AN EMPIRICAL STUDY OF EXAMPLE FORGETTING DURING DEEP NEURAL NETWORK LEARNING

Mariya Toneva[*†]
Carnegie Mellon University

Alessandro Sordoni[*]
Microsoft Research Montreal

Remi Tachet des Combes[*]
Microsoft Research Montreal

Adam Trischler
Microsoft Research Montreal

Yoshua Bengio
MILA, Université de Montréal
CIFAR Senior Fellow

Geoffrey J. Gordon
Microsoft Research Montreal
Carnegie Mellon University

ICLR 2019

Contents

- Motivation & Definition
- Experiments & Conclusions & Analysis

https://zhuanlan.zhihu.com/p/215734661

Motivation & Definition

Motivation

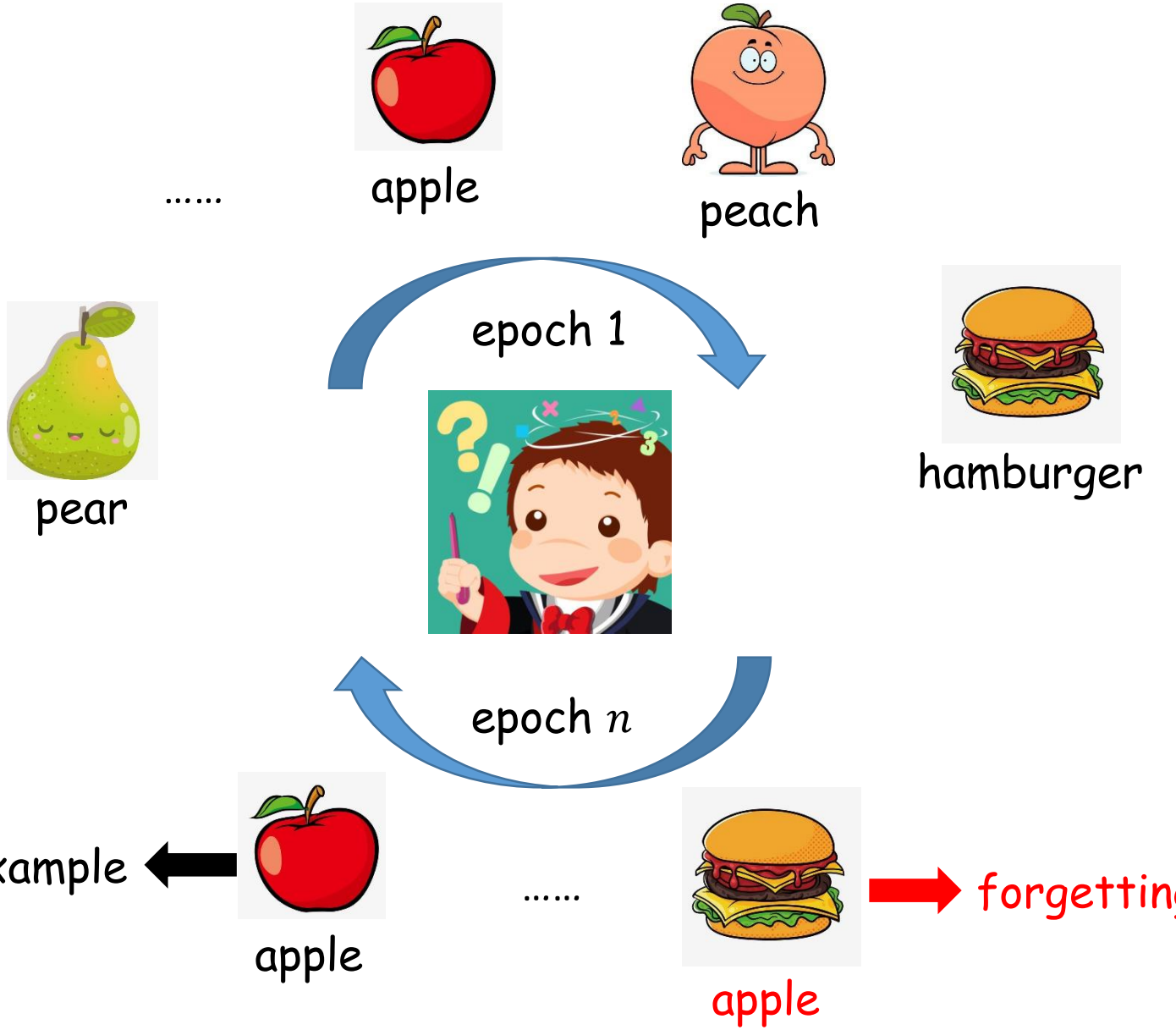
- Study of example forgetting during the learning process

Definition

- **"learnt"**: examples are correctly classified at some time t.
- **"first learnt"**: the epoch in which samples are "learnt" for the first time.
- **"last learnt"**: the epoch in which samples are "learnt" for the last time.

- **"forgetting event"**: examples that have been "learnt" at some time t in the optimization process are subsequently misclassified at a time t'. $(t' > t)$
- **"forgetting number"**: the number of "forgetting event" occurs in the learning process.
- **"unforgettable examples"**: examples that are never forgotten once "learnt".
- **"most forgettable examples"**: the example with the largest "forgetting number".

Intuitively



Experiments & Conclusions & Analysis

Conclusions

- There exist a large number of unforgettable examples in benchmark datasets, i.e., examples that are never forgotten once learnt, those examples are stable with different seeds and neural architectures.

- Examples with noisy labels are among the most forgotten examples, along with images with "uncommon" features, visually complicated to classify. (Noisy detection)

✓ Training a neural network on a dataset where a very large fraction of the least forgotten examples have been removed still results in extremely competitive performance on the test set.

Experiments & Conclusions & Analysis

Conclusion 1

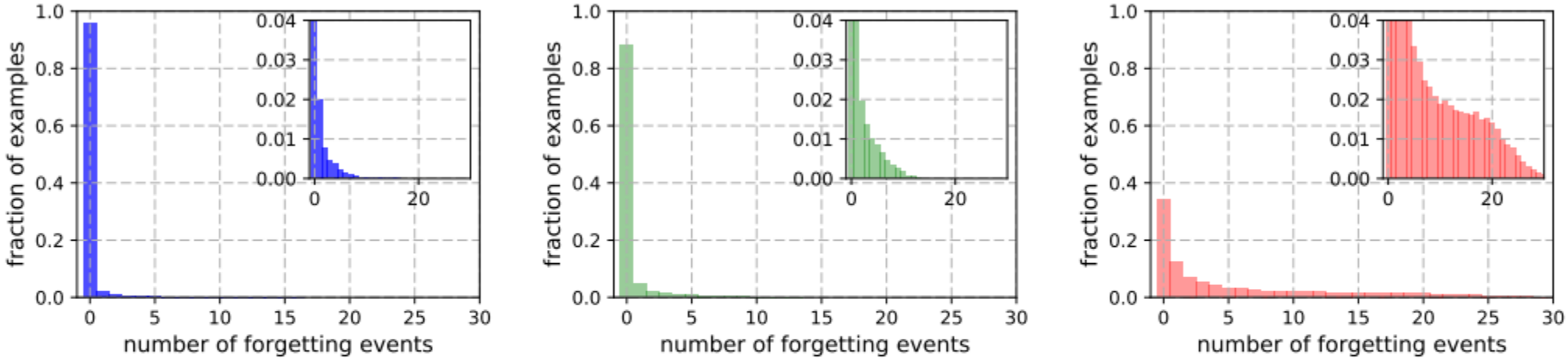


Figure 1: Histograms of forgetting events on (from left to right) *MNIST*, *permutedMNIST* and *CIFAR-10*. Insets show the zoomed-in y-axis.

Experiments & Conclusions & Analysis

Conclusion 2

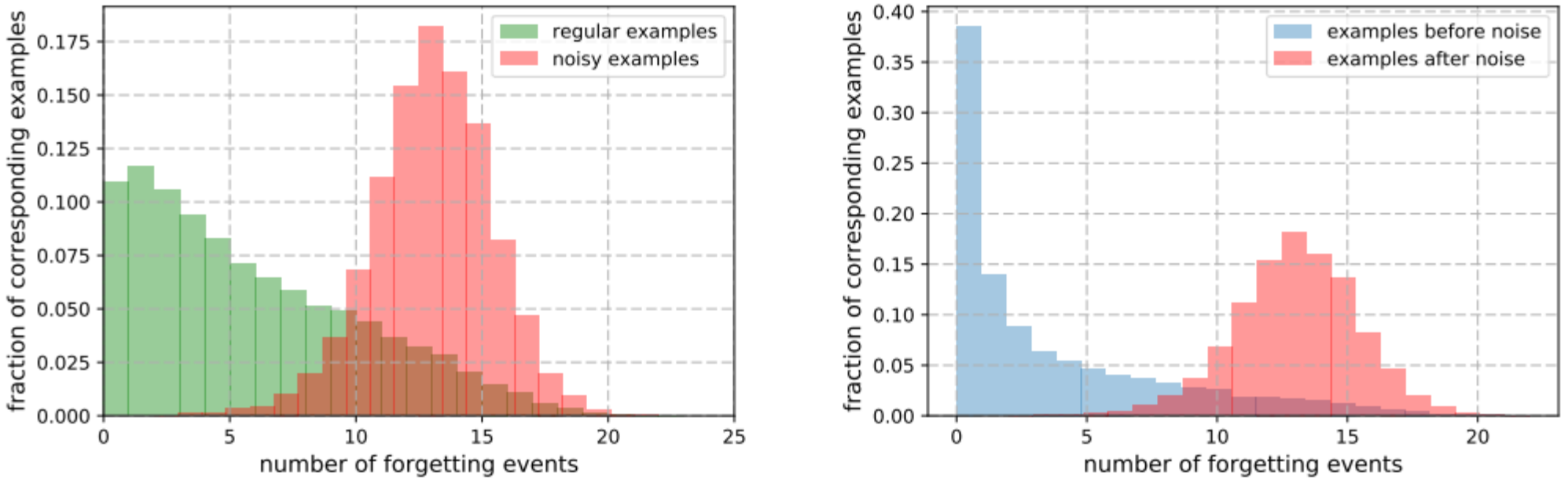
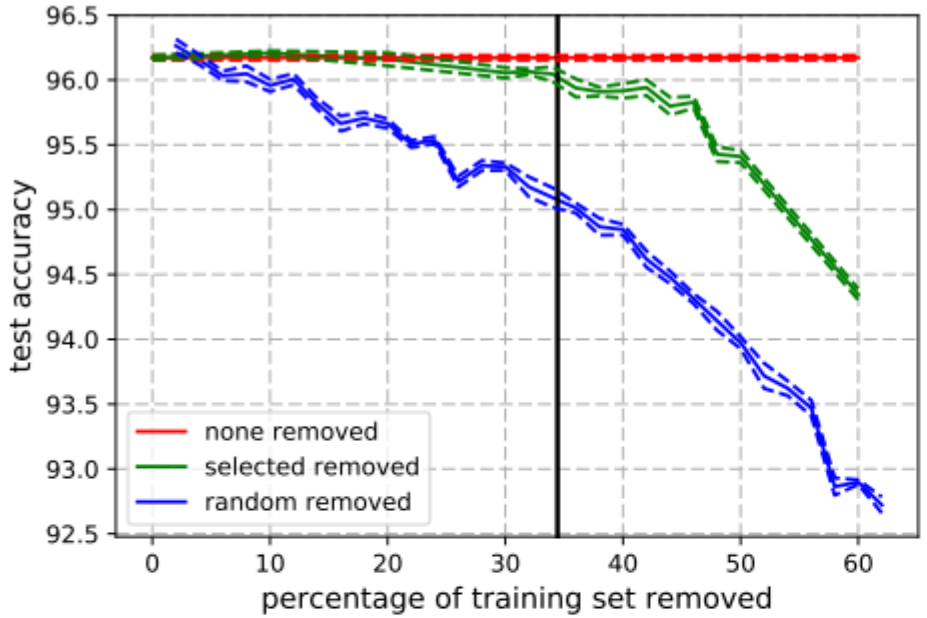


Figure 3: Distributions of forgetting events across training examples in *CIFAR-10* when 20% of labels are randomly changed. *Left.* Comparison of forgetting events between examples with noisy and original labels. The most forgotten examples are those with noisy labels. No noisy examples are unforgettable. *Right.* Comparison of forgetting events between examples with noisy labels and the same examples with original labels. Examples exhibit more forgetting when their labels are changed.

Experiments & Conclusions & Analysis

Conclusion 3



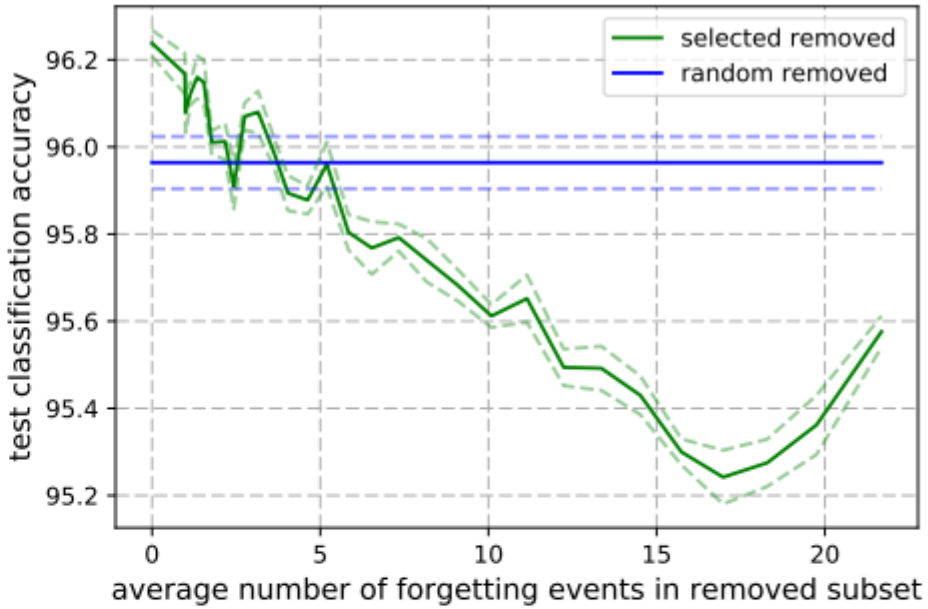


Figure 5: *Left* Generalization performance on *CIFAR-10* of ResNet18 where increasingly larger subsets of the training set are removed (mean +/- std error of 5 seeds). When the removed examples are selected at random, performance drops very fast. Selecting the examples according to our ordering can reduce the training set significantly without affecting generalization. The vertical line indicates the point at which all unforgettable examples are removed from the training set. *Right* Difference in generalization performance when contiguous chunks of 5000 increasingly forgotten examples are removed from the training set. Most important examples tend to be those that are forgotten the most.

Visualization

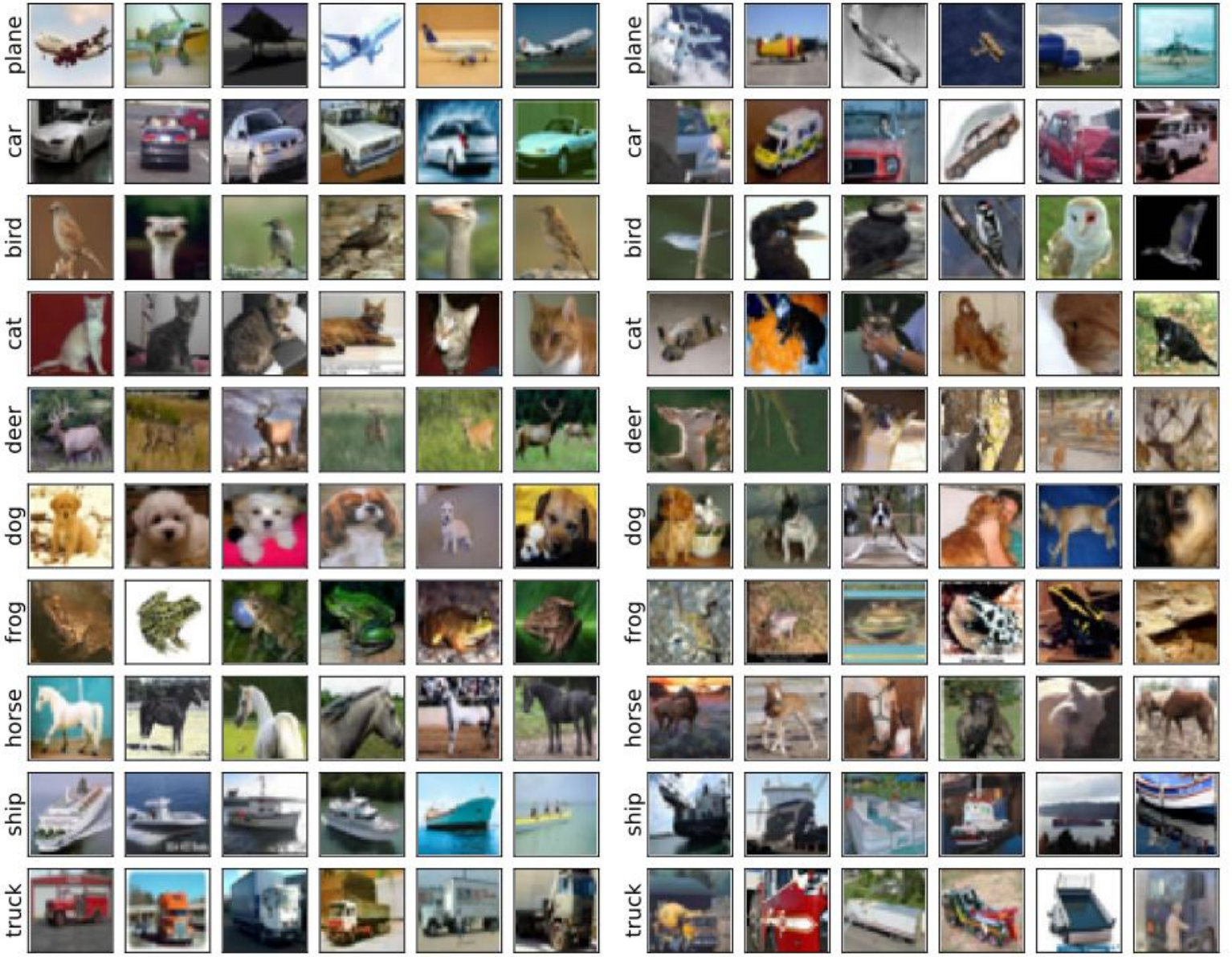


Figure 15: Additional pictures of the most unforgettable (*Left*) and forgettable examples (*Right*) of every *CIFAR-10* class, when examples are sorted by number of forgetting events (ties are broken randomly). Forgettable examples seem to exhibit peculiar or uncommon features.

Discussion

Noisy examples & "Hard" examples

- Noisy examples and "hard" examples are easily forgotten by models.
- From conclusion 2, no noisy examples are unforgettable, which means all unforgettable examples are clean examples.
- However, from conclusion 3, the performance of the model is basically unchanged with the most of "unforgettable" examples being removed. It shows that the "unforgettable" examples are not helpful to the model training. (有就行了，多无用)
- In other words, we only need "hard" examples. However, the noisy examples are also often "most forgettable", and we don't want these noisy examples.

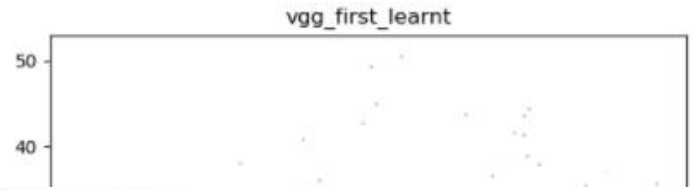
Discussion & Idea

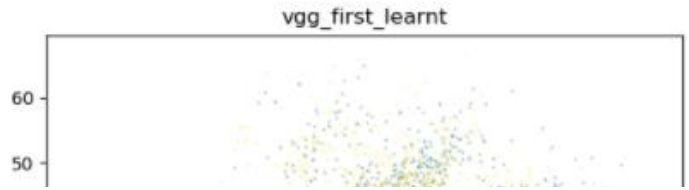
- Clean examples
  - Clean "easy" examples (Far away from the classification plane, strong correlation)
  - Clean "hard" examples (Close to the classification plane, weak correlation) (more important)

- Noisy examples
  - Noisy "easy" examples: clean "easy" examples + noisy labels
  - Noisy "hard" examples: clean "hard" examples + noisy labels

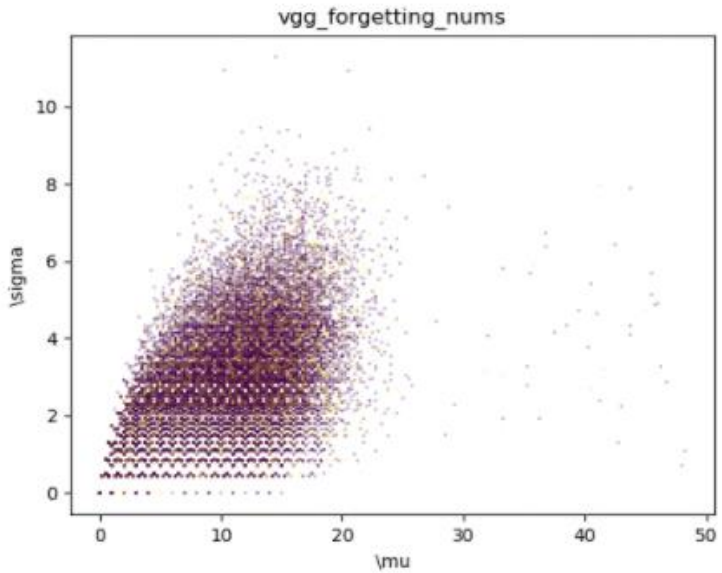
- Idea: when noisy "easy" examples are remembered by model, the clean "easy" examples may be forgotten more seriously. Can these phenomena be used to identify noisy samples?

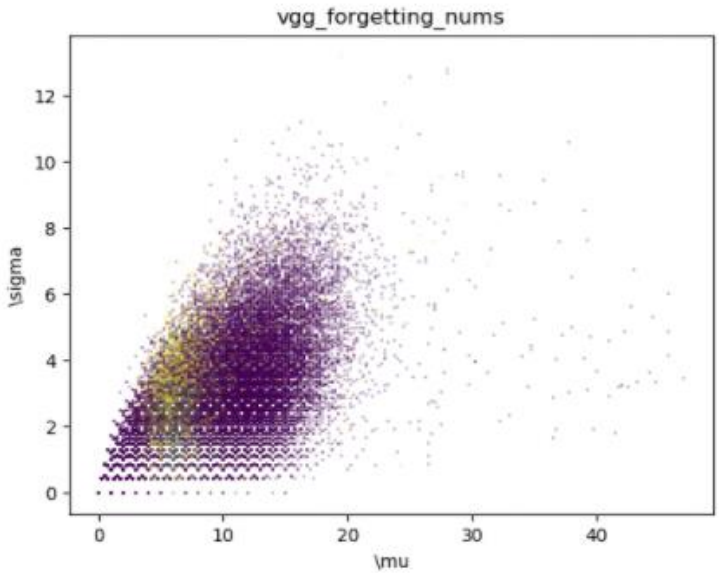
- **first-learnt**

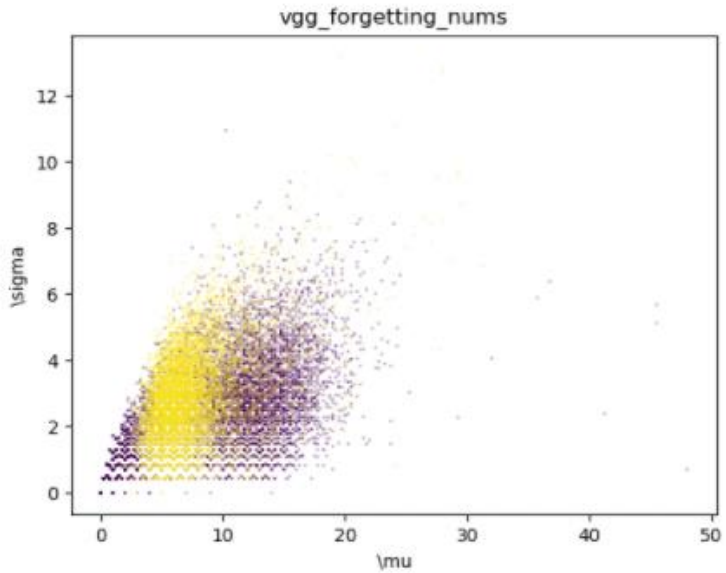


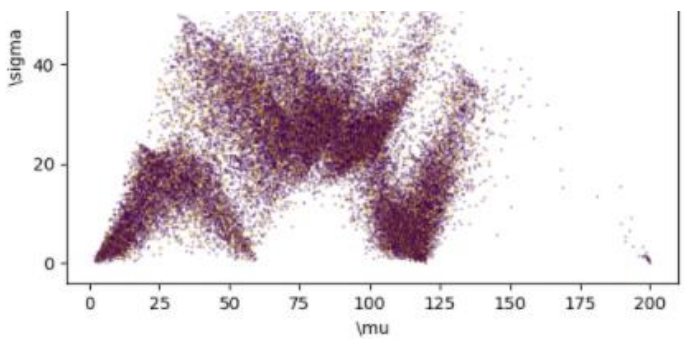


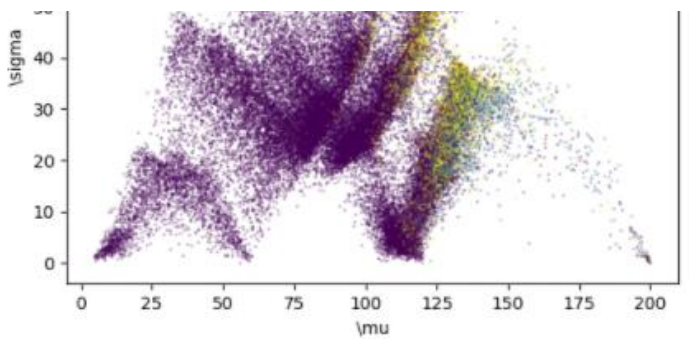
- **forgetting-numbers**

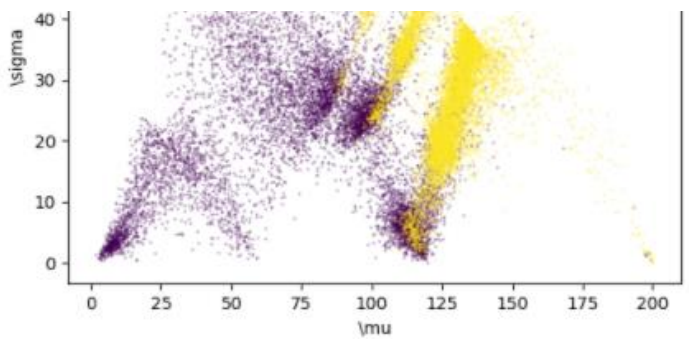






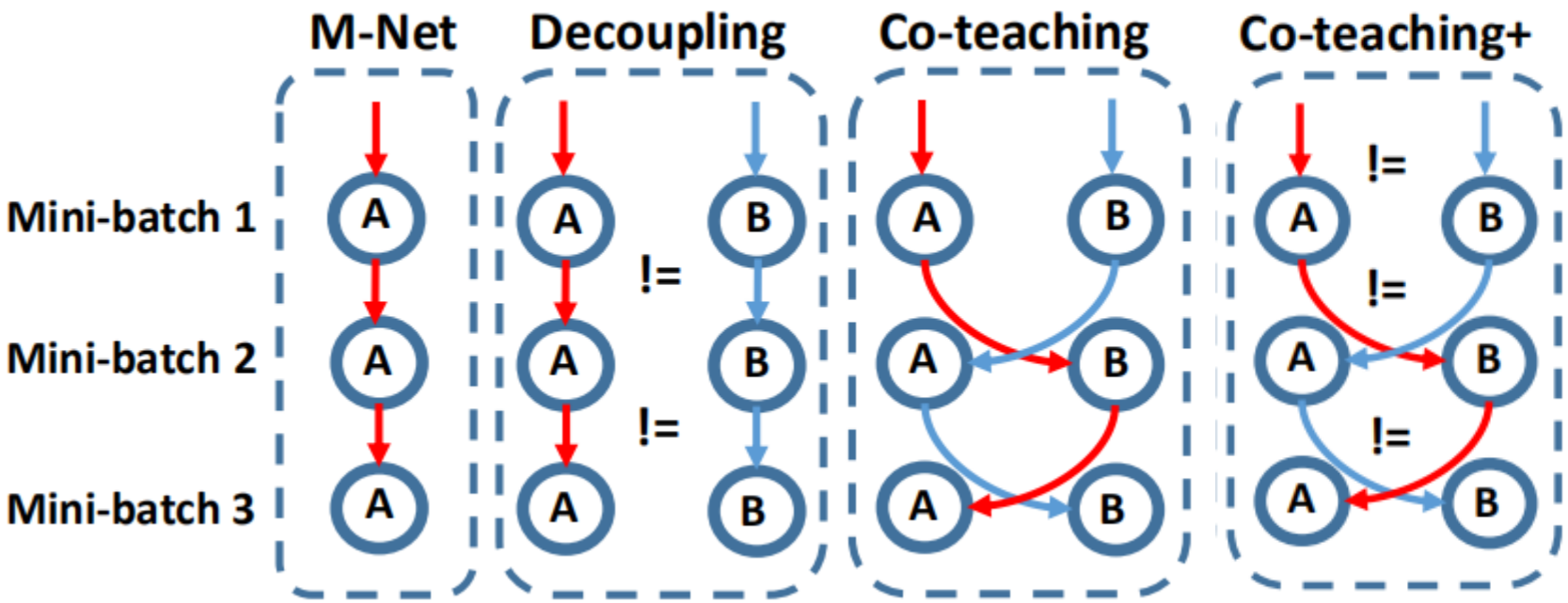






Co-Imitating

Co-Teaching



O2U

Co-teaching (NIPS, 2018) -> "peer review"

Co-teaching+ (ICML, 2019) -> "complementary peer learning"

Self Imitation Learning (ICML, 2018)

$$\mathcal{L}^{sil} = \mathbb{E}_{s,a,R}$$

$$\mathcal{L}^{sil}_{policy} = -\log$$

$$\mathcal{L}^{sil}_{value} = \frac{1}{2} \|(R$$

$$\max(\cdot, 0)$$

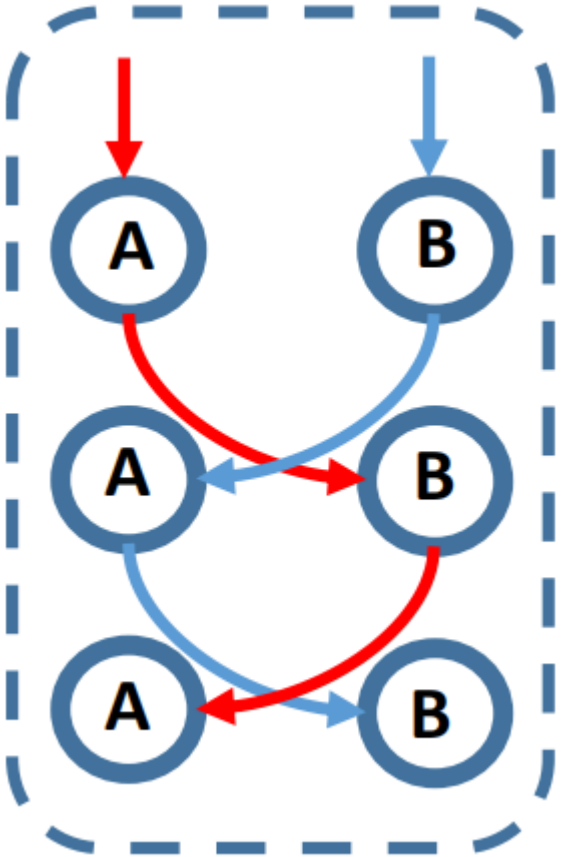
Algorithm 1 Actor-Critic with Self-Imitation Learning

Initialize parameter θ
Initialize replay buffer $\mathcal{D} \leftarrow \emptyset$
Initialize episode buffer $\mathcal{E} \leftarrow \emptyset$
for each iteration **do**
  # Collect on-policy samples
  for each step **do**
    Execute an action $s_t, a_t, r_t, s_{t+1} \sim \pi_\theta(a_t|s_t)$
    Store transition $\mathcal{E} \leftarrow \mathcal{E} \cup \{(s_t, a_t, r_t)\}$
  end for
  if s_{t+1} is terminal **then**
    # Update replay buffer
    Compute returns $R_t = \sum_k^\infty \gamma^{k-t} r_k$ for all t in $\mathcal{E}$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, R_t)\}$ for all t in $\mathcal{E}$
    Clear episode buffer $\mathcal{E} \leftarrow \emptyset$
  end if
  # Perform actor-critic using on-policy samples
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}^{a2c}$ (Eq. 4)
  # Perform self-imitation learning
  for $m = 1$ to M **do**
    Sample a mini-batch $\{(s, a, R)\}$ from $\mathcal{D}$
    $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}^{sil}$ (Eq. 1)
  end for
end for

Co-Imitating

- Two agents learn the demonstration from each other.

- Study:
  - How to select demonstrations
  - Diversity of agents



Thanks