Transfer Learning in Natural Language Processing Tutorial

Sebastian Ruder[1], **Matthew Peters**[2], **Swabha Swayamdipta**[3], **Thomas Wolf**[4]
[1] Insight Centre, NUI Galway & Aylien Ltd., Dublin
[2] Allen Institute for Artifical Intelligence
[3] Language Technologies Institute, CMU
[4] Huggingface Inc.
sebastian@ruder.io matthewp@allenai.org
swabha@cs.cmu.edu thomwolf@gmail.com

1 Introduction

The classic supervised machine learning paradigm is based on learning *in isolation*, a single predictive model for a task using a single dataset. This approach requires a large number of training examples and performs best for well-defined and narrow tasks. Transfer learning refers to a set of methods that extend this approach by leveraging data from additional domains or tasks to train a model with better generalization properties.

Over the last two years, the field of Natural Language Processing (NLP) has witnessed the emergence of several transfer learning methods and architectures which significantly improved upon the state-of-the-art on a wide range of NLP tasks (Peters et al., 2018a; Howard and Ruder, 2018; Radford et al., 2018; Devlin et al., 2018).

These improvements together with the wide availability and ease of integration of these methods are reminiscent of the factors that led to the success of pretrained word embeddings (Mikolov et al., 2013) and ImageNet pretraining in computer vision, and indicate that these methods will likely become a common tool in the NLP landscape as well as an important research direction.

We will present an overview of modern transfer learning methods in NLP, how models are pretrained, what information the representations they learn capture, and review examples and case studies on how these models can be integrated and adapted in downstream NLP tasks.

2 Description

The tutorial will start with a broad overview of transfer learning methods following Pan and Yang (2010). As part of this overview, we will also highlight connections to other related and promising directions of research such as meta-learning (Gu et al., 2018), multilingual transfer learning, and continual learning (Lopez-Paz and Ranzato, 2017).

We will then focus on the current most promising area, *sequential transfer learning* where tasks are learned in sequence. Sequential transfer learning consists of two stages: a *pretraining* phase in which general representations are learned on a *source* task or domain followed by an *adaptation* phase during which the learned knowledge is applied to a *target* task or domain.

Our discussion of the pretraining stage will review the main forms of pretraining methods commonly used today. We will try to provide attendants with an overview of what type of information these pretraining schemes are capturing and how pretraining schemes are devised.

In particular, we will review *unsupervised approaches* which aim to model the dataset itself, briefly presenting non-neural approaches (Deerwester et al., 1990; Brown et al., 1993; Blei et al., 2003) before detailing deep neural network approaches like auto-encoding/skip-thoughts models (Dai and Le, 2015; Kiros et al., 2015; Hill et al., 2016; Logeswaran and Lee, 2018) and the current trend of language model-based approaches (Dai and Le, 2015; Peters et al., 2018a; Howard and Ruder, 2018; Radford et al., 2018; Devlin et al., 2018). We will then describe *supervised approaches* which make use of large annotated datasets (Zoph et al., 2016; Yang et al., 2017; Wieting et al., 2016; Conneau et al., 2017; McCann et al., 2017) before turning to *distant supervision approaches* which use heuristics to automatically label datasets (Mintz et al., 2009; Severyn and Moschitti, 2015; Felbo et al., 2017; Yang et al., 2017).

Our review of distant supervision approaches will aim to provide attendants with a sense of how they can design heuristics that can automatically provide supervision in their own applications. Last

Proceedings of NAACL HLT 2019: Tutorial Abstracts, pages 15–18
Minneapolis, Minnesota, June 2, 2019.

but not least, we will highlight the use of multi-task learning for pretraining (Subramanian et al., 2018; Cer et al., 2018; Devlin et al., 2018).

This review of pretraining approaches will provide recommendations and discuss trade-offs of pretraining tasks based on our own experiments and recent studies (Zhang and Bowman, 2018; Anonymous, 2019).

We will then shed some light on what the learned representations can and cannot capture based on recent studies (Conneau et al., 2018; Peters et al., 2018b). We will discuss trade-offs between different modelling architectures and highlight the capabilities and deficiencies of individual models.

In the second part of the tutorial, we will focus on the second phase of sequential training, the *adaptation* phase as well as downstream applications. The adaptation phase involves a growing panel of methods:

Architecture modifications can range from a few additional embeddings to additional layers on top of the pre-trained to the insertion of intervening layers or modules inside the pre-trained model.

Optimization schedules for the adaptation phase can involve fine-tuning a varying portion of the pre-trained model (Long et al., 2015; Felbo et al., 2017; Howard and Ruder, 2018) with specifically designed regularization (Wiese et al., 2017; Kirkpatrick et al., 2017) or even fine-tuning in sequence a model on a series of datasets using several training objectives. We will summarize current trends in adapting pre-trained model to target tasks while highlighting best practices when they can be identified.

We will then focus on a selection of downstream applications such as classification (Howard and Ruder, 2018), natural language generation, structured prediction (Swayamdipta et al., 2018) or other classification tasks (Peters et al., 2018a; Devlin et al., 2018). This part will comprise hands-on examples designed around representative tasks and typical transfer learning schemes as detailed before. We will aim to demonstrate through practical examples how NLP researchers and practitioners can adapt these models to their own applications and provide them with a set of guidelines for practical usage.

Finally, we will present open problems, challenges, and directions in transfer learning for NLP.

3 Outline

This tutorial will be 3 hours long.

1. **Introduction** (15 minutes long): This section will introduce the theme of the tutorial: how transfer learning is used in current NLP. It will position sequential transfer learning among different transfer learning areas.

2. **Pretraining** (35 minutes): We will discuss unsupervised, supervised, and distantly supervised pretraining methods. As part of the unsupervised methods, we will also highlight seminal NLP approaches, such as LSA and Brown clusters.

3. **What do the representations capture** (20 minutes): Before discussing how the pretrained representations can be used in downstream tasks, we will discuss ways to analyze the representations and what properties they have been observed to capture.

4. **Break** (20 minutes)

5. **Adaptation** (30 minutes): In this section, we will present several ways to adapt these representations, feature extraction and fine-tuning. We will discuss practical considerations such as learning rate schedules, architecture modifications, etc.

6. **Down-stream applications** (40 minutes): In this section, we will highlight how pretrained representations have been used in different downstream tasks, such as text classification, natural language generation, structured prediction, among others. We will present hands-on examples and discuss best practices for each category of tasks.

7. **Open problems and directions** (20 minutes): In this final section, we will provide an outlook into the future. We will highlight both open problems and point to future research directions.

4 Prerequisites

- Machine Learning: Basic knowledge of common recent neural network architectures like RNN, CNN, and Transformers.

- Computational linguistics: Familiarity with standard NLP tasks such as text classification, natural language generation, and structured prediction.

5 Tutorial instructor information

Sebastian Ruder Sebastian Ruder is a research scientist at DeepMind. His research focuses on transfer learning in NLP. He has published widely read reviews of related areas, such as multi-task learning and cross-lingual word embeddings and co-organized the NLP Session at the Deep Learning Indaba 2018.

Matthew Peters Matthew Peters is a research scientist at AI2 focusing on large scale representation learning for NLP.

Swabha Swayamdipta Swabha Swayamdipta is a PhD student at the Language Technologies Institute at Carnegie Mellon University (currently a visiting student at University of Washington). Her primary research interests are developing efficient algorithms for structured prediction, with a focus on incorporating inductive biases from syntactic sources.

Thomas Wolf Thomas Wolf leads the Science Team at Huggingface, a Brooklyn-based startup working on open-domain dialog. He has open-sourced several widely used libraries for coreference resolution and transfer learning models in NLP and maintains a blog with practical tips for training large-scale transfer-learning and meta-learning models. His primary research interest is Natural Language Generation.

6 Audience size estimate

Due to the broad appeal and relevancy of the content of our tutorial, we expect a large audience, around 200 people.

References

Anonymous. 2019. Looking for ELMo's friends: Sentence-Level Pretraining Beyond Language Modeling.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022.

Peter F. Brown, Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, and Nicole Limtiaco. 2018. Universal Sentence Encoder.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. In *Proceedings of ACL 2018*.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised Sequence Learning. *Advances in Neural Information Processing Systems (NIPS '15)*.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Jiatao Gu, Yong Wang, Yun Chen, Kyunghyun Cho, and Victor O K Li. 2018. Meta-Learning for Low-Resource Neural Machine Translation. In *Proceedings of EMNLP 2018*.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. In *NAACL-HLT*.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of ACL 2018*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *PNAS*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-Thought Vectors. In *Proceedings of NIPS 2015*.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *Proceedings of ICLR 2018*.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning Transferable Features with Deep Adaptation Networks. In *Proceedings of ICML*, volume 37, Lille, France.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient Episodic Memory for Continuum Learning. In *Proceedings of NIPS 2017*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in Translation: Contextualized Word Vectors. In *Advances in Neural Information Processing Systems*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, pages 1003–1011.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, Wen-tau Yih, Paul G Allen, and Computer Science. 2018b. Dissecting Contextual Word Embeddings: Architecture and Representation. In *Proceedings of EMNLP 2018*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Aliaksei Severyn and Alessandro Moschitti. 2015. UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 464–469.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. In *Proceedings of ICLR 2018*.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. Syntactic Scaffolds for Semantic Structures. In *Proceedings of EMNLP 2018*.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural Domain Adaptation for Biomedical Question Answering. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards Universal Paraphrastic Sentence Embeddings. In *Proceedings of ICLR*.

Jie Yang, Yue Zhang, and Fei Dong. 2017. Neural Word Segmentation with Rich Pretraining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*.

Kelly W Zhang and Samuel R Bowman. 2018. Language Modeling Teaches You More Syntax than Translation Does: Lessons Learned Through Auxiliary Task Analysis. *arXiv preprint arXiv:1809.10040*.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of EMNLP 2016*.