Continual Learning with BERT for Questions and Answers Using Elastic Weight Consolidation

Di Xin dx489@nyu.edu
Jianan Gong jg6193@nyu.edu
Xingyu Wang xw913@nyu.edu
Zheyuan Zhang zz2498@nyu.edu

Abstract

The problem of catastrophic forgetting has become one of the main obstacles to achieving artificial general intelligence in the continual learning field. So far, many continual learning techniques have been developed to give models sequential learning capability. In this paper, we focus on a weight-based technique called Elastic Weight Consolidation (EWC) and apply it to the specific area of Question Answering in NLP based on BERT trying to experiment on whether the method could reduce catastrophic forgetting efficiently and further analyze its limit. We found that BERT in Question Answering suffers catastrophic forgetting but is not severe. And EWC can be an effective method to mitigate catastrophic forgetting with improvement in F1 score.

1 Introduction

Neural Network has achieved high performance on many different tasks. And continual learning, as a promising field in it, has been implemented on many NLP tasks in order to show how the model learns continually from a stream of data, built on what learned previously. However, humans can easily use the experience and knowledge they have learned before to quickly learn similar skills. For computers, such operations cannot be completed efficiently due to the problem of catastrophic forgetting (McCloskey and Cohen, 1989). It has become one of the main obstacles to achieving artificial general intelligence in continual learning.

"Catastrophic forgetting happens when a neural network loses the information learned in a previous task after training on subsequent tasks"(Serrà et al., 2018). It is a uniquely positioned task within neural network research fields, as well as natural language understanding. Recent work in avoiding catastrophic forgetting shows that a number of continual learning methods could solve these problems. In general, these methods either focus on data design, architecture design or weights of the neural network. They have all been proved effective yet the first two type of methods in many cases are inefficient regarding memory requirements and difficult for us to implement. In this paper, we focus on a weight-based technique called Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017), originally evaluated on MNIST dataset, and applied to the specific area of Question Answering in NLP based on BERT (Devlin et al., 2018). We hypothesize that using BERT as the baseline model and train it on SQuAD from MRQA datasets, and train it on NewsQA dataset. Then retest SQuAD and compare with the previous results. Additionally, we investigate Elastic Weights Consolidation (EWC) as the regularization terms on NewsQA tasks and compare with the previous results to see how effectively this technique can solve catastrophic forgetting. We discover that catastrophic forgetting is less evident among different QA tasks compared to text classification task and EWC is useful to alleviate catastrophic forgetting as we see performance boost on previous tasks compared to the situation without using EWC.

2 Related Work

There has been much existing work focusing on supervised learning (Parisi et al., 2019). In 2016, Li and Hoiem (2017) proposed an approach called Learning without Forgetting. It is an approach of changing weights to avoid catastrophic forgetting and implement it on Convolutional Neural Networks(CNN) and train it in Multi-tasks. By keeping the θ_0 which is the parameters for old tasks and optimize θ_s(shared parameters for all the tasks) and θ_n(parameters for new tasks) that are not going to change. It will make sure that the model still "remembers" its old parameters, for the sake of maintaining satisfactory performance on the pre-

vious tasks(Li and Hoiem, 2017). Lopez-Paz and Ranzato (2017) proposed a new method "Gradient Episodic Memory" to minimize catastrophic forgetting by efficiently using the subset of the observed examples(episodic memory) from the task and ensuring the loss of each previous task will not increase. However, it has intensive computational and memory costs. Chaudhry et al. (2018) modified GEM by averaging gradient episodic memory loss to alleviate the computational burden.

In biological field, it shows that learning different tasks would lead to "task related, branch specific Ca(2+) spikes" and cause "long-lasting potentiation of postsynaptic dendritic spines active at the time of spike generation"(Cichon and Gan, 2015), if the spines are erased by optical shrinkage, the ability learned is concomitantly forgotten (Hayashi-Takagi et al., 2015). In other words, there are specific spines critical for learning a specific related task. Inspired by this characteristic, EWC calculates the importance of the weights trained in old tasks using the Bayesian approach, viewing the importance as a posterior distribution, and controls their plasticity when learning new tasks hence important weights of a neural network are preserved and the model remembers how to solve old tasks. What's more, random matrix theory shows that when far away from the global extrema, the loss function of a neural network mainly has saddle points and the optimization process will get stuck in one of the many local minimal when approaching the global minimum (Choromanska et al., 2015) and this is actually good enough since you are already close to the global one. That is to say, making some weights relatively static in theory won't significantly hurt the network's ability to learn new tasks because the optimization process just ends in a different but still good result compared to setting no constraint on weights.

Some novel techniques mentioned above are tested in Question Answering and Text Classification to explore whether they can mitigate catastrophic forgetting by (de Masson d'Autume et al., 2019). Models are trained in a sequence of training datasets and evaluated on the union of the test datasets. They used F1 score to measure each model's performance including ENC-DEC (BERT-BASE) (F1 score: 53.1) as baseline and other models such as MbPA(F1 score:60.3), REPLAY(F1 score:57.9). Figure 1 is the performance of each model.

3 Methodology

3.1 Dataset

Dataset is collected from MRQA(Machine Reading for Question Answering) which provides specific tasks for the Question and Answering domain. We choose SQuAD v2.0 (Rajpurkar et al., 2018) and NewsQA(Trischler et al., 2016) as task A and task B respectively and collect both the training and testing dataset. SQuADv2.0 contains 150,000 questions and answers (or no answer) based on Wikipedia articles and NewsQA provides more than 100,000 question-answer pairs from over 10,000 news articles from CNN.

3.2 Language Model

We first preprocessed the dataset to convert examples to features. The parameters in BERT are frozen and we only fine tune the model-specific layer. The best hyperparameters are as follows: hidden_dropout prob of 0.1, hidden size of 768, max position embeddings of 512, num hidden layers of 12, max position embeddings of 512 and vocab size of 30522. We train our models for 2 epochs, using a batch size of 8.

3.3 Regularization Term

We introduced a regularization term called Elastic Weight Consolidation(EWC) which stands for regularizing parameters of a network trained on previous tasks by penalizing any change in them according to their importance. EWC will give us the optimized parameters of the overlap of all the tasks which we trained in our neural network. And penalized by the importance of the parameters which can be consolidated with the fisher information matrix.

3.4 Loss Function

To formulate the objective in a neural network, it is based on the posterior probability distribution, Following the Bayes Rule, given the previous tasks, and we will calculate the parameter θ by optimizing the following log-likelihood function:

$$\arg\max_{\boldsymbol{\theta}} \left\{ \ell(\boldsymbol{\theta}) = \log(p(\boldsymbol{\theta}|\boldsymbol{\Sigma})) \right\}$$

Σ is the combined dataset, including all the tasks A and B, and θ is the parameters that performed the best in this situation. So it will come up with the target objective function by the

Order	ENC-DEC	A-GEM	REPLAY	MBPA	MBPA$^{rand}_{++}$	MBPA++	MTL
i	57.7	56.1	60.1	60.8	60.0	**62.0**	67.6
ii	55.1	58.4	60.3	60.1	60.0	**62.4**	67.9
iii	41.6	52.4	58.8	58.9	58.8	**61.4**	67.9
iv	58.2	57.9	59.8	61.5	59.8	**62.4**	67.8
QA-avg.	53.1	56.2	57.9	60.3	59.7	**62.4**	67.8

Figure 1: Summary of results on question answering using F1 score. Adapted from (de Masson d'Autume et al., 2019)

tasks for task A and B.

$$\begin{aligned}&\log(p(\theta|\Sigma))\\ &= \log(p(B|A,\theta)) + \log(p(\theta|A)) - \log(p(B|A))\\ &= \log(p(B|\theta)) + \log(p(\theta|A)) - \log(p(B))\end{aligned}$$

$p(B|\theta)$ is the loss for current task B, $p(B)$ is the likelihood for B, and now posterior $p(\theta|A)$ for A becomes prior for B

In order to select the the importance of parameters, we used fisher information matrix I_A to evaluate the importance of parameters of task A, the simplified version can be seen as below:

$$\begin{aligned}\mathbb{I}_{\mathcal{A}} &= \mathbb{E}\left[-\frac{\partial^2(\log(p(\boldsymbol{\theta}|\mathcal{A})))}{\partial^2\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}^\star_{\mathcal{A}}}\right]\\ &= \mathbb{E}\left[\left(\left(\frac{\partial(\log(p(\boldsymbol{\theta}|\mathcal{A})))}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial(\log(p(\boldsymbol{\theta}|\mathcal{A})))}{\partial\boldsymbol{\theta}}\right)^\top\right)\Big|_{\boldsymbol{\theta}^\star_{\mathcal{A}}}\right]\end{aligned}$$

The final loss function would be the following

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

$\mathcal{L}_B(\theta)$ refers to the loss of task B without implementing EWC. It's worth pointing out that here F_i is not the same as the Fisher Information Matrix I_A. We do an approximation by only taking the value of the diagonal otherwise computing a matrix of size $N \times N$ (N is the sample size) will be very time consuming.

4 Experiments

The experiments are divided into two parts (Figure 2). For the baseline part, we first used BERT model to train on SQuAD v2.0 (Task A), and got F1 score after evaluation, and then we let the model learn over NewsQA (Task B), and evaluated back on the SQuAD (Task A) without further fine-tuning the model.

In the second part of the experiment, we implemented EWC over the BERT model so that the parameters could be regularized during the second learning phases and information that was learned from Task A could be preserved. Furthermore, we also evaluate NewsQA (Task B) to check if this kind of regularization undermines the performance of it since the parameters have been penalized. In order to achieve this, we used a pretrained BERTBASE mode as our guide, we also learned from an unofficial PyTorch implementation of DeepMind's paper (Kirkpatrick et al., 2017).

In the experiment, we used default BERT hyperparameters from pytorch_pretrained_bert, and we used λ=40 as EWC regularization parameter. During the training, We set the learning rate to 3e-5, which is the suggested learning rate for using BERT, according to preliminary experiments. We used a training batch size of 8 and used 6 GPU and 64GB of memory. We chose a relatively small batch size to prevent cuda from being overloaded and out of memory. And we used 2 epochs because in general cases this will give a reasonably acceptable score.

5 Results

Results for the experiments are listed in Table 1. The baseline shows without any regularization, NewsQA(Task B) gets F1 score 61.72. And based on NewsQA's parameters, SQuAD(Task A)'s F1 score drops from 84.66 to 73.62 which demonstrates the catastrophic forgetting really exists in the QA domain. Then we implemented EWC on our model, the performance of NewQA reduced by 5.48% with F1 score 58.34. This slight drop is resulted from taking the important parameters in SQuAD into penalty. SQuAD' s performance increases by 3.0% with F1 score 75.85. This indicates that EWC has a slight regularization effect. It alleviates catastrophic forgetting to some extent,

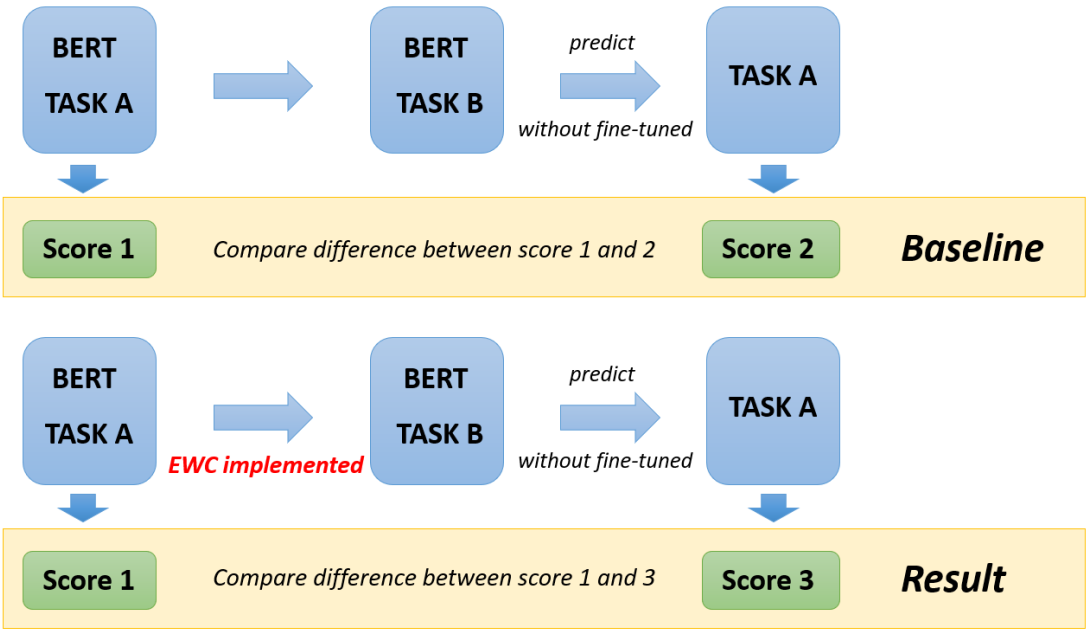

Figure 2: Expriment process

F1 Score	SQuAD	NewsQA	SQuAD (NewsQA's parameters)
Baseline	84.66	61.72	73.62
EWC Implementation	84.66	58.34	75.85

Table 1: Evaluation results for baseline and EWC implemented models

but the improvement is not significant.

In the meanwhile, we compare the severity of catastrophic forgetting between text classification and QA. BERT performs poorly on text classification when a new dataset is added with accuracy near 0. However, the baseline model performs much better in QA with a decrease by 13% which still has an F1 score over 70. This indicates BERT itself suffers less catastrophic forgetting in QA domain.

6 Conclusion and Discussion

In this research, we propose a continual learning approach EWC to solve the catastrophic forgetting problem in QA domain. Our research highlights that EWC can reduce catastrophic forgetting. Our research also shows BERT suffers less catastrophic forgetting in QA problems compared with text classification problems. Future work includes adding multiple tasks with different sequences to analyze the influence of sequence of tasks, and using different regularization methods to compare the performance of reducing catastrophic forgetting with EWC.

Collaboration Statement

Proposal: All
Partial draft: All
Literature Review: All
Baseline Model: Jianan & Zheyuan
EWC Understanding: Di & Xingyu
Code Implementation: Di & jianan & Zheyuan
Presentation: Di & Jianan & Zheyuan
Final Report Drafting: All
Final Report LaTeX Writing: Xingyu

Github Link

https://github.com/JasonZhangzy1757/NLU_Final_Project

References

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2018. Effi-

cient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. 2015. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204.

Joseph Cichon and Wen-Biao Gan. 2015. Branch-specific dendritic ca 2+ spikes cause persistent synaptic plasticity. *Nature*, 520(7546):180–185.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Akiko Hayashi-Takagi, Sho Yagishita, Mayumi Nakamura, Fukutoshi Shirai, Yi I Wu, Amanda L Loshbaugh, Brian Kuhlman, Klaus M Hahn, and Haruo Kasai. 2015. Labelling and optical erasure of synaptic memory traces in the motor cortex. *Nature*, 525(7569):333–338.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pages 6467–6476.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*, pages 13122–13131.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Joan Serrà, Didac Suris, Marius Miron, and Alexandros Karatzoglou. 2018. Overcoming catastrophic forgetting with hard attention to the task. *arXiv preprint arXiv:1801.01423*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.