Lifelong Learning Meeting 11/22

Sang Keun Choe

Topics

- Mariya Toneva et al., An Empirical Study of Example Forgetting during Deep Neural Network Learning, In ICLR, 2019

- Gradient updates by backpropagation and catastrophic interference.

An Empirical Study of Example Forgetting

- **Objective**: Study forgetting phenomenon when data *doesn't* undergo a significant distributional shift. (i.e. single task)
- **Tasks**: Image classification (MNIST/Permuted MNIST/CIFAR-10)
- **Findings**:
  - Certain examples are forgotten with high frequency while some not at all
  - (Un)forgettable examples generalize across various architectures
  - Omitting unforgettable examples still enables training without degrading performance

Definitions and Experimental Setting

1) $acc_i^t = 1$ if $data_i$ is classified correctly else 0
2) Learning event: $acc_i^t < acc_i^{t+1}$
3) Forgetting event: $acc_i^t > acc_i^{t+1}$
4) Unforgettable examples: Examples never forgotten during learning
5) Classification Margin:

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

   k : correct index

Algorithm 1 Computing forgetting statistics.

initialize 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 prev_acc$_i >$ acc$_i$ **then**
      $T[i] = T[i] + 1$
    prev_acc$_i$ = acc$_i$
  gradient update classifier on B
return T

Number of Unforgettable Events in each Dataset

MNIST: 91.7% / Permuted MNIST: 75.3% / CIFAR-10: 31.3%

Unforgettable events generalize across different random seeds

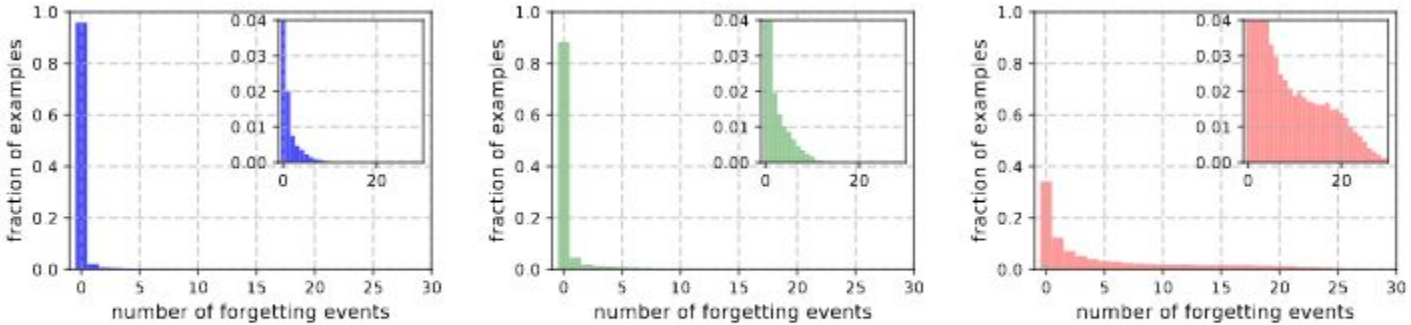
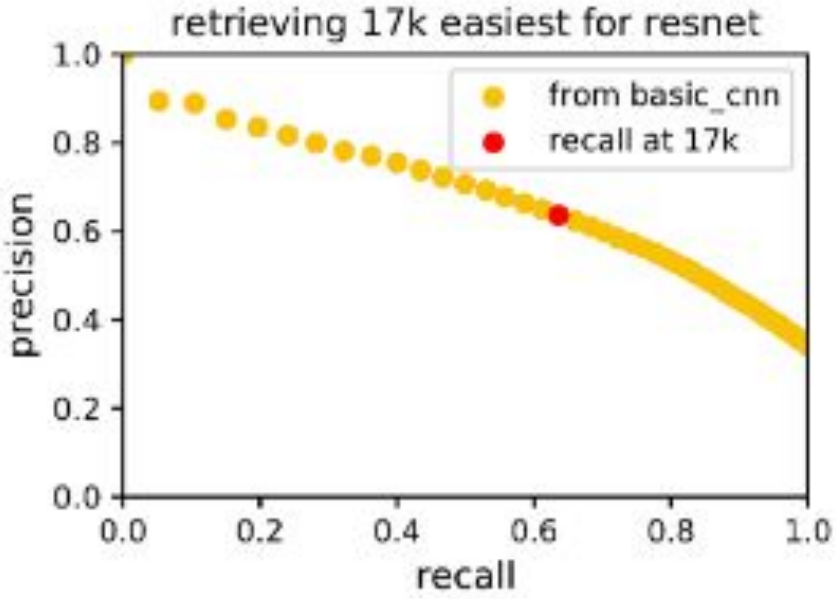
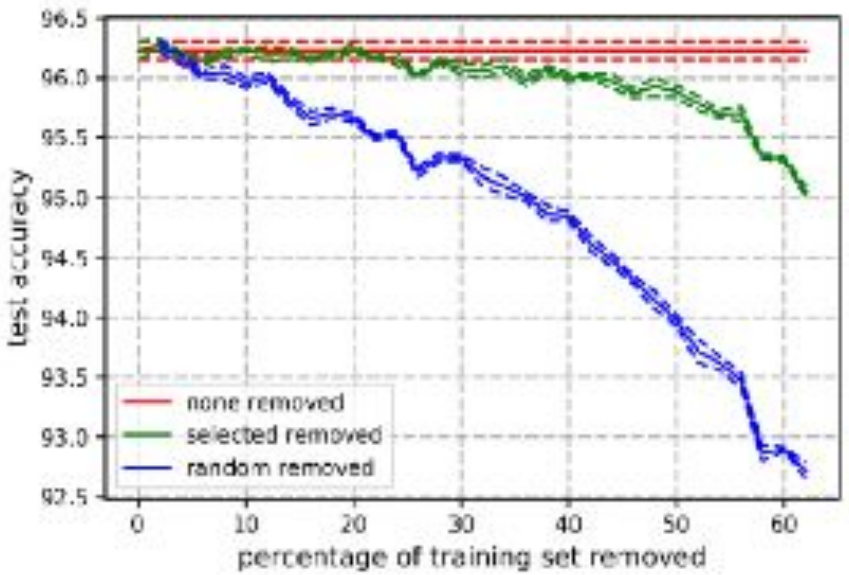


Figure 1: Histograms of forgetting events on (from left to right) *MNIST*, *permutedMNIST* and *CIFAR-10*. Insets show the zoomed-in y-axis.

Transferable Forgetting Events

Forgetting examples are transferable between different neural architectures.





Characteristics of Forgettable Examples

Hypothesis: Forgettable examples more likely to stay around decision boundaries.

Characteristics of Forgettable Examples

First learning event: In which epoch x_i is first correctly classified to y_i

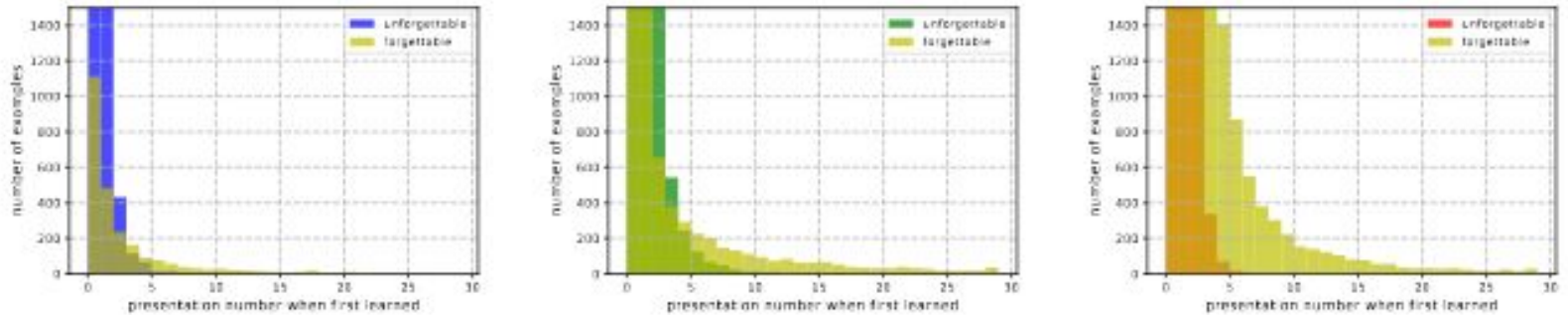
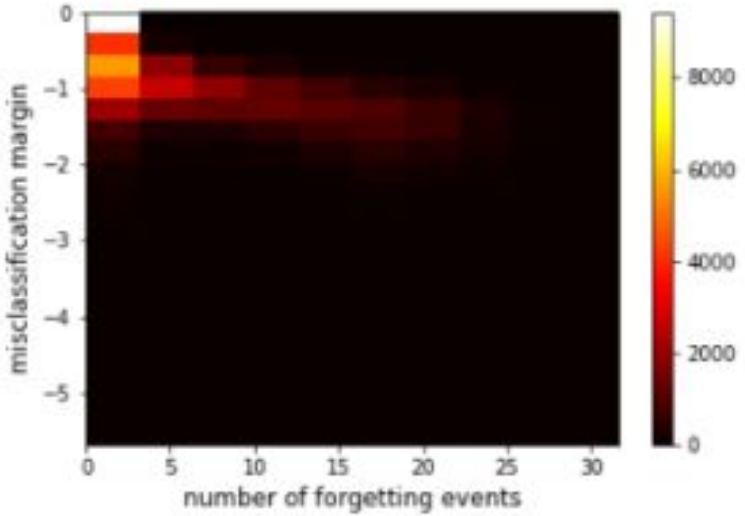


Figure 8: From left to right, distributions of the first presentation at which each unforgettable and forgettable example was learned in *MNIST*, *permutedMNIST* and *CIFAR-10* respectively. Rescaled view where the number of examples have been capped between 0 and 1500 for visualization purposes. Unforgettable examples are generally learnt early during training, thus may be considered as "easy" in the sense of Kumar et al. (2010), i.e. may have a low loss during most of the training.

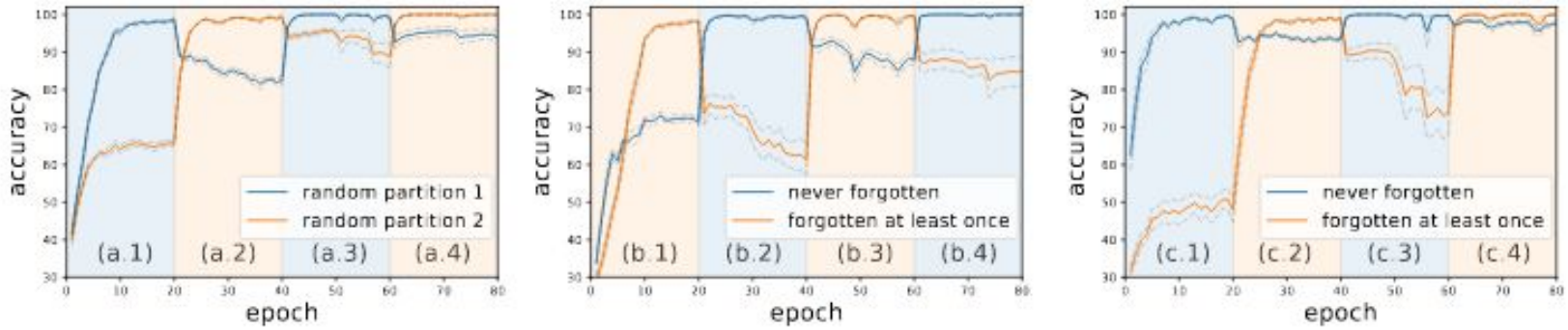
Characteristics of Forgettable Examples

Misclassification margin



Characteristics of Forgettable Examples

Iterative learning



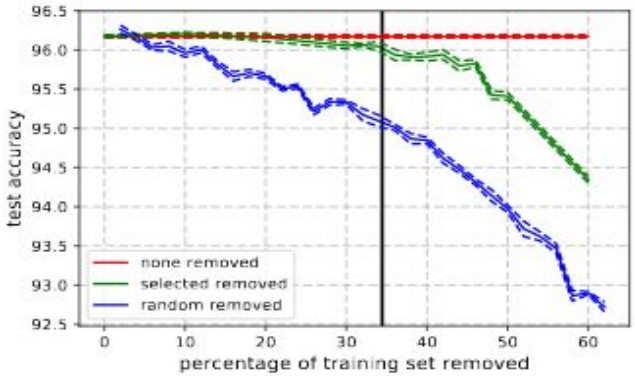
(a) random partitions

(b) partitioning by forgetting events

Figure 4: Synthetic continual learning setup for *CIFAR-10*. Background color in each column indicates the training partition, curves track performance on both partitions during interleaved training. Solids lines represent the average of 5 runs and dashed lines represent the standard error. The figure highlights that examples that have been forgotten at least once can "support" those that have never been forgotten, as shown in (c.2) and (b.3).

Characteristics of Forgettable Examples

Removing (un)forgettable examples



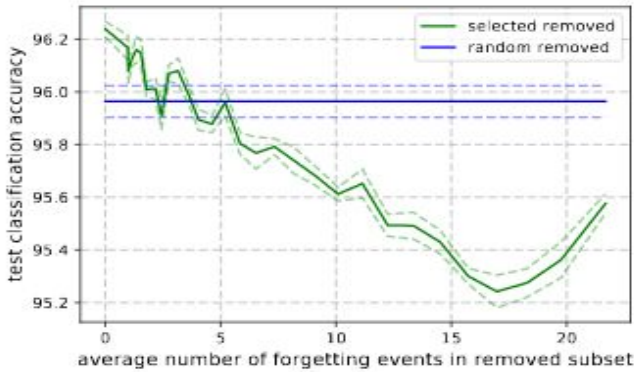


Figure 5: *Left* Generalization performance on *CIFAR-10* of ResNet18 where increasingly larger subsets of the training set are removed (mean +/- std error of 5 seeds). When the removed examples are selected at random, performance drops very fast. Selecting the examples according to our ordering can 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.

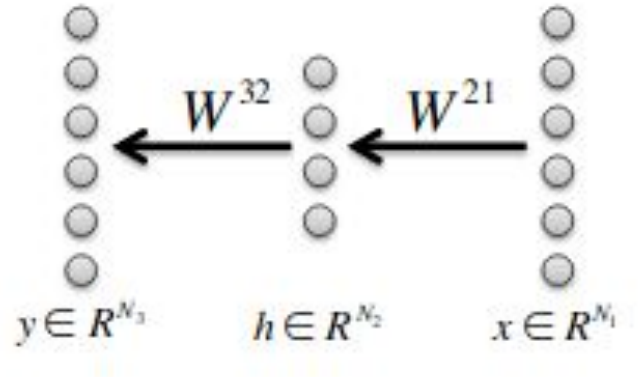
Characteristics of Forgettable Examples

Visual inspection

Figure 2: Pictures of unforgettable (*Top*) and forgettable examples (*Bottom*) of every *CIFAR-10* class. Forgettable examples seem to exhibit peculiar or uncommon features. Additional examples are available in Supplemental Figure 15.

Gradient Updates by Backpropagation

- Investigate how learning new tasks could affect previously learnt information
- Assume a same task, but change in data distributions (eg., MNIST-SVHN)
- Assume deep linear neural networks and squared error loss



$$\mathcal{L} = \sum_{\mu=1}^{P} \| y^{\mu} - W^{32} W^{21} x^{\mu} \|^2$$

Gradient Updates

$$\Delta W^{21} = \lambda \sum_{\mu=1}^{P} W^{32^T} \left(y^\mu x^{\mu T} - W^{32} W^{21} x^\mu x^{\mu T} \right)$$

$$\Delta W^{32} = \lambda \sum_{\mu=1}^{P} \left(y^\mu x^{\mu T} - W^{32} W^{21} x^\mu x^{\mu T} \right) W^{21^T}$$

$$\tau \frac{d}{dt} W^{21} = W^{32^T} \left(\Sigma^{31} - W^{32} W^{21} \Sigma^{11} \right), \qquad \tau \frac{d}{dt} W^{32} = \left(\Sigma^{31} - W^{32} W^{21} \Sigma^{11} \right) W^{21^T},$$

where, $\sum^{11} = \sum_{\mu=1}^{P} x^\mu x^{\mu T}$, $\sum^{31} = \sum_{\mu=1}^{P} y^\mu x^{\mu T}$

Possible Solution 1

Two datasets: (X, Y), (X’, Y’)

Assume X and X’ are whitened => Corr(X) = Corr(X’) = 1

Apply a linear transformation to (X, Y) => (AX, BY) so that

$$A\Sigma^{11}A^T = \Sigma'^{11} \text{ and } B\Sigma^{31}A^T = \Sigma'^{31}$$

The matrix A should be orthogonal

(Naive) The matrix B can be obtained by applying SVD to Sigma’^31

Then, gradient updates for two tasks would follow the same distribution.

Possible Solution (?) 2

Let Z = f(X), Z’ = g(X’) (f and g can be same)

If Z and Z’ share the same distribution, the above problem can be partly solved.

Various adversarial domain adaptation methods do similar things.