Script for Assignment 1 - Paper reading

Slide 1: Good morning sir, today I am going to present paper reading assignment where I learned very interesting topics related catastrophic forgetting, continual learning and active learning. I chose 2 papers which I term as base paper and additional paper. Base paper is on the topic which is greatly known in machine deep learning world called as catastrophic forgetting in continual learning and additional paper have some developments build upon catastrophic forgetting in active learning.

Slide 2: Why was I left with no other option than to chose 2 papers? it is because the paper I chose earlier was teh additional paper while it was relying much on the fundamentals of base paper. Thus it would be in-justice to the base paper if I had started with additional paper directly. And catastrophic forgetting is the topic which opened a new line of research after publication in 2018. So it was necessary to study the base paper to study additional developments in additional paper. We will study what is forgetting and what are forgettable or non-forgettable events in learning. In later parts of the presentation I will also present a methodology which is recently developed in the community which is called as label-dispersion. Authors from the additional paper have described label

dispersion as a tool which helps to sort the dataset and act as an acquisition function in Active Learning.

Slide 3: As mentioned, I will cover following content through the presentation.

Slide 4: An event is a transition of prediction of classes through the learning cycles. Examples that are tend to forgotten by the model in later stages of the learnning process, in other words, the predictions which were coming out to be correct in the earlier parts of the learning are getting incorrect in the later stages of the learning process. Such shift in the prediction is called as forgetting. In addition, when the examples which cannot be forgotten even in the later stages of the learning process are unforgettable. The informativeness is judged on the basis of examples which are forgettable. That is we need to identify that the exaples which are easily forgotten will be a good candidate for generalization. And on contrary we have the examples which are unforgettable throughout the training process are good candidate to be eliminated from the dataset because of being less informative. In the additional paper, author observed that with the active learning pipeline, most of the examples are getting different labels with susequent learning cycles. Therefor its a need to devise an optimization function which could modify the AL pipeline so that we could identify the

easy to difficult samples after each cycle.

Slide 5: A forgetting event is said to happen when accuracy of the example i decreases from time t to time t+∂. Formally, accuracy(of example i) at time t is greater than accuracy(of example i) at time t+∂. Classification margin The classification margin m is defined as the dif- ference between the logit of the correct class and the largest logit among the other classes. And unforgrttable examples are already defined in earlier slide. Now this diagram shows the complexity of the datasets from easy to hard in terms of learning. The MNIST dataset which contains example with reach features, diverisity in the examples and the most important they are least complex. This graph says that there are less forgetting events in the MNIST dataset or in other words, the MNIST dataset have more unforgettable examples. In contrary, the CIFAR-10 being the hardest dataset have large number of forgettable examples. As you can see in graph we still have significant number of forgetting events occuring in the later stages of the learning cycle.

Slide 6: Forgetting Statistics : The authors devised a technique that helps to make decisions for what examples you need to think 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. You initialize the empty variables namely previous accuracy and forgetting across the training samples which is an array. You start the loop till training stops, sample a minibatch B from the dataset D. Then, for the example i belong to minibatch B computer accuracy for the example i. If previous accuracy of the example i is greater than current accuracy of the example i then we need to worry about that and comnsider it as forgetting event. Then add the example to our forgetting array. You also need to update the running previous accuracy variable with this current accuracy. Now update the weights of the classifer on mini batch.

Slide 7 : 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.

Slide 8: Now to optimize the dataset, its claimed by the authors that we can safely remove the unforgettable examples. The graph shows the fallacy of the accuracies at the three different stages. First when no data is removed denoted by red line. Second, when you remove selected samples that are of no use to the learning algorithm that is unforgettable examples. and third, removing random examples. You can observe that only 2% change can be observed in the accuracy when you remove selected examples, where as graph falls 5% in the accuracy when you removed samples randomly.

Slide 9: After learning forgetting and learning dynamics, authors in the additional paper observed the entropy of the predicted labels. They call it as label dispersion. This terminology is nothing but measures the randomness of the predicted labels. They observed that the model changes its decision in almost every learning cycle / epoch. This randomness is quantized using label-dispersion acquistion function. Just need to remember that samples with high Label dispersion are the ones

where model is unsure about that example where are the sample with lowest label dispersion is the one where model is comfortable and have consistently resulted same class label.

Slide 10: we shall not confuse Prediction confidence with label-dispersion because, prediction confidence tells by what percentage the model claims that a particular example belongs to a particular class. Where as label dispersion is the quantity which measures the flutuations of the predicted labels. Now here we can see that among 4 examples, example a is getting classified as car consistently hence LD is negligible where as all rest of the 3 examples have high label dispersion.

Slide 11: How do we integrate Label dispersion in active learning pipeline? 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
The key idea is again informativeness which we saw in earlier part of the presentation that, the more label dispersion is, the more the mode is uncertain about the sample and the more is forgetting statisitcs and hence it will be the best candidate to

label. As I mentioned that forgettable examples are the most informative ones and unforgettable examples are the least informative, hence it would be a great advantage to query the uncertain sample (or forgettable sample) which will help to increase the generalization performance.

Slide 12: Limitation in the base paper (Toneva et al 2019)
Unlabelled dataset case was not considered. (Active Learning)
Causes of forgetting is but not limited to shift in training samples.
The results of the datasets after removing unforgettable samples from the dataset were not explained class wise.
Hypothesis space is loosely defined which makes it difficult to connect catastrophic forgetting with forgetting statistics.
Forgetting is explained profoundly, however it lacks with proper metric to quantify forgetting.
They have not explained about the use of noisy labels over noisy data and methods to perturb.
No mention of Batch Normalisation and Covariance shift!

Slide 13: Limitation in the additional paper (Bengar et al 2021)
This research only focuses on the score of label-dispersion, one step ahead of this work could hypothesise the underlying behaviour of calculated

LD scores. question to put is - Can we learn multiple class labels along with their LD scores such that you need not calculate the divergence or distance from all the classes.
In other words, for example, a cat is classified as frog with high PC and high LD. Now, can we just ignore rest of the classes and select first K labels such that a penalty is given for each wrong label.

Slide 14:
In simple terms, can I keep track of the labels along with their PC and LD for candidate examples and collect those examples from each class such that, we learn which class has the least possibility to be assigned with current example. for example below, if we assume cat and frog has such PC and LD score that it could not be anything apart from these 2 then our AL process can be relaxed.

Slide 15: can be skipped as of now

Sldie 16: Here are the key takeaways from the presentation!