Programming Assignment 1 Report
Advanced Machine Learning

An Empirical Study of Example Forgetting during Deep Neural Network Learning

By Sanyam Jain (P20QC001)

Terminologies and Definitions:

1. Catastrophic forgetting: Phenomenon where Neural Networks tend to lose the previously learned information when its trained for the new task. Here, each mini batch is a representation of the dataset with different features. If used mini-batch SGD, the learning model starts to forget the events learned with previous batches in continuous learning as each mini-batch can be considered as a new task.
2. Forgettable examples: An example which is learned in previous training representations but is being misclassified in later parts of the training.
3. Unforgettable examples: Examples which are learned by the model such that it is not misclassified after certain epochs of training.
4. Curriculum Learning: Learning where we start with easy training set and then increasing the capacity and complexity of the training samples for training. SGD with curriculum learning creates mini-batches w.r.t. hardness. That is, mini-batches are created on the basis of hardness by scoring the training samples from easy to hard. Then we say that can I use the easiest sample first and then train the model. Then we pick the second easiest sample but harder than the earlier one. This way we keep on doing.
5. Hardness: How difficult is the sample to be recognized by the model.
6. Scoring Function: Gives a score whether sample is easy or hard.
7. Pacing Function: Tells how do you want to increase the learning process. This is monotonically increasing function so that likelihood of the easier examples can only decrease. Samples which are seen in first batch should have less hardness than the samples which are seen in next batch. This helps to prevent redundancy.
8. Curriculum learning technique helps to decide which example can be safely removed from the dataset. In other words, as we have seen in the experiment that there are a significant number of unforgettable examples which are classified correctly in all the learning presentations. The model is not getting

any benefit from those samples in learning features. Hence comes the terminology of informative and less-informative.

9. The samples which are informative are forgettable in different learning epochs and samples which are less-informative are unforgettable. Thus it can be said that it is no harm to remove the unforgettable samples as they are not contributing towards learning important features for classification. Hence safe to remove unforgettable examples.
   For example,

(a)

(b)

(c)

(d)

Here in the figure above, it is clear and intuitive that (c) and (d) are the images of airplane thus they are unforgettable and more informative. (a) and (b) are having difficult to recognize features and not clearly to visualize as airplane. Thus forgettable and more informative.

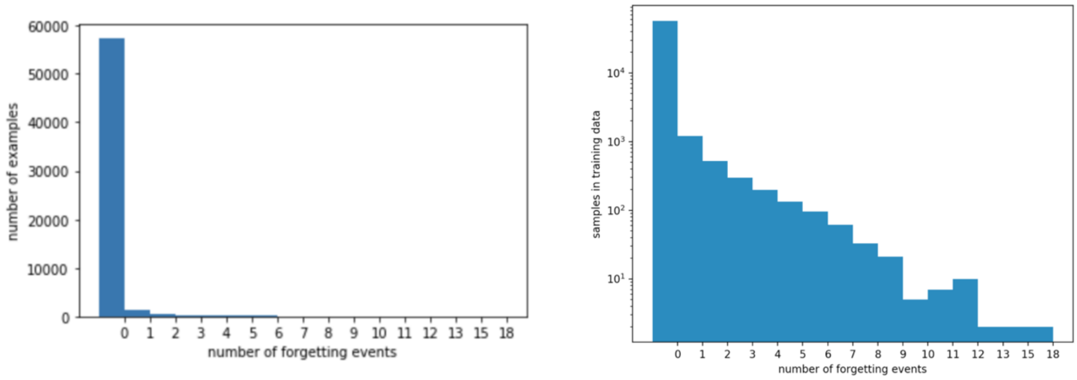
10. Misclassification margin: The difference between chances of getting classified as original label and chances of getting classified as other class(es) label.
11. The authors devised a technique that helps to make decisions for what examples you need to think about in earlier stages and then completely 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 consider 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 classifier on mini batch.

Procedure Solution:

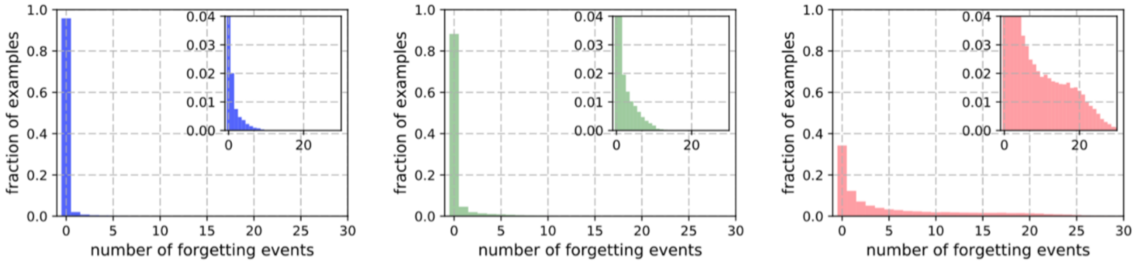
1. Train a classifier on a given dataset (MNIST/CIFAR)
2. Record forgetting events for each example. Note down which examples are getting misclassified over current mini-batch before SGD.
3. Assign value from 0 to 1 to each example and sort the examples based on the forgetting statistics.
4. Assign flag to the examples whether forgettable or unforgettable. And remove the unforgettable examples to reduce the dataset size.
5. Curriculum learning of the remaining examples.

Experimentation & Results: (Please check the detailed readme for stepwise python execution) Our expectations from the experimentation is to get information about mini-batch SGD & training examples in those mini-batches. In addition, to determine which examples are important to learn or more informative than examples which are less important to learn and less informative. (Forgettable vs unforgettable)

1. Number of forgetting events vs number of samples in MNIST



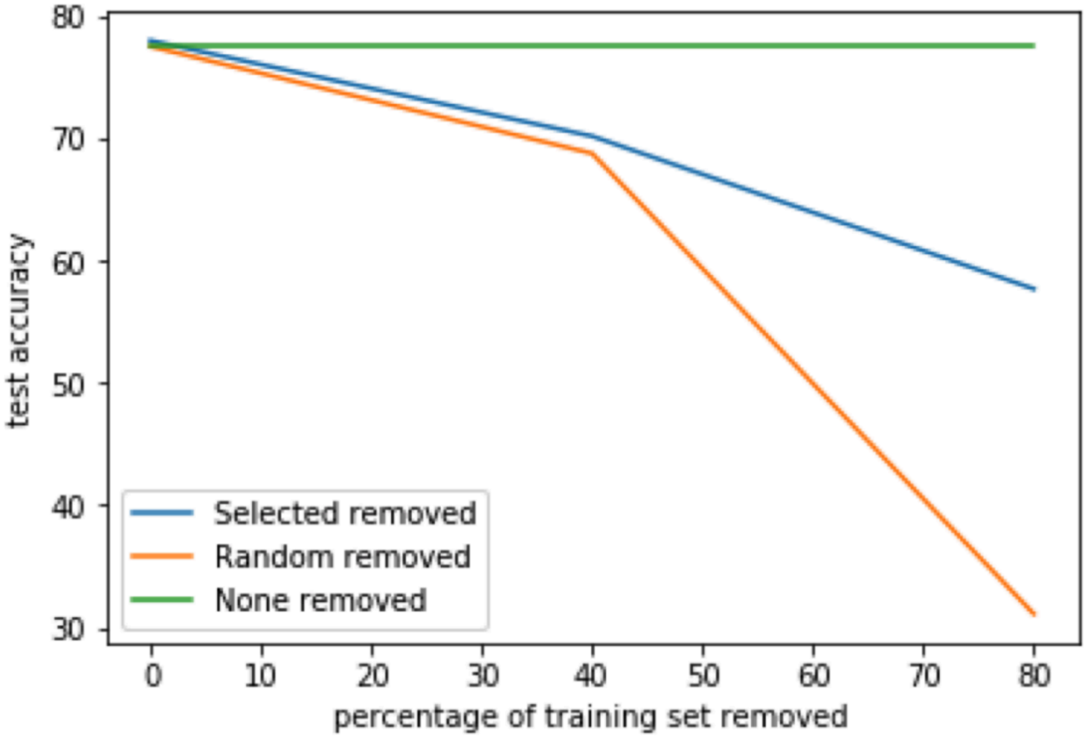
(output images MNIST)



(image from paper toneva et. al.)

Figure shows the number of forgetting events in training cycles of MNIST, Permuted-MNIST and CIFAR10 respectively. We can observe here that MNIST dataset is easy to learn by the deep learning classifier and hence number of forgetting events are very less. In other words, MNIST dataset does not have complex features and hence it is easier for classifier to learn. Thus decrease in the number of forgettable events and increase in number of unforgettable events. However in the last figure, where CIFAR dataset is used, it can be seen that the dataset is difficult to learn by the classifier and hence the forgetting events are more.

2. CIFAR10 with resnet18



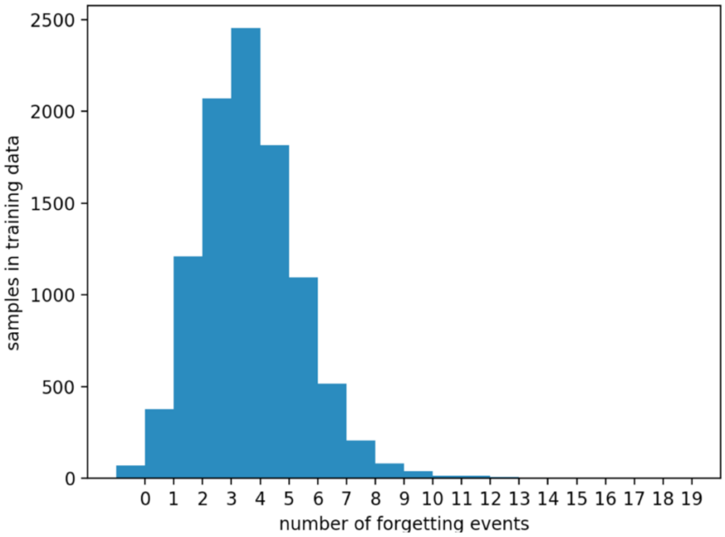
Test accuracy vs percentage of training set removed. That is green plot shows no change because nothing removed from the dataset. Orange curve shows random examples are removed from the dataset. In other words, you just subsample some examples randomly and throw them and check test accuracy. Blue curve shows how removing the selected samples meticulously and carefully prevents from bad testing accuracy. That is, when you have calculated forgetting statistics and have found what are the unforgettable examples, that can be removed safely without affecting much of the testing accuracy.

NOTE: The change between "none removed" and "selected removed" is produced larger than expected results because I followed 2 strategies to compensate the computational costs:

- *Reducing last 20% of the examples from each of the classes.*

- Reducing Epochs from 200 (as author have used 200 epochs for all their results) to 25.

3. The graph shows forgetting events for noisy CIFAR10 (resnet18). This shows that more the noisy dataset will be, more forgetting will be observed. In other words, examples with more noise are expected to be forgotten easily. Labels were changed randomly.



4. Comparison between datasets and investigating their complexity
    a. MNIST being easy to learn has less number of forgettable examples
    b. Permuted MNIST being moderate to learn have moderate number of forgettable examples.
    c. CIFAR10 and CIFAR100 being the most difficult to learn for the classifier has more number of forgettable examples.
    d. Examples which have undergone multiple noise and transformations are prone to forgetting event.
    e. Unforgettable examples can be removed safely along with some far-off forgettable examples in the feature space.
    f. Model learns most of the information from the forgettable examples