Report on the Test Task

Tobias Nauen

August 7, 2022

Experimental Setup

We trained a MobileNet-v2 with 10 classes on the first fold of the STL-10 dataset, using a LogSoftmax output, together with a negative log likelihood loss. This first fold of data only consists of 1000 labeled examples (100 for each of the 10 classes). This model is then used to gather more training data from the 100000 unlabeled images. This is done by using the models predictions as soft labels. Now, these labels are not perfect in any way; in fact, we expect them to be wrong for a lot of pictures. However, we can filter out the good labels by considering the models confidence in its prediction. This is especially important in this case, as the unlabeled images come from a different distribution that even has more classes then the training examples. To create a balanced dataset from these soft-labeled examples, for each class we take 1000 images that are soft-labeled as that class by the original model. These images are chosen on highest maximum of the LogSoftmax/Softmax output, as a measure of model confidence. The reason for this is that for these images, the model is confident in its prediction, making it likely that the prediction is actually correct, while the unlabeled images, that are not in any of the original classes are likely to get a low score, since the model has not seen similar images before, making the output essentially random. We then take this new dataset of 10000 images to train a new model (with the same architecture) and compare the resulting models on the STL-10 test set.

Results

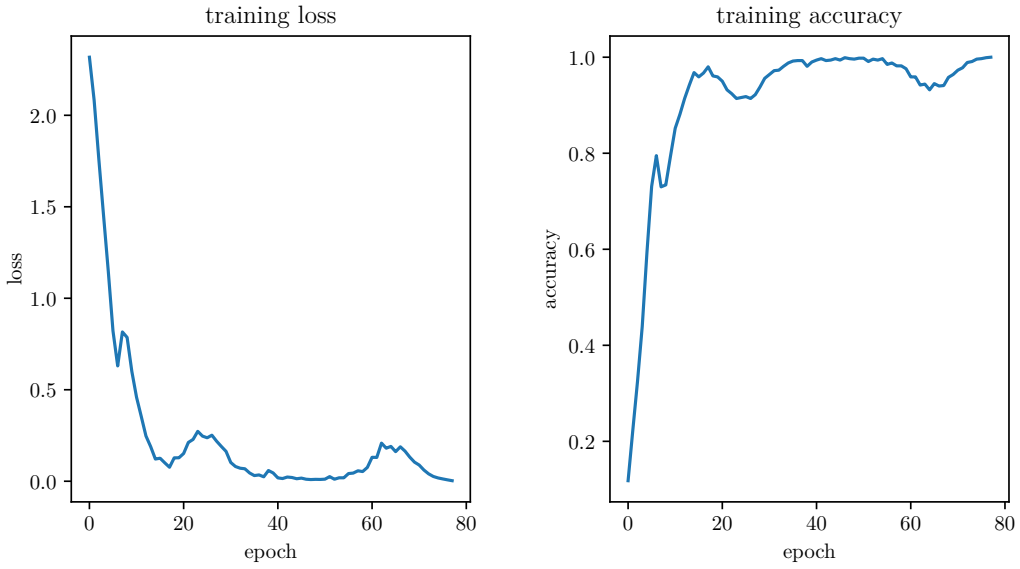

Figure 1: Training progress of the prior model.

Class	0	1	2	3	4	5	6	7	8	9
Relative size	10%	10%	10%	10%	10%	10%	10%	10%	10%	10%
Minimum confidence	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 1: Stats of the soft-labeled dataset of 10000 images.

When training the first model, the training (top-1) accuracy quickly surges to over 95% (as seen in Figure 1) and tops out at 100%. This is expected, as the model can quickly overfit to the training data. On generation of the new dataset from the unlabeled images, the minimum confidence in any of the classes still was 1.0, i.e. the model was sure in its predictions of all of these images.

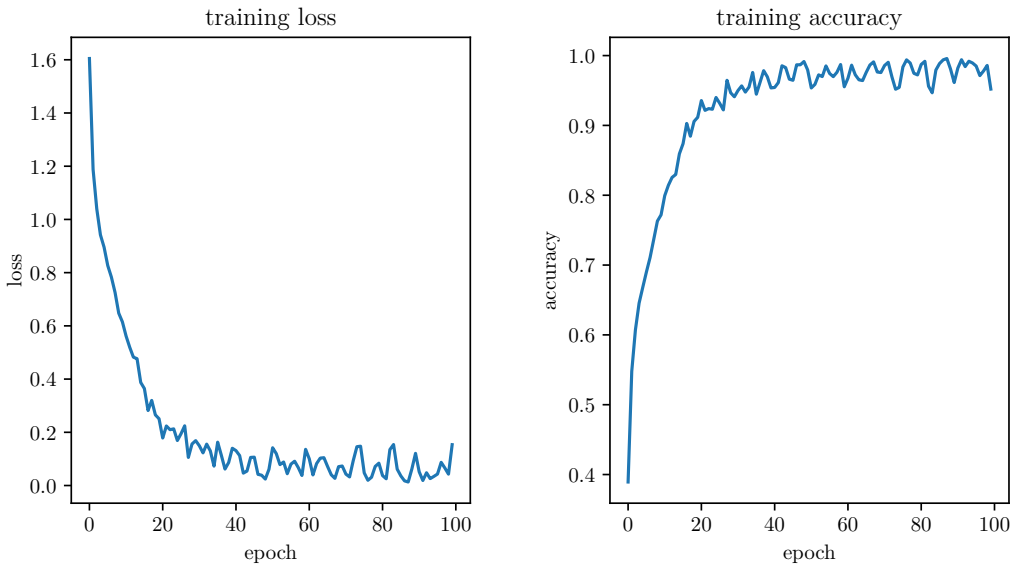

Figure 2: Training progress of the final classifier.

The model trained on the new dataset also obtains a training accuracy of more then 95%, but we expect it to generalize better to new, unseen images, since it was trained on a bigger dataset. This is shown to be true when evaluating on the test set. Here, the original model obtains an accuracy of 33.74%, while the second model has an accuracy of 37.88%. This is an accuracy boost of more then 10%.

Conclusion

Semi-supervised learning, that is machine learning, where some labeled examples (supervised) and a lot of unlabeled examples (unsupervised) are given is a useful tool in machine learning, especially when there is no dataset of labeled training examples, that is big enough for a given model architecture. It can generate a lot of training data, using a weaker (possibly smaller) model at first, which can then be used to train a larger classifier and therefore boost the classifiers accuracy. Semi-supervised learning relies on the fact, that examples without labels are much more common then labeled examples.
When fine tuning the original model on the new dataset, there would not be much of an improvement at all, since the model has no information to gain from the data. The examples are chosen, such that the model has a high confidence (close to 1.0). Therefore the loss on this dataset would be really small and no learning would happen. The new model, however, has not seen any images yet, and therefore can learn using the new dataset. The new model then is likely to generalize better, since it was trained on a larger dataset, given the labeling error was sufficiently small.