Understanding Forgetting in Artificial Neural Networks

An MSc Thesis Presentation

Dylan R. Ashley

2020-09-11

University of Alberta
Alberta Machine Intelligence Institute
Reinforcement Learning and Artificial Intelligence lab

What is all the fuss about artificial intelligence?

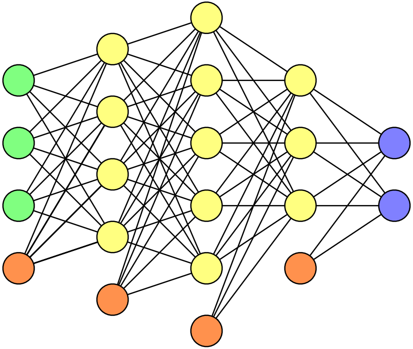
Artificial intelligence has the potential of solving many of the great challenges of our time. We can use it to

- provide personalized, on-demand medical care to everyone;
- reduce vehicular accidents; or
- decrease the social isolation that some individuals face.

Most of the recent breakthroughs in AI have involved artificial neural networks.

What are artificial neural networks?

Artificial neural networks (ANNs) are learning systems loosely based on the brain. They consist of networks of artificial neurons:



Each neuron takes some input $\mathbf{x}$, dot products it with some learned weights $\mathbf{w}$, then passes the result through an activation function $g : \mathbb{R} \to \mathbb{R}$.

Online Learning and ANNs

There are two categories of learning problems:

offline learning the full dataset is available to the learning system when learning starts (standard in supervised learning)

online learning the dataset becomes available to the learning system example-by-example

There are many examples of online learning problems we care about, but ANNs struggle in online learning problems because of catastrophic forgetting.

Research Questions

In this work, we try to answer five questions:

1. What is catastrophic forgetting?
2. How does forgetting in psychology relate to ideas in machine learning?
3. Does catastrophic forgetting exist in contemporary machine learning systems, and, if so, is it severe?
4. How can we measure how a system experiences catastrophic forgetting?
5. Are the current optimization algorithms we use to train ANNs adding to the severity of catastrophic forgetting?

Contributions

In the process of trying to answer the five questions, we make the following contributions:

1. We provide an analytical survey that looks at the concept of forgetting as it appears in psychology and connects it to various ideas in machine learning.
2. We give empirical evidence demonstrating the existence of catastrophic forgetting in some contemporary ANNs.
3. We provide a testbed that helps understand the degree to which some ANN-based learning systems suffer from catastrophic forgetting.
4. We give evidence that the choice of which modern gradient-based optimization algorithm is used to train an ANN has a significant impact on the amount of catastrophic forgetting that occurs during training.

Table of Contents

1. Forgetting: Psychology and Machine Learning
2. An Example of Catastrophic Forgetting
3. Building a Testbed
4. The Impact of Step-size Adaptation
5. Conclusions

Forgetting: Psychology and Machine Learning

What is forgetting?

According to the APA [9], forgetting is defined to be ”the failure to remember material previously learned.”

Forgetting has been primarily studied in biological systems but has become more important for machine learning in recent years due to the increasing complexity of our systems.

Forgetting is **both good and bad** as it frees up resources and encourages learning, but may do so at the cost of performance.

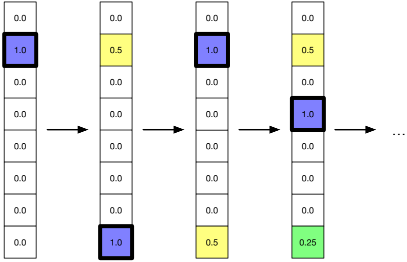
Measuring Forgetting

Hermann Ebbinghaus [3] conducted the earliest experimental study on forgetting in 1885. He would memorize lists of nonsense syllables, wait a bit, then see how much faster he could memorize them a second time.

The ratio between the time it takes to learn a task once and the time it takes to learn it a second time has been used to measure forgetting in ANNs. It is referred to as the relearning measure of forgetting.

Forgetting in Biological Systems

There are many competing theories for what causes forgetting in biological systems. Decay theory argues that learning leaves an impression on our brain which, without rehearsal, fades over time. **Eligibility traces**, as used in reinforcement learning, can be viewed as a model of forgetting under decay theory:



0.0	0.0	0.0	0.0
1.0	0.5	1.0	0.5
0.0	0.0	0.0	0.0
0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0
0.0	1.0	0.5	0.25

Experience replay can be said to overcome forgetting under decay theory by explicitly rehearsing recent events.

Interference Theory

In contrast to decay theory, Interference theory argues that it is the interference between memories which cause forgetting. It distinguishes two distinct kinds of interference:

retroactive interference when new learning causes us to forget things we previously learned

proactive interference when previous learning causes us to forget things we just learned

The original **1989** investigation McCloskey and Cohen [7] into catastrophic forgetting explicitly looked at retroactive interference.

Interference Theory and Transfer Learning

The psychology study by Barnes and Underwood [2] which McCloskey and Cohen referenced in their work sought to measure the effect of transfer, i.e., the effect of prior learning on future learning, on forgetting:

Expected Transfer	List 1 Example	List 2 Example	List 1 Recall
Y	den-red	den-angry	0.9
N	fu-green	fu-fast	0.3

A consequence of this is that the most common way of measuring forgetting in ANNs is to observe the ability to perform a first, previously mastered-task after mastering a second task, i.e., the retention.

Forgetting and Generalization

Forgetting is also intricately linked to generalization. French postulated that **the generalization ability of ANNs is the cause of catastrophic forgetting**[5]. He argued that catastrophic forgetting can be measured by looking at the degree of generalization in the network.

Abraham and Robins [1], however, found that contemporary neuroscience research suggests that the biological neural networks represent information in a neither wholly global nor wholly local fashion. This suggests that **the brain is able to deal with forgetting effectively despite its distributed representation**.

Summary

Here we

- looked at the concept of forgetting as it appears in psychology and connected it to various ideas in machine learning, and
- noted that **forgetting is a subtle, long-studied phenomenon** which is an integral part of many learning systems.

An Example of Catastrophic Forgetting

Why provide evidence of catastrophic forgetting here?

ANNs have changed significantly since McCloskey and Cohen's 1989 study. Thus, while much work has looked at catastrophic forgetting since then, it remains useful to validate that the results of McCloskey and Cohen's experiment would not have changed if done under contemporary practices. **Doing so would also serve as a useful demonstration of catastrophic forgetting.**

Additionally, this can serve as a good opportunity to validate and contrast later results that looked at the relearning metric.

Experimental Setup (1)

In this experiment, we construct **two, two-class classification problems from MNIST**. The first task is to classify images of 1s and 2s, and the second task is to classify images of 3s and 4s. In both tasks, samples are given to the learning system one-by-one.

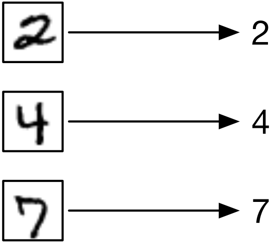


Figure 1: The MNIST dataset consists of pictures of handwritten digits paired with the digit the author was trying to write.

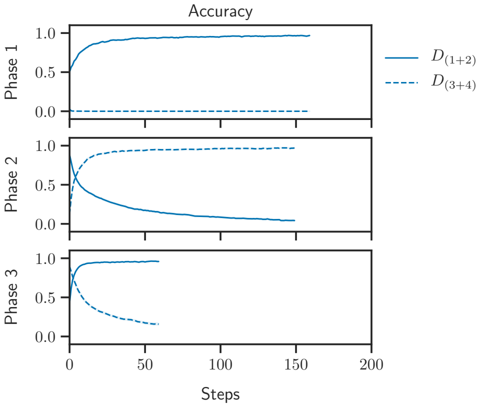
Experimental Setup (2)

Using these MNIST tasks, we

1. build a fully-connected feedforward ANN with one 100 ReLU hidden layer,
2. initialize the weights in this network by sampling from a Gaussian with mean 0 and a standard deviation of 0.1,
3. train this network with SGD and backpropagation on the 1s and 2s task until it achieves a **running accuracy of 90%**,
4. continue training this network on the 3s and 4s task until it achieves a running accuracy of 90% while simultaneously measuring its accuracy on the 1s and 2s task, and finally
5. continue training this network on the 1s and 2s task and compare how long it takes to achieve a running accuracy of 90% a second time as compared to the first time it tried to solve this task.

Results

While the learning system was able to solve both problems fairly easily, the retention following the second phase was low, and the relearning time was very short:



Summary

Here we

- gave empirical evidence demonstrating the existence of catastrophic forgetting in some contemporary ANNs, and
- showed that catastrophic forgetting cannot be effectively explained by only looking at retention-based metrics.

Building a Testbed

Why build a testbed?

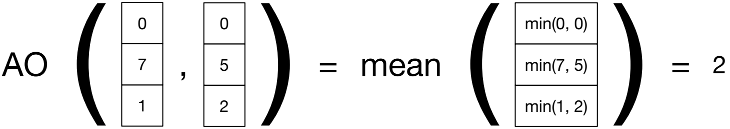
Some attempts have been made to build testbeds for catastrophic forgetting explicitly, but few have considered non-retention based metrics.

We earlier showed that **catastrophic forgetting is a subtle phenomenon**. So what we want to have is a testbed with

- several very different metrics for measuring catastrophic forgetting, and
- multiple fully-online learning tasks of which some have strong temporal-correlation in their data-stream.

Metrics (1)

Apart from the retention and relearning metrics, we also include the activation overlap metric. Activation overlap was proposed by French [4] in **1991** and uses the *similarity of the representation for two samples* as a measure of an ANN's susceptibility to catastrophic forgetting.



$$\text{AO}\left(\boxed{\begin{array}{c}0\\7\\1\end{array}}, \boxed{\begin{array}{c}0\\5\\2\end{array}}\right) = \text{mean}\left(\boxed{\begin{array}{c}\min(0, 0)\\\min(7, 5)\\\min(1, 2)\end{array}}\right) = 2$$

Figure 2: The activation overlap between two samples is defined to be the average of the element-wise minimum of activations the network produces on the samples.

Does activation overlap make sense?

Yes. A universal rule of distributed learning systems is that–all else being held constant–**the more local the representation a learning system employs, the less interference will occur during learning**.

However, modern thinking suggests using the dot product of representations rather than the average of the element-wise minimum. We refer to the result of this change as activation similarity to avoid confusion.

Metrics (2)

The fourth and final metric we include in our testbed is pairwise interference. Pairwise interference [8, 6, 5] measures the susceptibility of a learning system to catastrophic forgetting by directly observing how *the performance on one sample changes following an update induced by a second sample*. It can be written as follows:

$$PI(\theta_t; \mathrm{x}_t, \mathrm{x}_i) = J(\theta_{t+1}; \mathrm{x}_i) - J(\theta_t; \mathrm{x}_i)$$

Settings (1)

We retain the MNIST setting, but this does not address the issue that we require settings with strong temporal-correlation in their data-streams. So we additionally include mountain car under a fixed behaviour policy. This is thus a reinforcement learning prediction, or a **value estimation** problem.

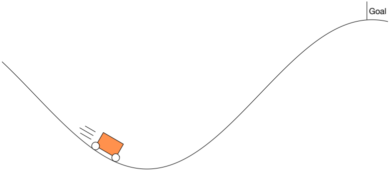


Figure 3: For mountain car, we use a policy that applies force in the direction of movement and no force if the velocity is currently zero.

Settings (2)

In addition to mountain car, we also include acrobot under a fixed policy in our testbed. Like mountain car, this is a value estimation problem.

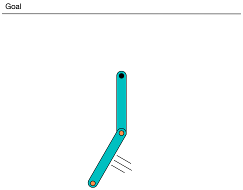


Figure 4: For acrobot, we use a policy that applies force in the direction of movement of the inner joint and no force if the outer joint has an absolute velocity more than ten times the inner joint. This overcomes instances of centripetal force leading to non-terminating episodes.

What does forgetting mean in a one-task setting?

Both mountain car and acrobot data-stream move smoothly through a state space. In such a situation, **learning from a sample in one area of the space can lead to forgetting about other areas of the space**. However, in a one-task setting, retention and relearning are not straightforward to measure and so we cannot apply them to the mountain car and acrobot settings.

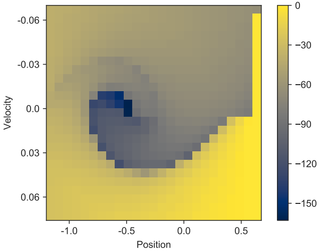


Figure 5: The state space in Mountain Car is defined by a position and velocity. Due to temporal correlation, catastrophic forgetting is possible while moving around the state space.

Summary

Here we

- constructed a testbed using MNIST, mountain car, and acrobot;
- included four metrics in our testbed: retention, relearning, activation similarity, and pairwise interference.

The Impact of Step-size Adaptation

What is the question we want to answer with the testbed?

We apply the testbed to answer whether or not the choice of which modern gradient-based optimization algorithm is used to train an ANN has a significant impact on the amount of catastrophic forgetting that occurs during training. We are especially interested in and experiment with Adam, SGD with Momentum, and RMSProp. We compare these to SGD.

Experimental Setup (1)

The testbed is designed for **ANN-based learning systems**. That means we require network architectures, initialization strategies, and optimization algorithms. We use the following network architectures and initialization strategies for all four optimizers:

MNIST fully-connected feedforward network with one hidden layer of 100 ReLUs and normal random initialization

Mountain car fully-connected feedforward network with one hidden layer of 50 ReLUs and Xavier initialization

Acrobot fully-connected feedforward network with two hidden layers of 32 then 256 ReLUs and He initialization

Experimental Setup (2)

Each of the optimizers has some additional hyperparameters:

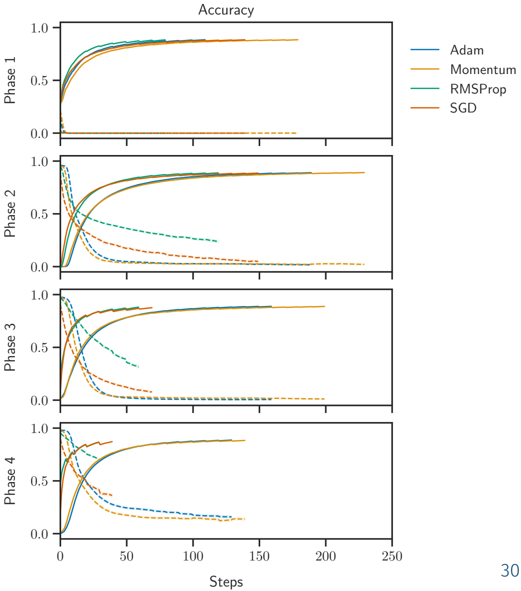
Optimizer	Hyperparameters
SGD	α
Adam	α, β_1, β_2, ϵ
SGD with Momentum	α, β_1
RMSProp	α, β_2, ϵ

We fix β_1, β_2, and ϵ to the normal values used in Adam: 0.9, 0.999, and 10^{-8}, respectively. We then select α for each optimizer and setting by following the typical strategy of sweeping over a range of values and selecting the value that minimizes the performance measure.

MNIST Retention and Relearning Results

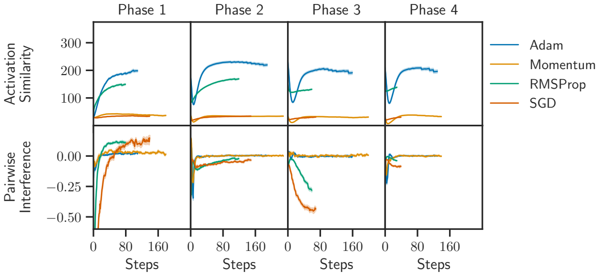
All of the optimizers were able to smoothly move through all four phases. Additionally:

- Adam and SGD with Momentum had the lowest retention and worst relearning
- **RMSProp had by far the highest retention**
- **SGD had by far the best relearning**



MNIST Activation Similarity and Pairwise Interference Results

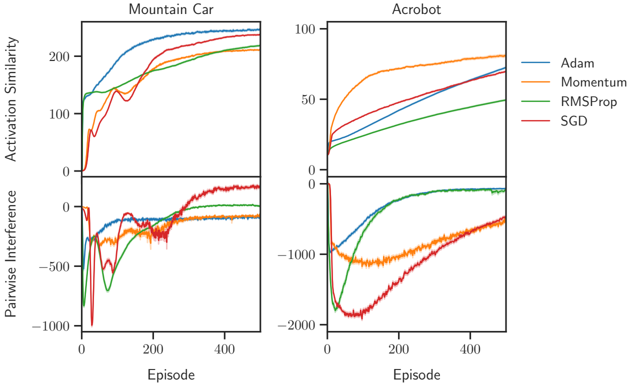
Again in MNIST, the optimizers showed widely different behaviour with respect to the new metrics:



Note that RMSProp displayed the second-highest quantity of activation similarity.

Mountain Car and Acrobot Results

In mountain car and acrobot, the optimizers showed widely different behaviour with respect to the new metrics:



Note that Adam had amongst the lowest pairwise interference in mountain car. This is the only instance of Adam being amongst the best in our results.

Summary

We applied our testbed to answer whether or not the choice of which modern gradient-based optimization algorithm is used to train an ANN has a significant impact on the amount of catastrophic forgetting that occurs during training. We concluded that:

- step-size adaptation does have a meaningful and large effect on catastrophic forgetting in ANNs,
- while more verification is needed, Adam seems to be more at risk from catastrophic forgetting than the other optimizers,
- the amount of catastrophic forgetting an algorithm experienced was highly dependent on the metric used as well as the setting, and
- omitted for brevity, α, β_1, and β_2 all had a pronounced but smooth effect in most cases.

Conclusions

Contributions

In this work, we tried to answer five questions. In the process of doing so, we

1. We provide an analytical survey that looks at the concept of forgetting as it appears in psychology and connects it to various ideas in machine learning.
2. We give empirical evidence demonstrating the existence of catastrophic forgetting in some contemporary ANNs.
3. We provide a testbed that helps understand the degree to which some ANN-based learning systems suffer from catastrophic forgetting.
4. We give evidence that the choice of which modern gradient-based optimization algorithm is used to train an ANN has a significant impact on the amount of catastrophic forgetting that occurs during training.

Implications

This work, among other things, suggests that

1. users should be wary of the optimization algorithm they use with their ANN in problems susceptible to catastrophic forgetting (especially when using Adam but less so when using SGD), and
2. individuals studying catastrophic forgetting should consider a holistic perspective on the phenomenon and not allow their work to be limited by looking at a single setting or a single metric.

Future Work

More work should be done to verify the conclusions reached here. Future directions for additional inquiry should seek to

- understand why Adam exhibited such a high degree of forgetting here, and
- observe how the testbed reports the degree to which other mechanisms in our contemporary learning systems are affecting catastrophic forgetting.

Acknowledgements (in alphabetical order)

A full list of acknowledgements appears in my thesis. I want to extend a special thanks to the following individuals and organizations:

Abhishek Naik, Alberta Machine Intelligence Institute, Banafsheh Rafiee, Chen Ma, Compute Canada, **Dr. Carrie Demmans Epp**, **Dr. Mark Ring**, Dr. Nuanyi Liang, **Dr. Patrick Pilarski**, **Dr. Richard S. Sutton**, Han Wang, Khurram Javed, Kris De Asis, My Family and Friends, Natural Sciences and Engineering Research Council of Canada, **Reinforcement Learning and Artificial Intelligence lab**, **Sina Ghiassian**, Tian Tian, University of Alberta, Vincent Liu, and many more

Questions?

References i

W. C. Abraham and A. Robins.
Memory retention–the synaptic stability versus plasticity dilemma.
Trends in Neurosciences, 28(2):73–78, 2005.

J. M. Barnes and B. J. Underwood.
“fate” of first-list associations in transfer theory.
Journal of Experimental Psychology, 58(2):97–105, 1959.

H. Ebbinghaus.
Memory: A contribution to experimental psychology.
Teachers College Press, 1885.
H. A. Ruger & C. E. Bussenius, Trans., 1913.

References ii

R. M. French.
Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks.
In *Proceedings of the Annual Conference of the Cognitive Science Society*, pages 173–178, 1991.

S. Ghiassian, B. Rafiee, Y. L. Lo, and A. White.
Improving performance in reinforcement learning by breaking generalization in neural networks.
In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, 2020.

V. Liu.
Sparse representation neural networks for online reinforcement learning.
Master's thesis, University of Alberta, 2019.

References iii

- M. McCloskey and N. J. Cohen.
  Catastrophic interference in connectionist networks: The sequential learning problem.
  24:109–165, 1989.

- M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro.
  Learning to learn without forgetting by maximizing transfer and minimizing interference.
  In *Proceedings of the International Conference on Learning Representations*. OpenReview, 2019.

- G. R. VandenBos.
  ***APA dictionary of psychology*.**
  American Psychological Association, 2 edition, 2015.