The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

MMTL: The Meta Multi-Task Learning for Aspect Category Sentiment Analysis

Guan-Yuan Chen[♠♣*] and **Ya-Fen Yeh**[♡*]
[♠]Telecommunication Laboratories, Chunghwa Telecom, Taoyuan, Taiwan
[♣]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
[♡]Industrial Technology Research Institute, Hsinchu, Taiwan
guanyuan@gapp.nthu.edu.tw, amilaok94@gmail.com

Abstract

Aspect Category Sentiment Analysis (ACSA), which aims to identify fine-grained sentiment polarities of the aspect categories discussed in user reviews. ACSA is challenging and costly when conducting it into real-world applications, that mainly due to the following reasons: 1.) Labeling the fine-grained ACSA data is often labor-intensive. 2.) The aspect categories will be dynamically updated and adjusted with the development of application scenarios, which means that the data must be relabeled frequently. 3.) Due to the increase of aspect categories, the model must be retrained frequently to fast adapt to the newly added aspect category data. To overcome the above-mentioned problems, we introduce a novel Meta Multi-Task Learning (MMTL) approach, that frame ACSA tasks as a meta-learning problem (i.e., regarding aspect-category sentiment polarity classification problems as the different training tasks for meta-learning) to learn an ideal and shareable initialization for the multi-task learning model that can be adapted to new ACSA tasks efficiently and effectively. Experiment results show that the proposed approach significantly outperforms the strong pre-trained transformer-based baseline model, especially, in the case of less labeled fine-grained training data.

Keywords: Aspect Category Sentiment Analysis, Meta-Learning, Multi-Task Learning

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) (Pontiki et al., 2014a,b,c) is an important fine-grained task in the field of sentiment analysis, that is considerable for grasping and understanding user comments in real-world applications. ABSA contains several sub-tasks, four of which are Aspect Term Extraction (ATE), Aspect Term Sentiment Analysis (ATSA), Aspect Category Detection (ACD), and Aspect Category Sentiment Analysis (ACSA). ATE extracts and identifies the corresponding Aspect Term from the sentences of user comments and ATSA aims to predict the polarity of the sentiment toward the identified aspect terms. ACD detects the aspect categories mentioned in review sentences, and ACSA classifies the sentiments of the detected aspect categories.

Since the ATE and ATSA aim to extract the aspect terms of sentences and to predict sentiments corresponding to the extracted aspect terms, this may encounter some problems when the aspect term is not explicitly mentioned or pointed out in the sentence. For example, ” 味道很棒, 很好吃” (Good-tasting). This is an example often seen in real internet reviews for a restaurant, which gives positive reviews on the taste of food but does not indicate the corresponding aspect term. To cope with the above problems, we mainly focus on the methods of ACD and ACSA (usually, the two will be combined and referred to as ACSA tasks), which dedicate to detects aspect categories of given sentences and classifying the sentiments polarities toward the detected aspect categories. For the above example, we can define suitable categories to conduct aspect-based sentiment analysis on user reviews by the ACD and ACSA approach, even the aspect term is not explicitly mentioned. For example, it may be detected as the taste of food category with positive reviews.

Since a user review may discuss more than

*denotes equal contribution

one aspect category and express different sentiments toward them, how to effectively detect various categories with their sentiment polarity at the same time is one of the most important research directions of ACSA. Wang et al. (2016) used the attention-based LSTM models for aspect-level sentiment classification. Cheng et al. (2017) proposed a HiErarchical ATtention (HEAT) network consisting of aspect attention and sentiment attention. Xue and Li (2018) introduced the Gated Convolutional Networks for ACSA and ATSA tasks with appropriate accuracy. Schmitt et al. (2018) used End-to-End Neural Networks which jointly model the detection of aspects and the classification of their polarity.

Recently, the transformer (Vaswani et al., 2017) based pre-trained language models such as BERT (Pre-training of Deep Bidirectional Transformers for Language Understanding) (Devlin et al., 2019), XLNet (Generalized Autoregressive Pretraining for Language Understanding) (Yang et al., 2019b), RoBERTa (A Robustly Optimized BERT Pretraining Approach) (Liu et al., 2019b), ELECTRA (Pre-training Text Encoders as Discriminators Rather Than Generators) (Clark et al., 2020) and DeBERTa (Decoding-enhanced BERT with Disentangled Attention) (He et al., 2020) have significantly improved the performance of many natural language processing (NLP) tasks on several benchmarks (Wang et al., 2019b,a; Xu et al., 2020).

In the ABSA field, some previous works have shown the promising of the pre-trained transformer models. Li et al. (2019) investigated the modeling power of contextualized embeddings from BERT to deal with End2End ABSA. Li et al. (2020) proposed a Multi-Instance Multi-Label Learning Network for ACSA tasks, and their experimental results showed that the BERT-based models significantly performed better than the non-BERT models (non-pre-trained transformer models) on the public datasets.

Despite previous studies that have demonstrated the success of deep learning models, especially, the pre-trained transformer models on the ABSA-related research and experiment setting, few works are studying and considering the crucial issues when conducting the deep ACSA models into real-world applications. In practical application, ACSA may be quite challenging and costly due to the following reasons: Firstly, Labeling the fine-grained ACSA data is often complicated and labor intensive (there may be so many aspect categories that need to detect and analyze). Secondly, the aspect categories may be dynamically updated and adjusted with the progress of application scenarios, which means that the data may need to be relabeled not infrequently. Thirdly, the model must be able to fast adapt to the newly added aspect category data, due to the increasing and changing of aspect categories.

In this paper, we propose a novel Meta Multi-Task Learning (MMTL) approach that considers ACSA tasks with various aspect categories as meta-learning and multi-task learning tasks (i.e., regarding aspect-category sentiment polarity classification problems as the training tasks for meta-learning and multi-task learning). Primary, we investigate the efficient and effective approaches for learning the well-conditioned and shareable initialization via the Model-Agnostic Meta-Learning algorithm (MAML) (Finn et al., 2017) and its variants (Nichol et al., 2018) for multi-task learning models. Different from previous MAML related works, in our case, the initialization learned through meta-learning must be shareable (parameter sharing) across the different polarity classification tasks of aspect categories with the same user review input. Because in actual applications, there will be a large number of aspect categories, and it is costly that different models are used to extract features for different aspect categories individually. Therefore, parameter (feature) sharing strategies such as multi-task learning is more appropriate. To achieve the above-mentioned goals, we introduce the new Meta Multi-Task Learning (MMTL) approach, which divides the model parameters into independent and shareable parts and uses different meta-learning objective functions for training on these two parts. For the part of parameter sharing, we employ the proximal regularization term in the objective function in the meta-learning inner loop training phase to encourage the model to learn parameters and

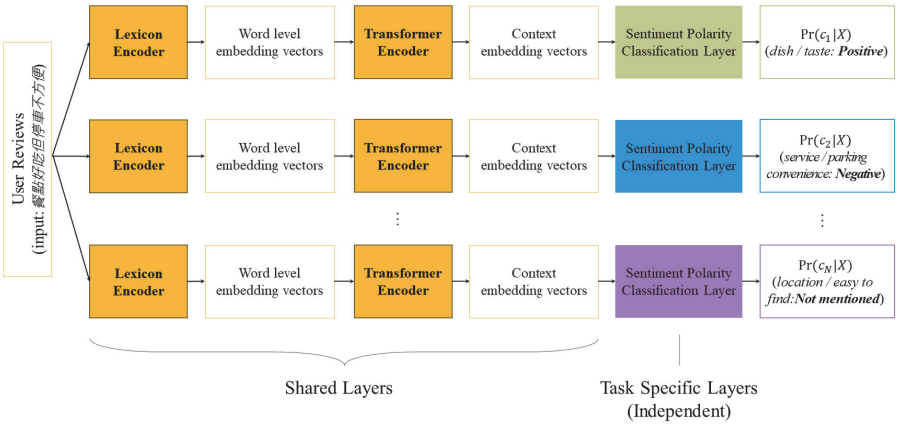

Figure 1: The architecture of the MMTL model for ACSA tasks. For each aspect category sentiment polarity classification task, most of the model parameters and features are shared, and only the parameters of the individual aspect category polarity classification layer (single layer neural network) are independent.

features that can be shared on different aspect categories tasks.

2 Related Work

Learning general representations of given text inputs for many tasks is the important goal for many Natural Language Processing (NLP) fields. The same is true for the Aspect Based Sentiment Analysis (ABSA) and its subtasks. Xue et al. (2017) proposed a multi-task learning model based on neural networks to solve the Aspect Category Classification and Aspect Term Extraction together. Yang et al. (2019a) introduced a Multi-task Learning Model for Aspect Polarity Classification and Aspect Term Extraction for Chinese-oriented tasks.

Since the transformer-based pre-trained language models have demonstrated their success in many NLP tasks, some works explored the potential of integrating the multi-task learning and pre-trained language models. Mainly, Liu et al. (2019a) presented Multi-Task Deep Neural Network (MT-DNN) learning representations by leveraging large amounts of cross-task data and obtaining state-of-the-art results on several NLU tasks.

However, there still exist some potential problems when adopting multi-task learning related algorithms into real-world applications. The most important is that multi-task learning may favor the tasks with more labeled data over the tasks with less labeled data ones. Inspired by Raghu et al. (2020); Rajeswaran et al. (2019) (they found that feature reuse is the dominant factor of the effectiveness of Model Agnostic Meta-Learning based algorithms, which means that meta-learning has the trend to learn features that can be reused in different tasks), we propose the Meta Multi-Task Learning (MMTL) approach to applying meta-learning algorithms for finding the well-conditioned and shareable initialization for multi-task learning models, such that the model can be significantly improved in the case of a small amount of data and can efficiently learn new tasks.

3 Proposed Approaches

The architecture of the Meta Multi-Task Learning (MMTL) model is shown in Figure 1. The proposed approaches are briefly described as follows. First, we treat different aspect categories of polarity classification tasks as different training tasks. Second, we apply the Model Agnostic Meta-Learning (MAML) based algorithms (Finn et al., 2017; Nichol et al., 2018) to finding the well-conditioned and shareable initialization for multi-task learning models for the different polarity classification tasks of aspect categories with the same review text. Finally, we using the multi-task learning approach with the shareable general representations and initialization to fine-tune the model on all aspect categories sentiment polarity clas-

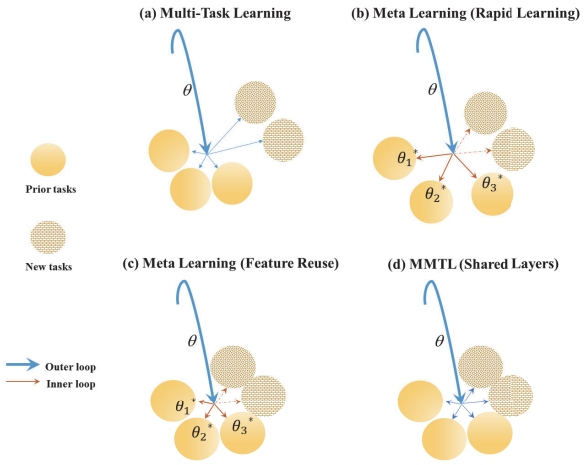

Figure 2: Differences in multi-task learning, rapid learning (meta-learning), feature reuse (meta-learning), and MMTL. (a) Multi-task learning can share the same parameter weights among different tasks, but it may favor the tasks with more data. (b) The Meta-Learning (Rapid Learning) obtains well-conditioned model initialization parameters through outer loop training, and inner loop updates result in significant task specialization. (c) The Meta-Learning (Feature Reuse) through the outer loop training to find the ideal initialization parameters of the model that can be feature reused. There are fewer differences in the updated parameters of different tasks during the inner loop training. (d) The MMTL utilizes the Meta-Learning (Feature Reuse) algorithm to find the ideal model initialization parameters that can be shared for different tasks. Then fine-tune the model through multi-task learning, so that the MMTL model can combine the advantages of multi-task learning and meta-learning, i.e. it can share most of the parameters on different tasks, and it can be adapted to new tasks with fewer training samples.

sification tasks. However, achieving the above goals is not trivial works. In particular, meta-learning and multi-task learning were regarded as two completely different methods in the past and few studies have discussed how to integrate the two methods and their respective advantages. Below, we will introduce details of the proposed method.

3.1 Meta-Learning and Multi-Task Learning

Since there are some obvious differences between meta-learning and multi-task learning, it is not trivial work to integrate these two learning algorithms with their advantages and characteristics. The differences between meta-learning and multi-task learning are shown in Figure 2. Multi-task learning trains different tasks together at the same time. This will cause multi-task learning to favor tasks with more annotated data and significantly worse performance for tasks with less annotated data. The main goal of the MAML algorithm (Finn et al., 2017) is to find good model initialization parameters such that the model can perform well to new tasks, even on tasks with fewer data. Even in many studies, it has been shown that the MAML algorithm can perform well in new tasks (especially in the case of a small amount of annotated data) (Finn et al., 2017; Gu et al., 2018; Nichol et al., 2018; Dou et al., 2019), why MAML has good learning ability in new tasks is still an issue to be analyzed. The effectiveness of MAML is mainly discussed in two different aspects (Raghu et al., 2020), 1.) Rapid Learning: There are large and effective changes in the representations, 2.) Feature Reuse: the meta-initialization containing high quality and reusable features. Since previous studies have found that MAML has the characteristics and capabilities of feature reuse, we explore ways to further impose training constraints on the model to encourage the MAML model to have the ability to share features for different tasks. Finally, we propose the novel Meta Multi-Task Learning (MMTL) algorithm to integrate meta-learning and multi-task learning algorithms. Experimental results show that the proposed MMTL algorithm can combine the advantages of meta-learning and multi-task learning, and is significantly outperform

the strong pre-trained language model baseline.

3.2 The Proposed Meta Multi-Task Learning (MMTL) Model

First, we regard the ACSA tasks of different s aspect categories as a set of tasks $\{T_1, T_2, ..., T_s\}$ for meta-learning. Given a model f_θ with parameters θ and a task distribution $p(T)$ over a set of tasks $\{T_1, T_2, ..., T_s\}$. We sample a batch of tasks $\{T_b\} \sim p(T)$, and update the model parameters by k gradient descent steps for each task $\{T_b\}$ for the inner loop training of meta-learning. Where, the $k \geq 1$ and the $p(T)$ is a uniform probability distribution. For the inner loop (task specific) training of meta learning, we use the following equation to update the model parameters θ:

$$\theta_b^{(k)} = \theta_b^{(k-1)} - \beta \nabla_{\theta_b^{(k-1)}} L_b \left(f_{\theta_b^{(k-1)}} \right)$$

Where L_b is the objective function (described as follows) and β is the learning rate (a hyperparameter) of the inner loop training.

To encourage the model to have the ability to share the parameters (feature reuse) for different tasks, we divide the model into the shared layers part and the task-specific layers part. For the shared layers part, we add a proximal regularization term in the inner loop training phase. Therefore, the definition of the objective function (loss function) of the shared layers part is as follows:

$$L_b = Loss(f_{\theta_b^{(k-1)}}) + \lambda \left\| \theta_b^{(k-1)} - \theta \right\|$$

And the definition of the objective function (loss function) of the task-specific layers part is as follows:

$$L_b = Loss(f_{\theta_b^{(k-1)}})$$

Where, the $Loss$ is the Cross-Entropy Loss calculated on the inner loop training task $\{T_b\}$, the λ is a hyperparameter, and the θ is the parameter of the model. Initially, θ is the weight of the pre-trained model and is updated by the training of the outer loop of the meta-learning.

Since the original MAML algorithm (Finn et al., 2017) needs to calculate the second derivatives, resulting in excessive calculation and memory usage, we use the Reptile (a first-order gradient-based meta-learning algorithm) (Nichol et al., 2018) to update the model parameters θ for the outer loop phase.

The equation of the Reptile is defined as:

$$\theta = \theta + \gamma \frac{1}{|\{T_b\}|} \sum_{T_b \sim p(T)} \left(\theta_b^{(k)} - \theta \right)$$

Where the γ is the learning rate (a hyperparameter) of the outer loop training.

Finally, we use the model parameters trained via meta-learning as the initialization parameters, and perform multi-task learning training (fine-tuning) on the data of ACSA tasks. Overall, the training process of MMTL mainly consists of three stages: 1.) the pre-training stage as in BERT or ELECTRA, 2.) the meta-learning stage, and 3.) the multi-task learning fine-tuning stage.

The model trained by the proposed MMTL algorithm is different from the multi-task learning model (that is shown in Figure 2). Attributable to the fact that we first use meta-learning and some constraints to make the parameters of the model can be shared on different tasks and perform ideally on new tasks, even if the new task only has a relatively small amount of training data. The MMTL model is also obviously different from the meta-learning model. The meta-learning model will eventually be fine-tuned to different weights on different tasks, and it is not possible to directly share parameters for different tasks. The MMTL model can share most of the model parameters between different tasks and has obvious computational advantages on ACSA tasks with a large number of categories.

4 Experiments

We conduct experiments on the AI Challenger 2018 Sentiment Analysis Dataset[1], the large-scale Chinese fine-grained sentiment analysis dataset for the Aspect Category Sentiment Analysis (ACSA) tasks. The dataset contains 105,000 training data, 15,000 validation data, and 15,000 testing data. And the data set contains 20 categories, each of which is composed of two layers (below we define

[1] https://github.com/AIChallenger/AI_Challenger_2018

these 20 categories in the form of "The first layer/The second layer"). The 20 aspect categories are respectively 1.) "location/traffic convenience", 2.) "location/distance from business district", 3.) "location/easy to find", 4.) "service/wait time", 5.) "service/waiter's attitude", 6.) "service/parking convenience", 7.) "service/serving speed", 8.) "price/price level", 9.) "price/cost-effective", 10.) "price/discount", 11.) "environment/decoration", 12.) "environment/noise", 13.) "environment/space", 14.) "environment/cleanness", 15.) "dish/portion", 16.) "dish/taste", 17.) "dish/look", 18.) "dish/recommendation", 19.) "others/overall experience", 20.) "others/willing to consume again". For each user review, the dataset provides the sentiment polarity label (the Positive or Neutral or Negative or Not mentioned) corresponding to the above 20 aspect categories. The goal of the model is to classify the sentiment polarity of different aspect categories.

Since the AI Challenger 2018 Sentiment Analysis Dataset does not provide annotation data for the test data, our experiment used the validation set of the original dataset to evaluate the quantitative performance of the model (as the test dataset for experiments), and we randomly split 15,000 data from the training set as the validation set.

To evaluate the performance of the model on new tasks and tasks with a small amount of data, we also perform some experimental settings on the dataset. We use the less frequently mentioned categories (also with the worst performance of the baseline models) in the dataset as new tasks ("location/distance from business district", "dish/look", "others/overall experience") and the other 17 categories are considered as prior tasks. Those are used to simulate the situation that the model encounters a new aspect category task. We also randomly sample 500, 1000, 2000 examples of the training data and test models' performance on these samples with a few-shot setting.

4.1 Model and Hyperparameter Setting

We compare our models with two strong baselines: 1.) the FastText model (Bojanowski et al., 2017; Joulin et al., 2017) and the ELECTRA model (Clark et al., 2020). For the FastText model, we used the publicly available code [2] for experiments. This code is mainly set for the AI Challenger 2018 Sentiment Analysis Dataset. Its performance is better than the SVM baseline model provided by AI Challenger 2018, and it is also more computationally efficient. For the ELECTRA model, we used the publicly available code[3] (Cui et al., 2020) for experiments. Although the ELECTRA model has larger architectures (large and base), in this experiment, we only consider the ELECTRA small model architecture. Since in actual application scenarios, transformer-based pre-training models will require more GPU computing resources, and larger models will increase the burden of computing resource costs. Therefore, we focus our experiments on smaller models that are more suitable for practical applications.

We implement our algorithms upon the ELECTRA-180g-small (Chinese) model[4]. We set the batch size to 32, the learning rate to 5e-5, and use the Adam optimizer to train the model. For the stages of meta-learning training (k is set to 5, b is set to 8, β is set to 1e-4, λ is set to 0.5 and γ is set to 1e-3) and multi-task learning fine-tuning, we train for 5 epochs individually.

5 Results

First, we use the proposed Meta Multi-Task Learning (MMTL) method to train the model on the AI Challenger 2018 Sentiment Analysis Dataset. Since the MMTL method involves three stages, 1.) the model pre-training stage (loading pre-trained model weights), 2) the meta-learning stage (using to find the optimal model initialization parameters), 3.) the fine-tuning stage of multi-task learning, we also compare the MMTL model with the results of using the pre-training model, meta-learning model, or multi-task learning model respectively.

Table 1 reports the experimental results on the test dataset (the experimental setup de-

[2] https://github.com/panyang/fastText-for-AI-Challenger-Sentiment-Analysis

[3] https://github.com/ymcui/Chinese-ELECTRA

[4] https://huggingface.co/hfl/chinese-electra-180g-small-discriminator/tree/main

F1 (macro)	FastText	ELECTRA	Multi-Task	Reptile	MMTL
Avg. of all 20 aspect categories	54.3	66.1	68.4	67.3	**68.9**
location/distance from business district	43.1	51.2	53.4	**56.6**	56.4
dish/look	43.4	54.4	55.3	57.6	**57.7**
others/overall experience	53.0	56.5	58.8	**60.3**	**60.3**

Table 1: The F1 (macro) results of the proposed models compare to the baseline. Multi-Task: The ELECTRA based model trained with the multi-task learning approach (share most of the parameters). Reptile: The ELECTRA based model trained with the mete-learning approach (no parameter sharing). MMTL: The ELECTRA based model trained with the MMTL approach (share most of the parameters). Note: the "location/distance from business district", "dish/look", and "others/overall experience" are the aspect categories that are less frequently mentioned by user reviews, and are also the aspect categories with the worst performance of the baseline models.

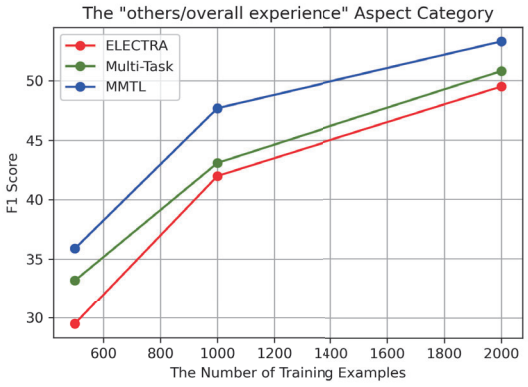

Figure 3: Results on settings for a small amount of training data (500, 1000, 2000 training samples). The target task is the sentiment polarity classification of the aspect category "others/overall experience".

tails are described in Section 4). As we can see, in general, MMTL achieves better performance than the strong baseline models. In addition, it is worth mentioning that the results of the multi-task learning and the meta-learning methods are better than pre-training models based on the same model architecture, but there are some differences in the effectiveness of the two methods.

Although multi-task learning can achieve a higher average f1 score of the 20 aspect categories than meta-learning, there is relatively little improvement in multi-task learning on categories that are less frequently mentioned by user reviews (categories with poor baseline model performance). The possible reason is that multi-task learning may favor categories with high-resource tasks over low-resource ones (Dou et al., 2019). The meta-learning model is different, it can have better performance in the above categories, but the average f1 score is lower than the multi-task learning model. In particular, the MMTL model integrates the advantages of multi-task learning and meta-learning. It performs well in both the average f1 score or the less frequently mentioned categories (more difficult categories).

Note that although it can be seen from Table 1 that the model based on meta-learning has a significant performance improvement on less-mentioned tasks, the model of meta-learning will eventually be fine-tuned to different weights for different aspect category tasks (no parameter sharing), hence it is more difficult applied to actual and real-time application scenarios (different categories require different model weights, e.g., 20 different weights are required on the AI Challenger 2018 Sentiment Analysis Dataset).

To evaluate models' performance with low-resource setting (to simulate the situation that the model encounters a new task with a new aspect category data), we also randomly sample 500, 1000, 2000 examples of the training data from the "others/overall experience" aspect category. Figure 3 shows that the proposed MMTL model significantly outperforms the multi-task learning model and the ELECTRA pre-training model when the amount of training data is small. This shows that the MMTL model that combines meta-learning and multi-task learning is helpful for new tasks (tasks with less data).

6 Conclusion

In this work, we proposed a learning approach called MMTL to combine meta-learning and multi-task learning methods for Aspect Cate-

gory Sentiment Analysis (ACSA) tasks. The experimental results show that the model based on the MMTL method overall outperforms the strong baseline models of pre-trained models, meta-learning models, and multi-task learning models. And when the amount of training data is small, compared with pre-trained models and multi-tasking learning models, the MMTL model also has relatively better performance. Compared with the meta-learning model (the model of meta-learning will eventually be fine-tuned to different weights for different aspect category tasks, i.e. no parameter sharing), as a result of the MMTL can share parameters between different aspect category tasks, it has better computing efficiency and less memory usage, thus it is more suitable for deployment in practical applications.

There are many future directions worthy of further exploration, especially in addition to ACSA, the Aspect-Based Sentiment Analysis field also contains many subtasks such as Aspect Term Extraction, Opinion Term Extraction, Multi-Aspect Sentiment Analysis, and Cross-domain Aspect-based Sentiment Analysis, how to effectively share model parameters in these subtasks and achieve better performance on new tasks with less data, these are important future research directions.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jiajun Cheng, Shenglin Zhao, Jiani Zhang, Irwin King, Xin Zhang, and Hui Wang. 2017. Aspect-level sentiment classification with HEAT (hierarchical attention) network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 97–106. ACM.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou, Keyi Yu, and Antonios Anastasopoulos. 2019. Investigating meta-learning algorithms for low-resource natural language understanding tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1192–1197, Hong Kong, China. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong,

China. Association for Computational Linguistics.

Yuncong Li, Cunxiang Yin, Sheng-hua Zhong, and Xu Pan. 2020. Multi-instance multi-label learning networks for aspect-category sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3550–3560, Online. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014a. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014b. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014c. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. 2020. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. 2019. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Martin Schmitt, Simon Steinheber, Konrad Schreiber, and Benjamin Roth. 2018. Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1109–1114, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*,

pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523, Melbourne, Australia. Association for Computational Linguistics.

Wei Xue, Wubai Zhou, Tao Li, and Qing Wang. 2017. MTNA: A neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 151–156, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Heng Yang, Biqing Zeng, Jianhao Yang, Youwei Song, and Ruyang Xu. 2019a. A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction. *CoRR*, abs/1912.07976.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019b. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.