Meta Learning and Its Applications to Natural Language Processing

Hung-yi Lee
National Taiwan University
hungyilee@ntu.edu.tw

Ngoc Thang Vu
University of Stuttgart
thangvu@ims.uni-stuttgart.de

Shang-Wen Li
Amazon AI
shangwel@amazon.com

1 Brief Description

Deep learning based natural language processing (NLP) has become the mainstream of research in recent years and significantly outperforms conventional methods. However, deep learning models are notorious for being data and computation hungry. These downsides limit such models' application from deployment to different domains, languages, countries, or styles, since collecting in-genre data and model training from scratch are costly. The long-tail nature of human language makes challenges even more significant.

Meta-learning, or 'Learning to Learn', aims to learn better learning algorithms, including better parameter initialization, optimization strategy, network architecture, distance metrics, and beyond. Meta-learning has been shown to allow faster fine-tuning, converge to better performance, and achieve outstanding results for few-shot learning in many applications. Meta-learning is one of the most important new techniques in machine learning in recent years. There is a related tutorial in ICML 2019[1] and a related course at Stanford[2], but most of the example applications given in these materials are about image processing. It is believed that meta-learning has excellent potential to be applied in NLP, and some works have been proposed with notable achievements in several relevant problems, e.g., relation extraction, machine translation, and dialogue generation and state tracking. However, it does not catch the same level of attention as in the image processing community.

In the tutorial, we will first introduce Meta-learning approaches and the theory behind them, and then review the works of applying this technology to NLP problems. Table 1 summarizes the content this tutorial will cover. This tutorial intends to facilitate researchers in the NLP community to understand this new technology better and promote more research studies using this new technology.

2 Type of the tutorial

The type of tutorial is **Cutting-edge**. Meta-learning is a newly emerging topic. The area of natural language processing has seen a growing number of papers about Meta-learning. However, there is no tutorial systematically reviewing relevant works at ACL/EMNLP/NAACL/EACL/COLING.

3 Tutorial Structure and Content

A typical machine learning algorithm, e.g., deep learning, can be considered as a sophisticated function. The function takes training data as input and a trained model as output. Today the learning algorithms are mostly human-designed. These algorithms have already achieved significant progress towards artificial intelligence, but still far from optimal. Usually, these algorithms are designed for one specific task and need a lot of labeled training data. One possible method that could overcome these challenges is meta-learning, also known as 'Learning to Learn', which aims to learn the learning algorithm. In the image processing research community, meta-learning has shown to be successful, especially few-shot learning. It has recently also been widely adopted to a wide range of NLP applications, which usually suffer from data scarcity. This tutorial has two parts. In part I, we will introduce several meta-learning approaches (**estimated 1.5 hours**). In part II, we will highlight the applications of the meta-learning methods to NLP (**estimated 1.5 hours**).

3.1 Part I - Introduction of Meta Learning

We will start with the problem definition of meta-learning, and then introduce the most well-known meta-learning approaches below.

[1] https://sites.google.com/view/icml19metalearning

[2] http://cs330.stanford.edu/

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 15–20, August 1st, 2021.

Table 1: Referrence of NLP tasks using different meta-learning methods.

	(A) **Learning to initialize**	(B) **Learning to compare**	(C) **Other**
Text Classification	(Dou et al., 2019) (Bansal et al., 2019)	(Yu et al., 2018) (Tan et al., 2019) (Geng et al., 2019) (Sun et al., 2019)	Learning the learning algorithm: (Wu et al., 2019)
Sequence Labelng	(Wu et al., 2020)	(Hou et al., 2020)	
Machine Translation	(Gu et al., 2018) (Indurthi et al., 2020)		
Speech Recognition	(Hsu et al., 2020) (Klejch et al., 2019) (Winata et al., 2020a) (Winata et al., 2020b)		Learning to optimize: (Klejch et al., 2018) Network architecture search: (Chen et al., 2020b) (Baruwa et al., 2019)
Relation Classification	(Obamuyide and Vlachos, 2019) (Bose et al., 2019) (Lv et al., 2019) (Wang et al., 2019)	(Ye and Ling, 2019) (Chen et al., 2019a) (Xiong et al., 2018) (Gao et al., 2019)	
Dialogue	(Qian and Yu, 2019) (Madotto et al., 2019) (Mi et al., 2019)		Learning to optimize: (Chien and Lieow, 2019)
Parsing	(Guo et al., 2019) (Huang et al., 2018)		
Word Embedding	(Hu et al., 2019)	(Sun et al., 2018)	
Multi-model		(Eloff et al., 2019)	Learning the learning algorithm: (Surís et al., 2019)
Keyword Spotting	(Chen et al., 2020a)		Network architecture search: (Mazzawi et al., 2019)
Sound Event Detection		(Shimada et al., 2020) (Chou et al., 2019)	
Voice Cloning			Learning the learning algorithm: (Chen et al., 2019b) (Serrà et al., 2019)

3.1.1 Learning to Initialize

Gradient descent is the core learning algorithm for deep learning. Most of the components in gradient descent are handcrafted. First, we have to determine how to initialize network parameters. Then the gradient is computed to update the parameters, and the learning rates are determined heuristically. Determining these components usually need experience, intuition, and trial and error. With meta-learning, those hyperparameters can be learned from data automatically. Among these series of approaches, learning a set of parameters to initialize gradient descent, or learning to initialize, is already widely studied.

Column (A) of Table 1 lists the NLP papers using learning to initialize. Learning to initialize is the most widely applied meta-learning approach in NLP today. The idea of learning to initialize spreads quickly in NLP probably because the idea of looking for better initialization is already widespread before the development of meta-learning. The researchers of NLP have applied lots of different transfer learning techniques to find a set of good initialization parameters for a specific task from its related tasks. Here we will not only introduce learning to initialize but also compare its difference with typical transfer learning.

3.1.2 Learning to Compare

Besides the gradient descent-based learning algorithm, the testing examples' labels are determined by their similarity to the training examples in some learning algorithms. In this category, methods to compute the distance between two data points are crucial. Therefore, a series of approaches have been proposed to learn the distance measures for the learning algorithms. This category of approaches is also known as metric-based approaches.

Column (B) of Table 1 lists the NLP papers using learning to compare. Natural language is intrinsically represented as sophisticated sequences. Comparing the similarity of two sequences is not trivial, and widely used handcrafted measures, such as, Euclidean distance, cannot be directly applied, which motivates the research of learning to compare in NLP.

3.1.3 Other Methods

Although the above two methods dominate the NLP field at the moment, other meta-learning approaches have also shown their potential. For example, besides parameter initialization, other gradient descent components such as learning rates and network structures can also be learned. In addition to learning the components in the existing learning algorithm, some attempts even make the machine invent an entirely new learning algorithm beyond gradient descent. There is already some effort towards learning a function that directly takes training data as input and outputs network parameters for the target task. Column (C) of Table 1 lists these methods.

3.2 Part II - Applications to NLP tasks

There is a growing number of studies applying meta-learning techniques to NLP applications and achieving excellent results. In the second part of the tutorial, we will review these studies. Here we summarize these studies by categorizing their applications. Please refer to Table 1 for a detailed list of studies we plan to cover in the tutorial.

3.2.1 Text Classification

Text classification has a vast spectrum of applications, such as sentiment classification and intent classification. The meta-learning algorithms developed for image classification can be applied to text classification with slight modification to incorporate domain knowledge in each application (Yu et al., 2018; Tan et al., 2019; Geng et al., 2019; Sun et al., 2019; Dou et al., 2019; Bansal et al., 2019).

3.2.2 Sequence Labeling

Using a meta-learning algorithm to make the model fast adapt to new languages or domains is also useful for sequence labeling like name-entity recognition (NER) (Wu et al., 2020) and slot tagging (Hou et al., 2020). However, the typical meta-learning methods developed on image classification may not be optimal for sequence labeling because sequence labeling benefits from modeling the dependencies between labels, which is not leveraged in typical meta-learning methods. Techniques, such as the collapsing labeling mechanism, are proposed to optimize meta-learning for sequence labeling problem (Hou et al., 2020).

3.2.3 Automatic Speech Recognition and Neural Machine Translation

Automatic speech recognition (ASR), Neural machine translation (NMT), and speech translation require a large amount of labeled training data. Collecting such data is cost-prohibitive. To facilitate the expansion of such systems to new use cases, meta-learning is applied in these systems for the fast adaptation to new languages in NMT (Gu et al., 2018) and ASR (Hsu et al., 2020; Chen et al., 2020b), and fast adaptation to new accents (Winata et al., 2020b), new speakers (Klejch et al., 2019, 2018), code-switching (Winata et al., 2020a) in ASR.

3.2.4 Relation Classification and Knowledge Graph Completion

The typical supervised learning approaches for relation classification and link prediction for knowledge graph completion require a large number of training instances for each relation. However, only about 10% of relations in Wikidata have no more than ten triples (Vrandei and Krtzsch, 2014), so many long-tail relations suffer from data sparsity. Therefore, meta-learning has been applied to the relation classification and knowledge graph completion to improve the performance of the relations with limited training examples (Obamuyide and Vlachos, 2019; Bose et al., 2019; Lv et al., 2019; Wang et al., 2019; Ye and Ling, 2019; Chen et al., 2019a; Xiong et al., 2018; Gao et al., 2019).

3.2.5 Task-oriented Dialogue and Chatbot

Domain adaptation is an essential task in dialog system building because modern personal assistants, such as Alexa and Siri, are composed of thousands of single-domain task-oriented dialog systems. However, training a learnable model for a task requires a large amount of labeled in-domain data, and collecting and annotating training data for the tasks is costly since it involves real user interactions. Therefore, researchers apply meta-learning to learn from multiple rich-resource tasks and adapt the meta-learned models to new domains with minimal training samples for dialog response generation (Qian and Yu, 2019) and dialogue state tracking (DST) (Huang et al., 2020).

Also, training personalized chatbot that can mimic speakers with different personas is useful but challenging. Collecting many dialogs involving a specific persona is expensive, while it is challenging to capture a persona using only a few conversations. Thus, meta-learning comes into play for learning persona with few-shot example conversations (Madotto et al., 2019).

4 Diversity

As the main applications of the meta-learning approaches are to find better metrics, model architec-

tures, or initializations such that the meta-trained model can generalize well in new tasks with limited data, the approach is often used at efficient knowledge transferring between domains and languages, and has seen many promising results. Meta-learning has the potential to democratize the progress of machine learning and NLP for different domains, languages, and countries in a scalable way.

5 Prerequisites for the attendees

The attendees have to understand derivatives as found in introductory Calculus and understand basic machine learning concepts such as classification, model optimization, and gradient descent.

6 Reading list

We encourage the audience to read the papers of some well-known meta-learning techqnieus before the tutorial, which are listed below.

- Learning to Initialize (Finn et al., 2017)
- Learning to Compare (Snell et al., 2017; Vinyals et al., 2016)
- Other Methods (Ravi and Larochelle, 2017; Andrychowicz et al., 2016)

7 Biographies of Presenters

Hung-yi Lee[3] received the M.S. and Ph.D. degrees from National Taiwan University (NTU), Taipei, Taiwan, in 2010 and 2012, respectively. From September 2012 to August 2013, he was a postdoctoral fellow in Research Center for Information Technology Innovation, Academia Sinica. From September 2013 to July 2014, he was a visiting scientist at the Spoken Language Systems Group of MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). He is currently an associate professor of the Department of Electrical Engineering of National Taiwan University, with a joint appointment at the Department of Computer Science Information Engineering of the university. His research focuses on machine learning (especially deep learning), speech processing and natural language processing. He owns a YouTube channel teaching deep learning (in Mandarin) with more than **5M** views and **60k** subscribers.

Ngoc Thang Vu[4] received his Diploma (2009) and PhD (2014) degrees in computer science from Karlsruhe Institute of Technology, Germany. From 2014 to 2015, he worked at Nuance Communications as a senior research scientist and at Ludwig-Maximilian University Munich as an acting professor in computational linguistics. In 2015, he was appointed assistant professor at University of Stuttgart, Germany. Since 2018, he has been a full professor at the Institute for Natural Language Processing in Stuttgart. His main research interests are natural language processing (esp. speech recognition and dialog systems) and machine learning (esp. deep learning) for low-resource settings.

Shang-Wen Li[5] is a senior Applied Scientist at Amazon AI. His research focuses on spoken language understanding, dialog management, and natural language generation. His recent interest is transfer learning for low-resourced conversational bots. He earned his Ph.D. from MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) in 2016. He received M.S. and B.S. from National Taiwan University. Before joining Amazon, he also worked at Apple Siri researching conversational AI.

8 Open access

We will allow the publication of our slides and video recording of the tutorial in the ACL Anthology.

References

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. 2016. Learning to learn by gradient descent by gradient descent. In *NIPS*.

Trapit Bansal, Rishikesh Jha, and Andrew McCallum. 2019. Learning to few-shot learn across diverse natural language classification tasks. In *arXiv*.

Ahmed Baruwa, Mojeed Abisiga, Ibrahim Gbadegesin, and Afeez Fakunle. 2019. Leveraging end-to-end speech recognition with neural architecture search. In *IJSER*.

Avishek Joey Bose, Ankit Jain, Piero Molino, and William L. Hamilton. 2019. Meta-graph:few shot link prediction via meta learning. In *arXiv*.

Mingyang Chen, Wen Zhang, Wei Zhang, Qiang Chen, and Huajun Chen. 2019a. Meta relational learning for few-shot link prediction in knowledge graphs. In *EMNLP*.

Yangbin Chen, Tom Ko, Lifeng Shang, Xiao Chen, Xin Jiang, and Qing Li. 2020a. An investigation of few-shot learning in spoken term classification. In *INTERSPEECH*.

Yi-Chen Chen, Jui-Yang Hsu, Cheng-Kuang Lee, and Hung yi Lee. 2020b. DARTS-ASR: Differentiable architecture search for multilingual speech recognition and adaptation. In *INTERSPEECH*.

[3] https://speech.ee.ntu.edu.tw/~hylee/index.html

[4] https://www.ims.uni-stuttgart.de/en/institute/team/Vu-00002

[5] https://scholar.google.com/citations?user=wFI97HUAAAAJ

Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C. Cobo, Andrew Trask, Ben Laurie, Caglar Gulcehre, Aäron van den Oord, Oriol Vinyals, and Nando de Freitas. 2019b. Sample efficient adaptive text-to-speech. In *ICLR*.

Jen-Tzung Chien and Wei Xiang Lieow. 2019. Meta learning for hyperparameter optimization in dialogue system. In *INTERSPEECH*.

Szu-Yu Chou, Kai-Hsiang Cheng, Jyh-Shing Roger Jang, and Yi-Hsuan Yang. 2019. Learning to match transient sound events using attentional similarity for few-shot sound recognition. In *ICASSP*.

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

Ryan Eloff, Herman A. Engelbrecht, and Herman Kamper. 2019. Multimodal one-shot learning of speech and images. In *ICASSP*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *AAAI*.

Ruiying Geng, Binhua Li, Yongbin Li, Xiaodan Zhu, Ping Jian, and Jian Sun. 2019. Induction networks for few-shot text classification. In *EMNLP*.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor O.K. Li. 2018. Meta-learning for low-resource neural machine translation. In *EMNLP*.

Daya Guo, Duyu Tang, Nan Duan, Ming Zhou, and Jian Yin. 2019. Coupling retrieval and meta-learning for context-dependent semantic parsing. In *ACL*.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. In *ACL*.

Jui-Yang Hsu, Yuan-Jui Chen, and Hung yi Lee. 2020. Meta learning for end-to-end low-resource speech recognition. In *ICASSP*.

Ziniu Hu, Ting Chen, Kai-Wei Chang, and Yizhou Sun. 2019. Few-shot representation learning for out-of-vocabulary words. In *ACL*.

Po-Sen Huang, Chenglong Wang, Rishabh Singh, Wen tau Yih, and Xiaodong He. 2018. Natural language to structured query generation via meta-learning. In *NAACL*.

Yi Huang, Junlan Feng, Min Hu, Xiaoting Wu, Xiaoyu Du, and Shuo Ma. 2020. Meta-reinforced multi-domain state generator for dialogue systems. In *ACL*.

Sathish Indurthi, Houjeung Han, Nikhil Kumar Lakumarapu, Beomseok Lee, Insoo Chung, Sangha Kim, and Chanwoo Kim. 2020. Data efficient direct speech-to-text translation with modality agnostic meta-learning. In *ICASSP*.

Ondřej Klejch, Joachim Fainberg, and Peter Bell. 2018. Learning to adapt:a meta-learning approach for speaker adaptation. In *INTERSPEECH*.

Ondřej Klejch, Joachim Fainberg, Peter Bell, and Steve Renals. 2019. Speaker adaptive training using model agnostic meta-learning. In *ASRU*.

Xin Lv, Yuxian Gu, Xu Han, Lei Hou, Juanzi Li, and Zhiyuan Liu. 2019. Adapting meta knowledge graph information for multi-hop reasoning over few-shot relations. In *EMNLP*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *ACL*.

Hanna Mazzawi, Xavi Gonzalvo, Aleks Kracun, Prashant Sridhar, Niranjan Subrahmanya, Ignacio Lopez Moreno, Hyun Jin Park, and Patrick Violette. 2019. Improving keyword spotting and language identification via neural architecture search at scale. In *INTERSPEECH*.

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

Abiola Obamuyide and Andreas Vlachos. 2019. Model-agnostic meta-learning for relation classification with limited supervision. In *ACL*.

Kun Qian and Zhou Yu. 2019. Domain adaptive dialog generation via meta learning. In *ACL*.

Sachin Ravi and Hugo Larochelle. 2017. Optimization as a model for few-shot learning. In *ICLR*.

Joan Serrà, Santiago Pascual, and Carlos Segura. 2019. Blow:a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion. In *NeurIPS*.

Kazuki Shimada, Yuichiro Koyama, and Akira Inoue. 2020. Metric learning with background noise class for few-shot detection of rare sound events. In *ICASSP*.

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *NIPS*.

Jingyuan Sun, Shaonan Wang, and Chengqing Zong. 2018. Memory, show the way:memory based few shot word representation learning. In *EMNLP*.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. 2019. Hierarchical attention prototypical networks for few-shot text classification. In *EMNLP*.

Dídac Surís, Dave Epstein, Heng Ji, Shih-Fu Chang, and Carl Vondrick. 2019. Learning to learn words from narrated video. In *arXiv*.

Ming Tan, Yang Yu, Haoyu Wang, Dakuo Wang, Saloni Potdar, Shiyu Chang, and Mo Yu. 2019. Out-of-domain detection for low-resource text classification tasks. In *EMNLP*.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *NIPS*.

Denny Vrandei and Markus Krtzsch. 2014. Wikidata: A free collaborative knowledge base. In *Communications of the ACM*.

Zihao Wang, Kwun Ping Lai, Piji Li, Lidong Bing, and Wai Lam. 2019. Tackling long-tailed relations and uncommon entities in knowledge graph completion. In *EMNLP*.

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, Peng Xu, and Pascale Fung. 2020a. Meta-transfer learning for code-switched speech recognition. In *ACL*.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, Peng Xu, and Pascale Fung. 2020b. Learning fast adaptation on cross-accented speech recognition. In *INTERSPEECH*.

Jiawei Wu, Wenhan Xiong, and William Yang Wang. 2019. Learning to learn and predict: A meta-learning approach for multi-label classification. In *EMNLP*.

Qianhui Wu, Zijia Lin, Guoxin Wang, Hui Chen, Börje F. Karlsson, Biqing Huang, and Chin-Yew Lin. 2020. Enhanced meta-learning for cross-lingual named entity recognition with minimal resources. In *AAAI*.

Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2018. One-shot relational learning for knowledge graphs. In *EMNLP*.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. In *ACL*.

Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. 2018. Diverse few-shot text classification with multiple metrics. In *ACL*.