cBERT: A Meta-Learning Approach to Contextual Representations for Continual Learning

William Huang
Courant Institute of
Mathematical Sciences
wh629@nyu.edu

Anhthy Ngo
Center for
Data Science
an3056@nyu.edu

Christine Shen
Center for
Data Science
zs1534@nyu.edu

Abha Sahay
Center for
Data Science
as13492@nyu.edu

Abstract

Transformer based pre-trained models such as BERT have produced state-of-the-art performance on many natural language understanding tasks. However, these models fail to learn in a continual learning setting, suffering from catastrophic forgetting. Meta-learning approaches have shown success in learning inductive biases robust against catastrophic forgetting. We hypothesize that further pre-training BERT with Online aware Meta-Learning will help learn contextual representations that are more suitable for continual learning. We refer to this further pre-trained model as continual BERT (cBERT). We compare cBERT and BERT's propensity for catastrophic forgetting by evaluating them on a continual learning question-answering curriculum of SQuAD and TriviaQA. We find that our model fails to learn our desired inductive bias with this meta-learning curriculum of question answering tasks.

1 Introduction

Continual learning is the process of constant development of complex tasks without end (Ring, 1995). For natural language processing (NLP), we define this as a sequence of tasks $\mathcal{T} = (T_1, T_2, \ldots, T_t, \ldots)$ that models should learn successively (Hu et al., 2019). Key attributes of continual learning methods should be their ability to perform reasonably well on previously learned tasks while reusing this information to quickly adapt to new tasks (Schwarz et al., 2018; Yogatama et al., 2019). We refer to these key attributes as general linguistic intelligence.

Large pre-trained Transformer (Vaswani et al., 2017) based models like BERT (Devlin et al., 2018) have achieved state-of-the-art (SOTA) performance on a wide array of natural language understanding (NLU) tasks. Ideally, successful NLP models can achieve human-like intelligence by learning from a continual stream of tasks. However, Yogatama et al. (2019) find that these models suffer from catastrophic forgetting and fail to achieve general linguistic intelligence. Javed and White (2019) have found success in using meta-learning to find inductive biases robust against catastrophic forgetting. We test this meta-learning technique, Online aware Meta-Learning (OML) (Javed and White, 2019), to further pre-train BERT to learn contextual representations that are more suitable for continual learning. We refer to this model as continual BERT (cBERT). For our meta-learning, we use other question answering tasks to learn cBERT weights. We find this approach fails to learn our desired inductive bias.

2 Related Work

Continual Learning for Large Pre-Trained Models Pre-trained models such as BERT (Devlin et al., 2018) use transfer learning to achieve super human-level performance on many NLU benchmarks. BERT is pre-trained using both masked language modeling and next sentence prediction objectives to learn contextual embeddings. Devlin et al. (2018) show that these embeddings only require small classification heads and limited fine-tuning steps to produce SOTA performance.

While these transfer learning methods show immense promise, Yogatama et al. (2019) find that BERT still lacks general linguistic intelligence. They show that the weights in BERT shift over the course of a continual learning curriculum, causing catastrophic forgetting. When training on a random mix of tasks from the curriculum, they find that their model is able to better retain information. However, this approach requires all tasks to be present at the beginning of training or to redundantly retrain all tasks after seeing a new task.

Prior work related to sequential and multi-task learning with BERT-like models show other NLU tasks help performance on downstream tasks. Liu et al. (2019) use multi-task learning on GLUE [1] to achieve better performance than single task fine-tuning. Phang et al. (2018) find similar success by intermediately training BERT on supervised tasks to boost their models' GLUE performance. Stickland and Murray (2019) use shared weights among related NLU tasks through multi-task training to achieve comparable GLUE performance with 7 times fewer weights. This begs the question can these models similarly benefit by learning NLU tasks in a continual manner?

Catastrophic Forgetting Catastrophic forgetting has been researched more broadly for neural networks with many of these approaches (Javed and White, 2019; Metz et al., 2019; Hu et al., 2019; Kirkpatrick et al., 2016) relying on meta-learning techniques to learn parameters robust against forgetting. Javed and White (2019) use an architecture with a representation learning network (RLN) and a task specific prediction learning network (PLN) along with their OML algorithm to learn robust RLN weights. OML takes into account the effects of continual learning through its completely online meta-learning updates. This includes catastrophic forgetting. Fine-tuning BERT readily fits the RLN and PLN framework. However, Javed and White (2019) freeze their RLN weights and learn a PLN network during continual learning.

Meta-Learning in NLP Meta-learning techniques have also shown promise in natural language processing and understanding. Wang et al. (2020) use gradient based meta-learning with auxiliary tasks to improve lexical relation classification. Dou and Yu (2019) use meta-learning for quick adaptation to new training data using Model Agnostic Meta-Learning (MAML). Through training their model on data similar to their final task, Dou and Yu (2019) are able to learn an inductive bias that speeds up their final fine-tuning.

While these methods exhibit the success of meta-learning approaches in NLP, they do not specifically focus on mitigating catastrophic forgetting. In our work, we adopt the OML algorithm from Javed and White (2019) to learn contextual embeddings appropriate for continual learning. We then test our model, cBERT, on the continual learning curriculum introduced by Yogatama et al. (2019). To stay true to Javed and White (2019)'s approach, we test BERT and cBERT weights with frozen embeddings.

3 Experiments

3.1 Methodology

We adopt the approach from Yogatama et al. of training BERT on a continual learning curriculum of SQuAD (Rajpurkar et al., 2016) → TriviaQA (Joshi et al., 2017). During the fine-tuning of TriviaQA, we perform zero-shot evaluations on SQuAD with the new BERT weights to track the model's catastrophic forgetting. This is depicted in the yellow sequence of Figure 1.

Subsequently, we use OML to further pre-train the original BERT weights to learn representations that are robust to catastrophic forgetting. We refer to this new pre-trained model as cBERT. In our meta-learning process, we train our model using several question-answering (QA) tasks in an online fashion. For our OML procedure, we perform meta-updates to the PLN with respect to the task specific objective. We then update the weights of the RLN, BERT, with respect to the OML objective (Javed and White, 2019):

$$\sum_{\mathcal{T}_i \sim p(\mathcal{T})} \sum_{S_k^j \sim p(S_k|\mathcal{T}_i)} \left[\mathcal{L}_{CLP_i} \left(U(W, \theta, S_k^j) \right) \right] \quad (1)$$

$\mathcal{T}_i$ is a task sampled from $\mathcal{T}$, S_k^j is a sampled stream of observations with length k, W is the weights of the PLN, and θ is the weights of the RLN. $U(W, \theta, S_k^j)$ represents meta-update function during the inner loop of the OML algorithm. Once we learn cBERT, we train the model on the same continual learning curriculum as BERT. This is described in the green sequence of Figure 1. We compare the two results, to evaluate how our meta-learning procedure affects catastrophic forgetting in BERT.

3.2 Data and Tools

We use the data from MRQA 2019[2] training tasks because of their standard formatting. Shared task data sets from MRQA are deployed with the goal of building robust NLP systems that generalize beyond seen training instances. We train cBERT using NewsQA (Trischler et al., 2016),

[1] https://gluebenchmark.com/

[2] https://mrqa.github.io/shared

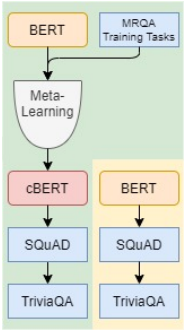


Figure 1: We compare two contextual embeddings, BERT and cBERT, on a continual learning curriculum of SQuAD → TriviaQA. To learn cBERT, we use OML to train BERT on a set of QA tasks.

HotpotQA (Yang et al., 2018), and NaturalQuestions (Kwiatkowski et al., 2019) during our meta-learning training. We then use SQuAD and TriviaQA data sets for our continual learning experiment. Part of the data standardization in MRQA ensures that answers to each question appear as a span of tokens in their contexts, as in SQuAD v1.1.

Our code is heavily adapted from Huggingface's `run_squad` example from the Transformers library (Wolf et al., 2019). For data loading, we reformatted the MRQA datasets to match the original SQuAD v1.1 data[3] and set all text to lower case. Despite this additional processing, Huggingface's data processor still ignores a portion of the training data. This corresponds to less than 1% for our continual learning tasks marked by (C) in Table 1. Our meta-learning tasks, marked by (M), corresponds to a larger portion of skipped data. However, we note that our meta-learning objective is not to train for high performance, but rather to learn an inductive bias robust against catastrophic forgetting. All validation data are kept.

	Training			Validation
Task	Total	Skipped		Total
(C) SQuAD	86,588	471	(0.5%)	10,507
(C) TriviaQA	61,688	560	(0.9%)	7,785
(M) NewsQA	74,160	11	(< 0.1%)	4,212
(M) HotpotQA	72,928	3,983	(5.4%)	5,904
(M) Natural Questions	104,071	3	(< 0.1%)	12,836

Table 1: Due to data formatting discrepancies, our data processing skips a portion of training samples.

[3] https://github.com/rajpurkar/SQuAD-explorer/tree/master/dataset

4 Results

4.1 BERT for Continual Learning

We first evaluate BERT on the continual learning curriculum of SQuAD → TriviaQA. Following Javed and White (2019)'s procedure, we freeze BERT weights and only train the layer used for QA. We first train the model on SQuAD. During this training, we train with a batch size of 24 and learning rate of 0.1 for 10,000 update iterations. We then used our best model and train on TriviaQA. Both tasks use Adam optimization (Kingma and Ba, 2015) with (0.9, 0.999) β's and linear learning rate decay over the number of training steps.

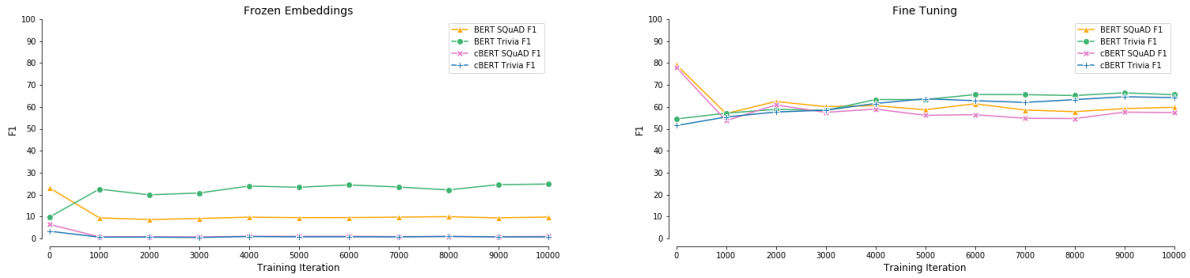
During the training on TriviaQA, we evaluate our model on SQuAD to measure BERT's catastrophic forgetting and report F1 scores as depicted in Figure 2a. Given the size of the QA layer, we find it difficult to train BERT on SQuAD and TriviaQA reaching top F1 scores of 23.1% and 24.9%, respectively (Table 2). Over 10,000 iterations our model improved on TriviaQA by about 15.1%. Meanwhile, performance on SQuAD declined dramatically by about 14.5%. These differences represent the margin between best to worst score on a single trajectory. This seems consistent with the findings from Yogatama et al. (2019), which report larger rates of forgetting over 100,000 iterations of fine-tuning.

4.2 cBERT

We use Javed and White (2019)'s implementation[4] of OML for meta-learning. For each meta-learning update, the algorithm first reinitializes the PLN. It then randomly samples 15 examples from 2 QA tasks and learns each task sequentially online with full stochastic gradient descent. These inner updates use a learning rate of 3e-3 and only change PLN weights. After 30 updates, we use another random sample of 30 examples from the same 2 tasks to evaluate our OML loss (Equation 1). Finally, we perform a single meta-update to the BERT embeddings using ADAM (Kingma and Ba, 2015) with learning rate 3e-5 and β's of (0.9, 0.999). We perform roughly 500 meta-updates.

Through this procedure, we fail to learn our desired inductive bias and instead learn weights that hinder learning on both SQuAD and TriviaQA. Compared to BERT, cBERT achieves no-

[4] https://github.com/khurramjaved96/mrcl



(a) Frozen Embeddings

(b) Fine-tuning

Figure 2: We train BERT on TriviaQA for 10,000 iterations and evaluate on SQuAD every 1,000 iterations to measure catastrophic forgetting. BERT exhibits catastrophic forgetting when training with both frozen embeddings and fine-tuning. cBERT fails to learn with frozen embeddings, but performs similarly to BERT with fine-tuning.

Model	Training	Task	Best F1	Worst F1	Change
BERT	Freeze	SQuAD	23.1%	8.7%	-14.5%
BERT	Freeze	TriviaQA	24.9%	9.8%	+15.1%
cBERT	Freeze	SQuAD	6.5%	0.8%	-5.7%
cBERT	Freeze	TriviaQA	3.4%	0.5%	-2.9%
BERT	Fine-tune	SQuAD	79.2%	56.9%	-22.3%
BERT	Fine-tune	TriviaQA	66.4%	54.5%	+11.9%
cBERT	Fine-tune	SQuAD	78.0%	53.7%	-24.3%
cBERT	Fine-tune	TriviaQA	64.6%	51.5%	+13.1%

Table 2: cBERT hinders training with frozen embeddings, but still performs similarly with fine-tuning while learning TriviaQA.

tably lower performance on both tasks (Figure 2b). As cBERT trains on TriviaQA, the model actually performs worse as the training loss decreases suggesting overfitting on the training data.

4.3 Effect of OML on Fine-Tuning

To investigate the impact of OML on BERT weights, we train both BERT and cBERT on the same continual learning curriculum using fine-tuning. We use a similar training procedure as the one in Section 4.1. We use a learning rate of 3e-5 to train on SQuAD and 5e-5 to train on TriviaQA. We find little impact to the embedding weights, with BERT and cBERT performing comparably on both tasks and catastrophically forgets at similar rates as shown in Figure 2b and Table 2.

5 Discussion

While BERT has accomplished SOTA performance on many NLU tasks, it still falls short of general linguistic intelligence and suffers from catastrophic forgetting in a continual learning setting (Yogatama et al., 2019). Learning tasks through a random sequence helps BERT retain information at the cost of knowing all tasks at the beginning of training. Training BERT-like models through multi-task and sequential learning on other NLU tasks seem to benefit specific downstream tasks (Liu et al., 2019; Phang et al., 2018; Stickland and Murray, 2019). Ideally, models like BERT can similarly benefit in a continual setting and learn new tasks as they are presented without catastrophically forgetting.

We follow Javed and White (2019)'s approach of using meta-learning with OML to learn an inductive bias robust against catastrophic forgetting. Using a sample of QA tasks, we train BERT to learn cBERT weights and evaluate both models on a continual learning curriculum of SQuAD $\rightarrow$ TriviaQA. We find that this method fails to learn the desired inductive bias and prevents cBERT from learning either QA task when training with frozen embeddings. However, when comparing both models through fine-tuning we see they achieve similar results. We release our code for this paper at `https://github.com/wh629/c-bert`.

Training BERT with frozen embeddings yields much lower performance on both continual learning QA tasks than training with fine-tuning (Table 2). This may be due to the simple QA linear layer used to find answer spans. Semnani et al. (2019) use more complex task specific layers to train BERT on SQuAD with frozen embeddings and achieve performance comparable to fine-tuning BERT parameters. Future work may incorporate these layers to better fit Javed and White (2019)'s OML framework. Additionally, altering the OML training procedure to update RLN weights during meta-training would better replicate fine-tuning and help learn an inductive bias more appropriate for BERT.

Collaboration Statement: William Huang contributed to writing the program and conducting experiment trainings. Anhthy Ngo contributed to the implementation of the code, running experiments, and literature review on continual learning. Christine Shen contributed to the implementation of the code, running experiments, and literature review on meta-learning. Abha Sahay contributed to conducting experiment trainings and literature review. All members contributed equally to the writing of the paper and presentation slides.

References

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

Zi-Yi Dou and Keyi Yu. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks.

Wenpeng Hu, Zhou Lin, Bing Liu, Chongyang Tao, Zhengwei, Dongyan Zhao, Jinwen Ma1, and Rui Yan. 2019. Overcoming catastrophic forgetting for continual learning via model adaptation.

Khurram Javed and Martha White. 2019. Meta-learning representations for continual learning.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

D. P. Kingma and L. J. Ba. 2015. Adam: A method for stochastic optimization.

James Kirkpatrick, Razvan Pascanu, Neil C. 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. 2016. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding.

Luke Metz, Niru Maheswaranathan, Brian Cheung, and Jascha Sohl-Dickstein. 2019. Meta-learning update rules for unsupervised representation learning.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *CoRR*, abs/1811.01088.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Mark Bishop Ring. 1995. Continual learning in reinforcement environments.

Jonathan Schwarz, Jelena Luketina, Wojciech M. Czarnecki, Razvan Pascanu Agnieszka Grabska-Barwinska, Yee Whye Teh, and Raia Hadsell. 2018. Progress compress: A scalable framework for continual learning.

Sina J. Semnani, Kaushik Ram Sadagopan, and Fatma Tlili. 2019. Bert-a: Fine-tuning bert with adapters and data augmentation.

Asa Cooper Stickland and Iain Murray. 2019. BERT and pals: Projected attention layers for efficient adaptation in multi-task learning. *CoRR*, abs/1902.02671.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Chengyu Wang, Minghui Qiu, Jun Huang, and Xiaofeng He. 2020. Investigating meta-learning algorithms for low-resource natural language understanding tasks.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Dani Yogatama, Cyprien Masson D Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence.