Multitask Learning in Natural Language Processing

Wenxuan Zhou
University of Southern California

Abstract

While deep neural networks have achieved promising results on various tasks, they often suffer from overfitting problem because of data sparsity. Instead of labeling more data, a more data-efficient method is to utilize knowledge from different but similar tasks, which is referred as multitask learning. It has been widely applied to the natural language processing. In this survey, we divide previous efforts into different groups by their network structures (parallel and hierarchical) and their sharing mechanism (manual and learning to learn). We further point out the limitations of current work and indicate a more realistic and promising direction.

1 Introduction

Deep learning provides a powerful mechanism for fitting large amounts of data, and has achieved great successes on wide range of tasks, such as natural language processing, computer vision and robotics. However, deep models usually suffer from data scarcity. Instead of simply labeling more data, another method is to transfer knowledge from other tasks, which is called multitask learning (MTL). Multitask learning acts like regularization. It changes the inductive bias of neural models, by encouraging them to prefer more universal features. It can also be motivated by how humans learn new tasks. For example, in the movie The Karate Kid, the sensei teaches the karate kid seemingly unrelated tasks like waxing a car, which turn out to help his karate skills as well.

Generally, multitask learning refers to all training methods with more than one loss functions, but in this survey, we only focus on learning from multiple datasets. Formally, given several datasets $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^M$, the goal is to train classifiers $f_i : \mathcal{X}_i \rightarrow \mathcal{Y}_i$, which share part of the model parameters. The traditional method of performing multitask learning is to manually select the related tasks and sharing scheme. Based on the sharing scheme, we can divide previous works into three classes, namely shared encoder, hierarchical sharing, and adaptive sharing, as illustrated in Figure 1.

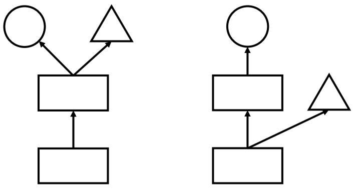

Figure 1: **Overview of the sharing schemes in multitask learning.** Left: Shared Encoder. Middle: Hierarchical Sharing. Right: Adaptive Sharing.

- **Shared Encoder** applies a common encoder to all tasks and add a task-specific decoder to the output of the encoder for each task. It is the simplest method for multitask learning, but its performance is often lower than single-task learning because of limited model capacity. Recently, this method revives after the introduction of universal encoders (Devlin et al., 2018), and has achieved state-of-the-art performance on Glue benchmark (Wang et al., 2018).

- **Hierarchical Sharing** arranges tasks to different layers. It assumes that some tasks are more fundamental and can provide knowledge for other tasks. For example, in natural language processing, part-of-speech (POS) is helpful for dependency parsing (DEP), which is an important feature for named entity recognition (NER). However, this method has a larger search space. We need to decide what hierarchy is optimal and which layer to add the decoders. Assume that we have 3 tasks and 5 layers, then the total number of combinations is $5^3 - 4^3 - 1 = 60$.

Paper	Tasks	Scheme
Liu et al. (2015)	Semantic Classification, Semantic Information Retrieval	Shared Encoder
Luong et al. (2015)	Machine translation, Image Captioning	Adaptive
Guo et al. (2016)	SRL, RE	Shared Encoder
Hashimoto et al. (2016)	POS, Chunk, DEP, Textual Entailmeng	Hierarchical
Strubell et al. (2018)	POS, DEP, SRL	Hierarchical
Keskar et al. (2019)	GLUE, MRC	Shared Encoder
Sanh et al. (2019)	NER, EMD, CR, RE	Hierarchical
Xu et al. (2019)	MRC (multiple datasets)	Shared Encoder
Liu et al. (2019)	GLUE	Shared Encoder + Hierarchical
Stickland and Murray (2019)	GLUE	Adaptive

Table 1: **Some works on applying multitask learning to NLP.** POS, DEP, SRL, RE, MRC, NER, EMD, CR stands for part-of-speech tagging, dependency parsing, semantic role labeling, relation extraction, machine reading comprehension, named-entity recognition, entity mention detection, coreference resolution, respectively.

- **Adaptive Sharing** assumes that some knowledge can be shared while some is private. For example, in semantic role labeling, "delicious food" and "disgusting food" have similar representations since they both belongs to the AM-MOD tag, while in sentiment analysis, they have totally different polarities. Despite its flexibility, the network design heavily relies on experiments.

Some efforts have been summarized in Table 1. In this line of work, the choice of tasks and sharing schemes are mainly decided by intuition and experiments, which requires lots of human labors. To solve this problem, some works study how to automatically search or learn the tasks and sharing scheme, which is also called learning to learn. This line of work can be divided into the following classes:

- **Automatic Weight Learning**. In multitasking, weights of different losses are required to balance the joint loss. They are hard to decide when more than two tasks present. This method automates this process by randomly initialize the weights and learns them during training. Kendall et al. (2018) considers the uncertainty of each task and derives the joint loss by maximizing the Gaussian likelihood of task-dependent uncertainty. Guo et al. (2019) models weights tuning as a multi-bandit problem.

- **Soft Sharing** resembles adaptive sharing, which also assumes some layers should be task-specific while others should be shared. However, soft sharing shares the layers "softly", where each task has its private subnet, but can choose to read the outputs of other subnets. Xiao et al. (2018) proposes a leaky CNN adaptor, which acts to merge outputs of layers in same level of other tasks. Ruder12 et al. (2017) adopts a similar method but further allows the classification layer to read from all layers. Meyerson and Miikkulainen (2017) allows the layers to have different execution orders for different tasks.

Limitations and Future Work. Although multitask learning has achieved impressive results, most of them only learn from few datasets (usually 2, some may use 8 datasets). Given that we have hundreds of datasets targeting for different tasks in NLP, this setting is too weak. Also, training instances are equally sampled from all datasets, even if some datasets are useless. This memory-inefficient property hinders multitasking learning to utilize large number of datasets. Based on these observations, we propose a "Open Multitask Learning" setting, in which the neural model has access to many datasets but is only required to perform well on few of them. We believe this setting is more data-efficent and realistic, while hasn't been well studied.

2 Conclusion

In this survey, we summarized previous efforts on multitask learning and classified them into different lines. We also indicated the limitations of their problem settings and proposed a more realistic setting, which is left as our future work.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2019. Autosem: Automatic task selection and mixing in multi-task learning. *arXiv preprint arXiv:1904.04153*.

Jiang Guo, Wanxiang Che, Haifeng Wang, Ting Liu, and Jun Xu. 2016. A unified architecture for semantic role labeling and relation classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1264–1274.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2016. A joint many-task model: Growing a neural network for multiple nlp tasks. *arXiv preprint arXiv:1611.01587*.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Unifying question answering and text classification via span extraction. *arXiv preprint arXiv:1904.09286*.

Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. 2015. Representation learning using multi-task deep neural networks for semantic classification and information retrieval.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Elliot Meyerson and Risto Miikkulainen. 2017. Beyond shared hierarchies: Deep multitask learning through soft layer ordering. *arXiv preprint arXiv:1711.00108*.

Sebastian Ruder12, Joachim Bingel, Isabelle Augenstein, and Anders Søgaard. 2017. Sluice networks: Learning what to share between loosely related tasks. *stat*, 1050:23.

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6949–6956.

Asa Cooper Stickland and Iain Murray. 2019. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. *arXiv preprint arXiv:1902.02671*.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. *arXiv preprint arXiv:1804.08199*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Liqiang Xiao, Honglun Zhang, Wenqing Chen, Yongkun Wang, and Yaohui Jin. 2018. Learning what to share: Leaky multi-task network for text classification. In *COLING*.

Yichong Xu, Xiaodong Liu, Yelong Shen, Jingjing Liu, and Jianfeng Gao. 2019. Multi-task learning with sample re-weighting for machine reading comprehension. 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 2644–2655.