AN EMPIRICAL STUDY OF EXAMPLE FORGETTING DURING DEEP NEURAL NETWORK LEARNING

Presented by: Qi Dang

2019.03.11

Questions

- Can we compress the dataset without compromising generalization accuracy?
- Can we use forgetting statistics to identify ”important” samples and detect outliers and exmples with noisy labels?

Introduction

- Neural networks cannot perform continual learning
- Catastrophic forgetting[1]: forget previously learnt information when trained on new task (input distribution changing)

- In this paper: Investigate the fotgetting process occurs in the same task

[1] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and Others. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, pp. 201611835, 2017

Terms definition

- Forgetting events: A sample is classified correctly in time step t but misclassified in time step t+1
- Learning events: A sample is misclassified in time step t but classified correctly in time step t+1
- Unforgettable examples: samples are learnt at some point and experience no forgetting after

Settings for forgetable examples calculation

- Standard image classification (MNIST, permuted MNIST, CIFAR-10)
- A neural network optimised by SGD
- 5 random seeds for each dataset

Experiment procedures

Algorithm 1 Computing forgetting statistics.

initialize $\text{prev\_acc}_i = 0, i \in \mathcal{D}$
initialize forgetting $T[i] = 0, i \in \mathcal{D}$
while not training done **do**
  $B \sim \mathcal{D}$ # sample a minibatch
  for example $i \in B$ **do**
    compute acc_i
    if $\text{prev\_acc}_i > \text{acc}_i$ **then**
      $T[i] = T[i] + 1$
    $\text{prev\_acc}_i = \text{acc}_i$
  gradient update classifier on B
return T

Number of forgetting events

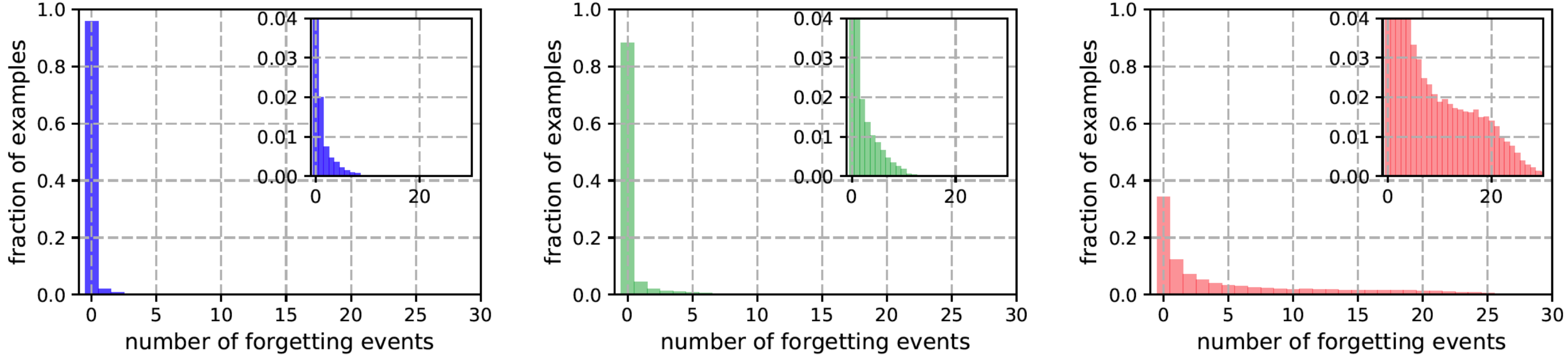


Figure 1: Histograms of forgetting events on (from left to right) *MNIST*, *permutedMNIST* and *CIFAR-10*. Insets show the zoomed-in y-axis.

- Number of forgetting events=0 means unforgettable
- Unforgettable samples across 5 seeds: 91.7%(MNIST), 75.3%(premuted MNIST), 31.3%(CIFAR-10)

Stablility across seeds

- They compute the number of forgetting events per example for 10 different random seeds. The average Pearson correlation is 89.2%
- When split the 10 seeds to 2 group of 5, cumulated number of forgetting events within those two sets shows a high correlation of 97.6%

Stablility across seeds

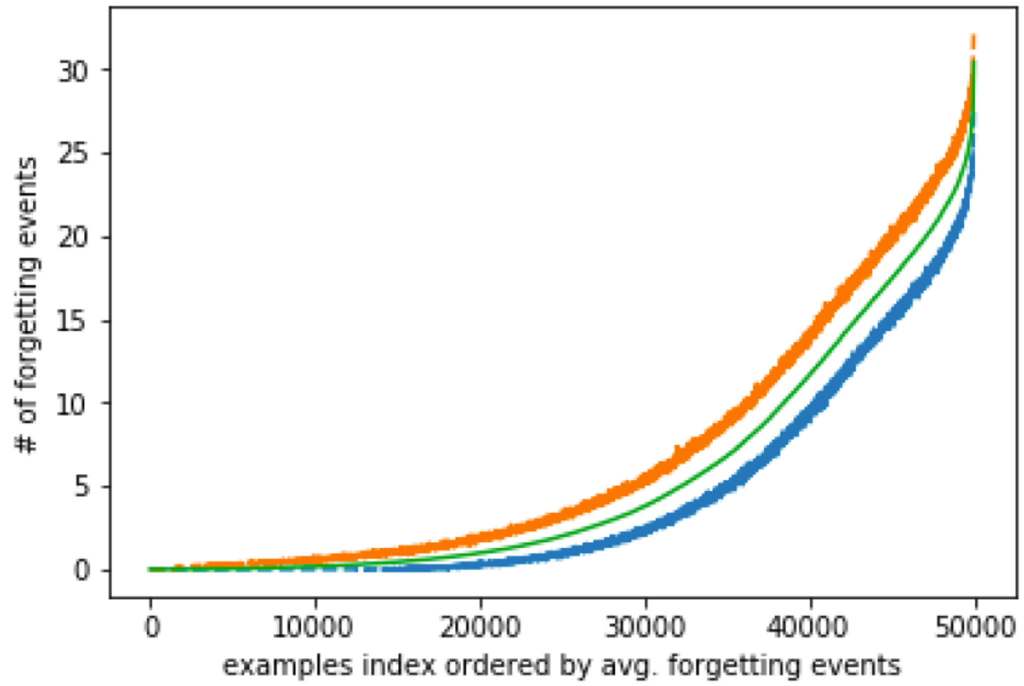


Figure 14: 95% confidence interval on forgetting events averaged over 5 seeds.

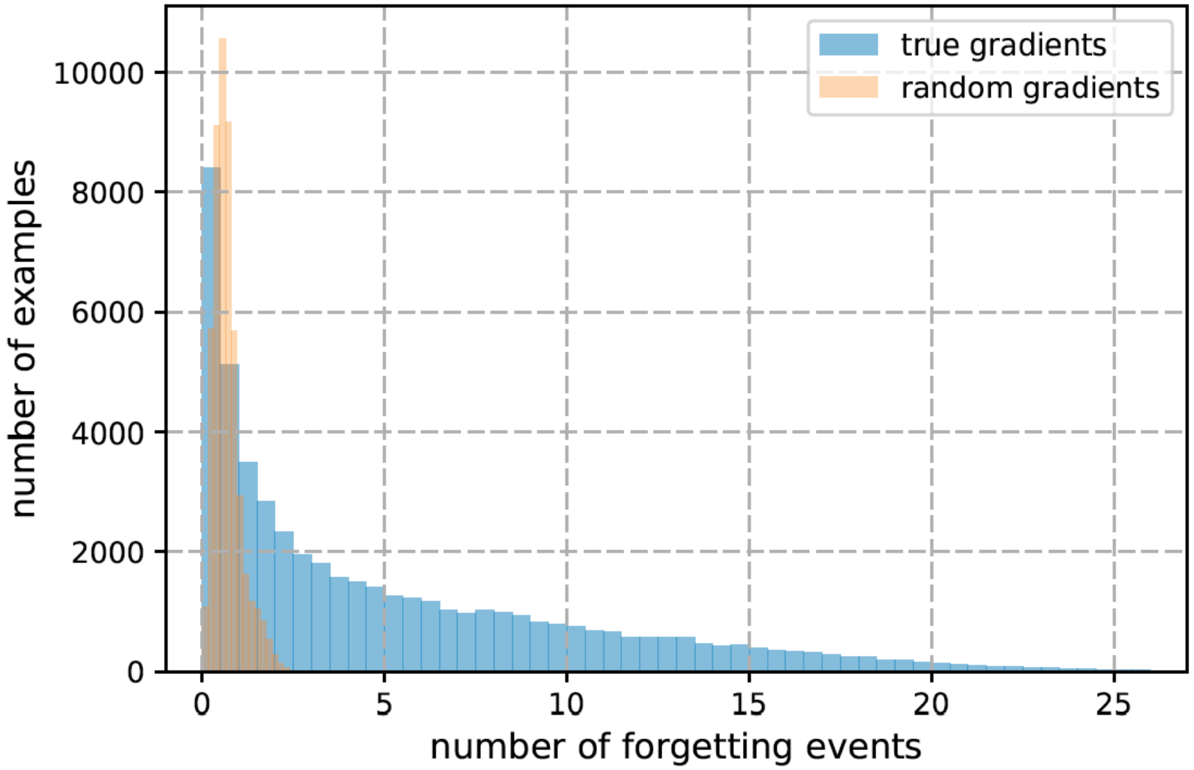
Forgetting by chance

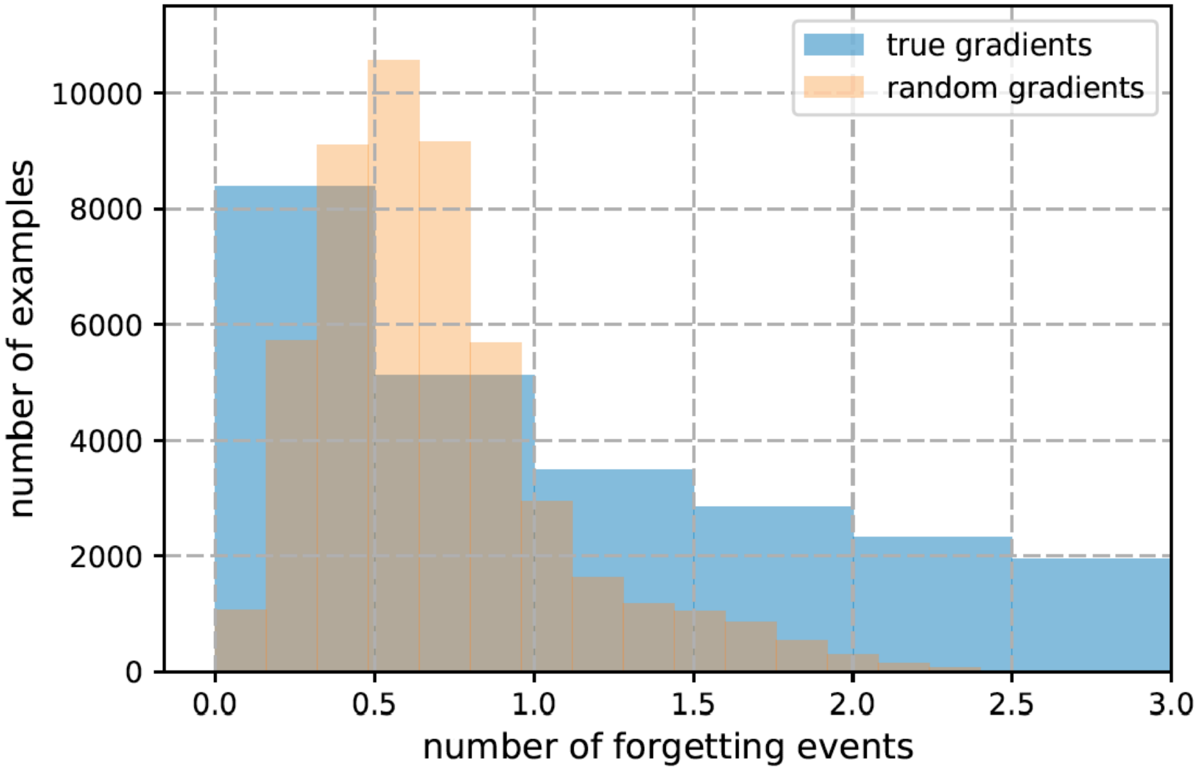
- compute the forgetting events of a classifier obtained by randomizing the update steps

?

1. Before the beginning of training, clone the "base" classifier into a new "clone" classifier with the same random weights.
2. At each training step, shuffle the gradients computed on the base classifier and apply those to the clone (the base classifier is still optimized the same way): this ensures that the statistics of the random updates match the statistics of the true gradients during learning.
3. Compute the forgetting events of the clone classifier on the training set exactly as is done with the base classifier.

Forgetting by chance

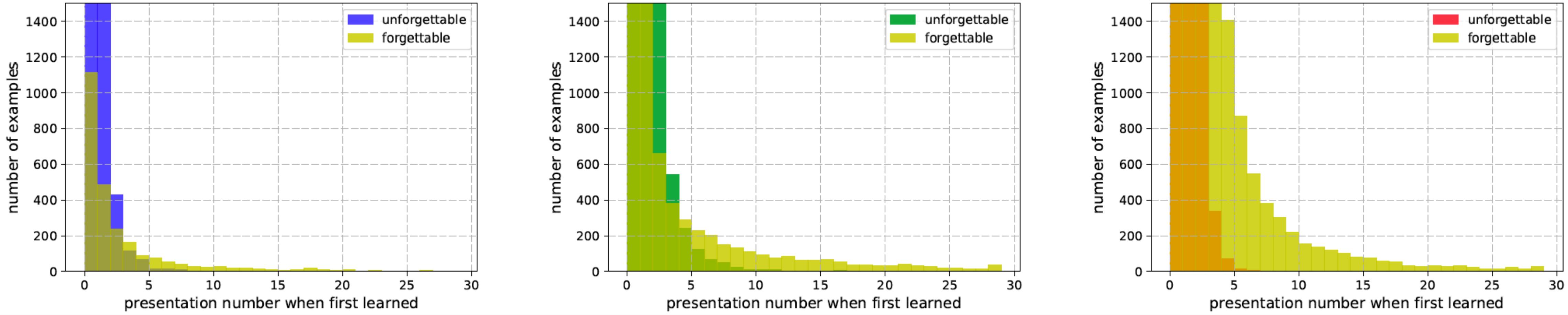




Zoomed-in

First learning events

- How many times dose a sample need to be presented to a model before it is learnt by the model learning?



MNIST   Permuted MNIST   CIFAR-10

Unforgettable examples are easier to learn

Missclassification margin

$$p(y_i|\mathbf{x}_i;\theta) = \sigma(\beta(\mathbf{x}_i))$$

$$m = \beta_k - \arg\max_{k' \neq k} \beta_{k'}$$

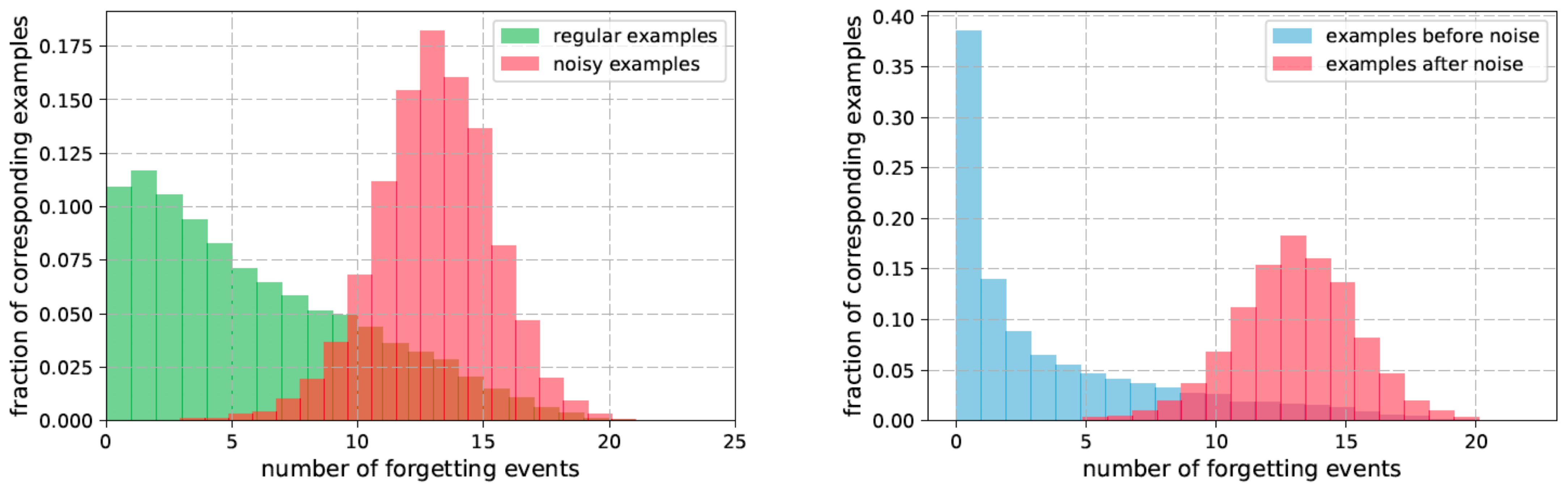
- m: missclassification margin
- β: logits of prediction
- σ: sigmoid(softmax) activation function
- k: index of correct class

The Spearman rank correlation between an example's number of forgetting events and its mean misclassification margin is -0.74

The examples that is easier to forget probably have larger missclassifation margin(are classified worse)

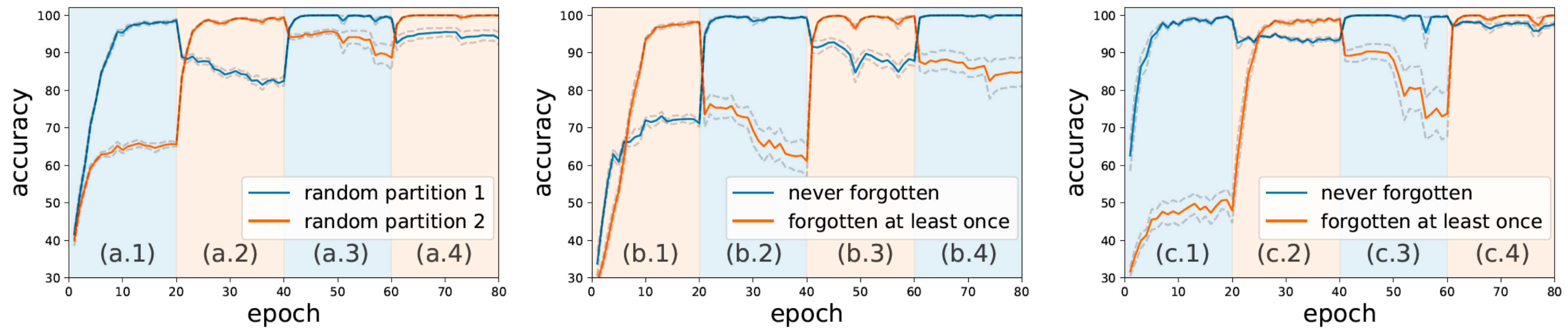
Visual inspection

Detection of noisy examples



- The label of 20% of CIFAR-10 samples are changed

Continual learning setup

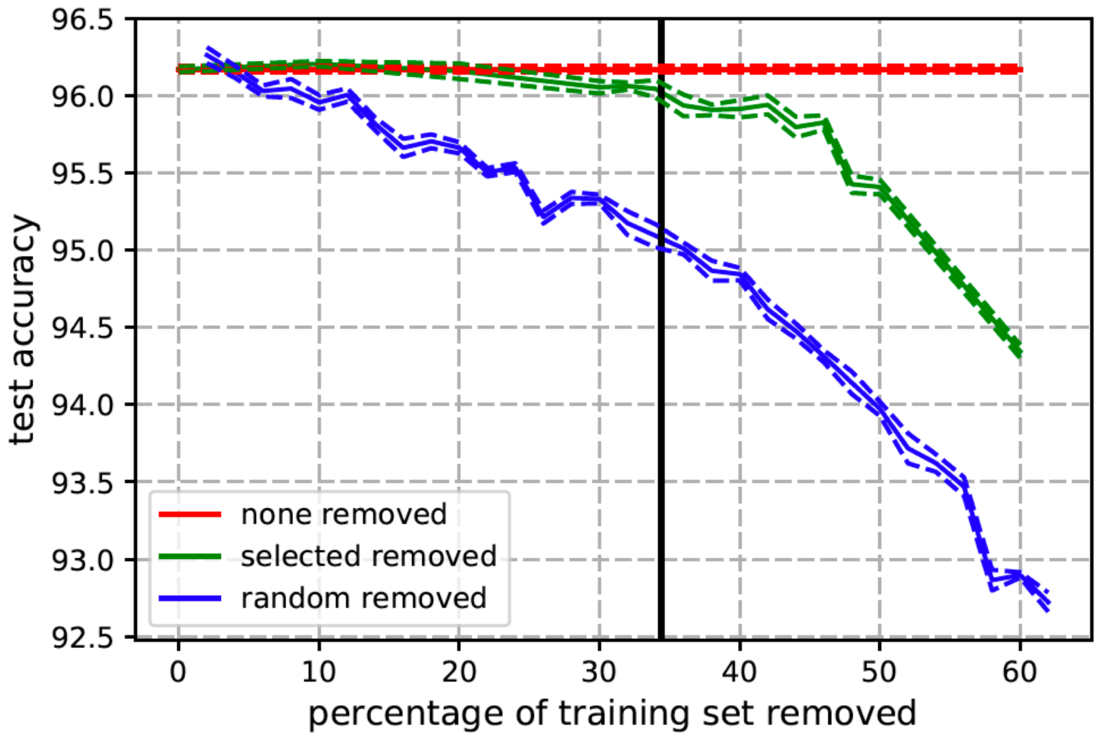


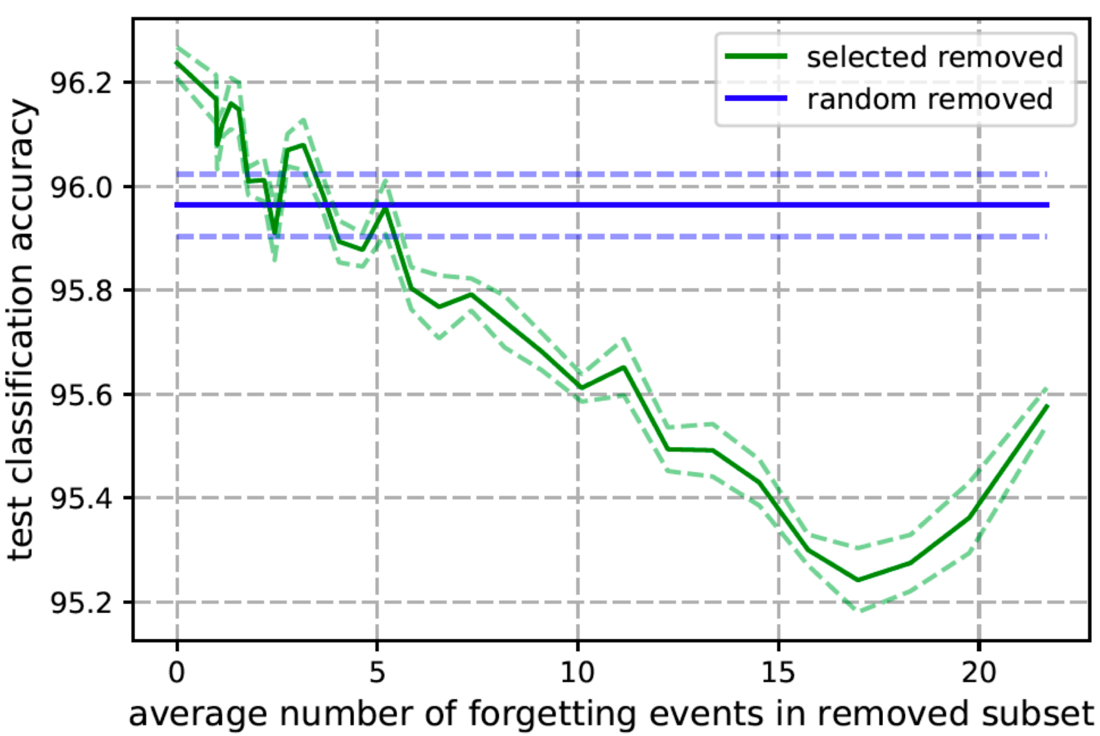
(a) random partitions

(b) partitioning by forgetting events

- (a) 10K examples are randomly sampled from CIFAR-10 and are divide to two groups(5k for each)
- (b) 5K examples are never forgotten, 5k examples are forgotten at least once

Removing unforgettable examples





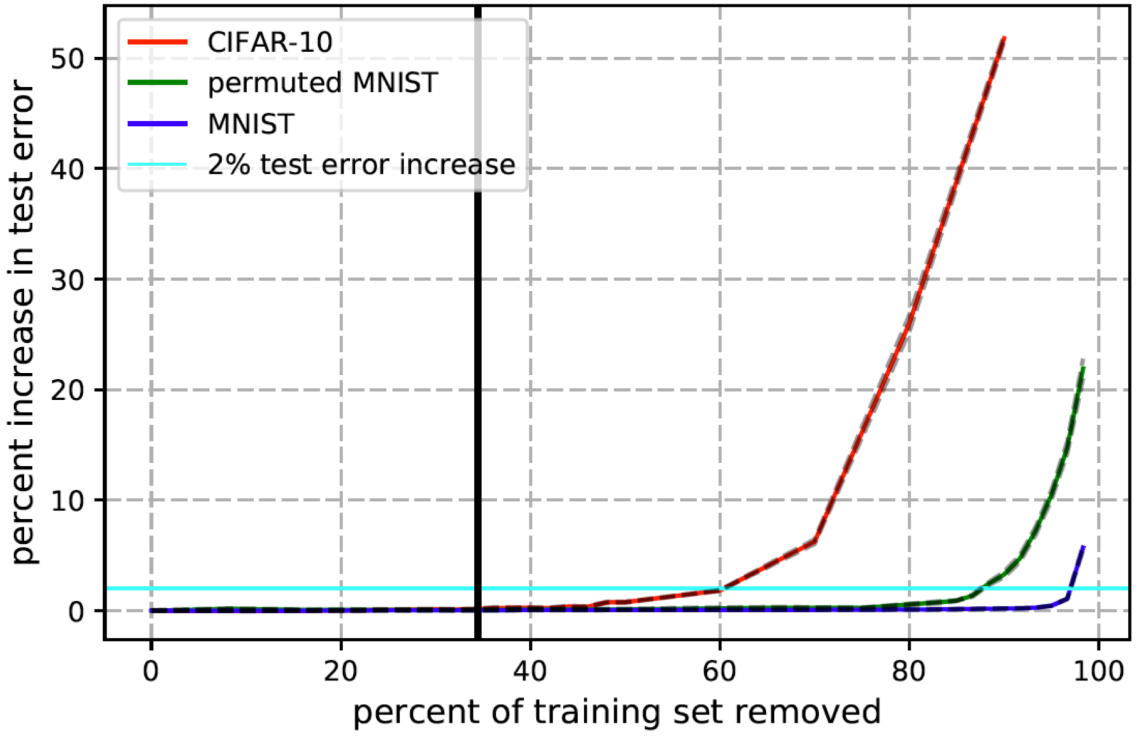
- Left: remove the most unforgettable examples step by step
- Right: remove 5k forgettable examples

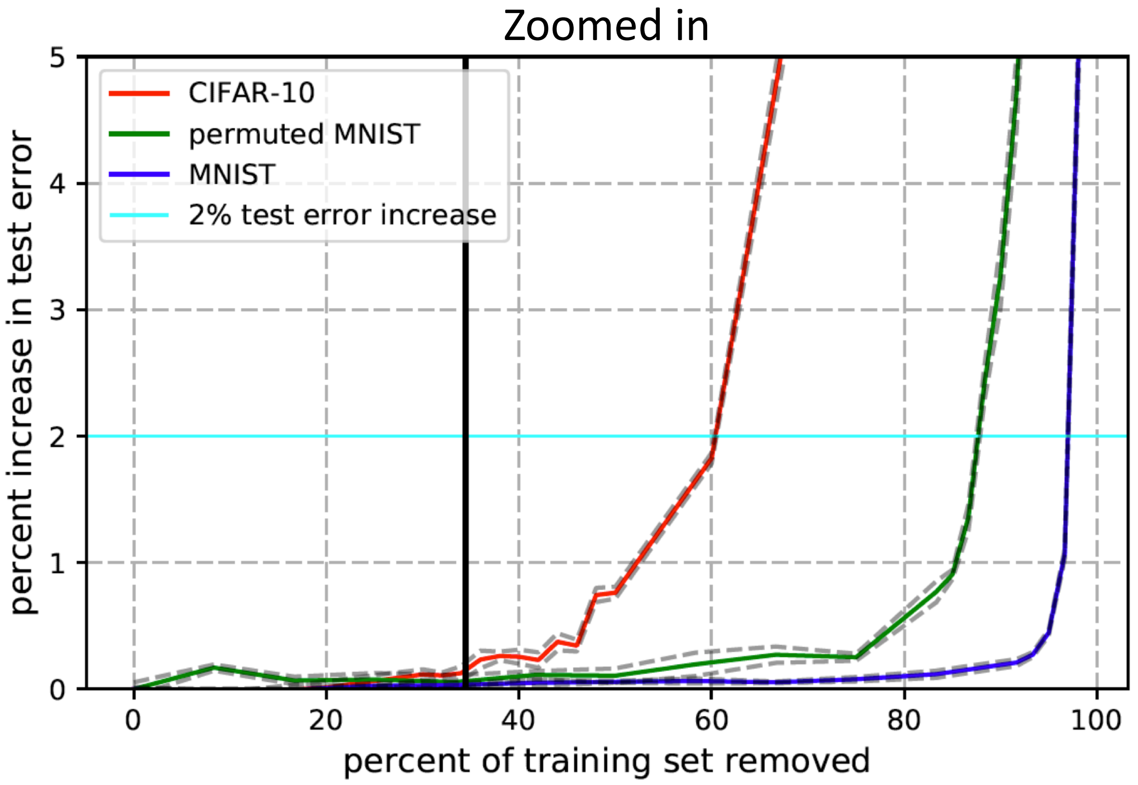
Analysis

- On separable data, a linear network will learn such a maximum margin classifier[1]
- Forgettable samples can be considered as support vectors (which are closer to decision boundry)

[1]Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The Implicit Bias of Gradient Descent on Separable Data.

Intrinsic dataset dimension

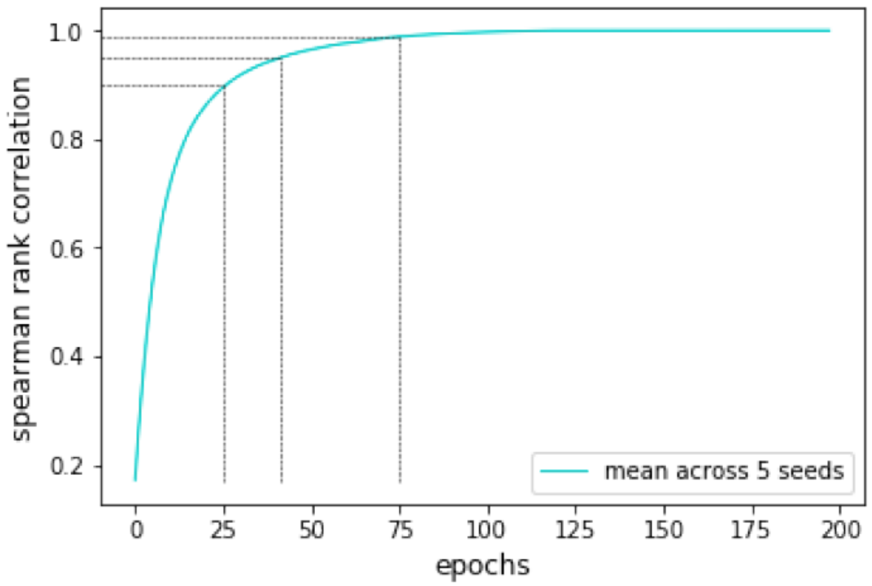


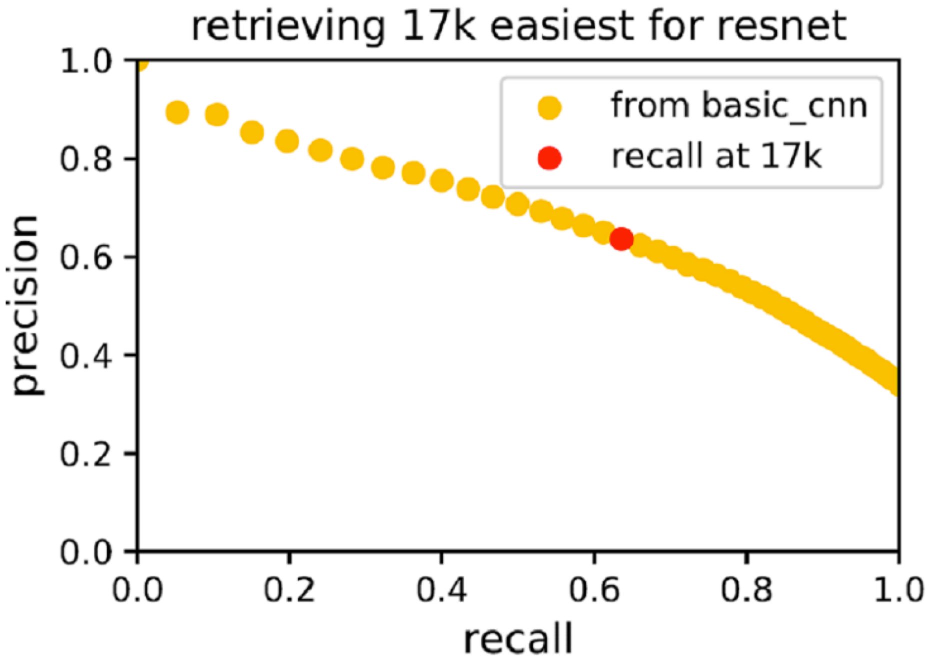


for a given architecture, the higher the intrinsic dataset dimension, the larger the number of support vectors, and the fewer the number of unforgettable examples.

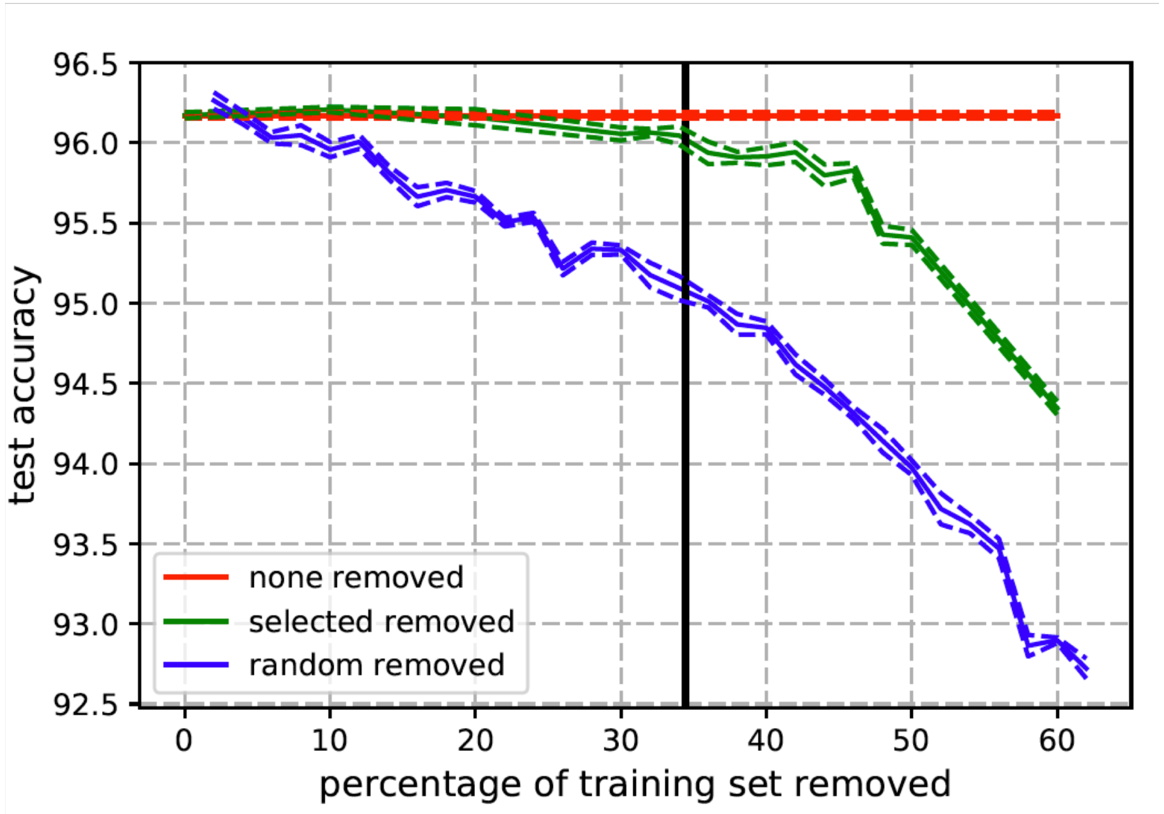
Transferable forgetting event

- Do we have similar statistics if we reduce the train epoch？
- Do we have similar result if we use the statics from one model but test in another model？

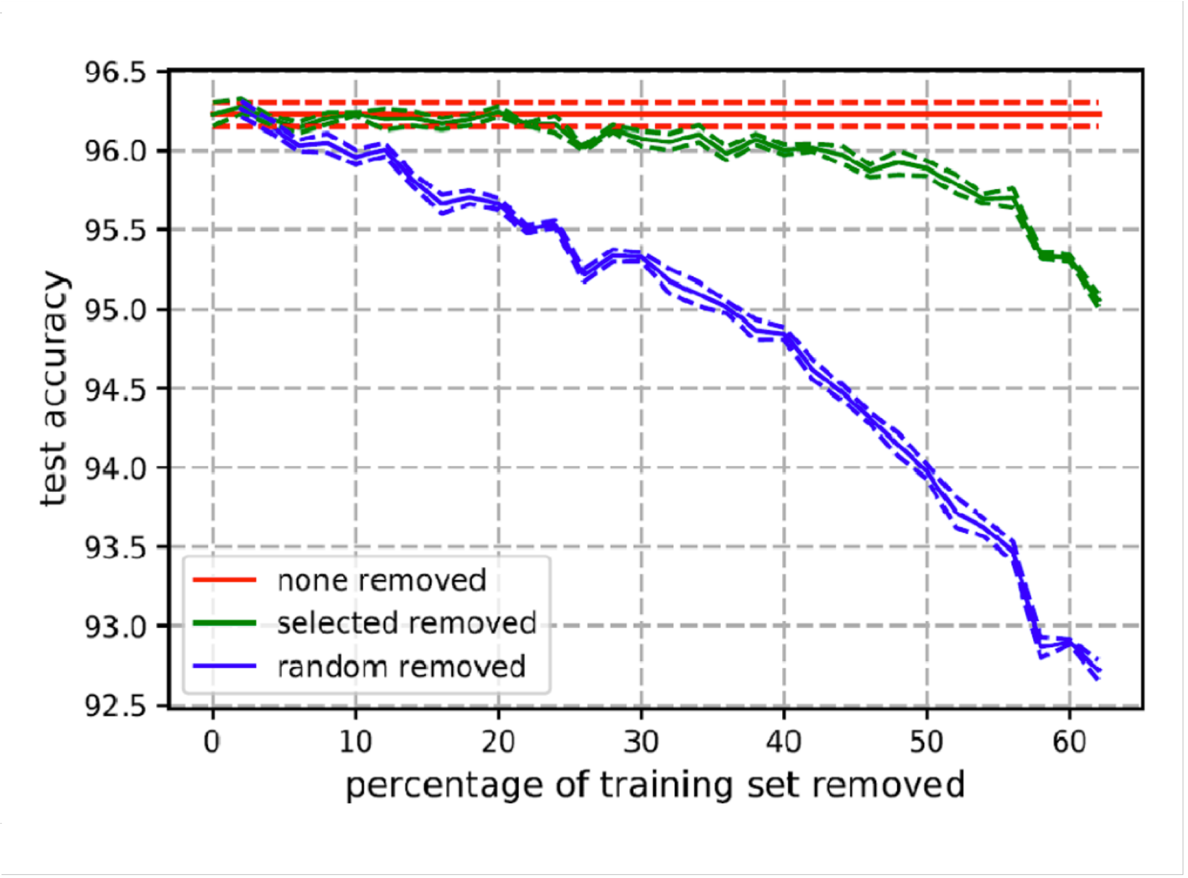




Transferable forgetting event



ResNet18



WideResNet

Conclusions

- Within a task, some examples are prone to be forgotten, while others are consistently unforgettable
- Forgeting statistics seem to be stable
- Unforgettable examples are not that important as they can be removed from training set without hurting generalization