Overcoming catastrophic forgetting in neural networks

James Kirkpatricka, Razvan Pascanua, Neil Rabinowitza, Joel Venessa, Guillaume Desjardinsa, Andrei A. Rusua, Kieran Milana, John Quana, Tiago Ramalhoa, Agnieszka Grabska-Barwinska a, Demis Hassabisa, Claudia Clopathb, Dharshan Kumarana, and Raia Hadsella

***Abstract*—In order to reduce catastrophic forgetting in a multi-task learning setup, the paper proposes a method in which the weight updates relevant to previous tasks are selectively constrained using elastic weight consolidation (EWC) to allow continual learning.**

I. Motivation for EWC

Task-specific synaptic consolidation : A proposed process for *continual learning* in the mammalian brain where the knowledge about previous tasks are encoded in the strengthened synapses, which are made less plastic over long time periods for avoiding catastrophic forgetting while acquiring knowledge about new skills.

II. Elastc Wight Consolidation

Motivated by *task-specific synaptic consolidation*, the authors propose a novel algorithm EWC which selectively slows down updates on certain weights depending on how relevant they are for previous tasks encountered as well as on the importance given to remembering the previous skills by constraining important parameters to stay close to their old values so that lower errors are experienced by previously learned tasks while learning new tasks. The parameters are optimized for their most probable values under the given data :

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

Consider 2 tasks A and B where A is learned first. Then while optimizing for B,

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

Hence the information about which parameters are important to task A must be encoded in the posterior distribution $p(\theta|\mathcal{D}_A)$. Since the true posterior probability is intractable, it is approximated as a Gaussian distribution with mean θ^*_A (parameters) and a diagonal precision is obtained from the diagonal of the Fisher information matrix ($\mathcal{F}$) which :

- is equivalent to second derivative of loss near a minimum
- can be computed using first-order derivatives (scalable to large models)
- is positive semi-definite

Hence, the EWC loss while learning task B is given by,

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

where $\mathcal{L}_B(\theta)$ is the loss for task B, subscript i denotes the parameters and λ controls the relative importance of the previous task. Hence EWC optimizes for task B without compromising much on task A by explicitly accounting for how important the weights are for task A. This can be extended to more number of tasks by either enforcing separate penalties for each task or summing up the penalties as a single penalty for all the previous tasks.

III. Supervised Learning Setup

Here, a fully connected multilayer neural network was trained on several supervised learning tasks (in sequence). Three different methods of training were evaluated :

1 Plain stochastic gradient descent
2 SGD + L2 regularisation on weights
3 Elastic Weight Consolidation

The analysis of results obtained showed that :

1 Performance on previous tasks degrades upon training for subsequent tasks
2 Little capacity spared for learning new tasks as all weights are constrained more or less equally
3 Retains performance on previous tasks (with modest errors) while having better capacity to learn new tasks.

Analysis of $\mathcal{F}$ showed that similar tasks depend on similar sets of weights whereas for more dissimilar tasks, the network allocates additional capacity.

IV. Reinforcement Learning Setup

Using DQN as the base network, their network comprised of a *task-recognition module*, and an *EWC penalty*. The task recognition is achieved through the *Forget Me Not algorithm* (the improvement in performance upon explicitly providing the agent with true task labels were modest) and the task context expressed as the latent variable of a Hidden Markov Model. $\mathcal{F}$ was computed at each task switch to incorporate the EWC penalty and the λ was decided through hyperparameter search. The network was trained using Double Q-learning algorithm. The network also had biases and per element multiplicative gains specific to each game. Here, although the agent learned multiple games in sequence while being much lesser prone to catastrophic forgetting, the scores on individual games were lesser than what was achieved using indivudual DQNs for each game. The authors suspect that this is due to under-estimation of parameter uncertainty due to approximation.

References

[1] James Kirkpatrick et al. “Overcoming catastrophic forgetting in neural networks”. In: *Proceedings of the national academy of sciences* (2017), p. 201611835.