Machine Learning HW13

Network Compression

ML TAs
ntu-ml-2022-spring-ta@googlegroups.com

Outline

- Task description
- Dataset
- Topics Intro
- Regulations
- Grading
- Guideline
- Reports

Links

- Kaggle Competition (with sample code)
- Colab
  - Colab sample code (not recommended)
  - Report Q3-1
    - pytorch network pruning tutorial
    - pruning sample code

Deadline

- Kaggle, Code(NTU COOL),Report(GradeScope):

  2022/06/17 23:59

Task Description

- Network Compression: Make your model smaller without losing the performance.
- In this task, you need to train a very small model to complete HW3 (image classification on the food-11 dataset)

Dataset - food-11

- Same as HW3.

Dataset - food-11

- The images are collected from the food-11 dataset classified into 11 classes.
- Training set: 9866 labeled images
- Validation set: 3430 labeled images
- Evaluation set: 3347 images
- DO NOT attempt to find the original labels of the testing set.
- DO NOT use any external datasets.

Intro

- **Knowledge distillation**
- **Architecture design**
- **Network pruning (Report Q3)**
- Parameter Quantization
- Dynamic Computation

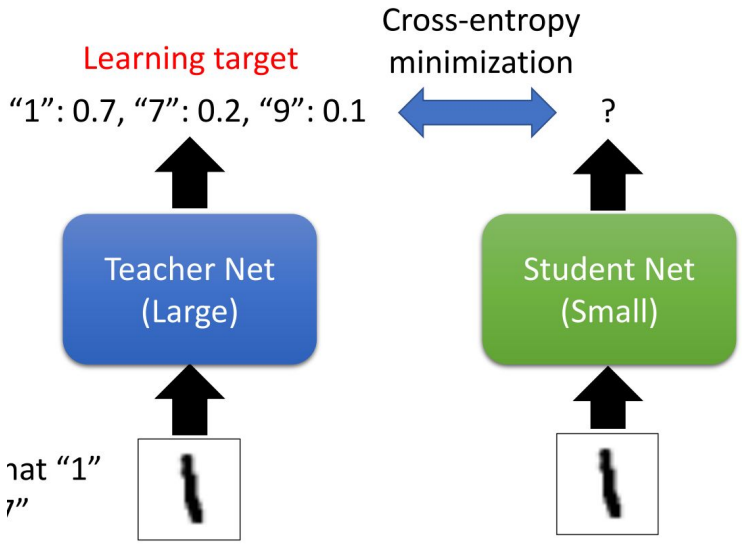
There are many different techniques in network compression. In this homework, we focus more on these three concepts.

video: 【機器學習2021】神經網路壓縮 (Network Compression) (一) - 類神經網路剪枝 (Pruning)
slides: Network Compression (ntu.edu.tw)

Intro - knowledge distillation

- When training a small model, distill some information from the large model (such as the probability distribution of the prediction) to help the small model learn better.
- We have provided a well-trained network to help you do knowledge distillation (test-Acc ≅ 0.899).
- Feel free to train your own teacher network. But you can not use any pretrained model or additional dataset.

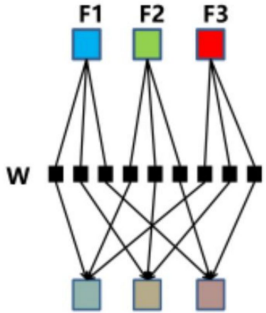
Knowledge Distillation. Knowledge distillation is model… | by Ujjwal Upadhyay | Neural Machine | Medium



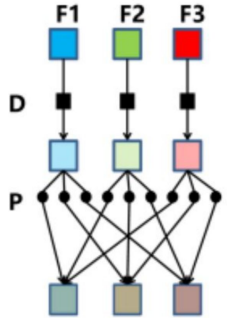
ref: Network Compression (ntu.edu.tw)

Intro - architecture design

- Depthwise & Pointwise Convolution Layer
  - You can consider the original convolution as a Dense/Linear Layer, but each line/each weight is a filter, and the original multiplication becomes a convolution operation. (input*weight → input * filter)
  - Depthwise: let each channel pass a respective filter first, and let every pixel pass the shared-weight Dense/Linear.
  - Pointwise is a 1x1 Conv.
- It is strongly recommended that you use similar techniques to design your model. (**IOkk vs Ikk+IO**)



(a) Standard Conv



(b) DSC

A Basic Introduction to Separable Convolutions | by Chi-Feng Wang | Towards Data Science

Regulations

- You should NOT plagiarize, if you use any other resource, you should cite it in the reference. (*)
- DO NOT share codes or prediction files with any living creatures.
- DO NOT use any approaches to submit your results more than 5 times a day.
- DO NOT search or use additional data.
- DO NOT search the label or dataset on the Internet.
- **DO NOT use pre-trained models on any image datasets.**
- Your final grade x 0.9 if you violate any of the above rules.
- Prof. Lee & TAs preserve the rights to change the rules & grades.

Special Regulations - 1

- **Make sure that the total number of parameters of your model should less than or equal to 100, 000.**
  - Please make sure to follow this rule before submitting kaggle / NTU COOL to prevent anyone from polluting the leaderboard.
  - If you don’t follow this rule, you’ll get 0 point in this assignment.
- **DO NOT USE TEST DATA FOR PURPOSES OTHER THAN INFERENCING.**
  - Because If you use teacher network to predict pseudo-labels of the test data, you can only use student network to overfit these pseudo-labels without train data. In this way, your kaggle accuracy will be as high as the teacher network, but the fact is that you just overfit the test data and your true testing accuracy is very low.
  - These contradict the purpose of this assignment (network compression); therefore, you should not misuse the test data.

*If you have any concerns, you can email us.

Special Regulations - 2

- Please use the **torchsummary** package to measure the number of parameters of your model. Note that non-trainable parameters should also be considered.

- Ensemble techniques / (or any other multi-model techniques) are allowed. But you need to **sum all numbers of the parameters and make sure this number is not exceed 100,000**.

Grading

- simple (public) +0.5 pts
- simple (private) +0.5 pts
- medium (public) +0.5 pts
- medium (private) +0.5 pts
- strong (public) +0.5 pts
- strong (private) +0.5 pts
- boss (public) +0.5 pts
- boss (private) +0.5 pts
- code submission +2 pts
- report +4 pts

Total : 10 pts

Grading -- Bonus

If your **ranking is in top 3 on the kaggle private leaderboard**, you can choose to share a bonus report to NTU COOL and get extra **0.5 pts**.

About the report

- Your *name* and *student_ID*
- Methods you used in code
- Reference
- in 200 words
- **Deadline of the bonus report is at 2022/6/20 23:59**
- Please upload to NTU COOL's discussion of HW13

Report template

Baseline Guide

- Simple baseline (acc > 0.44820, <1hour):
  - Just run the code and submit.
- Medium baseline (acc > 0.64840, <3hour):
  - Complete KL divergence loss function for knowledge distillation, control alpha & T and train longer.
- Strong baseline (acc > 0.82370, 8~12hour):
  - Modify model architecture with depth-wise and point-wise convolutions.
    - You can also take great ideas from MobileNet, ShuffleNet, DenseNet, SqueezeNet, GhostNet, etc.
  - Any method and techniques you learned in HW3. For example, stronger data augmentation, hyperparameters you used in HW3.

Baseline Guide

- Boss baseline (acc > 0.85159, unmeasurable):
  - Implement other advanced knowledge distillation
    - For example, FitNet, Relational KD, DML
  - Make your teacher much stronger
    - If your teacher is too strong, you can consider TAKD techniques.

Report Questions - 1

1-1. Please copy&paste your student model architecture code to the HW13 GradeScope (**0.5pts**).

1-2. Copy&Paste the *torchsummary* result of your student model to HW13 GradeScope. The *total params* should not exceed **100,000** (**0.5pts**).

Report Questions - 2

2-1. Please copy and paste your KL divergence loss function (**loss_fn_kd , choose alpha=0.5, temperature=1.0**) to HW13 GradeScope (**0.5pts**).

2-2. Which is true about the hyperparameter T (temperature) of KL divergence loss function for knowledge distillation (**0.5pts**)?

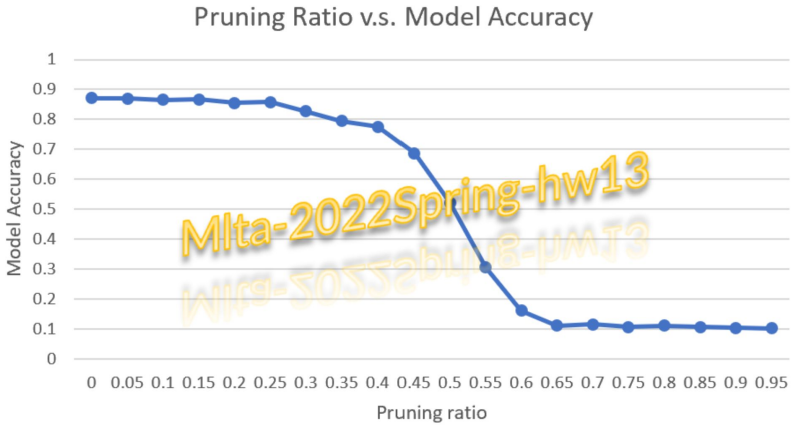
(a). Using a higher value for T produces a softer probability distribution over classes.

(b). Using a higher value for T produces a harder probability distribution over classes

ref: 1503.02531.pdf (arxiv.org)

Report Questions - 3-1

Please reference to this pytorch network pruning tutorial and the provided sample code to adopt network pruning on your teacher network (you can also use the provided one). Plot the graph like below to indicate the relationship between pruning ratio and accuracy (use validation set). **Please checkout HW13 GradeScope for more details about this question** (**1pts**).



Report Questions - 3-2

Continue to the previous question, do you think that the implementation of network pruning in the tutorial can speed up inference time? (If you have no idea, we encourage you to implement some inference-time measurement experiments by yourself) (**1pts**).

(a). Yes. After pruning, the neurons are removed so the amount of calculation is lower and the inference time is faster.

(b). No. Such pruning technique in the tutorial is just to add some masks to the modules. The amount of calculation is nearly the same so the inference time is also similar.

Code Submission

- NTU COOL
  - Compress your code and pack them into **.zip file**

    <student_ID>_hw13.zip

  - Your **.zip** file should include
    - **Code**: either **hw13.py** or **hw13.ipynb**
    - **student_best.ckpt** (Can reproduce the result of one of your chosen submission on the Kaggle competition.)
  - **Submit the code and the student_best.ckpt that corresponds to your chosen submission in Kaggle (One of the best)**

Report Submission

- Answer the questions on **HW13 GradeScope**.

If you have any question...

- NTU COOL (recommended)
  - HW13 disscussion board
- Kaggle discussion
- Email
  - mlta-2022-spring@googlegroups.com
  - The title should begin with "**[hw13]**"

Post-test Questionnaire(後測問卷)

教育部後測問卷

學生心得問卷