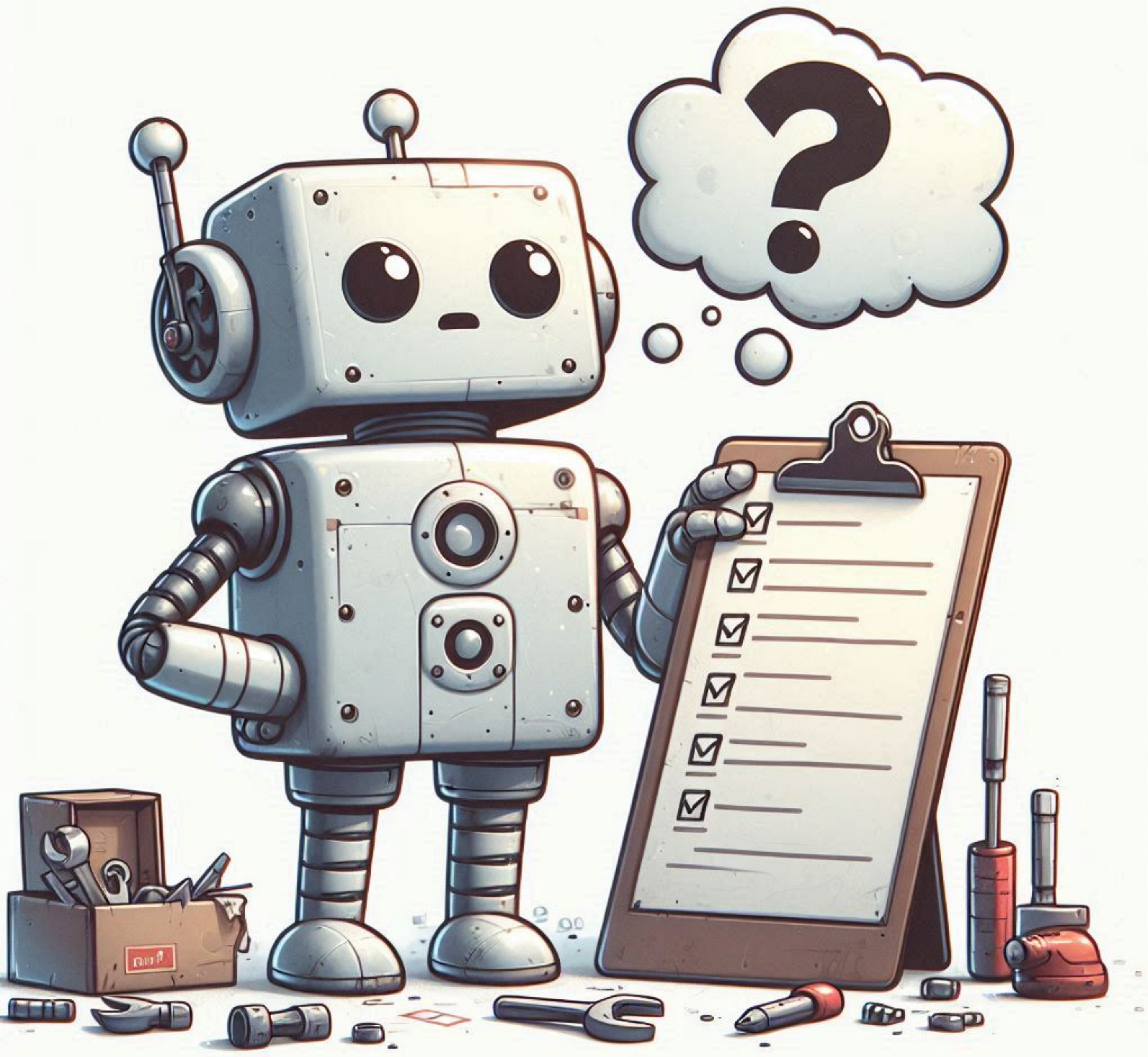


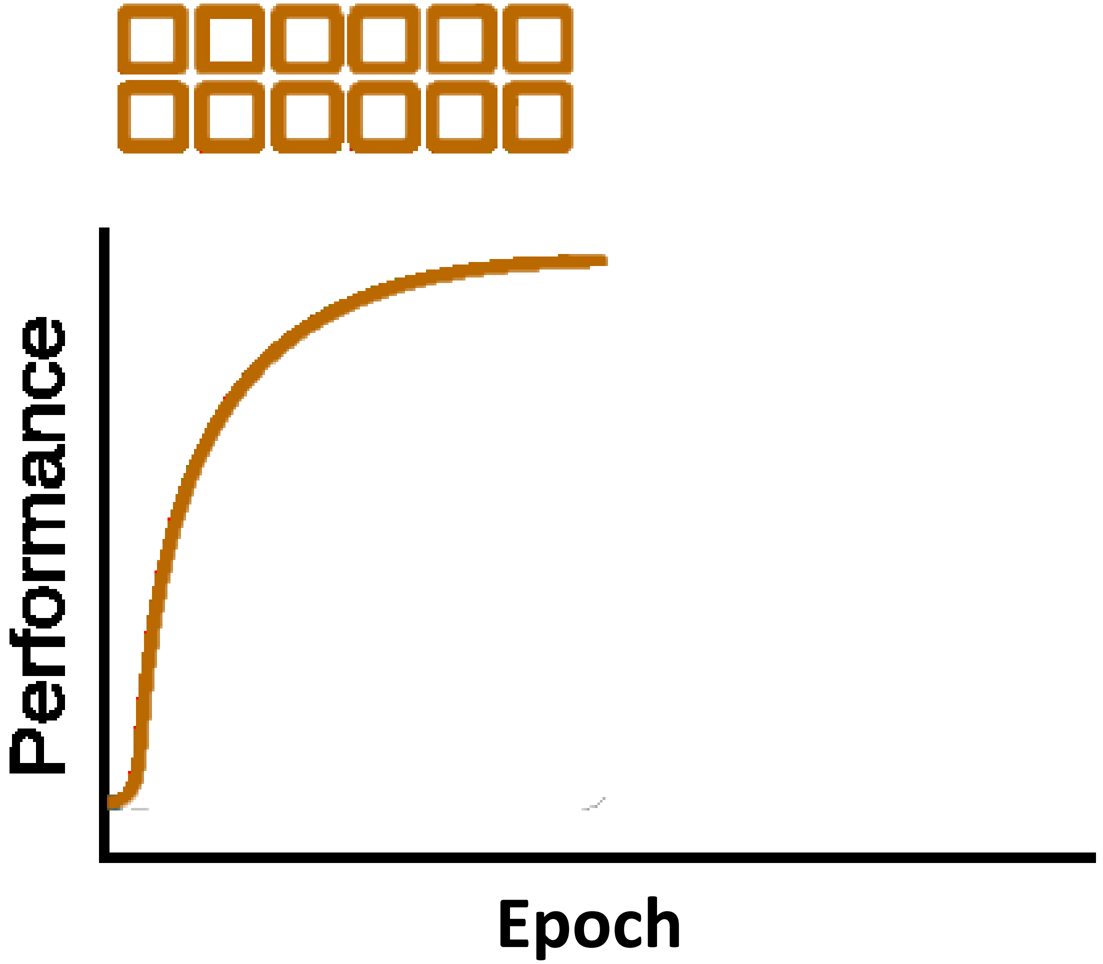
An empirical study of Example Forgetting during Deep Neural Network Learning

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, Geoffrey J. Gordon

Presented by: Itamar Salazar

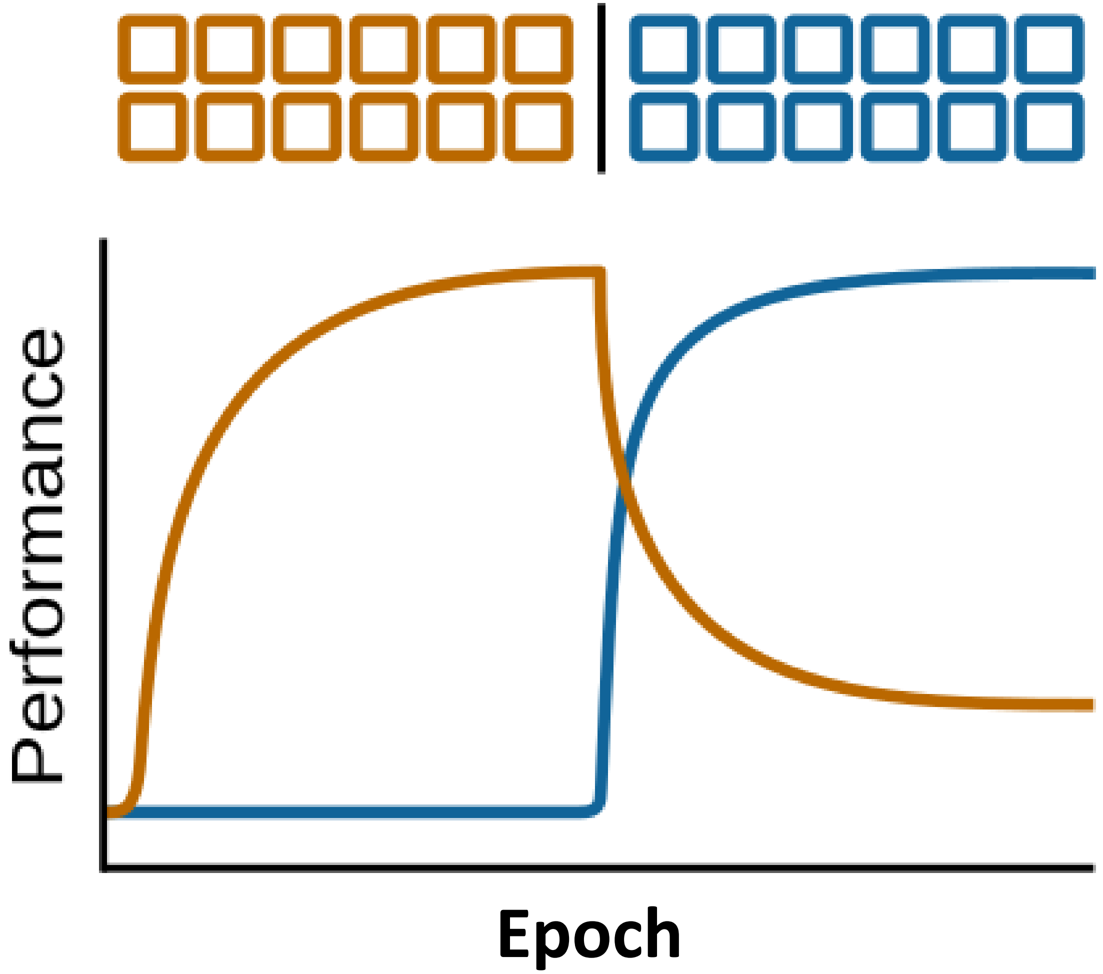
Deep Neural Network Training

- For training, the data is divided into smaller subsets called **batches**.
- An **epoch** is completed when the network has been trained with all batches of data.
- A DNN is trained for several epochs.



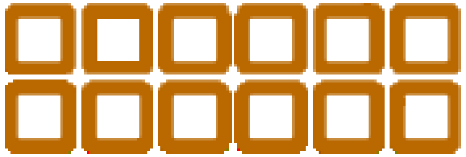
Catastrophic forgetting

- A model trained on a new task **forgets previous knowledge**, leading to significant drops in performance on earlier tasks.
- This is **attributed to out-of-distribution shifts** in the new task.



Example forgetting

- Each batch during training a neural network can be seen as a **“mini-task”**
- Is there a similar phenomenon when using “mini-task” instead of tasks?

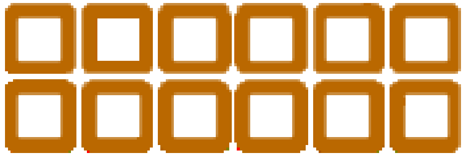


DEFINITIONS:

- **Forgetting event:** Example *i* has been correctly classified at step t but is misclassified at step t + 1 ($acc_i^t > acc_i^{t+1}$)
- **Learning event:** If $acc_i^t < acc_i^{t+1}$
- **Unforgettable examples:** If they are learnt at some point and experience no forgetting events during the whole course of training.
- **Forgettable examples:** Examples that have been forgotten at least once

Findings

1. Certain examples are forgotten with high frequency, and some not at all
2. A data set's (un)forgettable examples generalize across neural architectures
3. Based on forgetting dynamics: a significant fraction of examples can be omitted from the training data set while still maintaining state-of-the-art generalization performance.

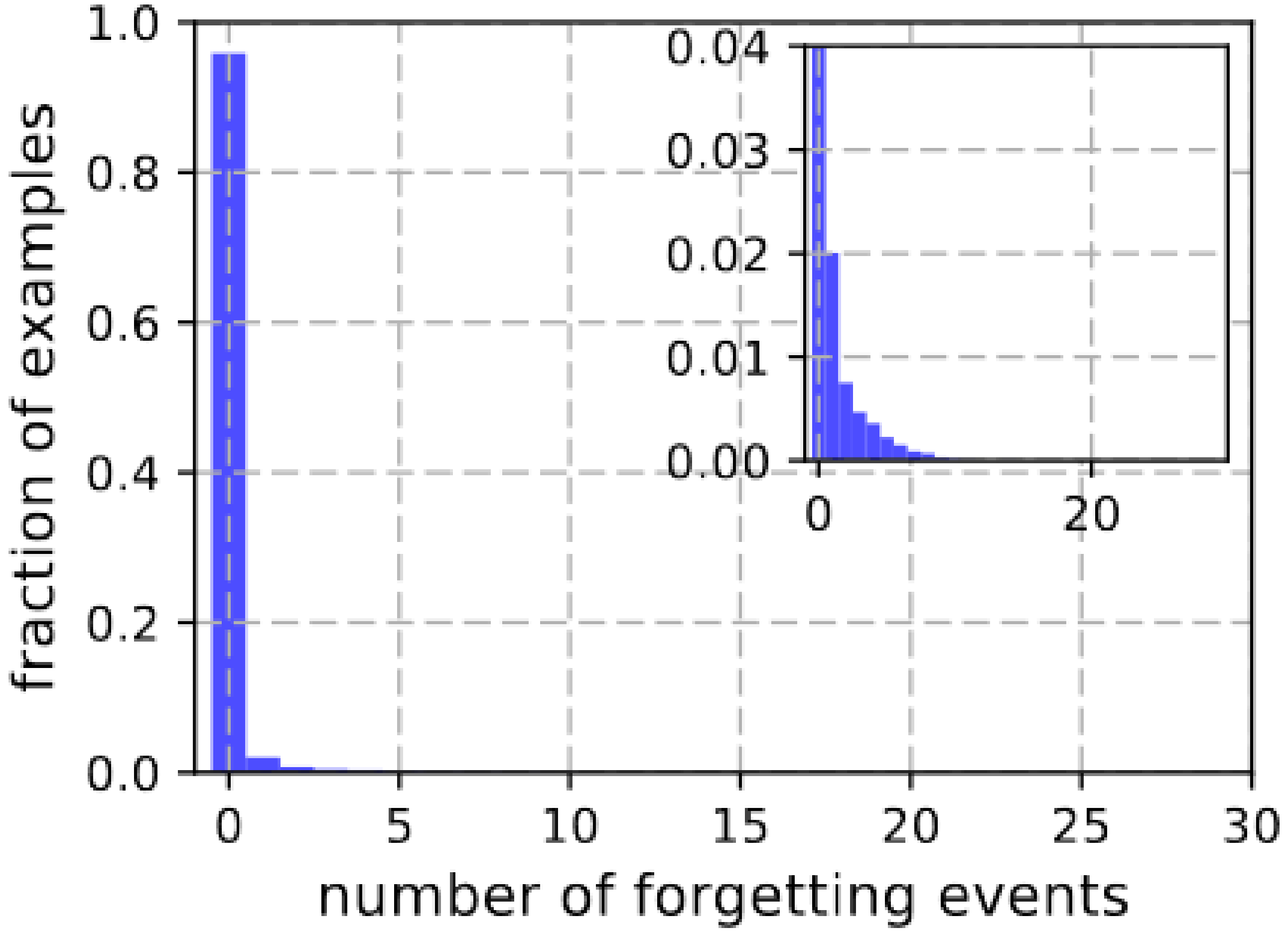


DEFINITIONS:

- **Forgetting event:** Example *i* has been correctly classified at step t but is misclassified at step t + 1 ($acc_i^t > acc_i^{t+1}$)
- **Learning event:** If $acc_i^t < acc_i^{t+1}$
- **Unforgettable examples:** If they are learnt at some point and experience no forgetting events during the whole course of training.
- **Forgettable examples:** Examples that have been forgotten at least once

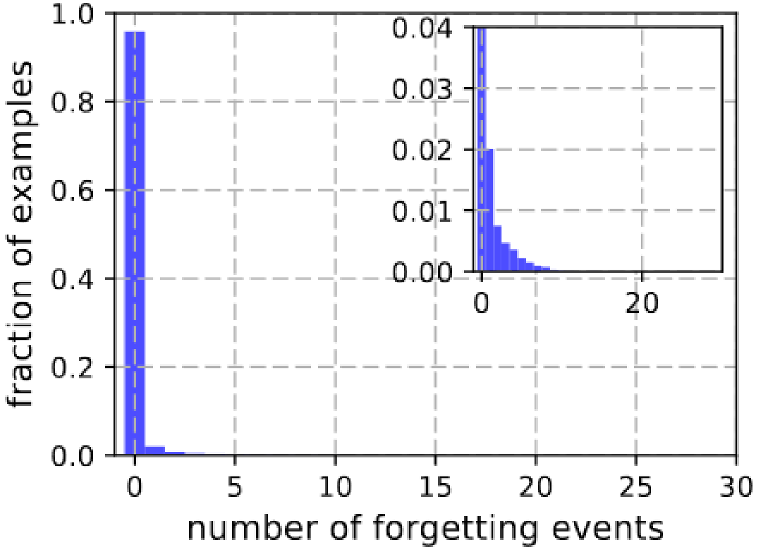
Results

- A big fraction of MNIST examples are unforgettable: once learned, they are not forgotten

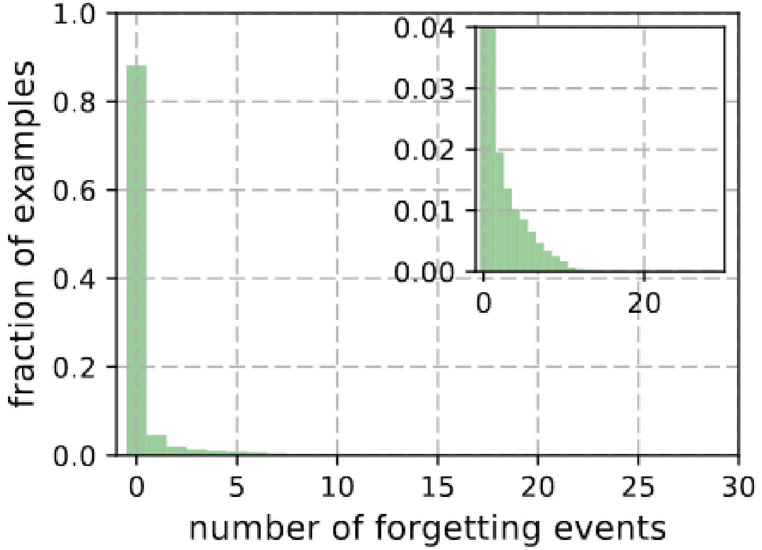


Results

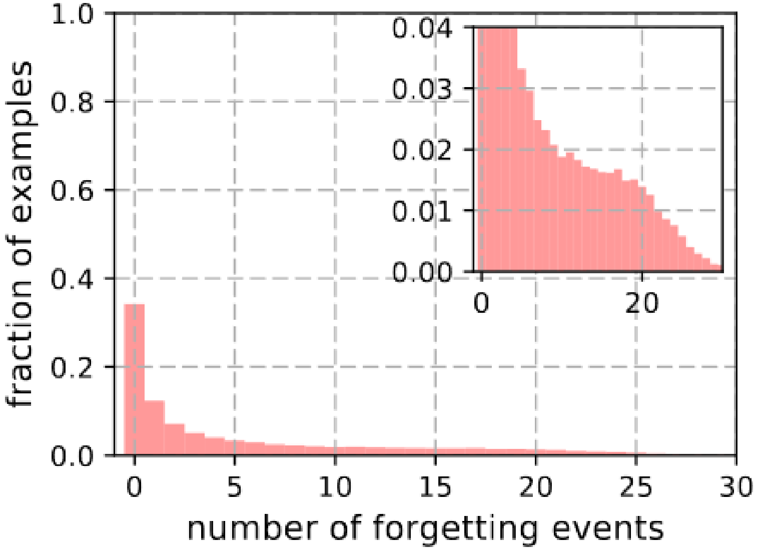
- Different datasets have different number of unforgettable examples.
- This finding seems to suggest a **correlation** between forgetting statistics and the intrinsic dimension of the learning problem.



Dataset 1
(MNIST)



Dataset 2
(Permuted-mnist)



Dataset 3
(CIFAR)

Results

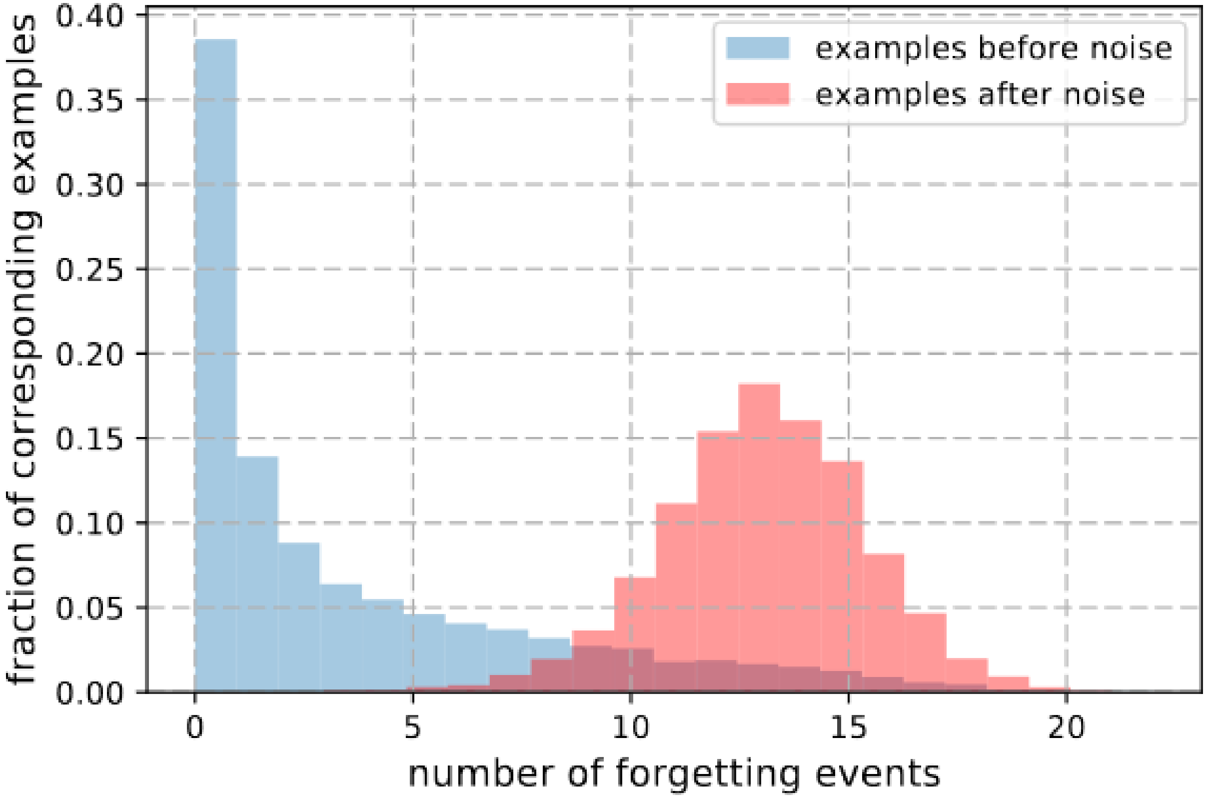
- Unforgettable samples: easily recognizable with the most **obvious class attributes** or centered objects
- Forgotten examples: exhibit more ambiguous characteristics that may not align with the learning signal **common to other examples** from the same class.

Results

- Experiment: Change the labels of 20% of CIFAR-10
- Examples with wrong labels (noisy): **Turn into forgotten examples**
- Examples with regular labels (green)

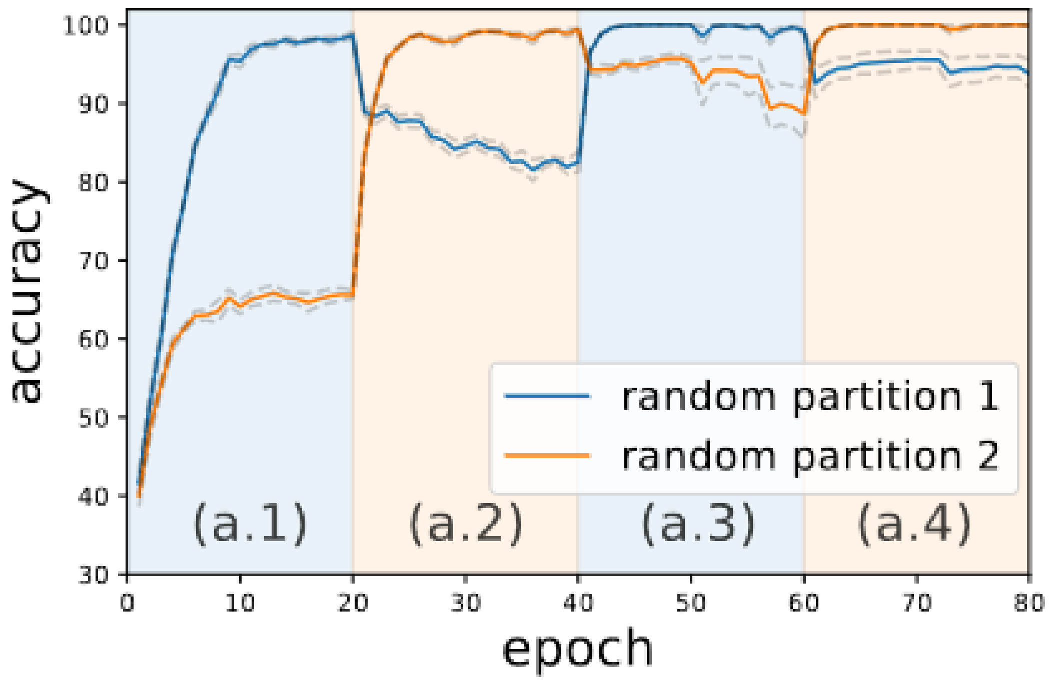
Results

- Experiment: Change the labels of 20% of CIFAR-10
- Examples with wrong labels (noisy): forgotten
- Same examples with regular labels (blue): **Recovers its unforgettable property**



Results

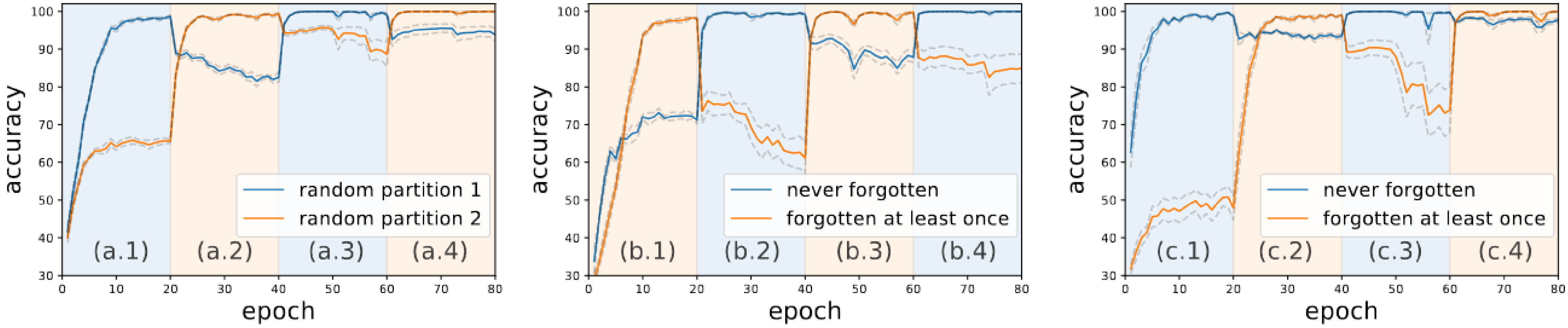
- Experiment: Split the dataset in two random partitions
- Background color: Training. Solid line: Test
- **Some forgetting of the second task when only train on the first task** (a.2)
- Surprising as the two tasks contain examples from the same underlying distribution.



(a) random partitions

Results

- Experiment: Partitioning the examples based on forgetting statistic
- Examples forgotten at least once suffer **more severe forgetting** than those in a random split (a.2 vs b.2)
- Examples from task (never forgotten) experience very **mild forgetting** when training on task (forgotten at least once) (b.3 and c.2).
- **Examples forgotten at least once may help "support" those never forgotten.**

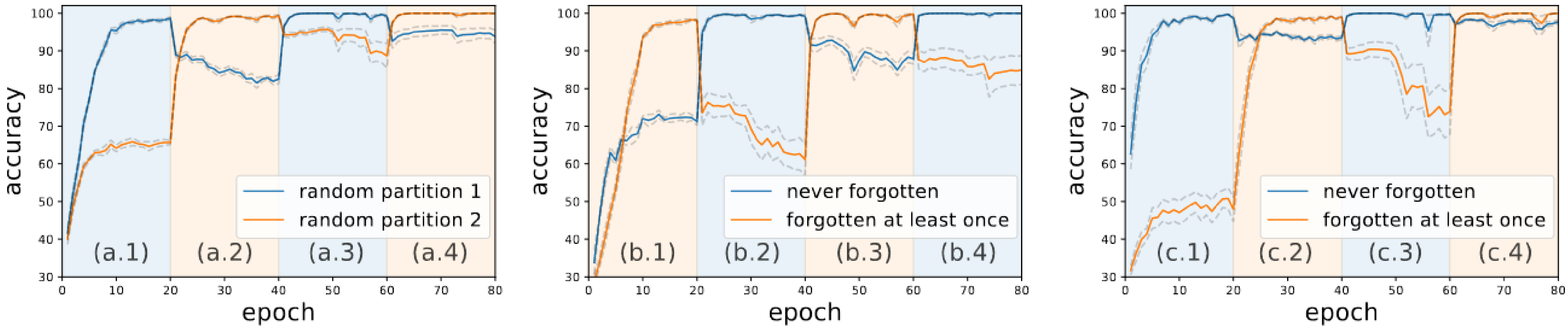


(a) random partitions

(b) partitioning by forgetting events

Results

- Experiment: Partitioning the examples based on forgetting statistic
- Examples forgotten at least once suffer **more severe forgetting** than those in a random split (a.2 vs b.2)
- Examples from task (never forgotten) experience very **mild forgetting** when training on task (forgotten at least once) (b.3 and c.2).
- **Examples forgotten at least once may help "support" those never forgotten.**

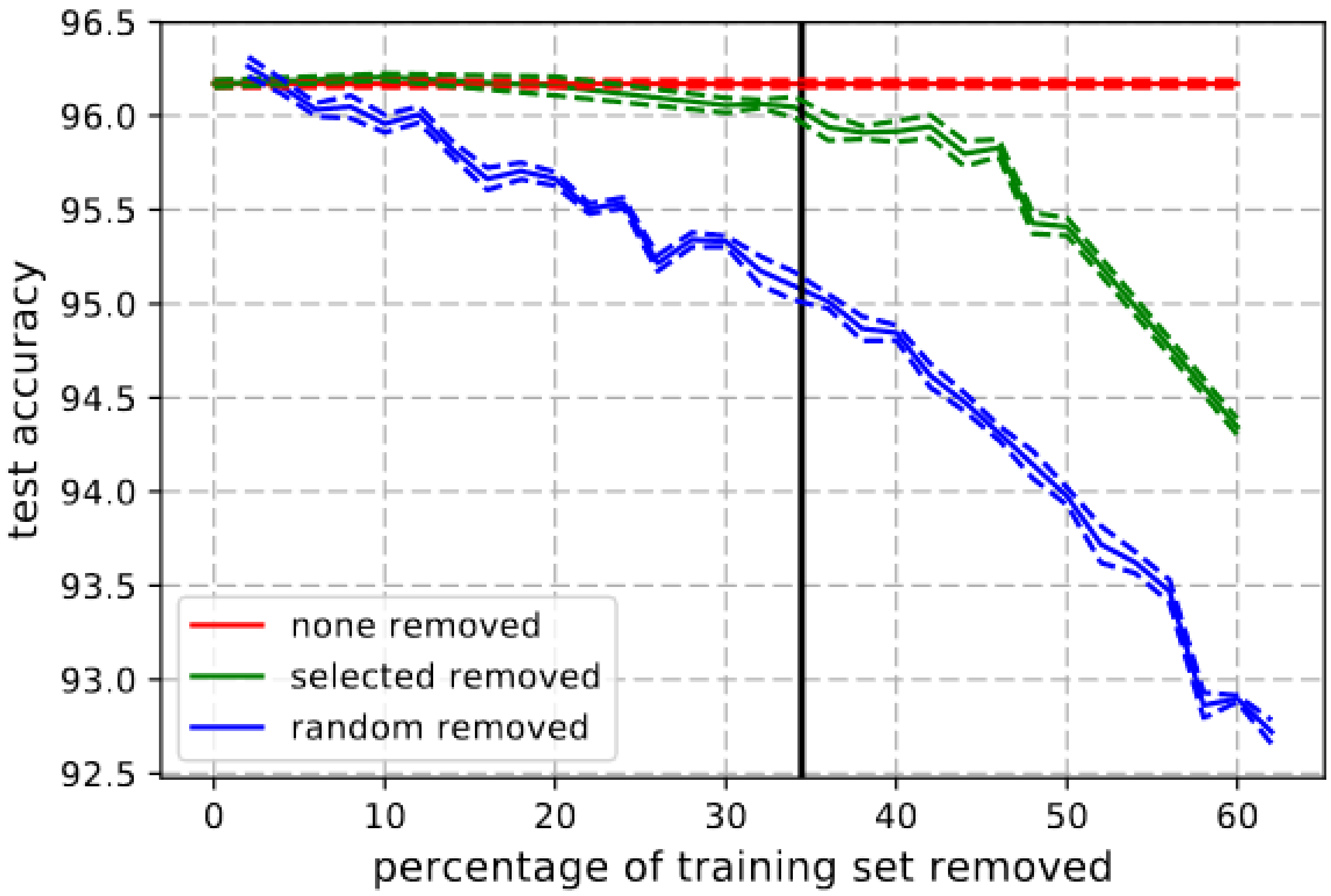


(a) random partitions

(b) partitioning by forgetting events

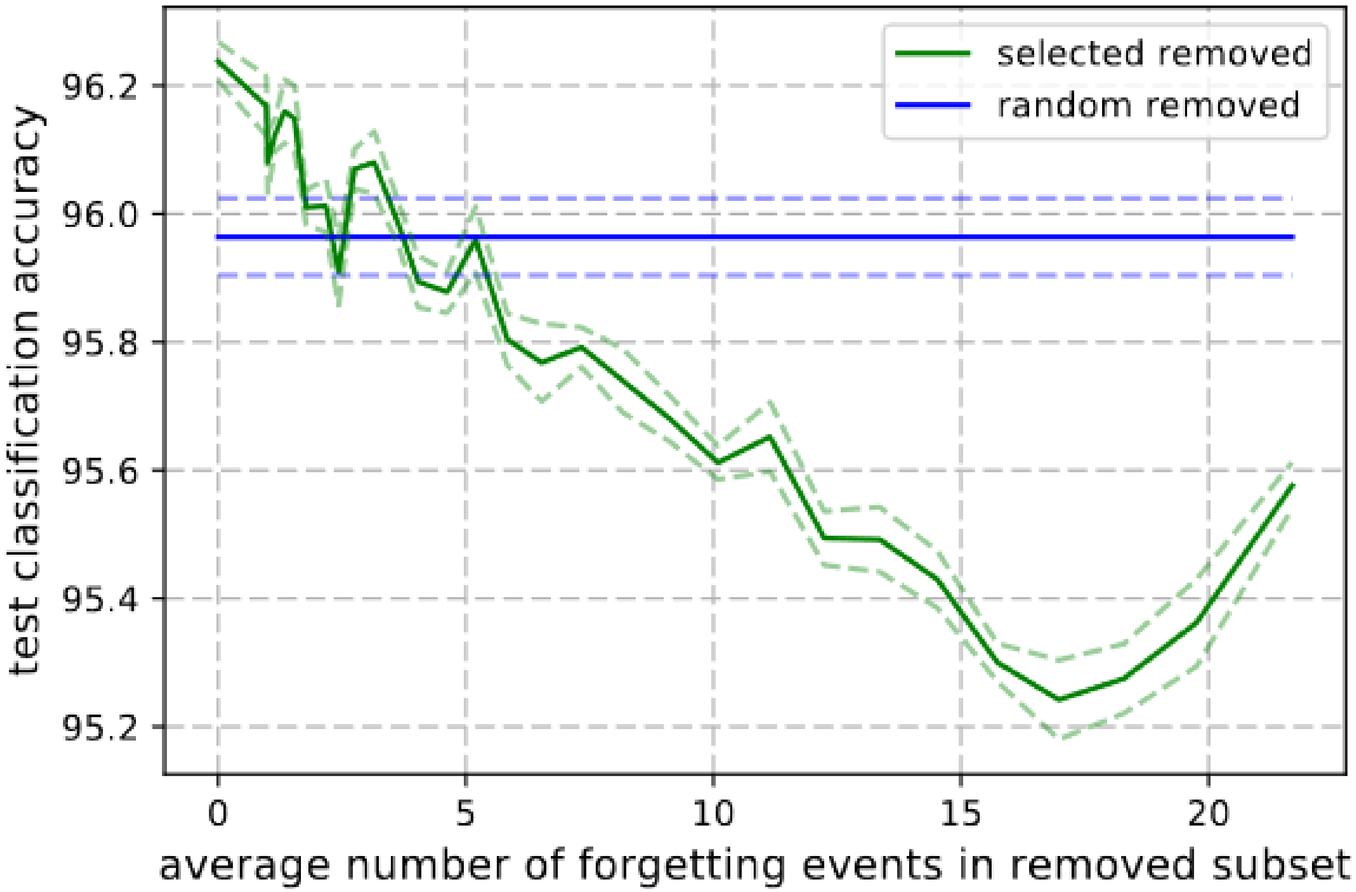
Results: Removing unforgettable examples

- Experiment: Remove examples from the training dataset by:
  - random
  - forgetting statistics.
- Removing a random subset of the dataset causes performance to rapidly decrease.
- Removing examples based on the number of forgetting events allows for 30% of the dataset to be removed while maintaining comparable generalization performance to the full dataset.
- Up to 35% of the dataset can be removed with only marginal degradation in performance (less than 0.2%)



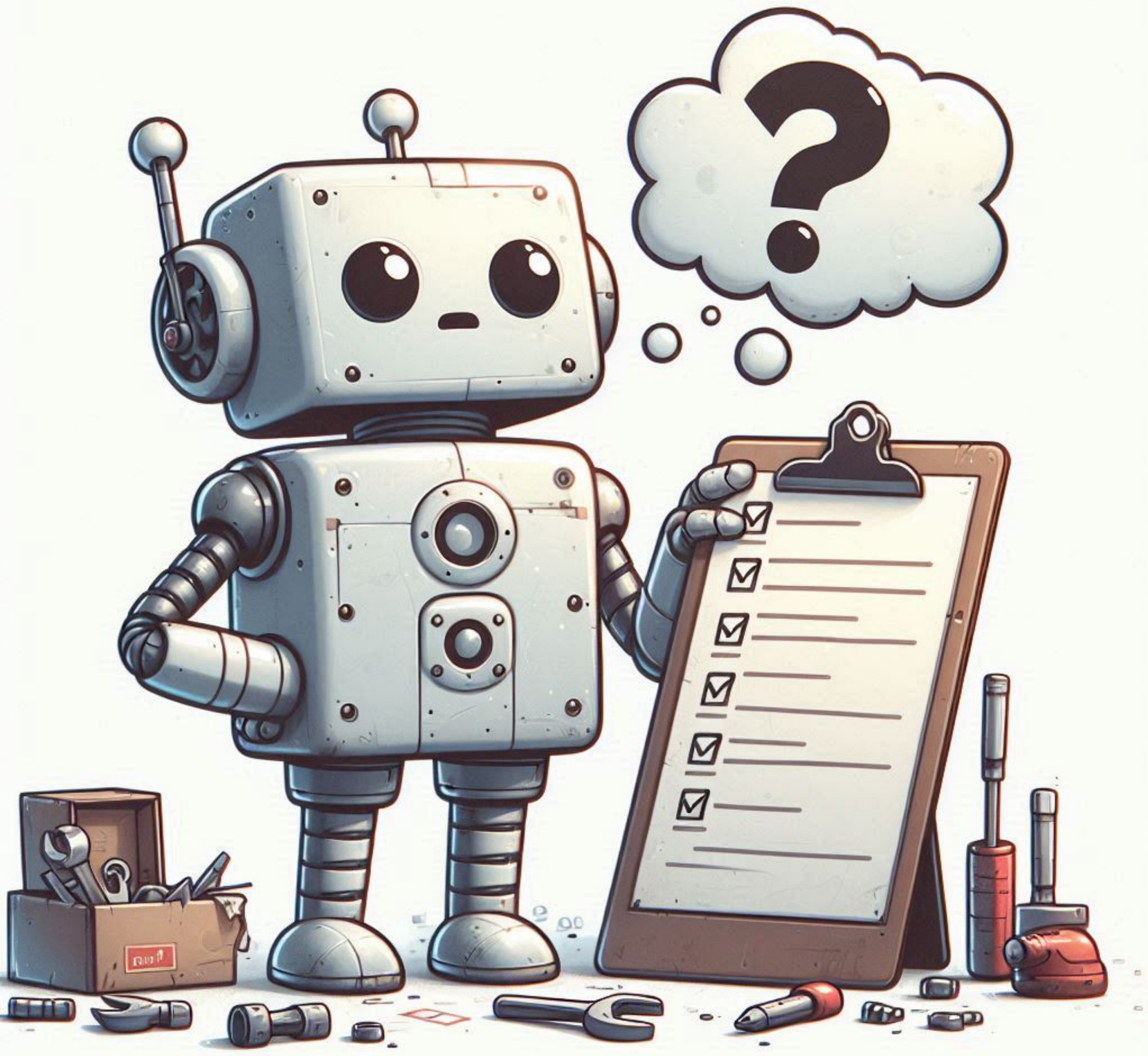
Results: Removing unforgettable examples (2)

- The generalization error is tracked when 5,000 The generalization error is tracked when 5,000 examples with increasing forgetting statistics are removed.
- Each point shows the error of a model trained on the full dataset minus 5,000 examples, based on the average number of forgetting events.
- **Worse generalization is seen when examples with more forgetting events are removed.**
- The rightmost part of the curve rises, indicating that **some of the most forgotten examples may actually hurt performance.**



Summary and Conclusions

1. The learning dynamics of neural networks in single classification tasks are investigated.
2. Catastrophic forgetting can be observed within what is conventionally considered a single task.
3. Some examples within a task are more susceptible to forgetting, while others remain consistently remembered.
4. The final performance of the classifier seems unaffected by the removal of these unforgettable examples from the training set, indicating their minimal impact on generalization.



Thanks

Presented by: Itamar Salazar