44873.pdf

- Regarding quantization, there are two good introductory white papers coming out of industry that are worth looking at:
  - Google: https://arxiv.org/pdf/1806.08342.pdf
  - Qualcomm: https://arxiv.org/pdf/2106.08295.pdf
- Distillation in the context of BERT: https://arxiv.org/abs/1903.12136

Lecture 3

The following references are recommended if you want to learn more about the topics presented:

- https://e-dorigatti.github.io/math/deep%20learning/2020/04/07/autodiff.html
- https://pytorch.org/tutorials/beginner/basics/autogradqs_tutorial.html
- https://www.jmlr.org/papers/volume18/17-468/17-468.pdf

Lecture 4

The mechanics of SGD are best examined through the lens of a simple NumPy implementation. You can find one based on fully connected deep neural networks in the code associated with Nielsen's deep learning online book *"Neural Networks and Deep Learning"* which is available from https://github.com/mnielsen/neural-networks-and-deep-learning/blob/master/src/network.py. (The book is freely available at http://neuralnetworksanddeeplearning.com/chap2.html.)

The following papers expand upon the discussions contained in this lecture:

- More on decreasing memory requirements of training can be found in Sohoni et al. (2019) https://arxiv.org/pdf/1904.10631.pdf

Lecture 5

References that provide additional context to this lecture include:

- https://siboehm.com/articles/22/CUDA-MMM
- https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html
- https://pytorch.org/tutorials/advanced/cpp_extension.html

Lecture 6

Additional information regarding topics from this lecture are available from:

- Chapters 3 and 5 of Vivienne Sze's 2020 textbook: "*Efficient Processing of Deep Neural Networks"* (Chapter 5 was also referenced in lecture 2 see above)
- If you are interested in reading on TPUs have a look at Jouppi et al. (2017) https://arxiv.org/pdf/1704.04760.pdf
- Additional details about the Groq processor can be found in Ross et al. (2020) https://groq.com/wp-content/uploads/2020/06/ISCA-TSP.pdf

Lecture 7

TBD.

Lecture 8

TBD.

Lecture 9

TBD.

Lecture 10

TBD.