Paper Reading Assignment – 1
Sanyam Jain (P20QC001)
Indian Institute of Technology, Jodhpur
September 2, 2021

Paper Summary

An Empirical Study Of Example Forgetting During Deep Neural Network Learning (2019)

&

When Deep Learners Change Their Mind: Learning Dynamics for Active Learning? (2021)

Looking towards catastrophic forgetting, the authors in the paper investigated more towards determining forgettable and unforgettable events (event is a transition of predictions of classes between different cycles). After achieving remarkable accuracy by the model, even in this case, the model tends to forget previously correctly classified examples or forgettable examples; however, they also investigated that there are some examples which are learned such that they are unforgettable even in different seeds and changing epochs. One of the inherent problems in continual learning is catastrophic forgetting; the feature space shifts when the model is presented with the new dataset.

The forgettable examples are characterized by their different feature sets or underlying representations which are distant from the features of the original task samples in the input space. Precisely, authors define forgettable events as when an example is classified correctly at a time (t) in one batch cycle, gets incorrectly classified at a time (t+μ) where μ is the time spent after the mini-batch. They hypothesized and then revealed that the examples which are consistently forgotten do not have anything common in feature space with the tasks which are classified correctly for that task. In addition, this helped in 2 tasks. First, in reducing the dataset size without losing generalization capabilities. Secondly, analysing the forgetting examples to check which sample is the most informative and which samples are least informative. Informative means that, from which example you can extract most of the features which could help better generalize the task. This helps in eliminating those examples which are of no use. According to the author, in simpler words, the examples having different features which lie far from the decision boundary but belong to the classification task, are most informative and important to generalize whereas the ones which are classified with the best confidence score are less informative.

Past work in curriculum learning where all the researchers majorly focused on the acquisition functions (scoring function and pacing function) rather than considering the fact that examples are getting forgotten after some epochs. One author claimed that empirically correctly classified examples or unforgettable examples can be removed without any problem. Further, it is proved that the more noisy the example will be, the more easily it will be forgotten by the model. The process of examining the forgetting of examples in the mini-batches was previously done inefficiently. However, the authors devised a technique that helps to make decisions for what examples you need to worry about in earlier stages and then completely eliminate the chances of forgetting in later stages or cycles of training. They call it forgetting-statistics. In simpler words, we calculate the number of forgetting examples in the multiple mini-batches. This will tell how many times an example has undergone forgetting. In addition, the dataset is then sorted according to the forgetting statistics of the examples in the mini-batch.

Authors further characterized the forgetting events on the basis of (1) stability across seeds where they concluded that small forgetting brings the confidence towards the correct classification of that example, (2) forgetting by chance, (3) first learning events, where some examples which are learned in the earlier part of the training tend to be forgettable and unforgettable examples, are learned in very later stages of the training process (4) misclassification margin that is the difference between chances of getting classified as original label and chances of getting classified as other class(es) label. Lastly, the author proved experimentally that the removal of unforgettable examples does not disturb the overall performance of the training accuracy and generalization capabilities.

Why did we learn catastrophic forgetting? To understand the core idea of Label Dispersion. To start with, we can say that the more predicted label fluctuations you see, the higher the label dispersion will be and vice versa. Hence, recall we mentioned that when the model starts to shift its predictions in later stages of the training, we say that it has undergone catastrophic forgetting. Therefore, in active learning when we equip this kind of acquisition function authors have obtained benchmark results.

By definition, AL is a learning technique that helps researchers and practitioners to start training the model with limited labelled data. This technique achieves significant performance when it is allowed to choose which sample to label. Whatever labelled pool you have, start with that. Afterward, allow the model to sample from an unlabelled pool using the acquisition function. Once labelled, the newly classified example is then sent to the labelled pool and re-train using an updated training set. This process is repeated till the budget exhausts.

However we can understand that labels are being "forgotten" but in active learning, this concept is new as the previous studies were done only in a supervised fashion, in other words, previous methods were limited to labelled data. And hence is unsupported to AL. Hence from that idea, authors started to investigate in the direction of forgettable events in AL. Pulling intermediate results from the model for specific samples while it's already been training, helps to better understand the variations in the predicted labels. Now, this uncertainty is measured and estimated; which is called Label Dispersion (LD). Samples with low LD have a high confidence score of the classification task and vice versa.

Many approaches to achieve the acquisition function of AL underlies analysing the certainty of the unlabelled data. Now when authors performed the training and select some predictions between the epochs, they found that the labels which are more fluctuated within the epochs are the most informative ones and appropriate candidates to be labelled. This can be categorised in the following box: Label-Dispersion(LD) and Prediction confidence (PC)

High PC, High LD (Rejected)	**Low PC, Low LD (Not welcomed)**
High PC, Low LD (Accepted)	**Low PC, High LD (Rejected)**

Low LD and Low PC - Predicting correct label with low confidence
Low LD and High PC - Predicting correct label with less confusion
High LD and Low PC - Predicting incorrect class label with high confusion High LD and High PC - Predicting incorrect label with high confidence

As mentioned earlier (in papers 1 and 2), unforgettable samples can be removed easily and forgettable samples are the ones that have most of the underlying information which can generalize better, and hence it is better to learn the forgettable examples. The catch is we don't have the ground labels for

the forgettable examples. This is where the new learning approach which uses LD as the acquisition function helps. It encountered that samples that change frequently are similar to those which are forgotten in continual learning.

The trick (Addition of LD in AL pipeline):
Step 1: Train the model with k labelled dataset
Step 2: Predict the unlabelled examples and save the labels.
Step 3: Now using the label-dispersion score, sort the examples. (Highest LD being the most uncertain example)
Step 4: Add to the labelled pool and remove from the unlabelled pool.
Step 5: Perform the AL cycle and repeat till convergence
Step 6: Evaluate