# Problem2

April 19, 2022

## 1 Problem2

This problem is based on two papers, by Mahajan et al. on weakly supervised pretraining and by Yalinz et al. on semi-supervised learning for image classification. Both of these papers are from Facebook and used 1B images with hashtags. Read the two papers thoroughly and then answer the following questions. You can discuss these papers with your classmates if this helps in clarifying your doubts and improving your understanding. However no sharing of answers is permitted and all the questions should be answered individually in your own words.

### 1.1 1

Both the papers use the same 1B image dataset. However one does weakly supervised pretraining while the other does semi-supervised . What is the difference between weakly supervised and semi-supervised pretraining ? How do they use the same dataset to do two different types of pretraining ? Explain. (2)

The difference between weakly supervised and semi-supervised pretraining is that weakly supervised pretraining use noisy labels or signals to pretrain the model whereas semi-supervised pretraining use the ground truth labels and unlabeled data to pretrain the model.

For **weakly supervised pretraining**, they use the hash tag to do the nosiy supervision.

For **semi-supervised pretraining** - They use the labelled data to initialize the teacher model - They use the predictions of this teacher model to rank the unlabeled images and pick top-K images to construct a new training data - They use this data to train a student model, which typically differs from the teacher model - Pre-trained student model is fine-tuned on the initial labeled data to circumvent potential labeling errors.

### 1.2 2

These questions are based on the paper by Mahajan et al.

#### 1.2.1 (a)

Are the model trained using hashtags robust against noise in the labels ? What experiments were done in the paper to study this and what was the finding ? Provide numbers from the paper to support your answer. (2)

The model trained on large-scale hashtag data is unexpectedly robust to label noise, and that the features learned allow a simple linear classifier to achieve state-of-the-art ImageNet-1k top-1 accuracy of 83.6% without any finetuning (compared to 84.2% with finetuning).

### 1.2.2 (b)

Why is resampling of hashtag distribution important during pretraining for transfer learning ? (2)

Using uniform or square-root sampling leads to an accuracy improvement of 5 to 6% irrespective of the number of ImageNet classes in the transfer task. Hashtag frequencies follow a Zipfian distribution, the reason can be the long-tail distribution of hashtags in the training data.

## 1.3 3

These questions are based on the paper by Yalzin et al.

### 1.3.1 (a)

Why are there two models, a teacher and a student, and how does the student model leverages the teacher model ? Explain why teacher-student modeling is a type of distillation technique. (2+2)

Benefits of teacher-student model:

- After the teacher network predicts the unlabeled data, it undergoes a top-K sampling, which ensures that the label noise in the new dataset is small.
- Due to the long-tailed distribution, some classes in the precision dataset have fewer samples, but there are usually many samples in the unlabeled dataset with a large sample size. The newly constructed dataset has the same number of samples per class, and the trained model works well for all classes, not just the main class.

Type of distillation technique: The student architecture is smaller than the teacher architecture, and the student architecture is trained on the teacher architecture's output. And student model are exposed to model data, so student model performs better than teacher model.

Distillation can be seen as a particular case of self-training, in that the teacher model makes prediction on unlabelled data, and the inferred labels are used to train the student in a supervised fashion.

### 1.3.2 (b)

What are the parameters K and P in stage 2 of the approach where unlabeled images are assigned classes using teacher network ? What was the idea behind taking $P > 1$ ? Explain in your own words. (2+2)

Sampling hyperparameter K: number K of images selected per class.

P-highest-score: For each image, they retain only the classes associated with the P highest scores, P being a parameter accounting for the fact that they expect only a few number of relevant classes to occur in each image.

P>1 is to allow some noisy labels and more than one supervision for some complex concepts. And $P > 1$ is a way to collect enough reliable examples for the tail classes that allow the model to learn something other than the predominant concept.

#### 1.3.3 (c)

Explain how a new labeled dataset is created using unlabeled images ? Can an image in this new dataset belong to more than one class ? Explain. (2+2)

After the P label for each images and N images for one class, the labeled images are dupliated P times in the dataset and we treat the new dataset as multi-class classification.

One image in this new dataset belongs to more than one class, but it is duplicated as the task is still multi-class classification, not multi-label classification.

#### 1.3.4 (d)

Refer to Figure 5 in the paper. Why does the accuracy of the student model first improves as we increase the value of K and then decreases ? (2)

The student model performance gets better first because we have more examples (varility and harness) for the training. The student model performance gets worse later because we give too much noise to the model training.