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 CIFAR-100 [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, when we test on five tasks, the optimal hyperparameter for CIFAR-100 is 1.0 while the optimal value for MNIST is 100.0. 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.

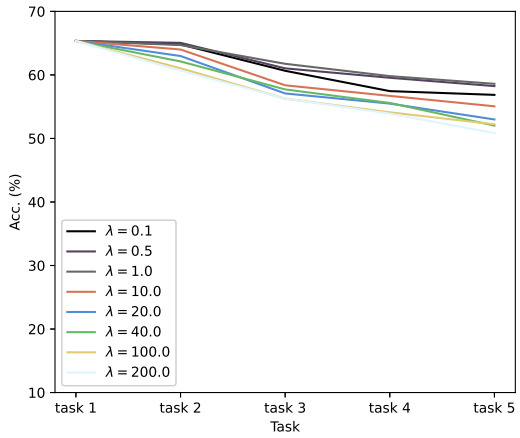
L2 regularization uses the following equation as its regularization term:

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

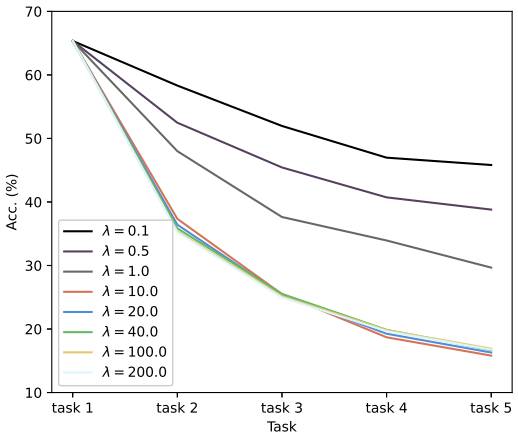
We can interpret that L2 regularization does not consider the importance of each weight. We use L2 regularization to validate whether the EWC's importance estimation is helpful or not. No regularization or vanilla training does not utilize L_{reg}.

3. Experiments

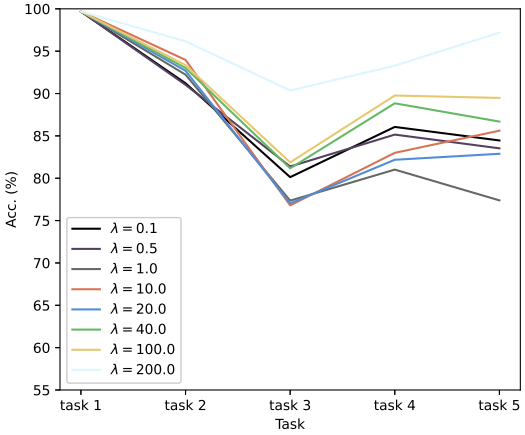
In this section, we describe the experiment settings and their results. First, we conduct experiments for finding the optimal value of λ used for balancing $\mathcal{L}_{CE}$ and $\mathcal{L}_{reg}$. In the second experiment, we compare the three methods described in the section 2. We follow the provided 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. We use five tasks for MNIST and twenty tasks for CIFAR-100.



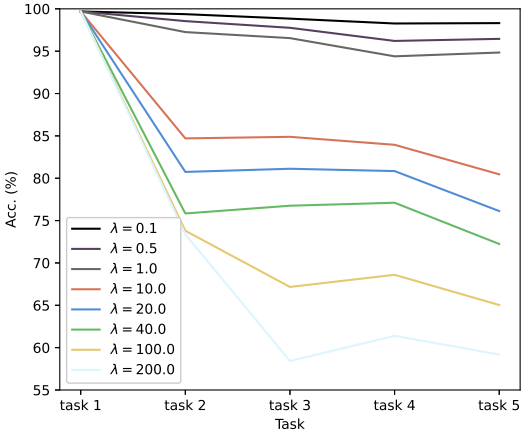
(a) EWC CIFAR-100



(b) L2 CIFAR-100



(c) EWC MNIST



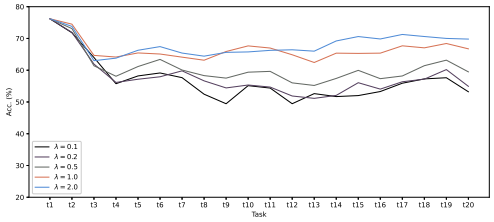
(d) L2 MNIST

Figure 1. The results of the hyperparameters searching experiments using five tasks. We conduct these experiments for CIFAR-100, too. The x-axis of each plot is the number of tasks learned so far, and the y-axis represents the average accuracy.

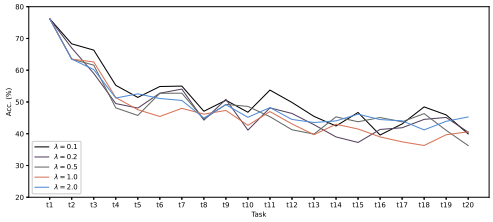
3.1. Hyperparameter Searching

In this experiment, we search the optimal λ for each regularization method. First, We use $\{0.1, 0.5, 1.0, 10.0, 20.0, 40.0, 100.0, 200.0\}$ for the hyperparameters searching scope when there are five tasks. Then, for CIFAR-100, we search the finer hyperparameters scope $\{0.1, 0.2, 0.5, 1.0, 2.0\}$. Figure 1 and Figure 2 show 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 in the five tasks experiment. 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. We further search the optimal hyperparameter using twenty tasks on the CIFAR-100 dataset. The results are shown in Figure 2. In this experiment, we find that 2.0 is optimal for both L2 and EWC. Therefore, we use these values for the following experiments.



(a) EWC CIFAR-100



(b) L2 CIFAR-100

Figure 2. The CIFAR-100 results of the hyperparameters searching experiments using twenty tasks. The x-axis of each plot is the number of tasks learned so far, and the y-axis represents the average accuracy.

3.2. Comparing Regularization Methods

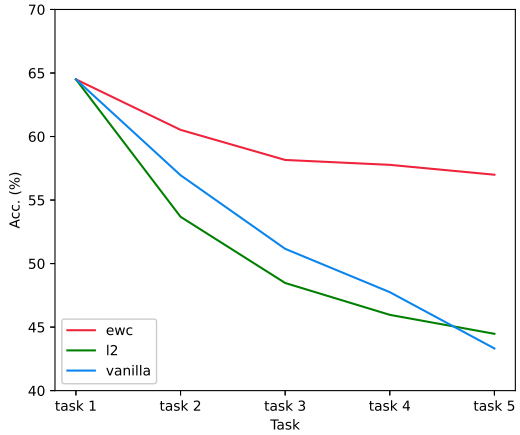
Using the optimal λ that we found in experiment 3.1, we compare the effectiveness of each regularization. As mentioned before, L2 regularization uses $\lambda = 0.1$ for both datasets, and EWC uses $\lambda = 1.0$ and $\lambda = 200.0$ for CIFAR-100 and MNIST, respectively. We vary the random seed from 0 to 6 and average the accuracies for removing randomness. Figure 3 and Figure 4 display the results of the experiments. As same as before, the x-axis represents the number of tasks learned so far, and the y-axis of each plot is the average accuracy.

For CIFAR-100 (Figure 3 (a)), EWC shows the best performance across all the stages with a high margin. Interestingly, L2 regularization shows slightly lower performance than vanilla learning until they reach the last task, which implies that L2 regularization is too hard for learning new tasks. In Figure 4, we observe the same results on the experimental results that used twenty tasks.

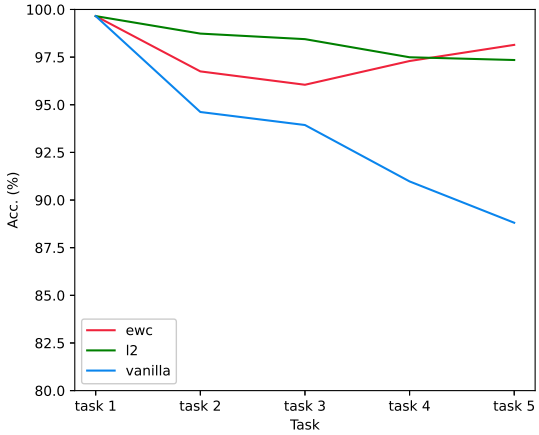
For MNIST (Figure 3 (b)), L2 regularization exhibits better performance than EWC except for the last stage. This observation implies that EWC can be less effective when using it on a simple dataset or problem. In this experiment, the average accuracy increases after task 3 while the accuracy consistently decreases on CIFAR-100. Vanilla training shows the worst performance as we can expect.

4. Conclusion

In this assignment, we implemented EWC, a pioneering work that prevents catastrophic forgetting in continual learning. Using MNIST and CIFAR-100 datasets, we conducted hyperparameter searching experiments for finding the optimal λ for each regularization. Using the found optimal λ, we conducted experiments for comparing the three methods (EWC, L2



(a) Comparing on CIFAR100



(b) Comparing on MNIST

Figure 3. The CIFAR-100 and MNIST results of the comparison experiments using five tasks. The x-axis and the y-axis represent the number of tasks learned so far and the average accuracy, respectively.

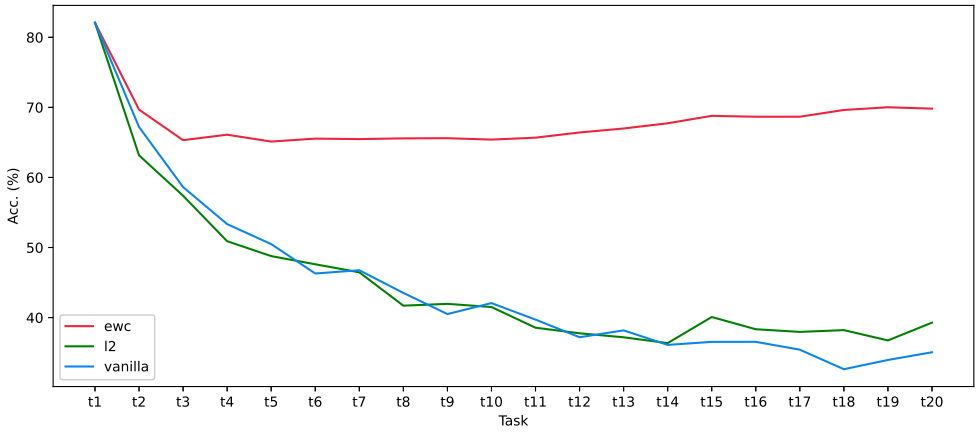


Figure 4. The CIFAR-100 results of the comparison experiments using twenty tasks. The x-axis and the y-axis represent the number of tasks learned so far and the average accuracy, respectively.

regularization, and vanilla training). We found that both L2 regularization and EWC can alleviate the catastrophic forgetting. However. L2 regularization was too hard, so it can hinder the model from learning new tasks. Moreover, we observed that the optimal hyperparameters for EWC vary depending on datasets. Finally, we observed that L2 regularization could be more effective than EWC on a simple dataset such as MNIST. These results are helpful to understand the actual effectiveness of EWC and other regularizations.

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