Data Science HW2 - Model Compression

姓名：游翔竣

ID: 311511056

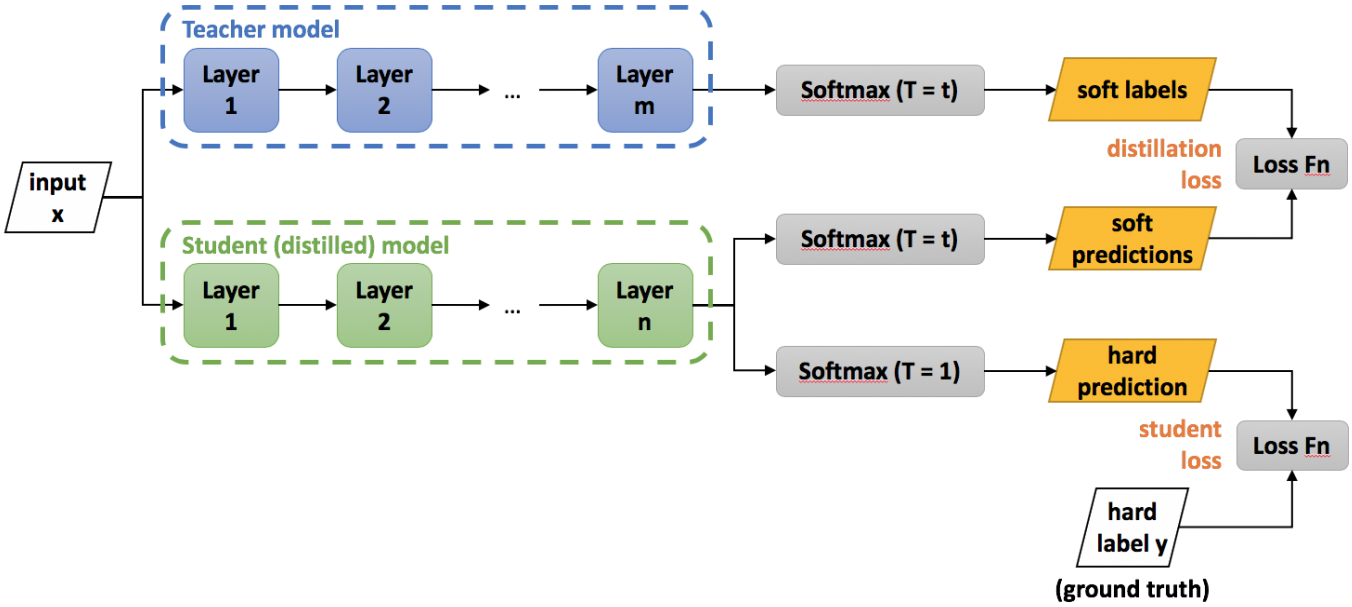
Torch summary output

```
==========================================================================
Layer (type:depth-idx)                   Output Shape              Param #
==========================================================================
├─Sequential: 1-1                        [-1, 16, 14, 14]          --
|    └─Conv2d: 2-1                       [-1, 16, 28, 28]          448
|    └─BatchNorm2d: 2-2                  [-1, 16, 28, 28]          32
|    └─ReLU: 2-3                         [-1, 16, 28, 28]          --
|    └─AvgPool2d: 2-4                    [-1, 16, 14, 14]          --
├─Sequential: 1-2                        [-1, 32, 7, 7]            --
|    └─Conv2d: 2-5                       [-1, 32, 14, 14]          4,640
|    └─BatchNorm2d: 2-6                  [-1, 32, 14, 14]          64
|    └─ReLU: 2-7                         [-1, 32, 14, 14]          --
|    └─AvgPool2d: 2-8                    [-1, 32, 7, 7]            --
├─Sequential: 1-3                        [-1, 64, 3, 3]            --
|    └─Conv2d: 2-9                       [-1, 64, 7, 7]            18,496
|    └─BatchNorm2d: 2-10                 [-1, 64, 7, 7]            128
|    └─ReLU: 2-11                        [-1, 64, 7, 7]            --
|    └─AvgPool2d: 2-12                   [-1, 64, 3, 3]            --
|    └─Dropout: 2-13                     [-1, 64, 3, 3]            --
├─Sequential: 1-4                        [-1, 10]                  --
|    └─Linear: 2-14                      [-1, 128]                 73,856
|    └─ReLU: 2-15                        [-1, 128]                 --
|    └─Dropout: 2-16                     [-1, 128]                 --
|    └─Linear: 2-17                      [-1, 10]                  1,290
==========================================================================
Total params: 98,954
Trainable params: 98,954
Non-trainable params: 0
Total mult-adds (M): 2.32
==========================================================================
Input size (MB): 0.01
Forward/backward pass size (MB): 0.34
Params size (MB): 0.38
Estimated Total Size (MB): 0.72
==========================================================================
```

Method

- I use 3-layer CNN and 2-layer fully connected.
- Furthermore, a hyperparameter α to trade of the distillation loss and student loss, where:
  - **distillation loss** L_{dis} : KL divergence between soft labels & soft predictions.
  - **student loss** L_{stu}: Cross entropy loss between hard predictions & true labels.
- Total loss:

$$L(x, y) = \alpha \cdot L_{dis} + (1 - \alpha) \cdot L_{stu}$$



Training detail

- Batch size : 128
- Learning rate : 3e-4
- Max iteration : 500
- Fix random seed : 8787
- Distilation temperature T : 5
- α : 0.3

Reference

[1]Distilling the Knowledge in a Neural Network (https://arxiv.org/pdf/1503.02531.pdf)
[2]Knowledge Distillation (https://intellabs.github.io/distiller/knowledge_distillation.html)