

CENG5030 Course Project Description

March 11, 2019

1 Overview

The main topic for course project is DNN compaction, i.e., compression and/or acceleration. Given a pre-trained model for a specific task, the objective is to compact the model with acceptable performance degradation. Any methodologies that are perceived useful can be selected and developed.

2 Requirement

- Target network structure:
VGG-16 or ResNet-56. Pre-trained models will be released soon.
- Task:
Image classification on CIFAR-10.
- Framework:
Caffe. Other deep learning frameworks are allowed upon approval in advance.

3 Evaluation

- Criteria:
compression ratio, acceleration ratio and performance degradation;
- Rule of score:
A reference performance (compression ratio / acceleration ratio / accuracy loss) will be provided for each network. Each submission will be compared with the reference performance and calculate the margin, and then ranked based on the margin. The scores will be normalized based on the quality ranking of all submissions.
- Bonus:
Doing for both networks are encouraged. The score is determined by the higher one.
- Duplicated submission will be marked as 0.

4 Q & A

Any inquiry and update will be posted on the Blackboard page.