Overcoming catastrophic forgetting in neural networks

presented by: Sarah Desouky
Reported by: Youmna Salah
May 1, 2017

1 GENERAL OVERVIEW

This presentation was about one of the drawbacks of Neural networks, which is forgetting. Neural networks depend souly on weight updates according to the error value calculated for each different input. It could be trained to learn a specific task for example classifying cars by training it on a given data set of different images of cars, yet if the same network that was trained to classify cars is tested on facial recognition it would fail miserably since it is not trained to do so and vice versa. Such a network has been known for it limitation to acquire more than one functionality and if it is trained on several functionalities it will only acquire the latest, each function overwrites the other. such a problem is known as catastrophic forgetting. In this presentation two papers were presented namely [KPR+16] and [FBB+17].

2 OVERCOMING CATASTROPHIC FORGETTING IN NNS

There were previous approaches to overcome catastrophic forgetting such as Multitask learning paradigm and System level consolidation. The Multitask learning paradigm have to have access to all the data for the tasks that the current network is trained to perform, which though it works, it is a drawback for such a technique. As for the System level consolidation, a network is trained on tasks in a sequential manner while maintaining previous training data and reusing them while training to prevent the network from forgetting, to do so a huge amount of memory is required, which is again a draw back.

In the first paper [KPR+16], a new technique namely the *Elastic weight consolidation,(EWC)* was developed, inspired by the way we as human beings learn several skills without having the skills overriding each other. They mapping such a biological technique by reducing the pace of the weight update when it comes to weights that were remarkable from previous tasks. So they tried to model that by slowing down training on weights that were considered important to previously learned tasks. The objective of such a technique is to minimize the loss function as well as the difference between the current weights and the previous ones. A so called *constraint* was used to avoid total change in the weights in the training process. The *constraint* of a network was controlled by the lambda Λ parameter, while the importance of features was determined by fisher's matrix F_i as in 2.1.

$$L(\theta) = L_B(\theta) + \Sigma_i \lambda/2 F_i(\theta_i - \theta^*_{A,i})^2 \tag{2.1}$$

Deep Q-network(DQN) with EMC was trained in a supervised manner and tested against other DQNs without EMC that were trained on each task individually and the results have shown that the latter outperformed the former due to the complete reliance on the fisher information matrix that might have mistaken important parameters for being unimportant.

The technique discussed in [KPR+16] was clearly explained, though this paper lacked a network diagram showing the structure of DQN with EMC and if they have the same structure as DQN, it was not mentioned explicitly.

3 PathNet: Evolution Channels Gradient Descent in Super Neural Networks

Again the second paper tackles the problem of catastrophic forgetting and multitasking. In this paper a new network architecture have been developed. Such a network architecture divides a network into a set of module, where each module is a NN in itself. The following three words were used throughout the paper:

- tournament selection
- Mutation
- Transfer learning

tournament selection was defined as the cycle of searching for the optimal path by mutating the modules throughout episodes till a certain accuracy is reached then the module values are fixed and the same is done to modules in the next layer. As for transfer learning, it is defined as the reuse the modules trained on one task as the base value to train on other tasks and that makes networks train faster without forgetting any of the previously trained tasks. A genotype was defined as matrix that tells which modules contribute to a pathway.genotypes compete to find the most efficient pathway. Such an architecture was tested on MNIST on a supervised

manner and the results have shown the network was able to learn in less time to achieve perfect score. It was also trained on Atari games using reinforcement learning, a speed up in the performance was noticed as well. So in a nutshell, Path-nets have solved the problem of catastrophic forgetting and also have reduced the time needed for learning.

The architecture discussed in such a paper is a bit similar to the network in network in the architecture, though the learning technique is different. A minor detail was missing and that is, the paper did not mention how was speedup in training measured? It might be the case that they were just using equipped PCs for training if not measured against network architectures with under the same environment.

REFERENCES

[FBB+17] Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A. Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *CoRR*, abs/1701.08734, 2017.

[KPR+16] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016.