Multi-Task Learning Using BERT With Soft Parameter Sharing Between Layers

Niraj Pahari
Department of Artificial Intelligence
Kyushu Institute of Technology
Fukuoka, Japan
nirajpahari@gmail.com

Kazutaka Shimada
Department of Artificial Intelligence
Kyushu Institute of Technology
Fukuoka, Japan
shimada@ai.kyutech.ac.jp

2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS) | 978-1-6654-9924-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SCISISIS55246.2022.10001943

***Abstract*—The ability to learn from one task and pass the knowledge to another task is a remarkable job that is possible because of multi-task learning (MTL). It is helpful in the scenario where data is scarce. In this paper, we study the use of MTL in natural language processing (NLP) using BERT for sentiment classification, category classification, and aspect-opinion sequence classification. The Restaurant-ACOS dataset is split into two subsets, namely the sentiment-category data subset, for the first two tasks and the sequence classification data subset, for the third task. The BERT layers are grouped into top, middle, and bottom layers. These layers are either frozen, shared, or kept unshared for individual tasks. Experiments are conducted to find the best combination for the overall layers of BERT. The experiments demonstrate that MTL can improve the performance on these tasks when compared as a baseline method that uses single task learning. One of the best configurations was based on freezing on the bottom layers, sharing on the middle layers, and keeping on the top layers, respectively. In addition, we investigate the role of each layer via the size of layers. From the result, the individual layers capture the task-specific patterns whereas shared layers capture the domain knowledge.**

***Index Terms*—multi-task learning, soft parameter sharing, BERT, aspect based sentiment analysis**

I. Introduction

The evolution of different deep learning methods has driven the Natural Language Processing (NLP) field forward massively by reaching high performances on various tasks. These methods are generally trained for a single task [1]. Moreover, the learning techniques usually require a large amount of data for training to achieve accurate results [2]. However, there is still a scarcity of that amount of data in many domains.

Multi-Task Learning (MTL) is one of the approaches which can help with the data insufficiency problem when multiple tasks are related to each other. MTL is the form of inductive transfer learning that improves the performance of the tasks learned together by introducing an inductive bias [3]. MTL has been successfully used in multiple domains [4]–[6]. There are two methods in MTL: hard parameter sharing and soft parameter sharing of hidden layers. In hard parameter sharing, the same model is used across the tasks, and the parameters are shared by all the tasks while task-specific models are kept separate. In soft parameter sharing, each task has its own model, and the parameters of the model are often regularized to encourage the similarity of the parameters [7]. Soft parameter sharing has been employed in this work. Despite the success of MTL, its effectiveness still remains open [8].

The introduction of BERT [9] has taken the NLP research field to another level by presenting state-of-the-art results in a wide variety of tasks [10]. Many researchers have studied the internal workings of BERT [11], [12]. MTL with BERT has also shown improvement in results in multiple domains in NLP [8], [10], [13], [14].

In this study, we explore the performance of MTL with soft parameter sharing on different groups of BERT layers. We group the 12 layers of BERT into three sub-layers for category classification, sentiment classification, and aspect-opinion extraction tasks in different configurations. Our main contributions are the following:

1) We study the effectiveness of MTL in three different Aspect-based sentiment analysis tasks.
2) We study the effective sharing strategy of layers within BERT encoders.

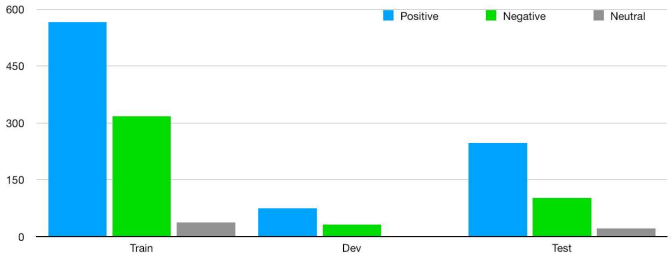
II. Related Work

Multi-task learning (MTL) has been studied for a long time now [3]. MTL uses the knowledge gained in one task to train different tasks. MTL utilizes the useful information contained in multiple tasks and helps to train a more accurate learner for each task [6]. This type of learning has already been used in various domains such as computer vision, natural language processing (NLP), bioinformatics, etc.

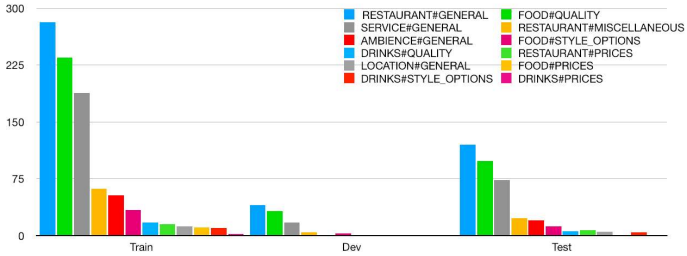
For NLP, MTL has been used in many applications. Balikas et al. [15] applied MTL to jointly learn the ternary and fine-grained classification and showed that one can improve the performance of others. Collobert et al. [16] have used MTL for joint learning of six different NLP tasks: semantic role labeling, named entity recognition, part of speech tagging, chunking, language modeling, and semantically related words. Their results showed that learning different tasks simultaneously can improve the generalization performance.

Similarly, different researchers have studied MTL approaches with BERT. Peng et al. [17] carried out an empirical study of MTL on BERT for Biomedical text mining. Their evaluation results show that fine-tuned MTL models can perform better than the state-of-the-art transformers models. Satapathy et al. [10] proposed a BERT-based MTL framework

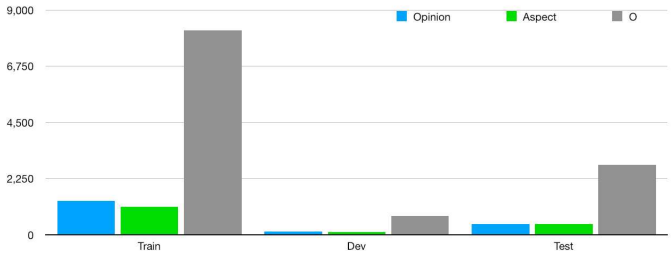
Authorized licensed use limited to: Stanford University. Downloaded on February 09,2023 at 06:03:59 UTC from IEEE Xplore. Restrictions apply.



(a) Sentiment Distribution



(b) Category Distribution



(c) Aspect, Opinion, and O tags distribution

Fig. 1: Distribution of classes in the dataset. In figure (a) and figure (b), the sentiment and category distribution are shown from the sentiment-category dataset, and in figure (c), the distribution for aspect, opinion, and O tokens are shown from the aspect-opinion dataset. BIO representations are used for representing aspect and opinion spans from the text. Frequency for beginning and inside tags for the aspect and opinion tags are combined and together with outside tags are shown in the graph.

that combined the sentiment and subjective detection and improved the result for the sentiment task by 15% and for the subjectivity task by 2-4% as compared with the single-task performance.

Even though BERT is very popular, it is not well-known for its inner workings. Several studies have been focused on studying how knowledge is distributed across different layers of BERT. Clark et al. [11] proposed the methods for understanding the attention mechanisms of different models and applied that to BERT. They show that linguistic knowledge is also present in the attention maps together with hidden states. Tenney et al. [12] explored different layers of the BERT network on how it can resolve syntactic and semantic structure within a sentence. They found that the BERT model represents steps of the traditional NLP pipeline. Each step follows the sequence of POS tagging, parsing, NER, semantic roles, and coreference.

Vig and Belinkar [18] studied the layers of the transformer language model and found that initial layers generally encode positional relations, middle layers mostly capture dependency relations, and later layers capture the unique patterns and structures. This same result was obtained and presented interactively for BERT models in exBERT by Hoover et al. [19].

Our work applies MTL with BERT and studies the effect of sharing BERT layers using a soft parameter sharing method. The work also studies the correct strategy for freezing, sharing, and keeping the BERT layers individual for each task.

III. Datasets

For the evaluation of the performance of the models, we use the Restaurant-ACOS dataset provided by Cai et al. [20]. This dataset is built on the SemEval 2016 Restaurant dataset [21] and its expansions [22], [23]. The original SemEval 2016 Restaurant dataset with the opinion annotations by [22], [23] was collected the annotations for implicit opinions were added.

In this study, we split the Restaurant-ACOS dataset[1] into two subsets. The Restaurant-ACOS dataset contains review sentences and their ACOS quadruples. Each ACOS quadruple has the aspect, category, opinion, and sentiment annotation. The aspect is an entity on which the opinion is targeted. It is a word or a phrase inside the review sentence. The category represents the predefined category within the domain. The opinion is the subjective statement towards the aspect. It is also a word or a phrase inside the review sentence. The sentiment is the semantic orientation towards the aspect. Hence in the dataset annotation for each ACOS quadruple, the category and sentiment belong to one of the predefined classes, whereas the aspect and opinion are the span annotation. Each review sentence can have one or more than one ACOS quadruple. The example of two different review sentences with a single quadruple and two quadruples are shown below:

e1) nice ambience, but highly overrated place.

```
1,2 AMBIENCE#GENERAL 2 0,1
6,7 RESTAURANT#GENERAL 0 5,6
```

e2) I'd go back there in a heartbeat.

```
-1,-1 RESTAURANT#GENERAL 2 -1,-1
```

The first example shows a single review with multiple quadruples. The first portion *"nice ambience"* on the sentence has a positive sentiment, the word *"nice"* represents an opinion word, and the word *"ambience"* represents the aspect. Finally the sentence falls in the category `AMBIENCE#GENERAL`. The second portion *"but highly overrated place"* has the negative sentiment polarity and the word *"overrated"* represents the opinion, and *"place"* represents the aspect. The sentence falls in the category `RESTAURANT#GENERAL`. In this case, a single sentence itself has different classes for sentiment and category for a single review.

The second example, on the other hand, has only one quadruple. The sentence has a positive sentiment polarity and

[1]Downloaded from https://github.com/NUSTM/ACOS

Authorized licensed use limited to: Stanford University. Downloaded on February 09,2023 at 06:03:59 UTC from IEEE Xplore. Restrictions apply.

TABLE I: Dataset statistics.

Dataset	Train	Dev	Test
ACOS-Restaurant	1530	171	583
Category-Sentiment Dataset	920	106	370
Aspect-Opinion Dataset	610	65	213

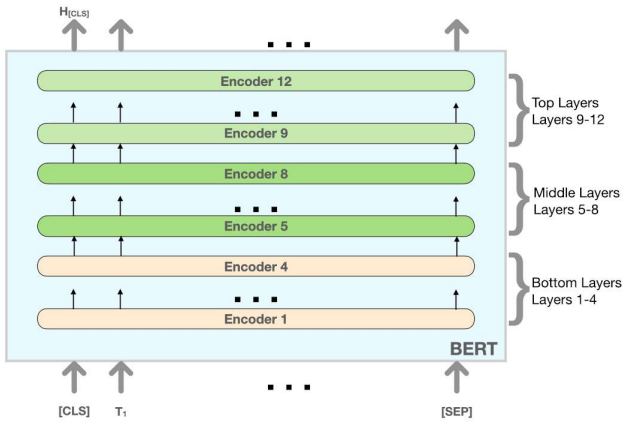


Fig. 2: Classification of BERT Layers into top, middle, and bottom.

belongs to the category `RESTAURANT#GENERAL`, and it does not contain any exact words for the opinion or aspect. From these observations in mind, the dataset is split into two subsets.

For each sentence in the original dataset, if the sentence has only one ACOS quadruple, then that sentence is added to a new category-sentiment classification dataset because the sentence only has a single category and a single sentiment. Hence this data can be useful for training a sentiment and a category classifier. On the other hand, if the sentence has more than one ACOS quadruples, more than one sentiment and category is present for that single sentence. Since the sentiment or category classifier should classify a single sentence to a single sentiment and a single category, the sentence having more than one sentiment and category does not help in the training of these classifiers. Hence that sentence is added to a new aspect-opinion sequence classification dataset. Furthermore, if a sentence has more than one ACOS quadruple, then there will be an explicit aspect or opinion in the sentence; hence it will be useful for sequence classification for aspect or opinion span tagging.

The statistics of the dataset after splitting the Restaurant-ACOS dataset into category-sentiment classification subset and sequence classification subset are shown in Table I.

The number of sentences belonging to each sentiment and category are shown in Fig 1(a) and Fig 1(b) from the category-sentiment dataset. For the aspect-opinion dataset, the count of tokens for aspect, opinion, and O tag from BIO tagging is shown in Fig 1(c).

IV. Tasks and Methods

In this section, we explain the methods and the tasks. We utilize BERT as the base model for our methods. BERT is the well-known and state-of-the-art approach in NLP. Then, we discuss three classification tasks in this paper. We also describe multi-task learning models based on BERT in Section IV-C. We prepare one baseline model and four configurations of MTL.

A. Base Model

Bidirectional Encoder Representations from Transformers (BERT) [9] is a language representation model which is designed to pre-train the bidirectional representations of the text using unlabeled text. BERT uses the famous attention mechanism called transformer [24] to learn the relations between the words and subwords. For English language, BERT has two models. The first one is $BERT_{BASE}$ model with 12 encoder layers, and the second one is $BERT_{LARGE}$ model with 24 encoders.

We use $BERT_{BASE}$ model as a base model for the study. The 12 encoder layers of BERT are divided into bottom, middle, and top layers. The layers 1-4 are the bottom layers, 5-8 are the middle layers, and 9-12 are the top layers, as shown in Fig 2. We compare different combinations of freezing, sharing, and isolating the layers for multiple tasks. In detail, we explain them in Section IV-B. The linear layers are added on top of BERT models for classification.

B. Tasks

In this study, we use three different classifiers to solve three different tasks as MTL. Three tasks are category classification, sentiment classification, and aspect-opinion extraction. Two of the tasks are multi-class classification problems and one of the task is a sequence task. The details of the tasks are described as follows:

1) Category Classification: The category classification task is to classify the review sentence into one of the 13 predefined categories. The BERT is the base model and a linear layer on top of BERT for a classification task. The category-sentiment dataset generated in Section III is used for the task. For example, when the review sentence (**e2**) in Section III is input, the output for this task should be `RESTAURANT#GENERAL`.

2) Sentiment Classification: The sentiment classification task is to classify the review sentence into one of the three predefined categories. The BERT is the base model and a linear layer on top of BERT for a classification task. The category-sentiment dataset in Section III is also used for the task. For example, when the review sentence (**e2**) in Section III is input, the output for this task should be `2`.

3) Aspect-Opinion Extraction: The aspect-opinion extraction task is to extract the aspect and opinion spans from the review sentence. The task is treated as a sequence labeling task. The BERT is the base model and a linear layer on top of the pooled output for a sequence labeling. The modified Begin-Inside-Outside (BIO) tagging scheme is used for span representation. The aspect-opinion dataset is used for the task. For example, when the review sentence (**e1**) in Section III is input, the output for this task should be `1,2` and `6,7` for the aspect spans and `0,1` and `5,6` for opinion spans.

Authorized licensed use limited to: Stanford University. Downloaded on February 09,2023 at 06:03:59 UTC from IEEE Xplore. Restrictions apply.

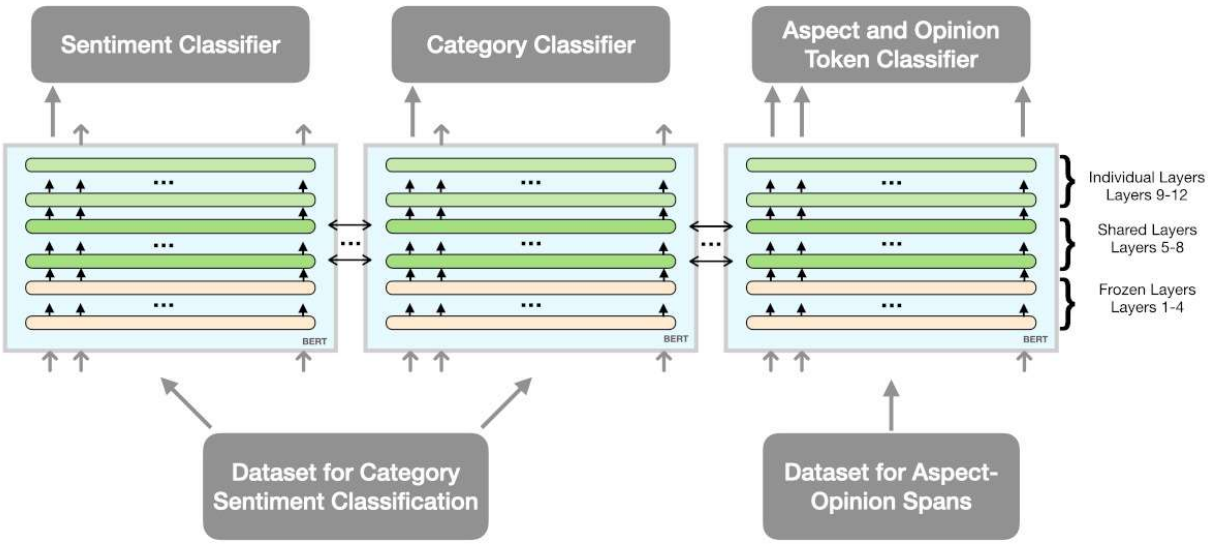


Fig. 3: F-S-I method; bottom layers are frozen, middle layers are shared and top layers are not shared for each task.

C. Experiment Models

In this paper, we propose four different combinations of multi-task learning. In this section, we describe a baseline model that is based on single-task learning first. Then, we describe four configurations of our MTL models. The outline of the MTL model is shown in Fig. 3. The frozen layers are the layers through which the sentences are run without updating the weights. Shared layers are the layers whose weights are updated based on all three tasks. Individual layers are the layers whose weights are updated based on its own individual task.

1) Baseline (Single task learning): The baseline model is based on single-task learning. We just use the BERT model for each classification task (sentiment classification, category classification, and aspect-opinion sequence classification) individually.

2) Freeze-Shared-Individual (F-S-I): As shown in Fig. 3, this model freezes the bottom four layers (1-4) of the BERT model. The middle layers (5-8) are shared across multiple tasks. The top four layers (9-12) are only updated for individual tasks. This model builds from the knowledge by Vig and Belinkor [18]; the initial/bottom layers generally encode the positional relations between the text, the middle layers encode the most dependency relations, and the top layers generally encode the unique patterns and structures.

3) Shared-Shared-Individual (S-S-I): Similar to the F-S-I method, this model also handles individual top layers (9-12), and not only the middle layers but also the bottom layers are shared across the multiple tasks. We expect that the shared learning of the bottom layers can learn more information about the task, and the top individual layers will be able to learn the task-specific details.

4) Freeze-Shared-Shared (F-S-S): This model freezes the bottom four layers (1-4), and the rest of the layers (5-12) are shared across the tasks. Our intuition is that the bottom layers have sufficient pre-trained knowledge of the sentence positional relations, and the model captures the semantic and task-specific information in the middle and top layers.

5) Shared-Shared-Shared (S-S-S): This model shares the parameters of all the BERT layers (1-12). We intend to study the effect on the results when all the parameters are updated by all the tasks involved. We expect to see dipping performance since all the layers are updated by the general knowledge of the corpus and no individual task knowledge is learned specifically.

V. Experiments

We evaluate all the models explained in Section IV-C. First we discuss the results for different combinations for multi-task learning. Then, for the best performance model, a further detailed study is conducted with different numbers of layers for each layer group, namely bottom, middle, and top.

A. Experimental Settings

In this experiment, we utilized the *bert-base-uncased* model from the huggingface transformers library[2]. We used the AdamW optimizer. The learning rate was set as 5e-5 for bert and 5e-3 for linear layers. The maximum length of each sentence is set to 100 to cover all the sentences from all three

[2]https://github.com/huggingface/transformers

Authorized licensed use limited to: Stanford University. Downloaded on February 09,2023 at 06:03:59 UTC from IEEE Xplore. Restrictions apply.

TABLE II: The experiment for different configurations of our MTL models.

	Category			Sentiment			Sequence Classification		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*
Baseline	0.16	0.19	0.17	0.52	0.57	0.54	0.60	0.53	0.51
F-S-I	0.32	0.28	**0.28**	0.53	0.55	0.54	0.78	0.72	**0.75**
S-S-I	0.26	0.26	0.24	0.52	0.57	0.55	0.79	0.71	**0.75**
F-S-S	0.14	0.18	0.15	0.60	0.60	**0.58**	0.74	0.61	0.63
S-S-S	0.13	0.16	0.14	0.53	0.58	0.56	0.74	0.61	0.63

TABLE III: The experiment about the number of layers in F-S-I. The original numbers are 4 Freeze, 4 Shared, and 4 Individual. For example, 2-4-6 denotes 2 Freeze, 4 Shared, and 6 Individual.

	Category			Sentiment			Sequence Classification		
	Precision	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*
F-S-I/ 4-4-4	0.32	0.28	0.28	0.53	0.55	0.54	0.78	0.72	0.75
2-4-6	0.30	0.28	0.27	0.52	0.58	0.55	0.79	0.74	**0.76**
4-6-2	0.27	0.25	0.24	0.52	0.57	0.55	0.78	0.68	0.71
6-4-2	0.24	0.23	0.22	0.53	0.58	0.55	0.77	0.68	0.71
8-2-2	0.25	0.23	0.22	0.51	0.56	0.53	0.75	0.64	0.67
2-6-4	0.30	0.27	0.27	0.61	0.58	**0.57**	0.78	0.71	0.74
1-6-5	0.33	0.28	0.27	0.55	0.59	0.56	0.79	0.73	0.75
1-7-4	0.40	0.30	**0.30**	0.52	0.57	0.55	0.80	0.73	**0.76**
1-8-3	0.25	0.26	0.24	0.64	0.58	0.56	0.80	0.72	0.75
1-5-6	0.30	0.28	0.27	0.52	0.57	0.55	0.80	0.72	**0.76**
1-4-7	0.32	0.29	0.28	0.54	0.58	0.56	0.80	0.71	0.75
1-3-8	0.34	0.25	0.25	0.53	0.57	0.54	0.80	0.73	**0.76**
2-2-8	0.37	0.27	0.27	0.52	0.57	0.54	0.79	0.73	**0.76**

splits of data. The batch size is set as 32. F1-Score are used as metric for evaluation due to the imbalance in the dataset.

B. Results and Discussion

1) MTL on different combinations: The results for all the models explained in Section IV-C about three different classification tasks are shown in Table II. The results are the average of 3 runs. The results show that almost all the multi-task learning models performed better than single task learning (Baseline in Table II).

S-S-S and F-S-S models obtained lower scores than the baseline model in category classification, although these models obtained higher scores in the other two tasks. The common characteristics between them are that both of them share the middle and the top layers. As compared with the F1-scores of F-S-I and S-S-I, the scores of F-S-S and S-S-S were slightly better in the sentiment classification task (0.54 and 0.55 vs 0.58 and 0.58). On the other hand, for the category classification and sequence classification tasks, F-S-I and S-S-I outperformed F-S-S and S-S-S (e.g., 0.28 and 0.24 vs. 0.15 and 0.14 for the category task on the F1-score). Since there are no task-specific layers on S-S-S because of all sharing, the models failed to learn the task-specific knowledge.

As compared with the baseline, F-S-I shows the improvement of 11% and 24% on the category classification and sequence classifications tasks. The performance of S-S-I is better by 7%, 1%, and 24% for the category classification, sentiment classification, and sequence classification respectively. Both the models utilized task-specific top layers: in other words, individual layers. These results support the claim by [18]; the upper layers of BERT generally learn the unique patterns in the task. Hence keeping top layers individual in multi-task learning is effective and appropriate. The difference between F-S-I and S-S-I is in the bottom layers. F-S-I contained frozen bottom layers, whereas S-S-I contained shared bottom layers. The performance of these models was not significantly different. Therefore this result suggests that sharing or freezing the bottom layers does not significantly change the performance. Regarding the top layers, it is better to keep these layers specific to individual tasks since individual patterns are learned on the top layers of BERT while the general knowledge is shared mostly across the middle layers.

2) The suitable number of layers: In the previous section, we compared the same size of layers for each model, namely 4 Freeze, 4 Shared, and 4 Individual. However, the size of layers is not always suitable. In this section, we compare several combinations of the size of layers.

We used the F-S-I model for the evaluation because it was the best model in the previous experiment. Table III shows the result. In the table, the three numbers on the left side denote the size of each layer. For example, a model using 2 layers as bottom layers, 4 layers as middle layers, and 6 layers as top layers is written as 2-4-6. 4-4-4 is the same as the result of F-S-I in Table II. The models with the small size of frozen layers, such as 1-X-X or 2-X-X, tended to be better for the three tasks. In other words, as the number of frozen layers increased, the performance tended to decrease. The decent F1-scores are obtained when six to eight layers are used in either middle or top layers.

For the sequence classification task, the large number of

Authorized licensed use limited to: Stanford University. Downloaded on February 09,2023 at 06:03:59 UTC from IEEE Xplore. Restrictions apply.

layers on top layers provides consistent high results, such as X-X-8. This result implies the sequence classification task requires more individual layers for capturing task-specific patterns. The best overall performance on the average of three tasks was the 1-7-4 combination. To sum up, the large number of shared layers leads to capturing the knowledge of the domain. In addition, individual layers contribute to capturing the knowledge, such as task-specific patterns, of each task in the domain. The combination of them has an important role in the tasks in this paper.

VI. Conclusion

We studied the effect of the multi-task learning model with BERT on three different classification tasks in the aspect-based sentiment analysis domain. In this paper, we compared different configurations for BERT encoder layers. The results suggested that the top layers should be kept specific to individual tasks to capture the individual task patterns. On the other hand, the bottom and middle layers mostly share the general domain knowledge. The best result was obtained when the bottom layers were frozen, the middle layers were shared, and the top layers were kept for individual tasks.

We also investigated the suitable numbers of layers for each layer group: bottom, middle and top. As a result, the performance tended to decrease when the number of frozen layers increased. The result said that the individual layers captured the task-specific patterns, whereas shared layers captured the domain knowledge.

In future work, other data and tasks from several domains will be studied to investigate the effectiveness of our MTL method. In addition, applying hard parameter sharing or other soft parameter sharing methods to another type of BERT is important future work.

References

[1] S. Vandenhende, S. Georgoulis, M. Proesmans, D. Dai, and L. Van Gool, "Revisiting multi-task learning in the deep learning era," *arXiv preprint arXiv:2004.13379*, vol. 2, 2020.

[2] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge & Data Engineering*, no. 01, 2021.

[3] R. Caruana, "Multitask learning," vol. 28, no. 1, pp. 41–75, 1997.

[4] E. Choi, D. Hewlett, J. Uszkoreit, I. Polosukhin, A. Lacoste, and J. Berant, "Coarse-to-fine question answering for long documents," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 209–220.

[5] A. Wilson, A. Fern, S. Ray, and P. Tadepalli, "Multi-task reinforcement learning: a hierarchical bayesian approach," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 1015–1022.

[6] Y. Zhang and Q. Yang, "An overview of multi-task learning," *National Science Review*, vol. 5, no. 1, pp. 30–43, 2018.

[7] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[8] S. Changpinyo, H. Hu, and F. Sha, "Multi-task learning for sequence tagging: An empirical study," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 2965–2977. [Online]. Available: https://aclanthology.org/C18-1251

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[10] R. Satapathy, S. R. Pardeshi, and E. Cambria, "Polarity and subjectivity detection with multitask learning and bert embedding," *Future Internet*, vol. 14, no. 7, p. 191, 2022.

[11] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? an analysis of BERT's attention," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 276–286. [Online]. Available: https://aclanthology.org/W19-4828

[12] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4593–4601. [Online]. Available: https://aclanthology.org/P19-1452

[13] W. Dai, T. Yu, Z. Liu, and P. Fung, "Kungfupanda at SemEval-2020 task 12: BERT-based multi-TaskLearning for offensive language detection," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 2060–2066. [Online]. Available: https://aclanthology.org/2020.semeval-1.272

[14] Q. Jia, J. Cui, Y. Xiao, C. Liu, P. Rashid, and E. Gehringer, "All-in-one: Multi-task learning bert models for evaluating peer assessments," in *Proceedings of The 14th International Conference on Educational Data Mining (EDM 2021)*. International Educational Datamining Society, 2021, pp. 525–532.

[15] G. Balikas, S. Moura, and M.-R. Amini, "Multitask learning for fine-grained twitter sentiment analysis," in *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, 2017, pp. 1005–1008.

[16] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 160–167.

[17] Y. Peng, Q. Chen, and Z. Lu, "An empirical study of multi-task learning on bert for biomedical text mining," in *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, 2020, pp. 205–214.

[18] J. Vig and Y. Belinkov, "Analyzing the structure of attention in a transformer language model," in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 63–76. [Online]. Available: https://aclanthology.org/W19-4808

[19] B. Hoover, H. Strobelt, and S. Gehrmann, "exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, Jul. 2020, pp. 187–196. [Online]. Available: https://aclanthology.org/2020.acl-demos.22

[20] H. Cai, R. Xia, and J. Yu, "Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 340–350.

[21] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq *et al.*, "Semeval-2016 task 5: Aspect based sentiment analysis," in *International workshop on semantic evaluation*, 2016, pp. 19–30.

[22] Z. Fan, Z. Wu, X. Dai, S. Huang, and J. Chen, "Target-oriented opinion words extraction with target-fused neural sequence labeling," 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)*, 2019, pp. 2509–2518.

[23] L. Xu, H. Li, W. Lu, and L. Bing, "Position-aware tagging for aspect sentiment triplet extraction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2339–2349.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

Authorized licensed use limited to: Stanford University. Downloaded on February 09,2023 at 06:03:59 UTC from IEEE Xplore. Restrictions apply.