Overcoming catastrophic forgetting in neural networks

s1424164

NIP Coursework 2

September 2, 2019

Describe the central problem the paper is attempting to address.

The central problem this paper is attempting to address is to overcome a historic weakness of neural networks, which is called catastrophic forgetting. Catastrophic forgetting is a phenomena whereby networks, once trained to perform well on domain specific tasks, are unable to take in information relating to other tasks and generalizing to them, without disrupting performance on distinct previous tasks. This means that neural networks cannot perform sequential learning across different tasks. This phenomena is an example of the stability-plasticity dilemma, which sees a trade-off between stability in light of new information and generalization capacity.

How does it build on previous work and evidence?

Catastrophic forgetting has existed as a recognized phenomena since the inception of neural networks in the 1980s, and naturally there have been various attempts which try to solve the problem. A few of the historic proposed solutions are briefly detailed here.

One approach [1], ensures that all data from all tasks are accessible during the training procedure, creating a joint estimation problem. This is analogous to the idea of an episodic memory system that records the tasks and then replays them to the network during training. However, this is an unconvenient solution as it requires all the data be accessible all the time.

Hetherington et al. [2] tries to circumvent the problem by looking at the problem through the lens of weight overlap destroying knowledge representation. Rather than using a fully connected network during training, hidden units are randomly connected to only a subset of input units, in a similar fashion to dropout, meaning that it is likely that different inputs are not encoded by the same hidden units and weights. Other work, done by Robert French, has found that creating orthogonal input representations that assist in sequential learning, for instance using -1 and 1 as input representations rather than 0 and 1, improve the generalization performance of the network [3]. Robert French also then goes on to try another approach [4]. Here a network is trained, then random inputs are fed into the network and the outputs known as pseudo-patterns are collected and then fed into the training data of the next task. These pseudo-patterns are theorized to give information about the error surface

of the task at hand. There has also been work to prevent catastrophic forgetting by using the local winner takes all property of neurons to achieve network dynamics that are modular [5]. Another approach [6], uses the so called novelty rule that tries to avoid weight overlap. Here, the network is presented with a novelty vector, which describes how the new stimulus has changed and thus the weights are changed in proportion to the novelty. However, this requires calculation of a novelty vector that may not be easy.

Previous work in the reinforcement learning domain has used a approach whereby several networks are trained on a task and then these networks are used to train a larger network that can perform all tasks [7].

Describe the methods used, focussing on aspects central to the results.

The paper introduces the framework of sequential learning by considering two tasks which follow in succession, task A followed by task B. The authors claim that the multiple realisability of performance in parameter space for task A, makes it 'likely' that a region of low error for task B can be found, close to that for task A. The key mathematical formalism used in this paper is a Gaussian distributed variational approximation of the posterior, which is manipulated in the log domain, $\log P(\theta|D)$ over the parameters of the neural network. The posterior over the parameters for task B includes the posterior of the parameters of task A. In order to then ascertain which parameters, or weights, affect the posterior induced from task A, a second order derivative matrix over the posterior with respect to the loss from task A is used, otherwise known as the Fisher information. This measures the curvature of the loss function when a particular parameter is modified.

This results in the following two equations:

$$\log p(\theta|D) = \log p(D|\theta) + \log p(D_B|\theta) + \log p(\theta|D_A) - \log p(D_B)$$

The plausibility of the data, given parameters (the likelihood) for task B then decomposes into two terms that need to be minimised. Here $L(\theta) = \log p(D|\theta)$

$$L(\theta) = L_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^\star_{A,i})^2$$

Here, λ is a hyper-parameter that determines the importance of the previous task over the current task. The term in the summation is large, when not only the parameters of task A and task B are different, but also the parameter is highly sensitive with respect to the loss over task A. This means that the posterior can now be computed, though the paper does not specify the time complexity of the computation of the log probability of the data, especially after it expands via the law of total probability. If a tertiary task C is introduced,

the optimization procedure can then include two penalties for both tasks.

Summarise the key finding(s) and advances presented in the paper.

EWC extends the memory lifetime for random patterns.

Here, the authors train networks to associate binary patterns with binary outcomes, in a similar vein to Hopfield neural networks. The authors note that the diagonal of the Fisher matrix, is proportional to the number of patterns observed. This is intuitive in the sense that as more patterns get added, the parameters become more sensitive to perturbation. Hopfield networks experience such a phenomena whereas the addition of noisy memories encoded into the network, eventually degrades the performance of the network entirely to the point of blackout catastrophe. The EWC algorithm also exhibits this phenomena.

The authors show that a network that uses gradient descent (GD) to optimise parameters to encode patterns performs worse than EWC in terms of the signal to noise ratio as the number of patterns encoded increases. After about 100 patterns, the new learning rule starts to clearly outperform GD, following a power law performance reduction, rather than an exponential reduction for GD. This means that GD can only encode about 400 patterns with full accuracy, whereas EWC encodes about 750 with full accuracy.

Analytic and numeric simulations show that EWC has can encode about five times as many patterns, for a signal strength above 10%, than GD, 10,000 rather than 2,300. EWC also has a reversal in the increase of noise as the number of patterns encoded increases, whereas GD has a very much monotonic increase in the noise as the number of patterns encoded into the network, increases.

EWC Allows Continual Learning in a Supervised Learning Context.

Here, an instantiation of a continual learning task is given in the MNIST domain with a deep neural network. Here, MNIST is modified such that subsequently learned digits have their pixels shuffled in random permutations. The specifics of the continual learning task were unclear from the paper as well the two papers it references to describe the task, which I feel is a significant weakness of this paper. I suspect the task consists of learning a group of MNIST number images to their class label pairing, followed by learning a group of identical number, but randomly permuted images to their class label pairing.

EWC is compared on a series of three tasks in order; A, B and C, to both SGD and an L2 regularised network. Here, there seemed to be two competing interpretations of the L2 regularisation; one where the weights had an L2 penalty applied to them if they deviated from the previously trained task (which made more analytical sense) and the other where it was simply a retrained network with L2 regularisation applied. I take the latter interpretation, in keeping with usual definition of L2 regularisation.

SGD, upon presented with a new task during training, unlearns - to an extent, whichever

task that it was previously trained on and then focusses on learning the new task. SGD when presented with only one task, consistently performs the best. The performance on the network trained with SGD over the other tasks generally degrades as multiple tasks are learned. A quadratic penalty over weights keeps performance strong for the first task for the entire training duration because it punishes weights that increase by much. I suspect that this is because it found an optima for the task A which happened to be near that for one for task B. EWC retains the best performance when tested on subsequent tasks. Whilst EWC loses a little performance over SGD on the third task C, it still retains the best performance on all experiments which involve more than one task. EWC performs well in general and is able to keep the fraction correct above 96% for 11 tasks, whereas SGD + dropout regularisation only manages about 79% over 11 tasks.

The authors then investigate the representation of the new patterns in the network and whether the same weights code for different tasks, through the Fisher overlap metric. This compares the Fisher matrices using the Fréchet distance, which is low when the same parameters are important to both tasks. The authors find that task similarity correlates with shared representation with respect to weights in the network. However, because the output labels are shared, they find that even dissimilar tasks do share some of the weights. The fact that the weights are shared, lends credence to the nontriviality of the approach the authors take. They also find that in early layers of the neural network, the fisher overlap is smaller when tasks are more different (larger degree of permutation)

EWC Allows Continual Learning in a Reinforcement Learning Context.

Here, the authors promote a more ambitious task in the reinforcement learning domain, specifically the DQN algorithm. The DQN algorithm uses neural networks to estimate the q-value function of an environment, which measures for a given state and an action, the quality so to speak of that pairing. The DQN algorithm in this case also has an experience replay buffer which allows for non-correlated updates to the state-action values. The authors also take inspiration from top down neural mechanisms that affect memory, action selection and sensory processing. They use a higher level idea, task context, to influence which quadratic constraints to update. Here, the Forget Me Not algorithm is used to perform unsupervised clustering based inference to know which task is being played. This information is then somehow internalised into the learning procedure, also allowing for certain network parameters to be game specific also. However, the specifics of this task contextual mechanism is not elaborated on in the paper. The authors find that the performance of the EWC algorithm is not as good as separate Deep Q-networks for all ten games tested on. To figure out why, they further investigate the Fisher matrix. The authors test whether the Fisher matrix actually gives a good estimate of parameter importance. By using the inverse Fisher matrix, they find that important parameters were generally important, in that modifying those weights created poor task outcomes, but often parameters that the Fisher matrix deemed unimportant were often important. This means they underestimate parameter uncertainty.

Discuss the biological plausibility if you have chosen a paper about artificial systems, or the insights that could benefit artificial systems if you have chosen a

biological paper.

There has been experimental evidence to suggest that the brain avoids catastrophic forgetting by protecting the information it already has [8]. This supports the biological plausibility of this paper. Here, experimental evidence finds that dendritic spines which receive excitatory inputs are strengthened and protected when a mouse learns a new skill. This is perhaps mathematically analogous to ensuring that task important parameters are less variable. However, there is an important difference. Protecting synapses when learning a new skill is different to waiting until a new task is learned before assessing parameter sensitivity.

The authors also link their construction to cascade models, which have been proposed to be biologically plausible. These model variable plasticity rates for different neurons resulting in improved task retention [9]. However, an important difference between cascade models and the model proposed here is that this model only focusses on one aspect of neuroplasticity i.e. memory retention rather than the forgetting of irrelevant information. Whether forgetting is an active process or simply an epiphenomena of weight change is not clear. Another difference is that cascade models consider potentiation and depression events, whereas this approach is more concerned with the degree to which synapses should change, without specifying the direction.
The authors also link their idea of neurons changing in accordance to their importance for a task, to a Bayesian understanding of neural computation, whereby each neuron has some understanding of the uncertainty surrounding its own weight, through the use of PSP variability [10]. However, they suggest that more work needs to be done to confirm these ideas.

Whilst the authors claim that memory retention is modelled using the Fisher matrix, it is clear from other work on the neurobiology of memory that there is considerable complexity to the way in which the brain stores memories, which are theorised to be stored in different attractor states. This paper gives no indication of how neural dynamics could give rise to a mathematical structure resembling the Fisher matrix. This is especially important as it requires global information with respect to the network to calculate it. In addition, a Gaussian approximation of the posterior is a likely naive and restricting stance. I suspect that in reality the brain probably uses a far more complex generative model. In [11], the diversity of memory mechanisms are emphasised. There is no account of how this work would fit into the different forms of memory and their neurobiological counterparts mentioned in [11]. Examples of this are across neuron signalling for short term memory and proposed positive feedback calmodulin based based molecular interactions within a synapse for long term memory. There is also no account of Lorente de N and Hebb's reverberating activity loops or how a continuum of memory states could occur or how ring attractor networks may form. All of the important, but perhaps cheap criticisms against the biological plausibility of deep learning based frameworks also apply, which relates aspects such as the lack of spikes and spike time data, or the use of backpropagation to implement learning, which inherently relies on global state to work.

References

[1] James L McClelland, Bruce L McNaughton, and Randall C O'reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995.

[2] McRae K Hetherington. Catastrophic interference is eliminated in pretrained networks. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, pages 723–728, 1993.

[3] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

[4] Robert M French and Nick Chater. Using noise to compute error surfaces in connectionist networks: A novel means of reducing catastrophic forgetting. *Neural computation*, 14(7):1755–1769, 2002.

[5] Rupesh K Srivastava, Jonathan Masci, Sohrob Kazerounian, Faustino Gomez, and Jürgen Schmidhuber. Compete to compute. In *Advances in neural information processing systems*, pages 2310–2318, 2013.

[6] Chris A Kortge. Episodic memory in connectionist networks. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, pages 764–771. Erlbaum, 1990.

[7] Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

[8] Guang Yang, Feng Pan, and Wen-Biao Gan. Stably maintained dendritic spines are associated with lifelong memories. *Nature*, 462(7275):920, 2009.

[9] Stefano Fusi, Patrick J Drew, and Larry F Abbott. Cascade models of synaptically stored memories. *Neuron*, 45(4):599–611, 2005.

[10] Laurence Aitchison and Peter E Latham. Synaptic sampling: A connection between psp variability and uncertainty explains neurophysiological observations. *arXiv preprint arXiv:1505.04544*, 2015.

[11] Rishidev Chaudhuri and Ila Fiete. Computational principles of memory. *Nature neuroscience*, 19(3):394, 2016.