Overcoming Forgetting in Continual Learning

Cristian D. Păduraru, Alexandru C. Sasu

Faculty of Mathematics and Computer Science, University of Bucharest

Motivation

Continual Learning

Continual learning (CL) is an important research topic in machine learning that aims to enable models to learn from a continuous stream of data over a long period of time. Two of the main challenges in CL are the problem of **catastrophic forgetting**, where models forget previously learned information when exposed to new data, along with the problem of **scalability**, where models need to be able to handle large amounts of data over long periods of time.

CLEAR-10 (Continual LEArning on Real-World Imagery) dataset

CLEAR is the first continual image classification benchmark dataset with a natural temporal evolution of visual concepts in the real world that spans a decade (2004-2014; it has a bucket of data for each year). CLEAR is built from existing large-scale image collections (YFCC100M) through an approach to visio-linguistic dataset curation. [1]

Goal

Preserve performance on old tasks when learning new ones.

Proposed Methods

Experience Replay

Experience replay is a technique used for training neural networks in reinforcement learning. It involves storing past experiences of an agent interacting with its environment in a replay buffer and then randomly sampling a mini-batch of experiences from the buffer to train the neural network.

However, saving large amounts of the training data from each bucket is not desirable. If we accumulate a large training set over time, then the new data that the model has to learn from would be in minority. This might lead to the model not improving on the new samples, since doing so could increase the loss on the older ones. We thus chose to only keep a few of the hardest samples from each bucket after the training was done on said bucket and tested multiple ways of choosing them.

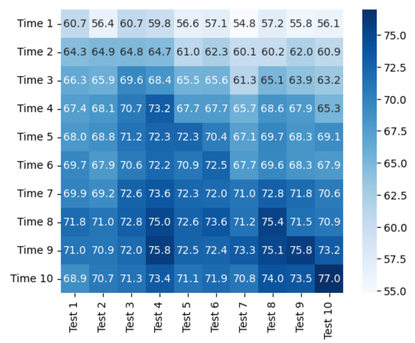
We ordered the samples according to 3 different metrics that should indicate their difficulty and only kept the hardest ones from each bucket. Let (x, y) be a train example, $\hat{y}$ the onehot encoding of y and z the logits (class probabilities) predicted for x. The 3 metrics we tested can be defined as follows:

- Error L2-Norm [2]: $EL2N(x, y) = \|z - \hat{y}\|_2$
- Entropy: $E(x, y) = H(z)$, where H is the Shannon entropy [3] of a random variable
- Margin [4]: $M(x, y) = z_y - max_{i \neq y}(z_i)$, where z_i is the probability of x being in class i

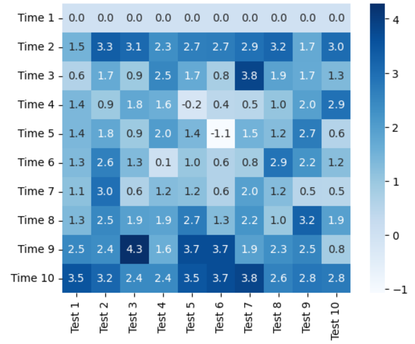
We also tested a random sampling strategy to compare against the proposed criteria.

Results

In a continual learning scenario with n different tasks, the accuracy matrix R is defined as follows: $R_{i,j}$ is the accuracy of the model on task j after training on the first i tasks.



(a) Average accuracy matrix for the continual learning setup used



(b) Average improvement over (a) brought by randomly sampling 5% of each bucket for Experience Replay

Figure 1

We observe in Figure 1 (a) that the accuracy decreases on each column starting from the bottom line on all the buckets, which shows that the model forgets older tasks as time goes on. For the first 4 buckets of data we can also see that the model actually has a better accuracy after being trained on other tasks. We believe that this is caused by the relatively small size of each bucket which makes the model prone to overfitting in the early stages of training. **In Figure 1 (b) we can see that our proposed method with random sampling improves the accuracy on older tasks in later stages of the training.** The fact that it also brings improvements for the last bucket of data also supports the idea that the model finds it hard to generalize on small buckets of data. We argue that, even in this case, our method helps to prevent catastrophic forgetting, since the improvement is larger for older tasks than for recent ones.

Based on the accuracy matrix R, the following metrics can be defined:

- Accuracy (Acc) $= \frac{\sum_{i \geq j}^{n} R_{i,j}}{\frac{n(n+1)}{2}}$
- In-Domain Accuracy $= \frac{\sum_{i=1}^{n} R_{i,i}}{n}$
- Next-Domain Accuracy $= \frac{\sum_{i=1}^{n-1} R_{i,i+1}}{n-1}$
- Backward Transfer (BwT) $= \frac{\sum_{i>j}^{n} R_{i,j}}{\frac{n(n-1)}{2}}$
- Forward Transfer (FwT) $= \frac{\sum_{i<j}^{n} R_{i,j}}{\frac{n(n-1)}{2}}$

Below, we present the average results obtained by running the experiments with 3 different seeds.

Samples saved from each bucket (%)	Metric used	In-domain acc(%)	Next-domain acc(%)	Acc(%)	BwT(%)	FwT(%)
1	el2n	72.06 ±0.2	68.94 ±0.12	72.24 ±0.08	72.28 ±0.06	65.98 ±0.25
	entropy	71.91 ±0.2	68.78 ±0.24	72.26 ±0.17	72.34 ±0.25	65.65 ±0.06
	margin	**72.39 ±0.14**	**69.27 ±0.38**	72.26 ±0.08	72.23 ±0.09	65.88 ±0.23
	random	72.02 ±0.31	69.21 ±0.48	**72.35 ±0.03**	**72.43 ±0.03**	**66.13 ±0.09**
5	el2n	72.36 ±0.53	68.57 ±0.56	72.7 ±0.12	72.78 ±0.11	65.92 ±0.27
	entropy	72.48 ±0.32	**69.56 ±0.2**	72.86 ±0.17	72.95 ±0.19	**66.32 ±0.06**
	margin	72.25 ±0.1	69.01 ±0.49	72.74 ±0.07	72.85 ±0.08	66.09 ±0.4
	random	**72.84 ±0.15**	69.27 ±0.48	**73.11 ±0.14**	**73.17 ±0.14**	66.17 ±0.33
Baseline cl		71.24 ±0.37	68.11 ±0.33	71.14 ±0.37	71.12 ±0.38	64.78 ±0.24
Baseline iid		80.55 ±0.96	-	-	-	-

Conclusions regarding the sampling strategies used:

- regardless of the sampling strategy used, we obtained an improvement in all the mentioned metrics
- randomly choosing the samples obtains better results on a majority of the metrics, which might mean that the hardest samples in a bucket are not necessarily the most helpful in preventing catastrophic forgetting
- future works could consider taking only the easiest samples, or the median ones in terms of difficulty

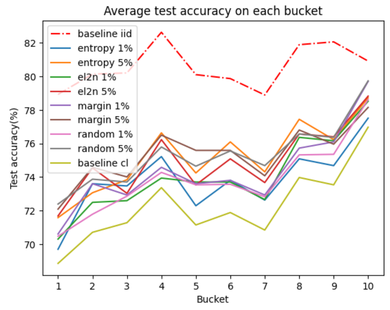


Figure 2. Average test accuracy after training on all the buckets

In Figure 2 we observe how all the sampling strategies tested surpass the test accuracy of the baseline scenario, but are also below the iid training setup where we train on all the buckets at once.

References

[1] Lin, Z., Shi, J., Pathak, D. and Ramanan, D., 2022. The CLEAR Benchmark: Continual LEArning on Real-World Imagery. arxiv: 2201.06289.

[2] Paul, M., Ganguli, S. and Dziugaite, G.K., 2021. Deep learning on a data diet: Finding important examples early in training. Advances in Neural Information Processing Systems, 34, pp.20596-20607.

[3] https://en.wikipedia.org/wiki/Entropy_(information_theory).

[4] Pleiss, G., Zhang, T., Elenberg, E. and Weinberger, K.Q., 2020. Identifying mislabeled data using the area under the margin ranking. Advances in Neural Information Processing Systems, 33, pp.17044-17056.