Active Learning Techniques

Albert Gregus

November 2021

1 Introduction

As current deep learning models are more and more data hungry, hence the current bottleneck for higher performances is to have large quantity, high quality training data. While gathering data is easier and cheaper, we need to label these for most machine learning tasks, which requires large amount of labour.

The main idea behind *active learning* is that instead of labelling all our data, we should just label the ones that offers the best model improvement. However, without labeling them, we do not have an exact measurement on how much particular datapoints would contribute. We have to rely on the properties of our model's preformance on unlabelled data to chose which ones are the most promising to label and include in the training set.

2 Experiments

I used cifar-10, a classification dataset for the experiments, which has 10 classes and consists of 50 000 training and 10 000 testing images. Every experiment is done with hyperparameter tuning and the the models with the best F1 scores on the test set were chosen to represent the performance of different active learning methods.

As a baseline, I ran an experiment with all the training data (*full_training_set*). This is the maximum KPI that could be reached with the given code if all the labelled data (50 000 images) are used.

2.1 Active learning techniques

To simulate an active learning scenario, the first 1000 image of the first batch file was used as 'already labelled', and another 5000 was selected from the remaining images in the training set as 'new labelings', using different algorithms for picking these. Finally, new models were trained on the original labelled plus the newly labelled data, and the methods were compared to each other.

To measure how much the newly labelled data has improved the model, a base experiment was run on the 1000 originally labelled (*original_label*) data. Compared to this, all the active learning techniques entail around 12% F1 score improvement (1).

2.1.1 Random pick

The easiest way to decide which unlabelled data to label is to pick it randomly (*random*). Besides being easy, it provides almost as much improvement as the other more elaborated methods (1).

2.1.2 Least certainty

The intuition behind least certainty picking is that if the model is not confident about the top prediction for an example, then the example is rather difficult or perhaps different from what the model has already seen [1]. Based on it, we pick and label the data where the model is the least confident (*least_certainty*). This gives 1% higher results than random sampling (1).

2.1.3 Least certainty margin

The idea behind least certainty margin sampling is that if the difference between the top two predictions is small, then the model must be confused between two categories. A good model should be able to separate categories as much as possible and thus, these examples are important to consider in the training data[1]. Based on this, we sample the data with the smallest confidence margin between their top two predictions (*least_certainty_margin*). Similarly to *least_certainty*, it provided 1% higher F1 scores than random sampling (1).

[1] https://towardsdatascience.com/active-learning-say-yeah-7598767806b2

2.1.4 Biggest entropy

Entropy can be defined 'information' or 'uncertainty'[2]. If the prediction probabilities have high entropy, that means the model is confused about an example[3]. Based on this, 5000 unlabelled samples with the highest prediction entropy were picked to label and use in the training (*entropy*). While the model trained on this provided better results, it only exceeded random sampling by 0.4% (1).

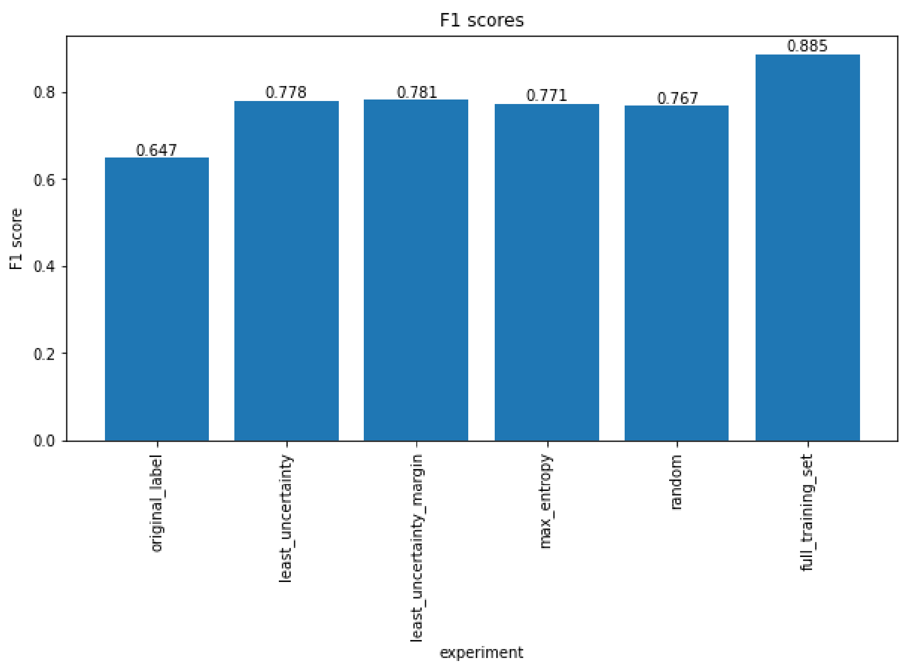

Figure 1: F1 scores for the different experiments

3 Conclusion

All selection methods provided around 12% improvement, halving the difference between the full train set baseline and the 1000 sample train. While all elaborated methods provided better results, there is only a 0.4-1% difference

[2] https://en.wikipedia.org/wiki/Entropy$_{(}information_{t}heory)$
[3] https://towardsdatascience.com/active-learning-say-yeah-7598767806b2

between their and random selection's performance. As a conclusion, more elaborated methods can provide better results, suggesting that with even more complex algorithms higher KPIs could be reached.