ECE-GY 7143 Advanced Machine Learning

Final Project Report

Explore the Concept of Catastrophic Forgetting & Implement Elastic Weight Consolidation

Name: Haoyu Wang ID: hw3256

1. Abstract

This project aimed to explore the concept of catastrophic forgetting in the field on continual learning and implement elastic weight consolidation (EWC) in BERT model to overcome this phenomenon. To test EWC efficiency, the BERT model is fine-tuned for the Emotion detection task, and then fine-tuned again to the sentiment analysis task but incorporates EWC in the training phase. The final test accuracy showed the model with EWC maintained high accuracy on the first task while training on the second task.

2. Background

Catastrophic forgetting is a phenomenon that occurs in the field of continual learning where a model trained on a set of tasks forgets how to perform previous tasks when it is trained on a new task. This occurs because the model's weights are updated during training to optimize performance on the new tasks, which can cause the model to overwrite the information learned during the previous tasks.

In other words, the model becomes so specialized on the new tasks that it "forgets" the information it learned during the previous tasks. This is a major challenge in the field of continual learning, as it limits the ability of models to learn continuously over time without forgetting previous knowledge.

To mitigate catastrophic forgetting, various methods have been proposed such as EWC [1], rehearsal-based approaches [2], and parameter isolation techniques [3]. These methods attempt to retain the previously learned knowledge while learning new tasks.

EWC was introduced by Kirkpatrick et al. in 2017[1]. It works by placing constraints on the network's weights during training to prevent the network from overwriting the information learned during previous tasks. It uses Bayesian inference to estimate the distribution of the weights of the network before and after training on a set of tasks. Then places constraints on the weights during subsequent training to prevent them from deviating too far from their original values, thus preserving the information learned during previous tasks. These constraints are implemented through a regularization term added to the loss function during the training.

One of the main advantages of EWC over other methods is that it is computationally efficient and easy to implement. It does not require any additional memory or computation during testing, which makes it suitable for real-world applications. Additionally, EWC can be used with any gradient-based optimizer, and it does not require any architectural modifications to the network.

Another advantage of EWC is that it can be used with a wide range of neural network architectures and datasets. It has been shown to work well on both shallow and deep networks, as well as on a variety of tasks such as image classification and reinforcement learning.

3. Method

During the implementation, I chose emotion detection as the first task and sentiment analysis as the second task for the BERT model. Before fitting the dataset with the model, I prepared and preprocessed the dataset in the tensor format, and then fine-tuned BERT model on the emotion detection dataset. The dataset was obtained from Kaggle [4], which contains sentences and emotion label for each sentence. One of the example in the training dataset is " i didnt feel humiliated;sadness". After the fitting, I stored the model's weights for later use.

To retain the knowledge learned from the first task, I estimated the importance of the model weights using Bayesian inference to determine the Fisher Information Matrix and stored the importance weights for each parameter. Next, the BERT model was fine-tuned on the sentiment analysis task. The dataset I used for sentiment analysis task was "imdb_reviews", which can be directly imported from TensorFlow Datasets. I trained the model with and without EWC and evaluated the model on both tasks at the end of the project.

4. Results

After training the model on two tasks and utilizing EWC regularization, I compared the test accuracy of the model and recorded them in the table below. We can see that the BERT model with EWC has relatively high accuracy for the first task after trained with the second task compared to the model without EWC. Due to the limited computing resources, the model only trained for one epoch with EWC regularization, which is the reason why the accuracy of the first task didn't achieve as high as expected. However, the test results reflected that EWC regularization effectively alleviated the problem of catastrophic forgetting in continual learning of the model, especially with more significant effects after multiple epochs of training.

BERT Model	Without EWC	With EWC
Emotion Detection 1st	93%	93%
Sentiment Analysis 2nd	93%	93%
Emotion Detection 3rd	14%	40%

Table 1

5. Future direction

Just like what Ke et al. in 2021 [5] argued, there are many approaches are proposed in recent years, but most of them only focused on overcoming catastrophic forgetting (CF) and no mechanism to encourage knowledge transfer (KT) (Ke, 2021). Recently, there are works

investigated in the forward KL, in other words, training on new tasks benefited from the knowledge gained on previous tasks. However, little attention is paid to the backward KL, which is meant to benefit knowledge of previous tasks based on what model is learning in the new tasks. We still have limited knowledge on this domain as it conflicts with the current views of many studies, which do not believe that learning new tasks can improve a model's prediction for previous tasks. However, in real life, such examples do happen to us, such as the calculus knowledge we learn in college gives us a more comprehensive and thorough understanding of multiplication and division when we were younger. So when models encounter new tasks that are similar to previous ones, they also update their knowledge base and achieve more accurate predictions for previous tasks.

6. Reference

[1] Kirkpatrick, J. N., Pascanu, R., Rabinowitz, N. C., Veness, J., Desjardins, G., Rusu, A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). *Overcoming catastrophic forgetting in neural networks*. Proceedings of the National Academy of Sciences of the United States of America, 114(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

[2] Yoon, J. (2021, June 2). *Online Coreset Selection for Rehearsal-based Continual Learning*. arXiv.org. https://arxiv.org/abs/2106.01085

[3] Zhao, Y. (2022, July 22). *Revisiting Parameter Reuse to Overcome Catastrophic Forgetting in Neural Networks*. arXiv.org. https://arxiv.org/abs/2207.11005

[4] Praveengovi. (2020). Classify Emotions in text with BERT. *Kaggle*. https://www.kaggle.com/code/praveengovi/classify-emotions-in-text-with-bert/notebook

[5] Ke, Z. (2021, December 5). *Achieving Forgetting Prevention and Knowledge Transfer in Continual Learning*. arXiv.org. https://arxiv.org/abs/2112.02706