Overcoming catastrophic forgetting in neural networks

Manan Tomar : ED14B023

Abstract

This work (Kirkpatrick et al., 2017) proposes one way to overcome the forgetting observed in neural networks when learning over a sequence of tasks. Specifically, learning over a task B leads to loss of information on how to solve a previously learnt task A. This is tested on classification tasks on the MNIST data set as well as on a series of Atari games.

1. Approach

The authors propose an algorithm called Elastic Weight Consolidation (EWC) which resembles the synaptic consolidation observed in animals, allowing them to preserve knowledge of a previously learnt task when learning a new one. Since multiple parameter configurations can lead to similar performance in artificial neural networks, EWC aims at storing the performance for a previously learnt task A by constraining the parameters to remain in a low error region when learning for task B. Shifting to a probabilistic viewpoint, optimizing for a given task involves finding the most probable parameter values θ that describe the task data $\mathcal{D}$.

$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}|\theta) + \log p(\theta) - \log p(\mathcal{D}) \tag{1}$$

Writing the likelihood $\log p(\mathcal{D}|\theta)$ as a sum of two tasks A and B,

$$\log p(\theta|\mathcal{D}) = \log p(\mathcal{D}_B|\theta) + \log p(\theta|\mathcal{D}_A) - \log p(\mathcal{D}_B) \tag{2}$$

The above ensures that the posterior $\log p(\theta|\mathcal{D}_A)$ captures all information regarding task A and therefore, constraining this term will result in protecting the learnt information from task A. This is done by modelling this distribution as a Gaussian with mean as θ^*_A and the diagonal of the Fischer Information matrix F providing the diagonal precision (MacKay, 1992). This approximation results in the following :

$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i}) \tag{3}$$

where λ decides the realtive importance of task A to task B and i indexes each parameter. Using the Fischer matrix is easy to calculate as it can be computed using first order derivatives. This allows scaling this approach to larger models.

2. Experiments

The authors show results on a series of MNIST tasks. They show that naive SGD will exhibit catastrophic forgetting when shifting from one task to other. Using an L2 penalty results in not learning task B efficiently since all parameters are constrained. However, using EWC penalty results in preserving performance across previous tasks after learning on a new task by selectively reducing the plasticity of the parameters that are most important to task A.

A similar behavior is observed in the Atari domain when a DQN (Mnih et al., 2015) is learnt over 10 different games using EWC penalty. They use a task recognition model for receiving context to the current task. The network is given a particular game for a finite time period and is repeated randomly after a sequence of other games follows. A simple SGD implementation results in learning only over one game, whereas using EWC allows learning around 8-9 tasks at human level performance. However, the authors also note that they observe underestimating the parameter uncertainty when approximating using the Fischer Information matrix.

References

Kirkpatrick, James, Pascanu, Razvan, Rabinowitz, Neil, Veness, Joel, Desjardins, Guillaume, Rusu, Andrei A, Milan, Kieran, Quan, John, Ramalho, Tiago, Grabska-Barwinska, Agnieszka, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, pp. 201611835, 2017.

MacKay, David JC. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A, Veness, Joel, Bellemare, Marc G, Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K, Ostrovski, Georg, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.