Meta-Learning for Natural Language Understanding under Continuous Learning Framework

Yong Fan
New York University
yf869@nyu.edu

Duo Jiang
New York University
dj1057@nyu.edu

Shiqing Li
New York University
sl7085@nyu.edu

Jiacheng Wang
New York University
jw5728@nyu.edu

Abstract

Neural network has been recognized with its accomplishments on tackling various natural language understanding (NLU) tasks. Methods have been developed to train a robust model to handle multiple tasks to gain a general representation of text. In this paper, we implement the model-agnostic meta-learning (MAML) and Online aware Meta-learning (OML) meta-objective under the continual framework for NLU tasks proposed by Javed and White(2019). We validate our methods on the GLUE benchmark(Wang et al., 2019) and [results to be completed].

1 Introduction

One ultimate goal of language modelling is to construct a model like human, to grasp general, flexible and robust meaning in language. From the past, NLU models have been building from specific task on given data domain but fail when dealing with out-of-domain data or performing on a new task. To combat this issue, several research areas in transfer learning including domain adaptation, cross lingual learning, multi-task learning and sequential transfer learning have been developed. However, transfer learning tends to favor high-resources tasks if not trained carefully, and it is also computationally expensive (Dou et al., 2019). Meta learning algorithm is appealing because it aims to introduce an agent that controls the model to learn in general knowledge that is useful in multiple tasks, or "learning to learn".

Specifically, model-agnostic meta learning (MAML) is an optimization method of meta learning that directly optimized the model by constructing an useful initial representation that could be efficiently trained to perform well on various tasks (Finn et al., 2017). However, there is still a potential problem of catastrophic forgetting, where a model trained with new tasks would start to perform worse on previous tasks. Continual learning is a framework that tries to alleviate catastrophic forgetting in a stream of data by incorporating it into its objectives. There are two objectives of designing a continual learning architecture, accelerate future learning where it exploits existing knowledge of a task quickly together with general knowledge from previous tasks to learn prediction on new samples; and to reduce interference in updates that overwrites old knowledge (Javed and White, 2019).

In this paper, we utilize algorithm derived from Jave and White (2019) which combines meta-learning approach with continuous learning framework, and our objective is to extend this framework in NLP field, specifically test on NLU tasks. By taking advantage of this model agnostic approach, meta-learning under continuous learning should be applicable on any exiting language model that is optimized by gradient descent. We hope to bring new research direction in NLP fields specifically focusing on such methods. The implementation of our code can be found at https://github.com/lexili24/NLUProject.

2 Background

To obtain a model that is robust enough to learn a general knowledge across various tasks, several algorithms have been introduced in Natural Language Understanding tasks, two specifically are meta learning and continual learning. Meta learning is a framework that enables model learning to excel in multiple domains with minimal resource and robust to training. Continual learning poses the potential problem of catastrophic forgetting. Plenty of research have been focused in these two areas and some efforts have succeeded in combining these two goals. One major contribution is achieve by the functionality in (Javed and White, 2019), a framework that introduces MAML-rep and OML objectives under the framework of con-

tinual learning. The rest of this section is dedicated to examine the implementation of methods solely or combined, in natural language and other fields, which leads us to develop our framework tackling Natural Language Understanding tasks.

2.1 Meta-Learning

There has been success in implementing MAML in NLU tasks (Dou et al., 2019). In their work, they explored the model-agnostic meta-learning algorithm (MAML) and its variants for low-resource NLU tasks and obtained impressive results on the GLUE benchmark(Wang et al., 2019). This proves that MAML can be applied to NLU tasks, and achieve comparable results on complex architectures like BERT and MT-DNN. However, this method does not address the potential problem of catastrophic forgetting. Another way of summarizing our job in a sentence is that we try to put their work in the context of continual learning setting.

In addition, meta learning is proved to excel on other natural language domains. Mi et al(2019) has shown promising results of incorporating MAML in natural language generation (NLG), a critical component of task-oriented dialogue system. NLG models, like many NLU tasks, are heavily affected by the domain they are trained on and are data-intensive. However, high annotation cost forces model to generalize well with low-resource data. Therefore, authors (2019) approach to generalize a NLG model with MAML to train on the optimization procedure and derive a meaningful initialization which could serve to adapt new low-resource NLG scenarios efficiently. In comparison, meta learning approach outperformed multi-task approach with higher BLEU score and lower error percentage. This is an indicator that given the constraints of current language tasks of low resource data, meta-learning could be beneficial to boost up model performance when scarce data source is scarce.

2.2 Continual Learning

Continual learning implemented in NLU tasks on top of transfer learning presented by Yogatama (2019) did not show generalization of the model. Yogatama et al followed the continual learning setup to train a new task on best SQuAD-trained BERT and ELMo model, and both architectures show catastrophic forgetting after TriviaQA or MNLI is trained, which degrades model performance on SQuAD dataset. Their work is an attempt to derive a generative language model, however they were trained on limited tasks that does not cover enough linguistic tasks that model would not be able to generalize. However their work provides a solid ground of continual learning in language modelling.

An implementation of meta-learning under continual framework is proposed in reinforcement learning by Alshedivat et al(2017) . In their paper, MAML is proved to be a complementary solution adding onto continuous adaption in reinforcement learning (RL) fields. Al-Shedivat et al(2017) considered nonstationary environments as sequences of stationary tasks for RL agents, which transferred nonstationary environment to learning-to-learning tasks. They developed a gradient-based meta-learning algorithm for quick adaption to continuously changing environment. They found that meta-learning is capable of adapting far more efficiently than baseline models in the few-shot regime. Although the implementation is outside the domain of Natural Language Processing, it is worth-noting that experts from different domains have implemented this method and sheds lights on authors to implement in NLU tasks.

To sum up, These researches show great success of both meta-learning and continual learning in several fields of study. In the meta learning ingredients, MAML have been applied on NLP tasks (Dou et al., 2019) (Mi et al., 2019) while not in a continual learning setting. In reinforcement learning, it can solve non-stationary environments (Al-Shedivat et al., 2017). Continual learning is also proven to enhance model performances on semantic textual similarity tasks (Liu et al., 2019). In next section, we rely on the work done by Javed and White (2019) and reiterate reasons that proved MAML to be useful for reaching the objective of continual learning (Javed and White, 2019).

3 Problem and Method

3.1 Problem Formation

As meta-learning presents significant outperforming results in solving the problem of sample inefficiency under domains/tasks from variety settings, Khurram et at(2019) proposed to implement meta-learning methodologies aiming at reaching continual learning objective. The representation learnt from existing knowledge by meta learning, accelerates the model to fit into new tasks

quickly. Moreover, the separation of neural network schemes approach the general prediction accuracy by only updating sparse parameters in the basic representation layers, so that the results are robust to forgetting in online updating for a series of continual tasks. Also, the learnt representations are complementary to other state-of-art continual learning strategies including knowledge retention methods, rehearsal methods, etc. Overall, the sparse representations demonstrates the efficiency on catastrophic interference.

Our model architecture strictly follows the architecture proposed in (Javed and White, 2019), where both MAML-Rep and OML objectives are tested in NLU tasks by training a pre-trained BERT model, we call models produced by these objectives MAML-Bert and OML-Bert, as BERT is a state-of-art language model that utilizes Transformer architectures (Devlin et al., 2018). Pre-trained BERT is chosen instead of an empty BERT because Yogatama et al (2019) have showed that training a BERT with supervised tasks instead of unsupervised tasks critically degrades model performance, and this paper focuses on supervised tasks only. To understand our training and evaluation methods, a brief overview of both objectives are introduced below.

A Continual Learning Problem consists of an unending stream of data

$$\mathcal{T} = (X_1, Y_1), (X_2, Y_2), ..., (X_t, Y_t), ...$$

for inputs X_t and targets Y_t. In our case, to put meta-learning into picture, we concatenate batches of data in order, each consisting of data from a task in the glue benchmark (Wang et al., 2019). In general, we treat the four high-resource tasks, namely SST-2, QQP,1 MNLI, and QNLI as learning tasks. The other four tasks: CoLA , MRPC , STS-B , and RTE are our tasks for testing. Original MAML proposed by Finn et al (2017) a task $\mathcal{T}_i$ is sampled from $p(\mathcal{T})$ during meta-training, the model is trained with K samples and feedback from the corresponding loss $L_{\mathcal{T}_i}$, and then tested on new samples from $\mathcal{T}_i$. Model is improved by looking at how test error on new data changes with respect to parameters.
For Meta-Training, we consider two meta-objective to minimize. (1) a MAML like objective and (2) OML objective. The OML objective

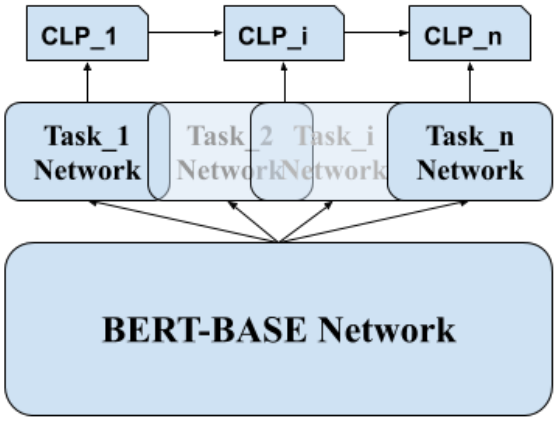

Figure 1: mBERT (modified BERT) model for continual learning problems (CLPs)

is defined as

$$\sum_{\mathcal{T}_i \sim p(\mathcal{T})} OML(W, \theta) = \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \sum_{S_j^k \sim p(S_k|\mathcal{T}_i)} \mathcal{L}(U(W, \theta, S_j^k))$$

where $S_k^j = (X_{j+1}^i, Y_{j+1}^i), (X_{j+2}^i, Y_{j+2}^i)$,... $, (X_{j+k}^i, Y_{j+k}^i)$ sampled from distribution $p(S_k|\mathcal{T}_i)$, and $U(W_t, \theta, S_k^j) = (W_{t+k}, \theta)$, with each ith update in U is taking parameters from (W_{t+i-1}, θ) with datapoint (X_{t+i}^i, Y_{t+i}^i) to obtain current step (W_{t+i}, θ). Specifically, we designed our *mBERT* (modified BERT model) inspired from (2019) into two parts: *BERT-Base Network* (Representation Learning Network, RLN) and *Task-specific Network* (Prediction Learning Network, PLN) illustrated in the Figure 1.

To minimize the objective above, our MAML-BERT model for continuous learning are elaborated in Algorithm 1. For OML-BERT continuous learning model, we proposed the same strategy except $step_8$ in *Algorithm 1*, which was changed to train the inner update step with only one sample of the training data at a time. Using this method, we tried to mimic the Stochastic Gradient Descent approach in order to avoid catastrophic forgetting.

3.2 Evaluation

For baseline results, we want to compare to the results by using MAML only and BERT on four tasks that are reserved for testing: CoLA, MRPC,

Algorithm 1: MAML-BERT Training

input : CLP problems under distribution $p(\mathcal{T})$

setting: Fine tuned parameters for PLN, RLN learning rates α, β

for T_i *in* $p(CLP)$ **do**
  Fetch T_i 's train dataset ;
  Initialize *BERT-Base Network* parameters θ ;
  while *not done* **do**
    Initialize *Task-specific Network* parameters W ;
    for *j in of inner update step* **do**
      Freeze *BERT-Base Network* ;
      Train for $Batch_j$ of query data to update W_j
    end
    Unfreeze *BERT-Base Network* ;
    Fetch T_i 's test dataset ;
    Update θ_i ;
  end
end

STS-B, RTE. Results obtained from (Dou et al., 2019). Note that different tasks have different metrics.

Tasks	MAML	BERT
CoLA	53.4	52.1
MRPC	89.5/85.8	88.9/84.8
STS-B	88.0/87.3	87.1/85.8
RTE	76.4	66.4

Additionally, We would apply our model in MRQA Shared Task 2019 containing 6 out of domain tasks BioASQ (BA), DROP (DP), DuoRC (DR), RACE (RA), RelationExtraction (RE), and TextbookQA (TQ), and compare our performance with state-of-art adversarial training model (2019).

1) hyper-parameter tunning on result 2) meta-training / meta-testing tasks split

4 Collaboration Statement

Shiqing Li and Yong Fan amended the dataloader to support GLUE datasets. Duo Jiang and Jiacheng Wang modified MAML and implemented OML. All contributed evenly to the partial draft.

Acknowledgments

References

Maruan Al-Shedivat, Trapit Bansal, Yuri Burda, Ilya Sutskever, Igor Mordatch, and Pieter Abbeel. 2017. Continuous adaptation via meta-learning in non-stationary and competitive environments. *CoRR*, abs/1710.03641.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. *arXiv preprint arXiv:1908.10423*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400.

Khurram Javed and Martha White. 2019. Meta-learning representations for continual learning. *CoRR*, abs/1905.12588.

Seanie Lee, Donggyu Kim, and Jangwon Park. 2019. Domain-agnostic question-answering with adversarial training.

Tianlin Liu, Lyle Ungar, and João Sedoc. 2019. Continual learning for sentence representations using conceptors. *CoRR*, abs/1904.09187.

Fei Mi, Minlie Huang, Jiyong Zhang, and Boi Faltings. 2019. Meta-learning for low-resource natural language generation in task-oriented dialogue systems. *CoRR*, abs/1905.05644.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy an, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1810.04805.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomás Kociský, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. 2019. Learning and evaluating general linguistic intelligence. *CoRR*, abs/1901.11373.

A Appendices

B Supplemental Material