Multi-task Peer-Review Score Prediction

Jiyi Li[1], **Ayaka Sato**[2], **Kazuya Shimura**[3] **and Fumiyo Fukumoto**[4]
University of Yamanashi, Kofu, Japan
{jyli[1],g17tk008[3],fukumoto[4]}@yamanashi.ac.jp,{t15cs027[2]}@gmail.com

Abstract

Automatic prediction of the peer-review aspect scores of academic papers can be a useful assistant tool for both reviewers and authors. To handle the small size of published datasets on the target aspect of scores, we propose a multi-task approach to leverage additional information from other aspects of scores for improving the performance of the target aspect. Because one of the problems of building multi-task models is how to select the proper resources of auxiliary tasks and how to select the proper shared structures, we thus propose a multi-task shared structure encoding approach that automatically selects good shared network structures as well as good auxiliary resources. The experiments based on peer-review datasets show that our approach is effective and has better performance on the target scores than the single-task method and naïve multi-task methods.

1 Introduction

Automatic prediction of the peer-review aspect scores (e.g. "clarity" and "originality") of academic papers can be a useful assistant tool for both reviewers and authors. On the one hand, because the number of submissions to AI-related international conferences has significantly increased in recent years, it is challenging for the review process. Rejecting some papers with evidently low quality can reduce the workload. On the other hand, suggesting the weak aspects to the authors can also help them improve their papers.

There are several existing works related to the paper review which concentrate on the quality of the review (De Silva and Vance, 2017; Langford and Guzdial, 2015). Huang (2018) et al. predicted the acceptance of a paper only based on a paper's visual appearance (Huang, 2018). Automatic essay scoring (Dong and Zhang, 2016; Dong et al., 2017; Amorim et al., 2018) can be regarded as a related sub-topic that mainly focus on the grammatical and syntactic features in short essays. PeerRead is the first public dataset of scientific peer reviews for research purposes (Kang et al., 2018), which can be used for paper acceptance classification and review aspect score prediction. It provides detailed peer-reviews including the final decisions, the aspect scores such as clarity and originality, and the review contents. It raises two NLP tasks, paper acceptance classification and review aspect score prediction. We focus on the later one in this paper. However, the dataset is relatively small; the set of papers for each review aspect can be different. To improve the performance of aspect score prediction, we propose a solution based on the multi-task learning that can leverage additional rich information from the resources obtained by other aspect scores. We treat the prediction of each aspect as a separate task. The multi-task model for each aspect score has a main-auxiliary manner.

Multi-task methods have been widely utilized in many NLP tasks, such as summarization (Isonuma et al., 2017; Guo et al., 2018), classification (Liu et al., 2017b; Shimura et al., 2019), parsing (Hershcovich et al., 2018), sequence labeling (Lin et al., 2018), and Entity and Relation (Luan et al., 2018). When building a multi-task model, there are two critical issues, i.e., which auxiliary resources (tasks) can be used for sharing useful information and how to share the information among the tasks. In these previous studies, researchers always select specific auxiliary resources, and design handcrafted shared structure in the model for a particular NLP topic.

However, for different datasets and tasks, there may exist other better auxiliary resources and shared structures. We thus propose an approach selecting the shared structures automatically as well as the auxiliary resources that are more beneficial

Proceedings of the First Workshop on Scholarly Document Processing, pages 121–126

https://doi.org/10.18653/v1/P17

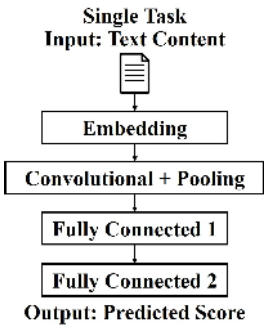

Figure 1: Basic model CNN

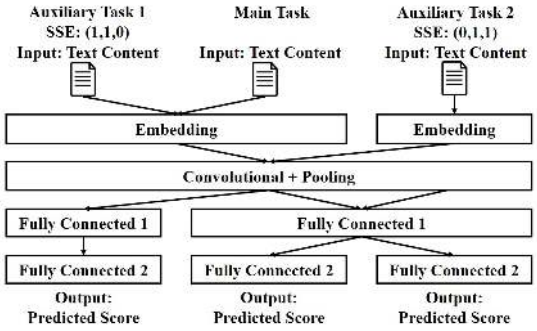

Figure 2: Example of Multi-task CNN with Shared Structure Encoding (SSE)

for the main task. There are diverse parameter sharing manners in the multi-task methods for deep neural networks (Ruder, 2017). How to define the exploration space for automatic selection is a problem. Our approach encodes the multi-task shared structures in the manner of hard parameter sharing and defines the exploration space. We also propose a strategy to search the optimal structures and auxiliaries from the candidate models. It is also flexible to add more auxiliary tasks.

Our approach can be integrated with hyperparameter optimization methods (Snoek et al., 2012) or network architecture search methods (Zoph and Le, 2016) for searching. Furthermore, our method is capable for not only review score prediction but also some other NLP tasks such as text classification. Our main contributions can be summarized as follows. (1). We address an application that predicting the peer-review aspect scores of papers which can be a useful assistant tool for both reviewers and authors. (2). We propose a multi-task shared structure encoding method which automatically selects good shared network structures as well as good auxiliary resources. (3). The experiments based on real paper peer-review datasets show that our approach can build a multi-task model with effective structures and auxiliaries which has better performance than the single-task model and naïve multi-task models.

2 Our Approach

2.1 Preliminary

Peer-review aspect score prediction is a regression problem with text data. We can utilize existing text classification methods (Kim, 2014; Liu et al., 2017a) based on deep neural network for this problem by changing the loss function from cross-entropy for classification to mean squared error for regression. Without loss of generality, we use the basic CNN-based text classification model (Kim, 2014) as the example to facilitate the description of our multi-task approach. Figure 1 shows the architecture of this model for predicting the aspect score. It includes the embedding layer, convolutional and pooling layer, and fully connected layers. The multi-task approach we propose is not limited to be adapted with this model. It can be integrated with similar neural network structures in this example, e.g., XML-CNN (Liu et al., 2017a) and DPCNN (Johnson and Zhang, 2017).

We have n single tasks (i.e., aspect scores) and assume that they have the same network structures with k layers. For each task, we regard it as the main task and search the proper shared structures and auxiliary tasks.

2.2 Multi-task Shared Structures

To automatically search the proper shared structures and auxiliary tasks, we need to define the exploration space. Because it is difficult to mix diverse parameter sharing manners proposed in various multi-task methods (Ruder, 2017), we utilize the typical manner of hard parameter sharing as the starting point to implement our idea. Other manners of parameter sharing will be addressed in future work.

Figure 2 shows an example of the shared structure encoding (SSE) that we propose with three tasks (one main task and two auxiliary tasks). Given a main task t_0, for each auxiliary task t_i, if the jth layer of t_i is shared with t_0, then we encode this shared structure as $l_{ij} = 1$; if the jth layer is not shared, then $l_{ij} = 0$. We do not encode the shared structures among auxiliary tasks to decrease the complexity of the model. It is flexible to add more auxiliary tasks to a model. There are two special cases of this SSE. One is $l_{ij} = 1$ for all aux-

iliary tasks. The corresponding model is equivalent to one single model for all tasks. Another is $l_{ij} = 0$ for all auxiliary tasks. It is equivalent to a single-task model for the main task. In other words, in the search stage, these models are also included. Lu et al. (2017) adaptively generate the feature sharing structure by splitting the network into branches without merging. Its exploration space is a subset of our approach.

Our multi-task approach utilizes a main-auxiliary manner, rather than a manner which equally treats all tasks. The later manner makes a sum of the weighted losses of all tasks and requires a trade-off among the tasks (Sener and Koltun, 2018), which may not be able to reach optimal results for a specific task. In our approach, we thus use every single task as the main task respectively and other tasks as the candidates for auxiliary tasks. It is flexible for us to define all candidate shared structures in the exploration space and decrease the size of the exploration space.

2.2.1 Shared Structure and Auxiliary Task Search

In our search strategy, we denote the number of auxiliary tasks in a model as m, $m \leq n-1$. There are $\binom{n-1}{m}$ combinations of the auxiliary tasks. For each combination of auxiliary tasks, we search the shared structures and select the one with minimized loss. For the selection criterion, because the dataset is too small, we use the loss on both the training set and validation set rather than only using the loss of validation set.

After selecting the shared structures for all combinations of the auxiliary tasks, we select the combination of which the average loss of all candidate shared structures is minimum. For a main task, the number of candidate multi-task models is $\mathcal{N}_m = \binom{n-1}{m} \times 2^{km}$. When $m = n-1$, i.e., using all other tasks as the auxiliary tasks, this number is $\mathcal{N}_{n-1} = 2^{k(n-1)}$. If $m \ll n-1$, then $\mathcal{N}_m \ll \mathcal{N}_{n-1}$.

If $\mathcal{N}_m$ is small, we can explore all candidates. Otherwise, we need to refer some other methods to search in the exploration space, for example, the hyperparameter optimization methods based on Bayesian optimization (Snoek et al., 2012); the network architecture search (NAS) methods based on reinforcement learning (Zoph and Le, 2016; Zoph et al., 2018; Liu et al., 2018). Random search is also possible to be used.

Dataset	Aspects	Train	Valid	Test
ICLR	Clarity	65	8	6
	Originality	72	11	5
	Correctness	64	6	4
	Comparison	27	6	2
	Substance	38	7	2
	Impact	51	9	4
ACL	All six	137	7	7

Table 1: Statistics of Datasets

Settings	CNN	XMN-CNN
Input word vectors	fastText	fastText
Embedding Dimension	200	200
Stride size	1	2
Filter region size	2	2
Feature maps (m)	64	64
Pooling	max pooling	dynamic max pooling
Activation function	ReLu	ReLu
Hidden layers	1024	512
Batch sizes	8	8
Dropout rate 1	0.25	0.25
Dropout rate 2	0.5	0.5
Optimizer	Adam	Adam
Loss function	MSE	MSE
Epoch	40	40

Table 2: Settings of basic models CNN and XML-CNN: Dropout rate 1 is for the embedding layer, and Dropout rate 2 is for the fully connected layers.

3 Experiments

3.1 Experimental Settings

We use the ICLR and ACL datasets in the PeerRead Dataset (Kang et al., 2018) because they provide the scores of the peer-review aspects. Table 1 shows the statistics of these datasets. We utilize the papers which have the scores in some of the six aspects ($n = 6$), i.e., Clarity (cla), Originality (ori), Correctness (cor), Comparison (com), Substance (sub) and Impact (imp). The scale of these scores is from 1 to 5. We utilize the dataset splitting provided by PeerRead. Because not all papers contain all six aspects in the ICLR dataset, the number of papers for each aspect are diverse. For the ground truth, we use the mean score of multiple reviews which is the general method of multiple score aggregation without considering the review bias. Analyzing the review bias among different reviewers is out of the scope of this paper.

Note that although PeerRead contains both paper text and review text, we only used the paper text because the purpose of this work is to predict the aspect scores before review progress. Moreover, because in the PeerRead (Kang et al., 2018) article, the authors utilized the first 1,000 tokens because the paper text was extremely long; and we used

full paper text with our own text pre-processing in the experiments, the results obtained by our experiments and that reported in PeerRead are thus not exactly comparable.

We remove the stop words and use stemming to the words in the papers. The initial word embeddings in the models are pre-trained by fastText (Bojanowski et al., 2016; Joulin et al., 2016) from each dataset. The hyperparameters of the CNN structures for the approaches refer to the common ones used in exiting work (Shimura et al., 2018). Table 2 shows the parameter settings of CNN and XML-CNN, which are used as basic models of the proposed multi-task approach in the paper.

The baselines are as follows.

Single task model: It is equivalent to the case that SSEs of all auxiliary tasks are "000". It uses one network for one aspect score like the models in (Dong and Zhang, 2016; Dong et al., 2017).

All-in-one (Ain1): It builds a single model that the main task and m auxiliary tasks use same network like the models in the PeerRead (Kang et al., 2018). It is equivalent to treating the prediction of all aspects as one task or as a multi-task that SSEs of all auxiliary tasks are "111".

Average performance of all explored Multi-Task models (AMT): It is equivalent to the expectation of the performance if randomly selecting a multi-task model from all candidates.

We select the aspect of Clarity, which has most test data as the main task for the evaluation in this paper. The evaluation metric is the Root Mean Square Error (RMSE). We first verify our approach by using CNN (Kim, 2014) as the basic model. We set $m \in [1, 2, n-1]$. When $m = n-1$, the $\mathcal{N}_m = 8^5$ is very huge. We use random search method by exploring 1000 candidate models and evaluate the mean performance of five times.

3.2 Experimental Results

We first verify whether our SSE method can select a good shared structure for a given combination of auxiliary tasks. Table 3.(a) shows the results in the case of $m = 1$. It shows that our method successfully builds a better model than the single task model and the model in which all tasks completely share with each other. The comparison result with AMT shows our method can select a better shared structure from all candidate structures.

Table 3.(b) shows the results in the case of $m = 2$. Our method can select a better shared structure from all candidate structures. But it cannot always be better than the single task model this time. It is because that the corresponding combinations of auxiliaries are not proper. After using our search strategy to select the combinations of auxiliaries, in 2nd row of Table 4, our method can select the auxiliaries and structures with better performance. In addition, in Table 4, the performance for $m = 2$ is better than $m = 1$, it shows that increasing m is possible to improve the performance. However, a large m results in a large $\mathcal{N}_m$. In the case of $m = 5$, although it is possible to obtain a better model than $m = 1$ or 2 if exploring all $\mathcal{N}_5 = 8^5$ candidate models, only exploring a subset ($\mathcal{N}'_5 = 1000$) cannot reach better performance even though $\mathcal{N}'_5$ has been larger than $\mathcal{N}_2$. Without a better search method, using a small m (e.g., $m = 2$) rather than a large m (e.g., $m = 5$, all

Auxiliary	Our (SSE)	AMT	Ain1
ori	***0.801*** (001)	0.931	1.027
cor	***0.839*** (111)	0.951	0.858
com	***0.792*** (100)	0.913	0.908
sub	***0.782*** (100)	0.916	0.981
imp	***0.831*** (100)	0.924	0.970

(a). $m = 1$

Auxiliaries	Our (SSEs)	AMT	Ain1
ori,cor	*0.881* (001,110)	0.957	1.036
ori,com	*0.946* (111,101)	0.976	1.136
ori,sub	*0.849* (001,101)	0.971	1.211
ori,imp	***0.853*** (001,100)	0.977	1.046
cor,com	0.996 (111,101)	*0.965*	1.226
cor,sub	***0.761*** (101,001)	0.967	1.143
cor,imp	***0.799*** (101,001)	0.965	1.189
com,sub	*0.892* (001,001)	0.979	1.243
com,imp	***0.732*** (101,101)	0.981	0.918
sub,imp	*0.932* (001,101)	0.969	1.087

(b). $m = 2$

Table 3: Results (Performance and SSEs) of shared structure selection for each combination of auxiliary tasks. Main task: "Clarity"; basic model: CNN; dataset: ICLR; metric: RMSE; performance of single task model: 0.849. **Bold** marks the best performance (including performance of the single task model). *Italic* marks the better one between "Our" and "AMT".

m	Our			AMT
	$\mathcal{N}'_m$	Selected (SSEs)	RMSE	
1	40	com (100)	***0.792***	0.927
2	640	ori, imp (101,101)	***0.732***	0.971
5	1000	All (5 times)	***0.841***	1.001

Table 4: Results of selecting both shared structures and auxiliary tasks. Main task: "Clarity"; basic model: CNN; dataset: ICLR; performance of single task model: 0.849. **Bold** marks the best performance. *Italic* marks the better one between "Our" and "AMT".

Changed Settings	m	Our			AMT	Single
		$\mathcal{N}_m$	Selected (SSEs)	RMSE		
Basic model: XML-CNN	1	40	cor (111)	***0.939***	1.144	0.976
	2	640	ori,cor (100,100)	***0.842***	1.201	
Main Task: Originality	1	40	sub (101)	***0.725***	1.032	1.004
	2	640	com,imp(111,001)	***0.887***	1.017	
Dataset: ACL	1	40	cor (101)	***1.296***	1.414	1.332
	2	640	cor,sub (001,100)	***1.237***	1.455	
Embedding: Wikipedia	1	40	com (101)	***1.151***	1.272	1.241
	2	640	com,sub (101,001)	***0.992***	1.280	

Table 5: Results of selecting both shared structures and auxiliary tasks, by changing four settings respectively

other aspects as auxiliaries) is recommended.

Furthermore, we also respectively change the following four settings while keeping other settings unchanged to verify our approach in different conditions, (1). basic model: one of the SOTA text classification methods XML-CNN (Liu et al., 2017a); (2). main task: Originality, besides the clarity aspect, we also show the results when another aspect is the main task; (3). dataset: ACL. (4). embedding: the pre-trained embeddings by fastText are initialized by the embeddings trained from Wikipedia data.

Table 5 shows that our approach can robustly generate better results in different settings. Table 4 and 5 also show that the selected auxiliary tasks and shared structures are diverse in different settings. It would be better to automatically select them rather than manually decide them. For the underlying characteristics of review aspects in this dataset, there is no apparent observation that one aspect is exactly related to the main aspect and must be the auxiliary. Finally, from the results of "originality" aspect in Table 5, it shows that "substance", "comparison" and "impact" support "originality", the selected aspects by SSEs is reasonable and fit human intuitions.

4 Conclusion

In this paper, we focus on the peer-review score prediction for papers. We propose a multi-task shared structure encoding approach which automatically selects good shared network structures as well as good auxiliary resources. There are some issues in the future work, e.g., trying search methods such as network architecture search and finding evidences of the score predictions.

Acknowledgments

This work was partially supported by KDDI Foundation Research Grant Program.

References

Evelin Amorim, Marcia Cançado, and Adriano Veloso. 2018. Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 229–237, New Orleans, Louisiana. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Pali UK De Silva and Candace K Vance. 2017. Preserving the quality of scientific research: peer review of research articles. In *Scientific Scholarly Communication*, pages 73–99. Springer.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Soft layer-specific multi-task summarization with entailment and question generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 687–697, Melbourne, Australia. Association for Computational Linguistics.

Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2018. Multitask parsing across semantic representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 373–385, Melbourne, Australia. Association for Computational Linguistics.

Jia-Bin Huang. 2018. Deep paper gestalt. *arXiv preprint arXiv:1812.08775*.

Masaru Isonuma, Toru Fujino, Junichiro Mori, Yutaka Matsuo, and Ichiro Sakata. 2017. Extractive summarization using multi-task learning with document classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2101–2110, Copenhagen, Denmark. Association for Computational Linguistics.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 562–570, Vancouver, Canada. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

John Langford and Mark Guzdial. 2015. The arbitrariness of reviews, and advice for school administrators. *Communications of the ACM*, 58(4):12–13.

Ying Lin, Shengqi Yang, Veselin Stoyanov, and Heng Ji. 2018. A multi-lingual multi-task architecture for low-resource sequence labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 799–809, Melbourne, Australia. Association for Computational Linguistics.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. 2018. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017a. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124. ACM.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017b. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–10, Vancouver, Canada. Association for Computational Linguistics.

Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogério Schmidt Feris. 2017. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1131–1140.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, pages 525–536.

Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto. 2018. HFT-CNN: Learning hierarchical category structure for multi-label short text categorization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 811–816, Brussels, Belgium. Association for Computational Linguistics.

Kazuya Shimura, Jiyi Li, and Fumiyo Fukumoto. 2019. Text categorization by learning predominant sense of words as auxiliary task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1109–1119, Florence, Italy. Association for Computational Linguistics.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959.

Barret Zoph and Quoc V Le. 2016. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. 2018. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710.