The Report of the Assignment 1: Implementing and Reproducing Elastic Weight Consolidation

Seungmin Lee
Seoul National University
dltmdals14@snu.ac.kr

1. Introduction

Elastic Weight Consolidation (EWC) [2] is a pioneering work that tackles continual learning problems. Even though many researchers have already validated EWC in their papers, we implement and reproduce EWC and compare it with L2 regularization. We conduct two experiments using CIFAR100 [3] and MNIST [4]: (a) hyperparameter searching, (b) comparing the effectiveness of EWC, L2 regularization, and no regularization. In these experiments, we learn the following three lessons. First, both regularization can alleviate *catastrophic forgetting*. Second, the optimal hyperparameters can be highly different between datasets. For example, the optimal hyperparameter for CIFAR100 is 1.0. On the other hand, the optimal value is 100.0 for MNIST. Third, on simple datasets such as MNIST, EWC can be less effective than L2 regularization.

2. Methods

In this section, we summarize the methods that we implement for this assignment. Because our primary concerns are regularization-based continual learning methods, the loss functions have the same form as follows:

$$\mathcal{L}(\theta) = \mathcal{L}_{CE}(\theta) + \lambda \mathcal{L}_{reg}(\theta, \theta_T) \tag{1}$$

where θ and θ_T denote the current model weights and the previous weights after training on task T, repectively, and $lambda$ is the hyperparameter that balances the classification loss $\mathcal{L}_{CE}$ and the regularization loss $\mathcal{L}_{reg}$. The regularization loss term $\mathcal{L}_{reg}$ makes the difference between EWC, L2 regularization, and no regularization.

EWC alleviates catastrophic forgetting by preventing θ far away from the important weights of θ_T because θ_T encodes the knowledge for solving the previous tasks. EWC estimates the importance of each weight by calculating Fisher Information Matrix. Therefore, the regularization term of EWC [2] is

$$\mathcal{L}_{reg}(\theta, \theta_T) = \sum_i F_i(\theta_i - \theta_{T,i})^2 \tag{2}$$

where F denotes the Fisher information matrix, i is an index of each parameters. For efficiency, EWC adopts the diagonal approximation of the Fisher information matrix.

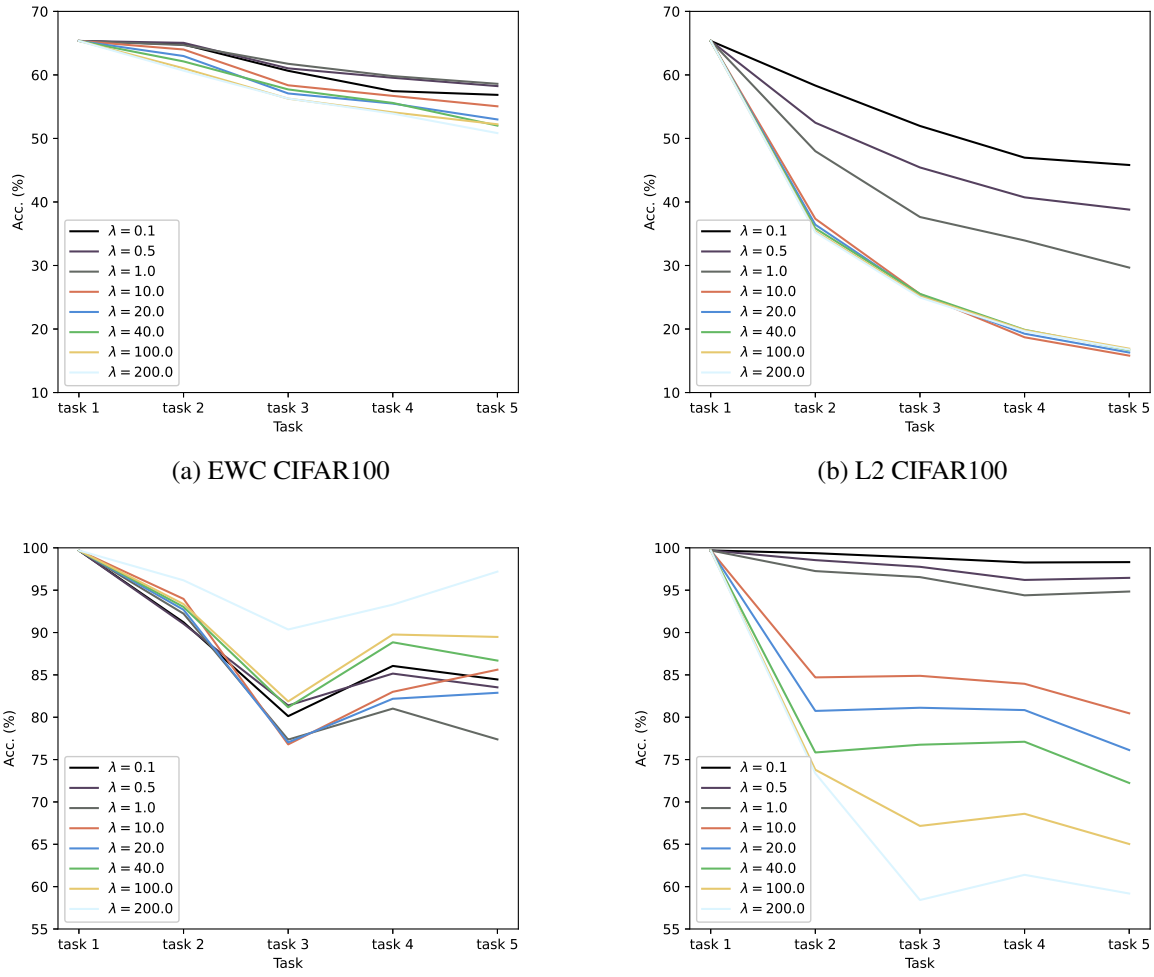
L2 regularization uses the following equation as its regularization term:

$$\mathcal{L}_{reg}(\theta, \theta_T) = \sum_i (\theta_i - \theta_{T,i})^2. \tag{3}$$

We can interpret that L2 regularization does not consider the importance of each weight. We use L2 regularization to validate whether the EWC's importance estimation is helpful or not. No regularization or vanilla training does not utilize L_{reg}.

3. Experiments

In this section, we describe the experiment settings and their results. First, we conduct experiments for finding the optimal value of λ used for balancing $\mathcal{L}_{CE}$ and $\mathcal{L}_{reg}$. In the second experiment, we compare the three methods described in the section 2. We follow the provided guideline for the experiments. We use Adam optimizer [1], 1e-3 as the learning rate, and batch size 256 for all the experiments. We use 20 and 60 epochs for MNIST [4] and CIFAR-100 [3], respectively.



(a) EWC CIFAR100

(b) L2 CIFAR100

(c) EWC MNIST

(d) L2 MNIST

Figure 1. The results of the hyperparameters searching experiments.

3.1. Hyperparameter Searching

We use $\{0.1, 0.5, 1.0, 10.0, 20.0, 40.0, 100.0, 200.0\}$ for the hyperparameters searching scope. Figure 1 shows the results of the experiments. The x-axis of each plot is the number of tasks learned so far, and the y-axis represents the average accuracy. We fix the random seed as 0 for replacing another randomness. As we can observe, the smallest λ shows the best average accuracy for L2 regularization. We can also observe that bigger λ has detrimental effects on learning new tasks. On the other hand, the optimal λ for EWC varies depending on datasets. For example, the optimal λ for CIFAR-100 is 1.0, but the optimal value for MNIST is 200.0. The large gap between the two values implies that the complexity of given tasks can affect the importance of the regularization term. Furthermore, we can observe that EWC is less sensitive to the hyperparameter.

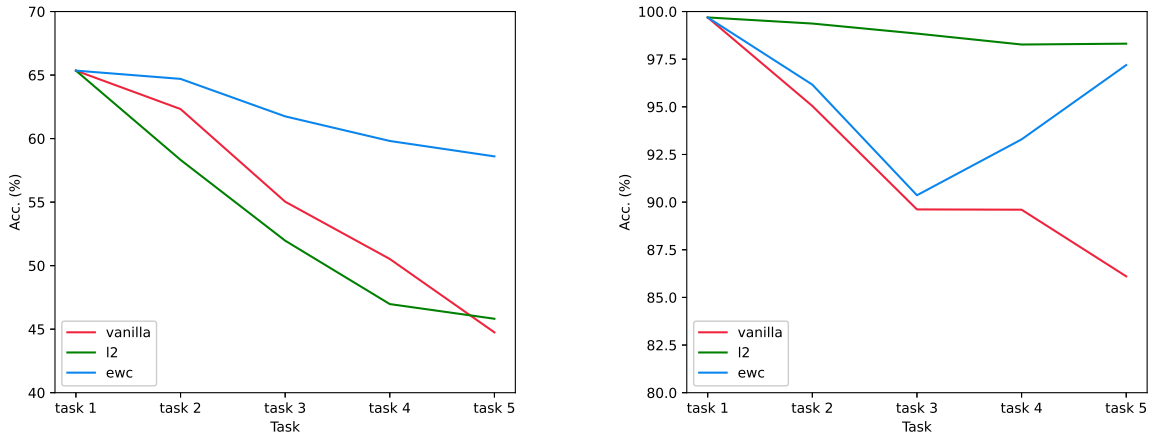
References

[1] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[2] James Kirkpatrick, Razvan Pascanu, Neil 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. In *PNAS*, 2017. 1

[3] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). 1

[4] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. 1



(a) Comparing on CIFAR100

(b) Comparing on MNIST

Figure 2. some captions