Zero- and Few-Shot NLP with Pretrained Language Models

Iz Beltagy[†] **Arman Cohan**[†*] **Robert L. Logan IV**[‡] **Sewon Min**[*] **Sameer Singh**[‡]
[†]Allen Institute for AI, Seattle, WA [‡]University of California, Irvine
[*]Paul G. Allen School, University of Washington, Seattle, WA
{beltagy, armanc} @allenai.org
{rlogan, sameer} @uci.edu
sewon@cs.washington.edu

1 Introduction

The ability to efficiently learn from little-to-no data is critical to applying NLP to tasks where data collection is costly or otherwise difficult. This is a challenging setting both academically and practically—particularly because training neutral models typically require large amount of labeled data. More recently, advances in pretraining on unlabelled data have brought up the potential of better zero-shot or few-shot learning (Devlin et al., 2019; Brown et al., 2020). In particular, over the past year, a great deal of research has been conducted to better learn from limited data using large-scale language models.

In this tutorial, we aim at bringing interested NLP researchers up to speed about the recent and ongoing techniques for zero- and few-shot learning with pretrained language models. Additionally, our goal is to reveal new research opportunities to the audience, which will hopefully bring us closer to address existing challenges in this domain.

The detailed content of the tutorial is described in Section 2. The tutorial will start by motivating the challenge of learning from limited data, and providing an overview of historical few-shot NLP techniques. The tutorial will then start mainly focusing on recent few-shot learning methods using language models. It will cover methods from manual engineering, better inference algorithms to better tuning methods. We will then discuss the impact of different pretraining objectives, and meta-training strategies. Lastly, we will survey the current landscape of evaluation benchmarks, and their limitations. We will conclude the tutorial by suggesting open questions, and providing coding examples and web-based demonstrations instructing attendees how to easily use these methods using public resources.

2 Tutorial Content and Outline

This tutorial covers methods for zero- and few-shot learning with pretrained language models (LMs). The tutorial will be 3 hours long. Tutorial materials will be made available at: https://github.com/allenai/acl2022-zerofewshot-tutorial.

Introduction - (10 minutes) We will start by motivating why zero- and few-shot learning are important. In many situations, labelled data may be costly or otherwise difficult to procure. Language model finetuning, the predominant training paradigm in use today, exhibits poor performance in low-data regimes (Dodge et al., 2020). Furthermore, as LMs continue to grow in size, so do the associated costs of training and storing separate weights for each downstream task. Recent work on zero- and few-shot learning with pretrained language models can provide a potential solution.

Earlier work - (15 minutes) In the second section, we will review well-established methods for zero- and few-shot learning that do not necessarily use LMs, including data augmentation, semi-supervised learning, consistency training and co-training (Miyato et al., 2017; Clark et al., 2018; Xie et al., 2020; Chen et al., 2020).

Language models as few-shot learners - (20 minutes) In the third section, we will focus on few-shot approaches using LMs without any tuning. The fundamental observation in this section is that, by reformulating tasks as complete-the-sentence problems and potentially including training examples in-context, large pretrained language models can be used to solve NLP tasks without having to resort to finetuning. We will survey a few key papers, notably GPT-3 (Brown et al., 2020), and follow up work demonstrating the limitations of in-context learning (Perez et al., 2021). We will also discuss alternative approaches for calibrating and

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, pages 32 - 37
May 22-27, 2022 ©2022 Association for Computational Linguistics

scoring LM outputs (Zhao et al., 2021; Holtzman et al., 2021; Min et al., 2021).

Prompt-based finetuning - (25 minutes) In the next section, we will discuss prompt-based finetuning, which relaxes the restriction that the LM weights cannot be updated. We will introduce the technique of pattern exploiting training (Schick and Schütze, 2021a,b; Le Scao and Rush, 2021, PET) which utilizes manually written cloze style prompts in conjunction with language model finetuning to attain higher accuracy and improved stability over the finetuning approach proposed by Devlin et al. (2019). We will then discuss a variety of related works that seek to streamline PET (Tam et al., 2021; Logan IV et al., 2021). In particular we will cover methods that try to automate the task of prompt-construction, either in the vocabulary space (Shin et al., 2020; Gao et al., 2021b), or the embedding space (Li and Liang, 2021; Lester et al., 2021; Zhong et al., 2021; Qin and Eisner, 2021). We will contrast these methods with non-tuning methods covered in the previous section, in terms of their performance, memory and computation requirement, amount of required engineering, and more.

Pretraining - (20 minutes) The following section will focus on the factor underlying the success of these methods—language model pretraining. First, we will provide a review of popular language model pretraining objectives and architectures. Topics will include: causal (Radford et al., 2019) vs. masked (Devlin et al., 2019) pretraining, encoder-only (Devlin et al., 2019; Liu et al., 2019) vs. decoder-only (Radford et al., 2019) vs. encoder-decoder architectures (Lewis et al., 2020; Raffel et al., 2020), and the impact of training data (Aghajanyan et al., 2021; Saxton et al., 2019; Gao et al., 2021a).

Meta-training - (25 minutes) Next we will discuss meta-training approaches that train the LM to adapt to zero- and few-shot use cases. A variety of work has demonstrated that transfer learning is extremely effective when trained on a diverse set of tasks and prompts (Wei et al., 2021; Sanh et al., 2021). Furthermore, recent papers propose to learn from *instructions* where the model is given instructions that humans would often read when performing a new task, e.g., in a crowdsourcing task (Efrat and Levy, 2020; Mishra et al., 2021).

Evaluation benchmarks - (25 minutes) We will then discuss few-shot evaluation benchmarks such as FLEX (Bragg et al., 2021), FewNLU (Zheng et al., 2021), The BIG-Bench (BIG-bench collaboration, 2021) and CrossFit (Ye et al., 2021). We will discuss the problems in existing evaluations and how new few-shot evaluation benchmarks were carefully designed to measure a variety of scopes in generalization. We will also cover benchmarks specifically for instruction learning (Efrat and Levy, 2020; Mishra et al., 2021).

Open questions and future work - (20 minutes) The future work section will discuss open questions and future research directions like the need for multilingual evaluation data, challenges in evaluation, reducing engineering efforts and variance and more.

Coding example - (20 minutes) Finally, we will demonstrate code examples for representative few-shot methods using the most widely-used libraries/APIs at the time of the event, such as the Transformers library. This will help audience to easily use publicly available resources for real-world few-shot applications.

3 Type of the Tutorial

This tutorial will cover **cutting-edge** research in zero- and few-shot learning with pretrained language models. This topic has not been previously covered in *CL tutorials.

4 Breadth

The tutorial covers a diverse set of topics related to zero- and few-shot learning including pretraining, prompting, finetuning, evaluation, open research questions, etc. The tutorial also briefly discusses pre-language models work but not in depth. Note that most of the work we will cover is not authored by the presenters.

5 Diversity Considerations

The methods and techniques we are going to present are language-agnostic and can be easily applied to non-English data and tasks. Zero- and few-shot learning can be relevant for low-resource languages and tasks (assuming there exist unlabeled resources to build a pretrained model). The tutorial covers work from diverse groups, both geographically (America, Europe, Asia) and gender.

For instructors, three are senior and two are junior NLP researchers, one is female, and they represent two universities and one industry research lab.

6 Prerequisites

We assume attendees are familiar with:

- Machine Learning: Basic knowledge of common recent neural network architectures, particularly Transformers.

- Computational linguistics: Familiarity with the concept of pretrained language models, as well as standard NLP tasks such as text classification, natural language generation, and question answering.

7 Reading List

Reading the following papers is nice to have but not required for attendance.

- Language Models are Few-Shot Learners (Brown et al., 2020)

- It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners (Schick and Schütze, 2021b)

- Finetuned Language Models are Zero-Shot Learners (Wei et al., 2021)

- FLEX: Unifying Evaluation for Few-Shot NLP (Bragg et al., 2021)

8 Instructors

In alphabetical order,

Iz Beltagy Iz Beltagy is a Research Scientist at AI2 focusing on language modeling, transfer learning, summarization, explainability and efficiency. His research has been recognized with a best paper honorary mention at ACL 2020 and an outstanding paper award at AKBC 2021. He was a co-instructor of the tutorial on "Beyond Paragraphs: NLP for Long Sequences" (NAACL-HLT 2021). He worked as a Teaching Assistant at the University of Texas at Austin teaching computer science.
Email: beltagy@allenai.org
Homepage: beltagy.net

Arman Cohan Arman Cohan is a Research Scientist at AI2 and an Affiliate Assistant Professor at University of Washington, focusing on representation learning and transfer learning methods, as well as NLP applications in specialized domains and scientific text. His research has been recognized with a best paper award at EMNLP 2017, an honorable mention at COLING 2018, and Harold N. Glassman Distinguished Doctoral Dissertation award in 2019. He was a co-instructor of the tutorial on "Beyond Paragraphs: NLP for Long Sequences" (NAACL-HLT 2021).
Email: armanc@allenai.org
Homepage: armancohan.com

Robert L. Logan IV Robert L. Logan IV is a Ph.D. student at the University of California, Irvine, advised by Sameer Singh and Padhraic Smyth. His research focuses on problems at the intersection of information extraction and language modeling, and encompasses recently published work on language model prompting that is relevant to this proposal. He has presented invited talks at the SoCal NLP Symposium (2019), the CHASE-CI Workshop (2019), and the UCI Center for Machine Learning Seminar (2021).
Email: rlogan@uci.edu
Homepage: rloganiv.github.io

Sewon Min Sewon Min is a Ph.D. student in the Paul G. Allen School of Computer Science & Engineering at the University of Washington, advised by Hannaneh Hajishirzi and Luke Zettlemoyer. Her research focuses on natural language understanding, question answering, and knowledge representation. She was a co-instructor of the tutorial on "Beyond Paragraphs: NLP for Long Sequences" (NAACL-HLT 2021), and was a co-organizer of the 3rd Workshop on Machine Reading for Question Answering (EMNLP 2021), Competition on Efficient Open-domain Question Answering (NeurIPS 2020), and Workshop on Structured and Unstructured KBs (AKBC 2020, 2021).
Email: sewon@cs.washington.edu
Homepage: shmsw25.github.io

Sameer Singh Sameer Singh is an Associate Professor of Computer Science at the University of California, Irvine and an Allen AI Fellow at the Allen Institute for AI. He is working on large-scale and interpretable machine learning models for NLP. His work has received paper awards at ACL 2020, AKBC 2020, EMNLP 2019, ACL 2018, and KDD

2016. Sameer has presented a number of tutorials, many relevant to this proposal, such as Deep Adversarial Learning Tutorial at NAACL 2019, Mining Knowledge Graphs from Text Tutorial at WSDM 2018 and AAAI 2017, tutorial on Interpretability and Explanations in NeurIPS 2020 and EMNLP 2020, and tutorial on Robustness in NLP at EMNLP 2021. Sameer has also received teaching awards at UCI.
Email: sameer@uci.edu
Homepage: http://sameersingh.org/

9 Ethical Statement

This tutorial covers work that extensively uses large (up to hundreds of billions of parameters) language models, which are associated with substantial financial and environmental costs (Strubell et al., 2019), as well as other harms (Bender et al., 2021).

References

Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2021. HTLM: Hyper-Text Pre-Training and Prompting of Language Models. *ArXiv preprint*, abs/2107.06955.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

BIG-bench collaboration. 2021. Beyond the imitation game: Measuring and extrapolating the capabilities of language models. *In preparation.*

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. FLEX: Unifying Evaluation for Few-Shot NLP. *ArXiv preprint*, abs/2107.07170.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-supervised sequence modeling with cross-view training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv preprint*, abs/2002.06305.

Avia Efrat and Omer Levy. 2020. The Turking Test: Can Language Models Understand Instructions? *ArXiv preprint*, abs/2010.11982.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2021a. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv preprint*, abs/2101.00027.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021b. Making pre-trained language models better few-shot learners. 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)*, pages 3816–3830, Online. Association for Computational Linguistics.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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)*, pages 4582–4597, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv preprint*, abs/1907.11692.

Robert L. Logan IV, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2021. Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. *ArXiv preprint*, abs/2106.13353.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2021. Noisy Channel Language Model Prompting for Few-Shot Text Classification. *ArXiv preprint*, abs/2108.04106.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-Task Generalization via Natural Language Crowdsourcing Instructions. *ArXiv preprint*, abs/2104.08773.

Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. 2017. Adversarial training methods for semi-supervised text classification. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True Few-Shot Learning with Language Models. *ArXiv preprint*, abs/2105.11447.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *ArXiv preprint*, abs/2110.08207.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. 2021. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models Are Zero-Shot Learners. *ArXiv preprint*, abs/2109.01652.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Yanan Zheng, Jing Zhou, Yujie Qian, Ming Ding, Jian Li, Ruslan Salakhutdinov, Jie Tang, Sebastian Ruder, and Zhilin Yang. 2021. FewNLU: Benchmarking State-of-the-Art Methods for Few-Shot Natural Language Understanding. *ArXiv preprint*, abs/2109.12742.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online. Association for Computational Linguistics.