Assignment 1 - Paper Reading

Advanced Machine Learning Course

Paper titles:

An Empirical Study Of Example Forgetting During Deep Neural Network Learning (2019)

&

When Deep Learners Change Their Mind: Learning Dynamics for Active Learning? (2021)

Why 2 papers?

- The base paper (Toneva et. al. 2019) provides a fundamental background to support the claims in additional paper (Bengar et. al. 2021)
- The base paper focus on building concepts of forgettable and unforgettable events in continual learning along with the dataset optimisation.
- The additional paper introduces "label dispersion" which quantifies the uncertainty of predictions in multiple training cycles. Further this technique is used in AL paradigm as acquisition function to provide remarkable results.

Contents

1. Motivation.
2. Key Algorithmic details.
3. Limitations based on the understanding.
4. Key takeaways

Motivation

- Forgettable and unforgettable events (catastrophic forgetting)
- Informativeness of the examples
- Fluctuating predicted labels
- Optimise the AL acquisition function such that we get good generalisation performance with low budget.

Key Algorithmic details

Analysing Example Forgetting & Statistics

- What are forgetting and learning events?

- Classification margin

- Unforgettable examples

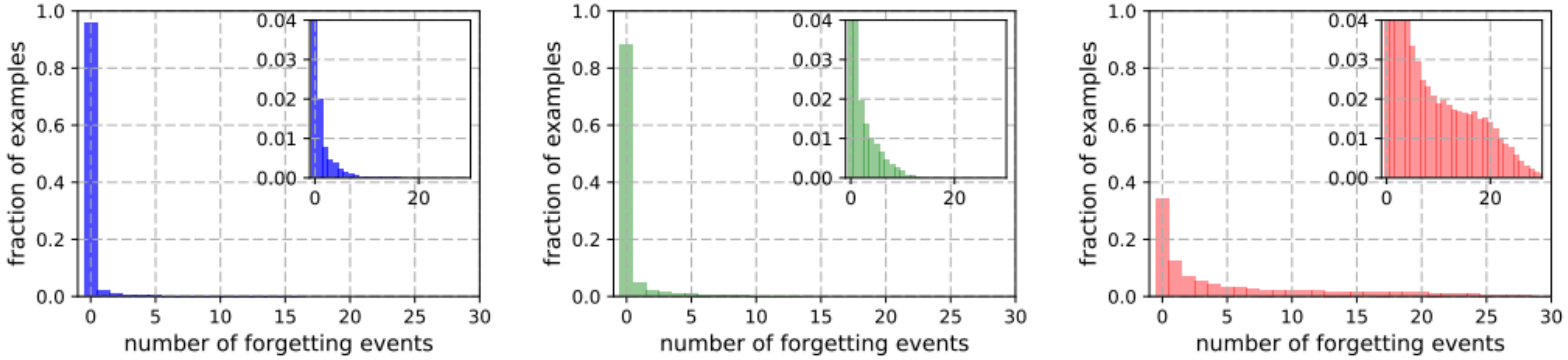


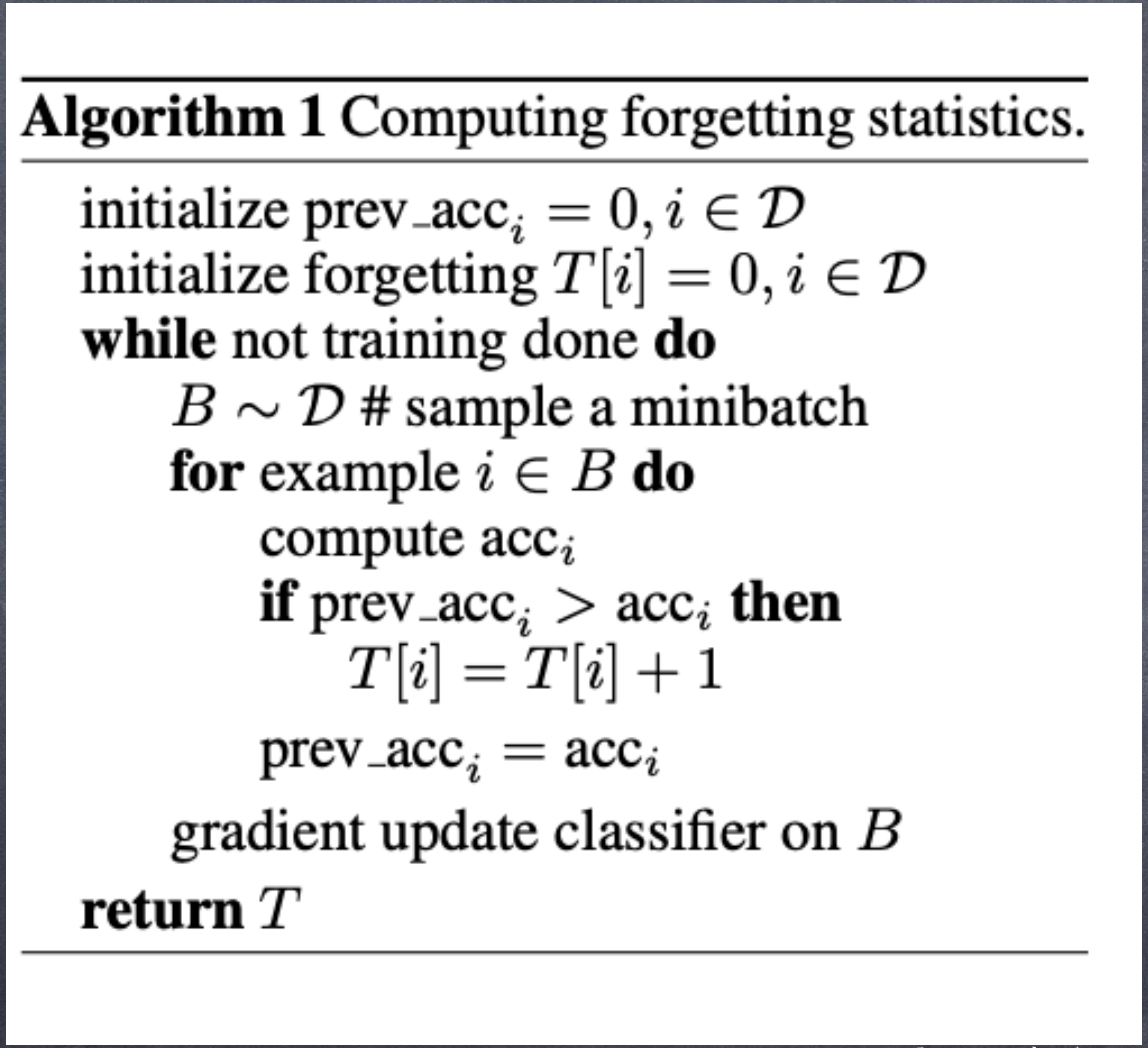
Figure 1: Histograms of forgetting events on (from left to right) *MNIST*, *permutedMNIST* and *CIFAR-10*. Insets show the zoomed-in y-axis.

source: Toneva et. al. 2019

Experimental details

Key Algorithmic details

Forgetting Statistics



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 $T$
```

source: Toneva et. al. 2019

Key Algorithmic details

Different kinds of forgetting events observed:

- Number of forgetting events
- Stability across seeds
- Forgetting by chance
- First learning events
- Misclassification margin
- Visual Inspection
- Detection of noisy examples

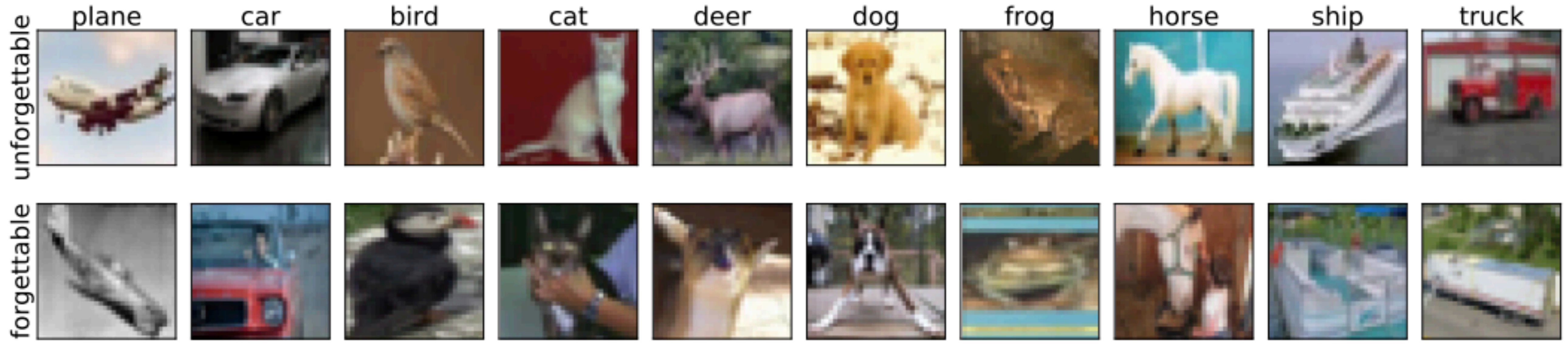
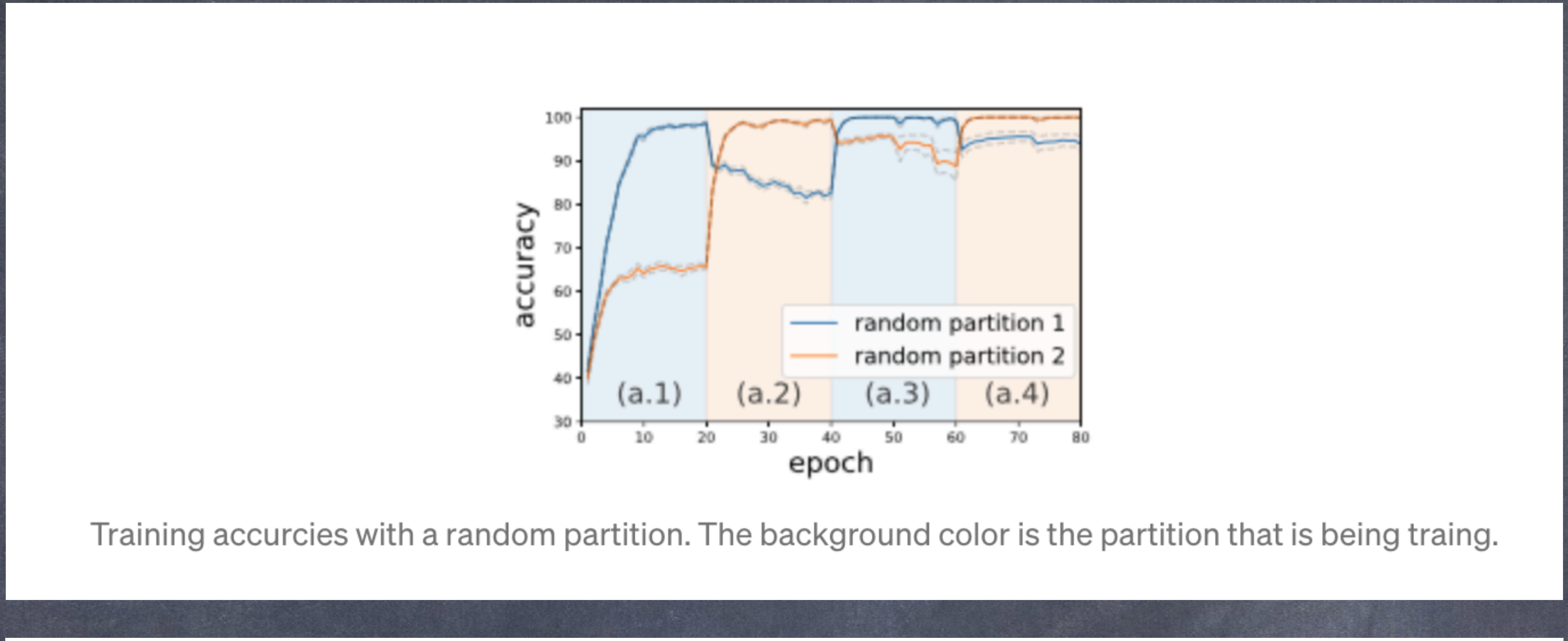


Figure 2: Pictures of unforgettable (*Top*) and forgettable examples (*Bottom*) of every *CIFAR-10* class. Forgettable examples seem to exhibit peculiar or uncommon features. Additional examples are available in Supplemental Figure 15.

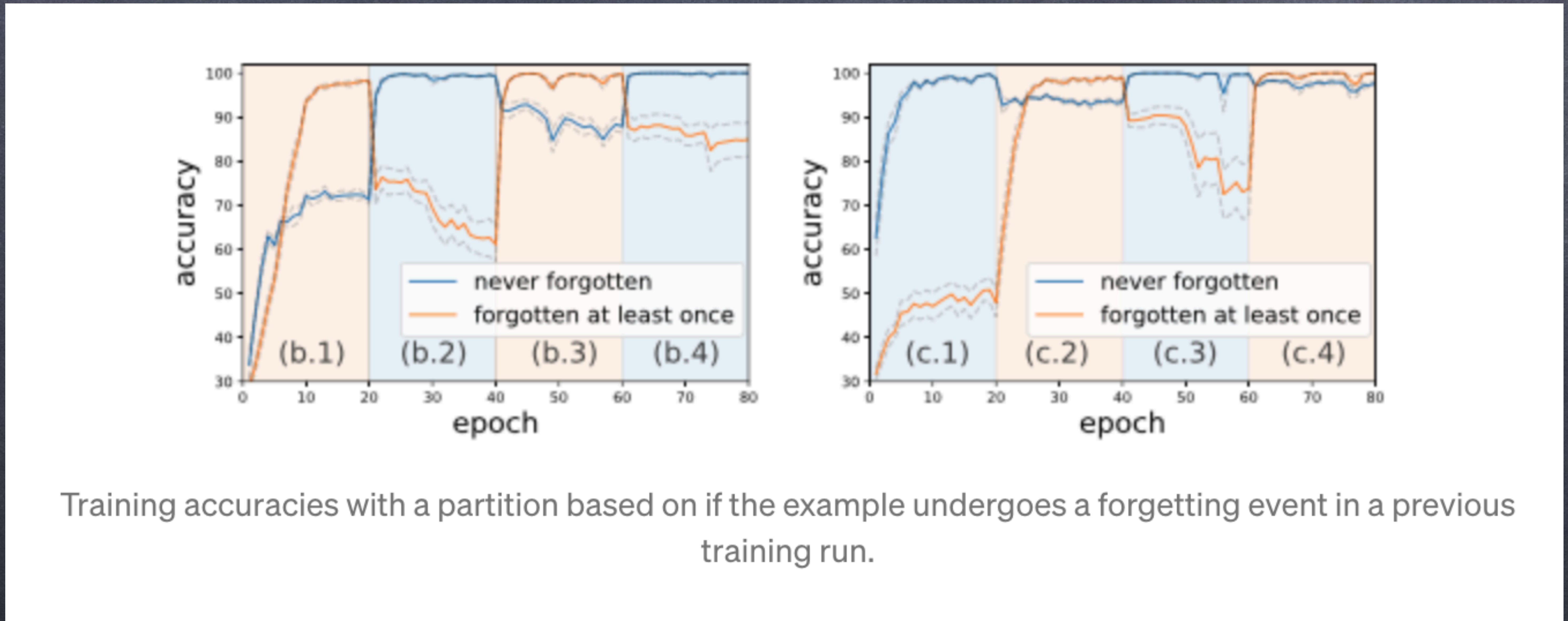
source: Toneva et. al. 2019

Continual Learning...

Key Algorithmic details



Training accurcies with a random partition. The background color is the partition that is being traing.

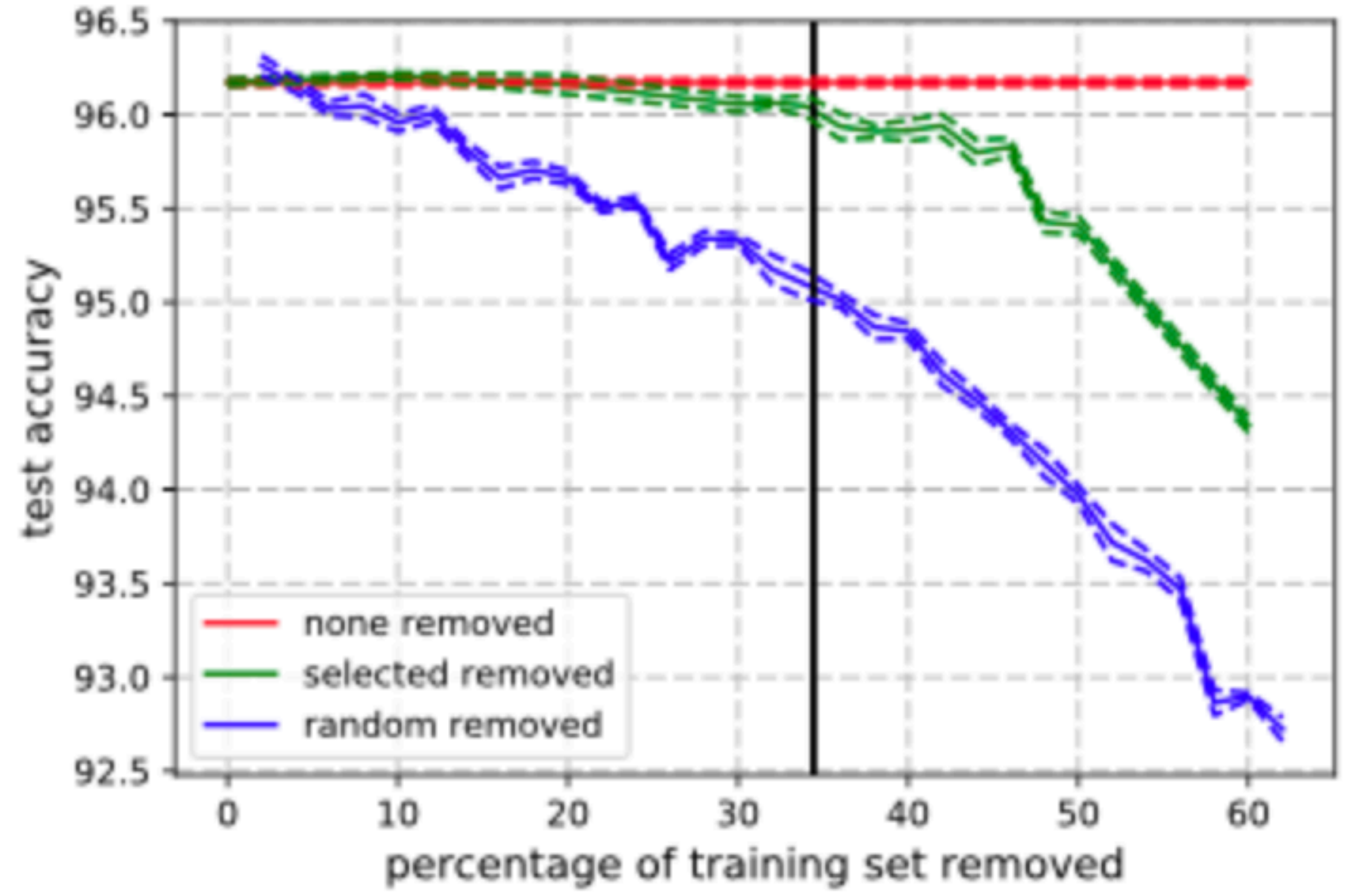


Training accuracies with a partition based on if the example undergoes a forgetting event in a previous training run.

source: Toneva et. al. 2019

Key Algorithmic details

Experimental outcomes of removing unforgettable examples:



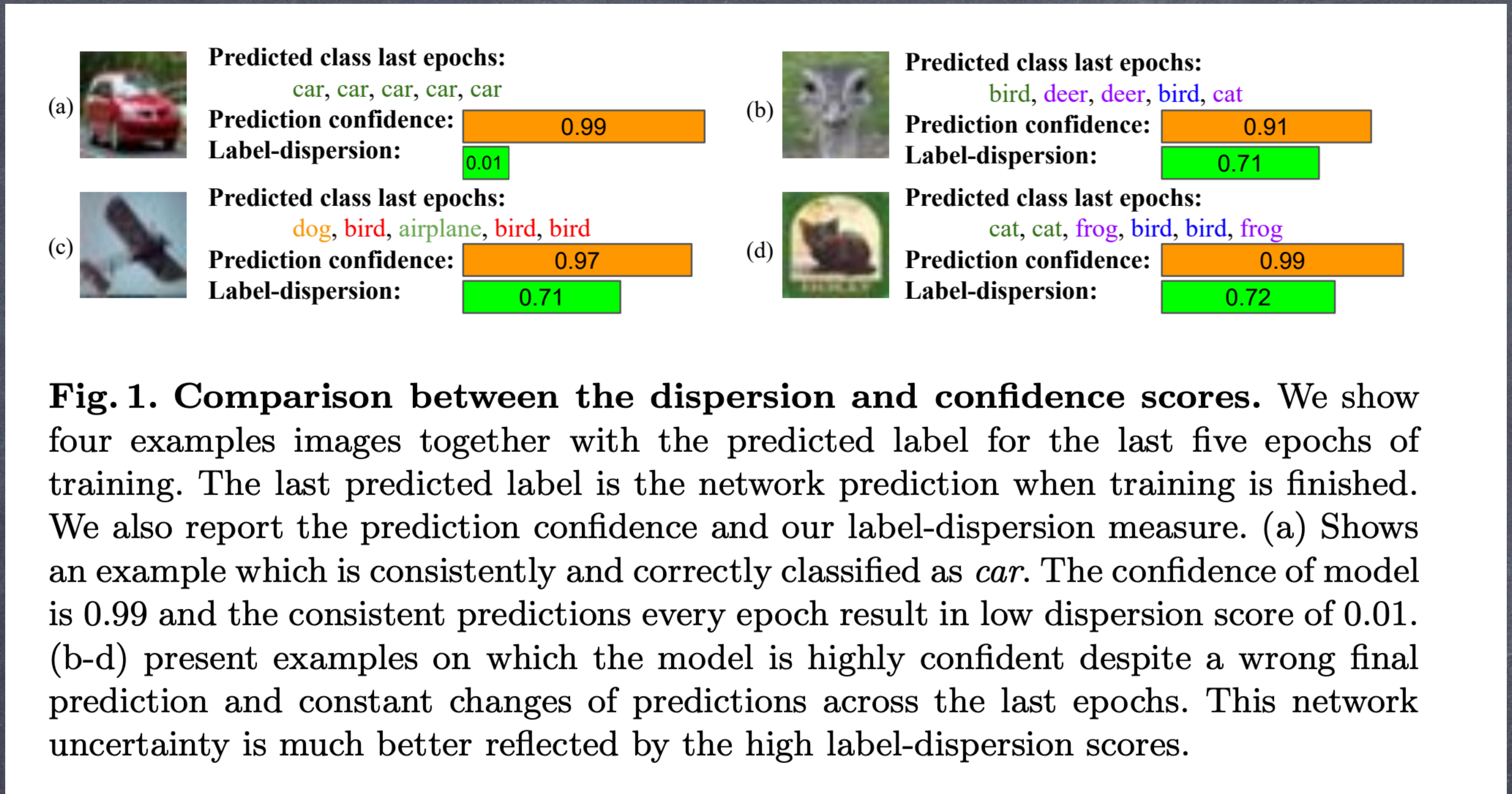
Effect of removing a percentage of the training set of CIFAR-10 when trained using ResNet18. The green line represents removing unforgettable examples. The blue line represents removing examples randomly. The vertical line indicates when all unforgettable examples have been removed. Note that the y-axis is the test accuracy.

source: Toneva et. al. 2019

Key Algorithmic details

- What’s Active Learning
- Variations in labels predicted by the model
- Label Dispersion Acquisition function

Key Algorithmic details



(a) Predicted class last epochs: car, car, car, car, car
Prediction confidence: 0.99
Label-dispersion: 0.01

(b) Predicted class last epochs: bird, deer, deer, bird, cat
Prediction confidence: 0.91
Label-dispersion: 0.71

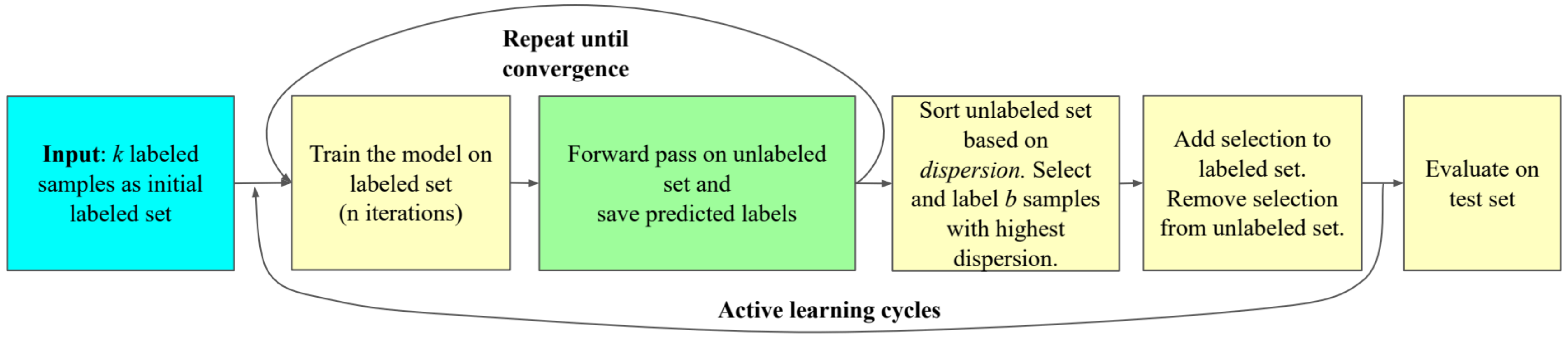
(c) Predicted class last epochs: dog, bird, airplane, bird, bird
Prediction confidence: 0.97
Label-dispersion: 0.71

(d) Predicted class last epochs: cat, cat, frog, bird, bird, frog
Prediction confidence: 0.99
Label-dispersion: 0.72

Fig. 1. Comparison between the dispersion and confidence scores. We show four examples images together with the predicted label for the last five epochs of training. The last predicted label is the network prediction when training is finished. We also report the prediction confidence and our label-dispersion measure. (a) Shows an example which is consistently and correctly classified as *car*. The confidence of model is 0.99 and the consistent predictions every epoch result in low dispersion score of 0.01. (b-d) present examples on which the model is highly confident despite a wrong final prediction and constant changes of predictions across the last epochs. This network uncertainty is much better reflected by the high label-dispersion scores.

source: Bengar et. al. 2021

Key Algorithmic details



Input: *k* labeled samples as initial labeled set

Repeat until convergence

Train the model on labeled set (n iterations)

Forward pass on unlabeled set and save predicted labels

Sort unlabeled set based on *dispersion*. Select and label *b* samples with highest dispersion.

Add selection to labeled set. Remove selection from unlabeled set.

Evaluate on test set

Active learning cycles

Fig. 2. Active learning framework using Dispersion. Active learning cycles start with initial labeled pool. The model trained on labeled pool is used to output the predictions and compute dispersion for each sample. The samples with highest dispersion are queried for labeling and added to labeled set. This cycle repeats until the annotation budget is exhausted.

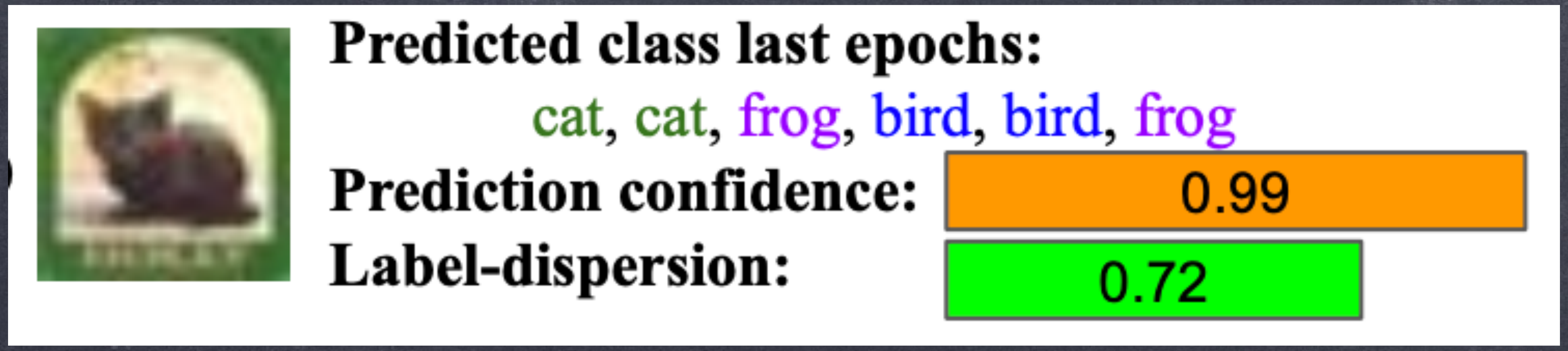
source: Bengar et. al. 2021

Limitations

- Limitation in the paper (Toneva et al 2019)
  - Unlabelled dataset case was not considered. (Active Learning)
  - Causes of forgetting is but not limited to shift in training samples.
  - The results of the datasets after removing unforgettable samples from the dataset were not explained class wise.
  - Hypothesis space is loosely defined which makes it difficult to connect catastrophic forgetting with forgetting statistics.
  - Forgetting is explained profoundly, however it lacks with proper metric to quantify forgetting.
  - They have not explained about the use of noisy labels over noisy data and methods to perturb.
  - No mention of Batch Normalisation and Covariance shift!

Limitations

- Limitation in the paper (Bengar et al 2021)
  - This research only focuses on the score of label-dispersion, one step ahead of this work could hypothesise the underlying behaviour of calculated LD scores. question to put is - Can we learn multiple class labels along with their LD scores such that you need not calculate the divergence or distance from all the classes.
  - In other words, for example, a cat is classified as frog with high PC and high LD. Now, can we just ignore rest of the classes and select first K labels such that a penalty is given for each wrong label.



Predicted class last epochs:

cat, cat, frog, bird, bird, frog

Prediction confidence: 0.99

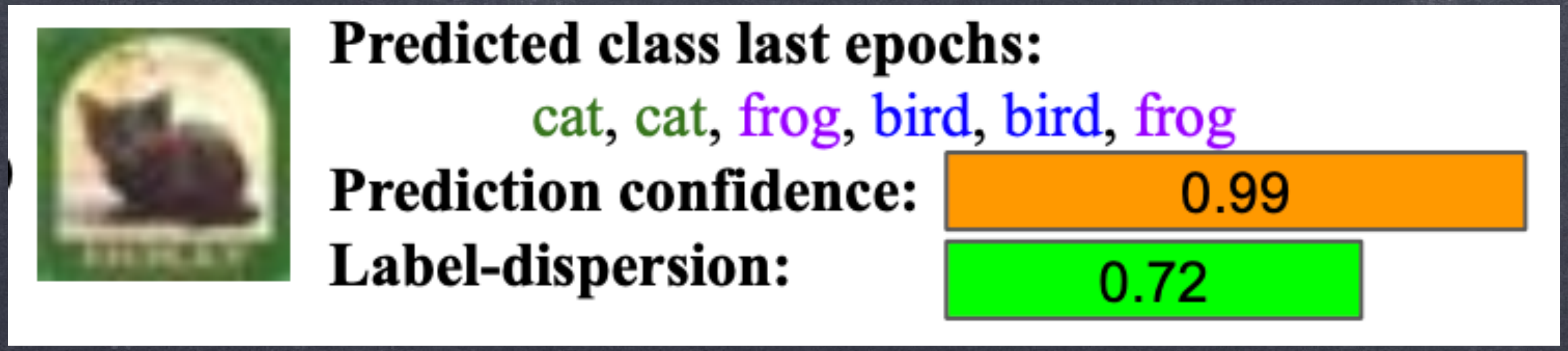
Label-dispersion: 0.72

Limitations

...cont

In simple terms, can I keep track of the labels along with their PC and LD for candidate examples and collect those examples from each class such that, we learn which class has the least possibility to be assigned with current example.

for example below, if we assume cat and frog has such PC and LD score that it could not be anything apart from these 2 then our AL process can be relaxed.



Predicted class last epochs:

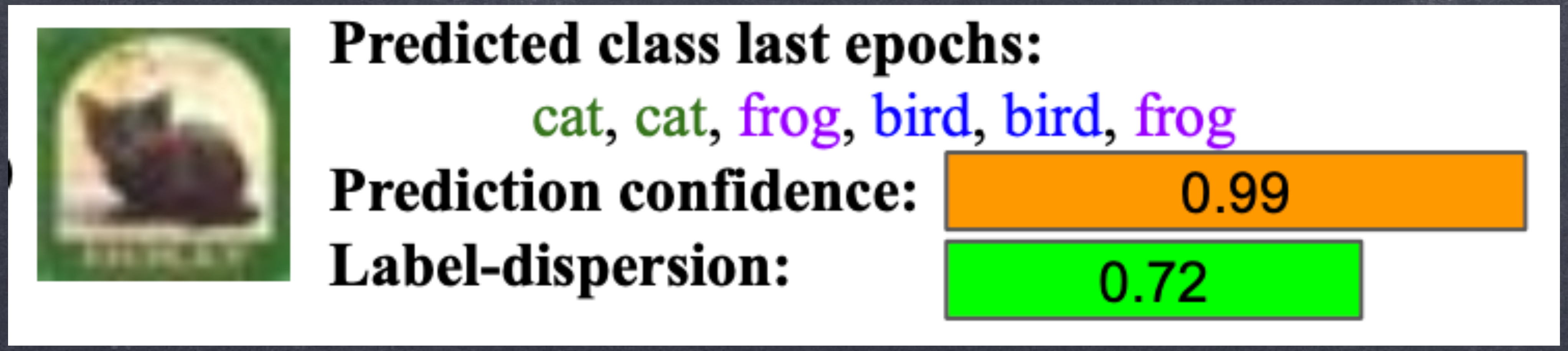
cat, cat, frog, bird, bird, frog

Prediction confidence: 0.99

Label-dispersion: 0.72

Limitations

Since the author compares this kind of occurrence of label shift, with catastrophic forgetting, can we compare the forgetting-statistics, PC and LD metrics with the example of unforgettable example. And we can try to visualise our most confused example with unforgettable example in the input space. However curse of dimensionality will remain a major issue in that case.



Predicted class last epochs:

cat, cat, frog, bird, bird, frog

Prediction confidence: 0.99

Label-dispersion: 0.72

Key Takeaways

1. After achieving remarkable accuracy by the model, even in this case, the model tends to forget previously correctly classified examples or forgettable examples.
2. There exist some examples which are unforgettable.
3. Unforgettable examples are least informative and are very close to decision boundary. Those examples could be eliminated safely without affecting much of the test accuracy.
4. Noisy examples (in terms of label) are easily forgotten by the model.
5. Forgetting statistics tell how many times an example has undergone forgetting.
6. Some examples which are learned in the earlier part of the training tend to be forgettable and unforgettable examples, are learned in very later stages of the training process.
7. Label Dispersion is the key metric define the variations of predicted class labels in intermediary batches.
8. In the AL pipeline, after predicting unlabelled examples, use LD score to sort the dataset from low LD to high LD (recall curriculum learning - easy to difficult). Now add those labelled ones back to pool and repeat till budget exhausts.