Title: Deep Learning on Edge Devices by Deep Compression

Subtitle: a case study on medical image analysis on mobile phones

Content and Purpose

Design a U-Net with related algorithm and make it learn from medical images. Then, tune and train it to obtain the model which is capable of analyzing medical images.

Further work is to compress this model. Because deep neural networks are not only computationally intensive but also memory intensive, mobile devices with limited hardware resources are confronted with difficulties when running models of deep neural networks. Therefore, to enable the target model to be deployed on mobile devices, the model should be compressed to reduce required storage and energy and in the meanwhile, model's accuracy should not be affected.

Requirements

1. Skilled in using Python, Docker, CI process

- Python: as the main programming language, including ML frameworks, like Keras, Tensorflow;
- Docker: build ML development environment with docker, build images and run them as containers, manage clusters with Kubernetes;
- CI process: Git as VCS, GitLab as code hosting platform, Jenkins as automation server, all of them implement the whole CI process;

2. Familar with U-Net

The whole network is based on the U-Net which is an outstanding network proposed in 2015 that can be trained on very few training data sets, especially useful in the biomedical field for thousands of training images are usually beyond reach. To get the well-performing model and compress it, the student on this project should well know about U-Net architecture and be able to use it skillfully.

3. Knowledge about model compression

The core part of this project is compressing the trained model to enable it running on edge devices like mobile phones. To achieve a satisfactory compression ratio, the student on this project ought to have a grasp of knowledge about mainstream model compression methods.

References(till now)

1. Olaf Ronneberger, Philipp Fischer, Thomas Brox: U-Net: Convolutional Networks for Biomedical Image Segmentation. 2015.
2. Karttikeya Mangalam, Mathieu Salzamann: On Compressing U-net Using Knowledge Distillation
3. Song Han, Huizi Mao, William J. Dally: Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding
4. P. Grunwald, M. A. Pitt and I. J. Myung (eds.), Advances in Minimum Description Length: Theory and Applications，M.I.T. Press (MIT Press), April 2005, ISBN0-262-07262-9
5. Jason Wang, Luis Perez: The Effectiveness of Data Augmentation in Image Classification using Deep Learning