Homework Assignment #2

Advanced Artificial Intelligence Techniques

This homework aims at exploring the following concepts:

- Performing network distillation:
    - Distil a larger network into a smaller one
    - Distil the same architecture into itself
- Learning from unlabelled data using pseudo labels

The dataset that will be used is **Imagewoof** (**not Imagenette** like for the NN homework) 160 px https://github.com/fastai/imagenette:

- The scripts provided by OpenAI are used to generate noisy-labels
    - Create your own script that will split your data in labelled and unlabelled data using some factor X (i.e. generate variations with 1% labelled data, 5% labelled data, etc.)
        - The splitting script should account for the original class distribution and try to keep the splits as close as possible to it. Provide normalised bar charts for each split for the class distribution.
    - You will be manipulating only the training set. The validation set will be the one originally provided with the dataset.

Task 1 - Distillation:

- Use a resnet18 as the default "large" network and train it for N epochs
- Use a 5 layer CNN as the "small" network (use the convolution blocks with normalisation and activations from the NN homework)
- You will be using a training set with 100% labelled data
- Compare the following experiments:
    - Distilling the pretrained resnet18 into a small network (the small network is not pretrained)
        - Train the resnet18 for N epochs to get resnet18 baseline
        - Train the small network for N epochs to get small network baseline
        - Compare these 2 baseline with the distilled small network
    - Distilling the small network into itself (train the small network for M epochs, create a new copy of the small network and distil the pretrained one into it for M epochs, repeat the process X times)
        - 2M * X should be as close as possible to N
        - For the same M and X values, compare vanilla distillation to non-maximum-permutation distillation (you permute all teacher output values except for the maxim class score label; i.e: if teacher outputs [0.2, 0.8, 0.05, 0.05] then the distillation target could be [0.05, 0.8, 0.2, 0.05])
        - Compare over 3 sets of M and X values for vanilla distillation
        - Compare the best set of M and X with the resnet18 distilled network
        - Compare an ensemble made from all the intermediary networks obtained from the best set of M and X with the resnet18 distilled network

Task 2 - Pseudo Labelling:

- Generate the following training splits: **100%, 75%, 50%, 25%, 10%, 5%, 1% labelled** data (the 100% split will not need to be tested if you choose one of the architectures used in the previous task)
- Choose any architecture that you want and compare a pseudo-labeling strategy over all dataset splits for N epochs
  - An epoch is considered 1 full pass through the labelled and unlabeled datasets. You can choose any weaving strategy you want (i.e, for the 10% split : 1 labelled image for every 10 unlabelled images / go through all the labelled images first, then through the unlabelled ones / go through the unlabelled ones first, then through the labelled ones etc.). The chosen weaving strategy should be kept consistent throughout all experiments.
  - You can choose any pseudo-labeling strategy you want. We recommend that you opt for FixMatch as it's easy to implement and it serves the purpose of this homework. The minimum requirement for the pseudo-labelling strategy is that it is some variation of UDA-like methods (i.e. MixMatch, ReMixMatch, Noisy Student training, etc.):
    - You are performing some sort of augmentation on the unlabeled samples
    - You are applying some sort of consistency criterion w.r.t the final label of the unlabeled samples (augmented should perform the same as non-augmented / label is the average label of multiple augmented views etc.)

Comparison guidelines:

- Plot overlapping line charts only with the train and validation set accuracy. We are not interested in loss values. Use dotted lines for training values and solid lines for validation values.
  - Plot the accuracy at each epoch N
- Plot confusion matrices after final training epochs when performing comparisons, i.e:
  - resnet18 confusion matrix vs. small network confusion matrix vs. distilled small network confusion matrix.
  - pseudo-labelling confusion matrices for all training splits
- Choose N such that the loss curve for the small network resembles the logarithm function on the 100% labelled dataset. (small network baseline)