Recent Advances in Pre-trained Language Models: Why Do They Work and How Do They Work

Cheng-Han Chiang
National Taiwan University
dcml0714@gmail.com

Yung-Sung Chuang
CSAIL, MIT
yungsung@mit.edu

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

1 Brief Description

Deep learning-based natural language processing (NLP) has become mainstream research in recent years and has shown significant improvements over conventional methods. Among all deep learning methods, fine-tuning a self-supervisedly pre-trained language model (PLM) on downstream tasks of interest has become the standard pipeline in NLP tasks. Ever since ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) were proposed in 2018, models fine-tuned from PLMs have dominated numerous leader-boards in various tasks including question answering, natural language understanding, natural language inference, machine translation, and sentence similarity. Aside from applying PLMs on various downstream tasks, many have been delving into understanding the properties and characteristics of PLMs, including the linguistic knowledge encoded in the representations of PLMs, and the factual knowledge the PLMs acquire during pre-training. While it has been three years since PLMs were first proposed, there is no sign of decay in the research related to PLMs.

There were two tutorials focusing self-supervised learning/PLMs: a tutorial in NAACL 2019 (Ruder et al., 2019) and one in AACL 2020[1]. However, given the ever-evolving nature of this realm, it is conceivable that there have been significant progress in the study of PLMs. Specifically, compared with PLMs back in 2019, when they are mostly held by tech giants and used in scientific research, the PLMs nowadays have become more widely adopted in various real-world scenarios by users with different hardware infrastructures and amount of data, and thus posing problems that have never arisen before. Substantial progress, including possible answers to the effectiveness of PLMs and new training paradigms, have been made to allow PLMs better deployed in more realistic settings. Hence, we see it necessary and timely to inform the NLP community about the recent advances in PLMs through a well-organized tutorial.

This tutorial is divided into two parts: **why do PLMs work** and **how do PLMs work**. Table 1 summarizes the content this tutorial will cover. This tutorial intends to facilitate researchers in the NLP community to have a more comprehensive view of the advances in PLMs during recent years, and apply these newly emerging techniques to their domain of interest. As self-supervised learning and PLMs are very popular in these days, we expect our tutorial to have at least **100 attendees**.

Type of the tutorial The type of this tutorial is **Cutting-edge**. We will cover the cutting-edge advances in PLMs which have been flourishing in the NLP community **since 2020**. No tutorial has systematically reviewed any topics that we aim to cover (as listed in Table 1) at ACL/EMNLP/NAACL/EACL/AACL/COLING.

2 Tutorial Structure and Content

Pre-trained language models are language models that are pre-trained on large-scaled corpora in a self-supervised fashion. Traditional self-supervised pre-training tasks mostly involve recovering a corrupted input sentence, or auto-regressive language modeling. After these PLMs are pre-trained, they can be fine-tuned on downstream tasks. Conventionally, these fine-tuning protocol includes adding a linear layer on top of the PLMs and training the whole model on the downstream tasks, or formulating the downstream tasks as a sentence completion task and fine-tuning the downstream tasks in a seq2seq way. Fine-tuning PLMs on downstream tasks often yield exceptional performance gain, which is why PLMs have become so popular.

In the first part of the tutorial (**estimated 40**

[1] https://www.youtube.com/watch?v=Okgeff7PN14

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts, pages 8–15
November 20-21, 2022.

Part	Sub-Category	References	
(I) Why	(A) Empirical	Sinha et al. (2021); Aghajanyan et al. (2021); Chiang and Lee (2022); Sanh et al. (2022); Abdou et al. (2022)	
	(B) Theoretical	Saunshi et al. (2020); Zhang and Hashimoto (2021); Lee et al. (2021); Xie et al. (2022)	
(II) How	Pre-training	(C) Improving existing methods	Micheli et al. (2020); Zhang et al. (2021); Chiang et al. (2020); Izsak et al. (2021); Tay et al. (2022); Wettig et al. (2022); Gao et al. (2022); Hou et al. (2022)
		(D) New methods	Meng et al. (2021); Gao et al. (2021b); Su et al. (2021); Meng et al. (2022); Giorgi et al. (2021); Yan et al. (2021); Chuang et al. (2022); Du et al. (2022); Jiang et al. (2022); Jiang and Wang (2022); Zhang et al. (2022); Jian et al. (2022)
	Fine-tuning	(E) Parameter-efficient fine-tuning	**Adapter/Prefix tuning** Houlsby et al. (2019); Lester et al. (2021); Zhong et al. (2021); Qin and Eisner (2021); Zaken et al. (2021); Li and Liang (2021); Hambardzumyan et al. (2021); Hu et al. (2022); Mahabadi et al. (2021); He et al. (2022); Webson and Pavlick (2022)
		(F) Data-efficient fine-tuning	**Semi-supervised learning** Schick and Schütze (2021a,b); Mi et al. (2021); Lang et al. (2022) **Few-shot learning** Brown et al. (2020); Zhao et al. (2021); Gao et al. (2021a); Vu et al. (2021); Le Scao and Rush (2021); Min et al. (2022b); Cui et al. (2022); Min et al. (2022a); Zheng et al. (2022) **Zero-shot learning** Brown et al. (2020); Sanh et al. (2022); Wei et al. (2022); Xu et al. (2022); Aghajanyan et al. (2022)
		(G) Cross-task transfer	**Inter-mediate task fine-tuning**: Wang et al. (2019); Pruksachatkun et al. (2020); Vu et al. (2020); Phang et al. (2020); Chang and Lu (2021); Vu et al. (2022) **Multi-task learning**: Pilault et al. (2020); Chen et al. (2022)

Table 1: Works in the past three years (from 2020 to 2022) related to our tutorial, to list just a few.

mins), we will summarize some findings that partially explain why PLMs lead to exceptional downstream performance. Some of these results have helped researchers to design better pre-training and fine-tuning methods. In the second part (**estimated 2 hrs 20 mins**), we will introduce recent progress in how to pre-train and fine-tune PLMs; the new techniques covered in this part have been shown to bring significant efficiency in terms of hardware resource, training data, and model parameters while achieving superb performance.

2.1 Part I: Why Do PLMs Work

We will introduce several results that partially explain the effectiveness of PLMs from two aspects: empirical and theoretical.

2.1.1 Empirical Explanations

Many researchers have conducted empirical experiments to show what PLMs have learned during pre-training that aids downstream performance. They mostly construct a special pre-training dataset to examine the transferability of the PLM and draw connect the transferability of the PLM with the characteristic of the pre-training dataset. Block (A) in Table 1 lists the relevant works in recent years.

2.1.2 Theoretical Explanations

Some researchers aim to understand the effectiveness of PLMs by rigorous mathematics, as shown in block (B) in Table 1. Their results range from using statistical models to what PLMs are learning during pre-training, or bounding the generalization errors of the downstream tasks.

2.2 Part II: How Do PLM Work

In this part, we will introduce some new techniques in pre-training and fine-tuning PLMs.

2.2.1 Pre-training

Improving Existing Pre-training Methods Language model pre-training is a resource-hungry task when PLMs were first proposed, requiring a large amount of data, high-end hardware equipment, and lengthy pre-training time. To mitigate the above issues, some research aims to mitigate the above issues, as listed in block (C) in Table 1. Some of these works provide answers about the sufficient amount of data and time to pre-train a PLM that is good enough for downstream tasks, and others provide implementation optimization solutions to cut down the high-end requirement on hardware resources.

New Pre-training Methods Aside from improving existing pre-training methods, there have also been new pre-training methods designed for specific downstream tasks. One of the important topics we aim to cover is applying contrastive learning on language model pre-training. Contrastive learning has been widely applied to pre-training models in computer vision, and we will introduce how contrastive learning has improved PLMs recently. Relevant works are listed in block (D) in Table 1.

2.2.2 Fine-tuning

In this part, we will go through several important fine-tuning protocols that have emerged recently. We categorize them based on the scenario in which the fine-tuning method is used.

Parameter-Efficient Fine-tuning PLMs are enormous, often having millions or even billions of numbers of parameters. In the traditional fine-tuning method, fine-tuning each distinct downstream task produces a fine-tuned model that is are bulky as the original PLM. To reduce the number of parameters for fine-tuning PLMs on downstream tasks, there has been a surge of research on parameter-efficient fine-tuning in NLP, as listed in block (E) in Table 1.

Data-Efficient Fine-tuning A large amount of labeled data is not always available for all downstream tasks, and it is thus important to find a way to apply the PLMs on downstream tasks with limited labeled data. These endeavors are included in block (F) in Table 1. We will discuss how to apply PLMs under different levels of labeled data scarcity.

In case we have a large amount of unlabeled data, **semi-supervised learning** fine-tuning protocols provide effective ways to utilize those unlabeled data and can boost the downstream performance. If those few labeled data are the only thing available, then we must harness the knowledge that the PLM possesses to aid the performance of **few-shot learning**. When we have no labeled data, **zero-shot learning** is still possible in certain cases, if you use the PLM correctly. We will discuss how to make a PLM able to perform well in the zero-shot setting.

Cross-Task Transfer When we have a target task of interest, it is canonical to fine-tune the PLM on the target task. While transferring from PLMs leads to exceptional performance gain, sometimes we want more. This can be achieved by transferring from the PLMs *and* additional guidance from other auxiliary tasks in the form of **intermediate task fine-tuning** or **mutitask learning**. Relevant works are listed in block (G) in Table 1. We will discuss how can cross-task transfer improve the downstream performance together with the power of PLMs.

3 Diversity

PLMs have shown promising results on different domains and have boosted the performance of low-resource languages on many tasks. The **why part** covered in this tutorial has the potential to help individuals of different groups to pre-train their own PLMs more efficiently. The **how part** covered in this tutorial specifically focuses on how to apply PLMs under different real-world scenarios with data scarcity and restricted model parameters, which will enable individuals of different groups to apply PLMs on the domains of interest in a more realistic setting. We see this tutorial to benefit diverse groups in the community.

The tutorial instructors are also diverse: Chuang is a PhD student in the USA, and Lee and Chiang are researchers in Taiwan. Also, Chuang and Chiang are currently Ph.D. students familiar with precise implementations, while Lee is a senior researcher with ten years of experience in human language processing research. This diversity in members enables our team to provide a thorough and detailed yet comprehensive and unified view on PLMs.

4 Prerequisites for Attendees

We expect the attendees to have basic machine learning concepts such as gradient descent and model optimization. The attendees will need to have basic knowledge in linear algebra and calculus to understand some contents in block (B) in Table 1. The attendees should also have minimal knowledge about PLMs and transformer models.

5 Reading List

We encourage attendees to read the following emblematic papers on PLMs and transformer model architectures:

- Transformer model: Vaswani et al. (2017)
- PLMs: Radford et al.; Devlin et al. (2019); Raffel et al. (2019)

6 Biographies of Presenters

Cheng-Han Chiang[2] is a PhD student in National Taiwan University. His research focuses on natural language processing and self-supervised learning, and he has published several papers analyzing PLMs. He has experiences in giving lectures on machine learning topics: he gave a lecture on BERT in AI Summer School 2020[3], and his two lectures on graph neural network (in Mandarin) has received over **68k** views on Youtube [4][5]. He has also served as reviewers in EMNLP 2021, ICLR 2022, NeurIPS 2022, EMNLP 2022, and AAAI 2023.

Yung-Sung Chuang[6] is a PhD student in Electrical Engineering and Computer Science at MIT CSAIL, where he works with Dr. James Glass. His research focuses on learning representations for natural language which helps downstream tasks such as natural language understanding, natural language generation, question answering. He has published several paper in this direction in EMNLP, ACL, NeurIPS, and NAACL. He also has served as reviewers in NeurIPS 2021, ICLR 2022, ICML 2022, NeurIPS 2022, EMNLP 2022, and AAAI 2023.

Hung-yi Lee[7] is 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 deep learning, speech processing, and natural language processing. He owns a YouTube channel teaching deep learning (in Mandarin) with more than **8M** views and **100k** subscribers. He gave tutorials at ICASSP 2018[8], APSIPA 2018, ISCSLP 2018, INTERSPEECH 2019[9], SIPS 2019, INTERSPEECH 2020, ICASSP 2021, ACL 2021. He is the co-organizer of the special session on "New Trends in self-supervised speech processing" at INTERSPEECH (2020), the workshop on "Self-Supervised Learning for Speech and Audio Processing" at NeurIPS (2020), the workshop on "Meta Learning and Its Applications to Natural Language Processing" at ACL (2021), and the workshop on "Self-Supervised Learning for Speech and Audio Processing" at AAAI (2022). He will give the tutorial, "Self-supervised Representation

[2] https://d223302.github.io/

[3] https://ai.ntu.edu.tw/?p=3534

[4] https://www.youtube.com/watch?v=eybCCtNKwzA&ab_channel=Hung-yiLee

[5] https://www.youtube.com/watch?v=M9ht8vsVEw8&ab_channel=Hung-yiLee

[6] https://people.csail.mit.edu/yungsung/

[7] https://speech.ee.ntu.edu.tw/~hylee/index.php

[8] The tutorial has the most participants among the 14 tutorials in ICASSP 2018.

[9] The tutorial also has the most participants among the 8 tutorials in INTERSPEECH 2019.

Learning for Speech Processing" with other researchers at ICASSP 2022 and NAACL 2022. He is the lead guest editor of IEEE JSTSP Special Issue on Self-Supervised Learning for Speech and Audio Processing, member of the Speech and Language Technical Committee (SLTC) of IEEE Signal Processing Society (SPS), SPS Education Center Editorial Board member, and Associate Editor for the SPS Open Journal of Signal Processing.

7 Open Access

We will allow our slides and video recording of the tutorial published in the ACL Anthology. All the slides and videos used in the tutorial, along with the reading lists related with the tutorial, will be updated at this tutorial website[10].

References

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. Word order does matter and shuffled language models know it. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6907–6919, Dublin, Ireland. Association for Computational Linguistics.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. 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 7319–7328.

Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2022. Htlm: Hyper-text pre-training and prompting of language models.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ting-Yun Chang and Chi-Jen Lu. 2021. Rethinking why intermediate-task fine-tuning works. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 706–713, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shuxiao Chen, Koby Crammer, Hangfeng He, Dan Roth, and Weijie J Su. 2022. Weighted training for cross-task learning.

[10] https://d223302.github.io/AACL2022-Pretrain-Language-Model-Tutorial/

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. Pretrained language model embryology: The birth of ALBERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6813–6828, Online. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2022. On the transferability of pre-trained language models: A study from artificial datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. 2022. Prototypical verbalizer for prompt-based few-shot tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7014–7024, Dublin, Ireland. 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.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Jiahui Gao, Hang Xu, Xiaozhe Ren, Philip LH Yu, Xiaodan Liang, Xin Jiang, Zhenguo Li, et al. 2022. Autobert-zero: Evolving bert backbone from scratch. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. 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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. 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 879–895.

Karen Hambardzumyan, Hrant Khachatrian, and Jonathan May. 2021. WARP: Word-level Adversarial ReProgramming. 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 4921–4933, Online. Association for Computational Linguistics.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning.

Le Hou, Richard Yuanzhe Pang, Tianyi Zhou, Yuexin Wu, Xinying Song, Xiaodan Song, and Denny Zhou. 2022. Token dropping for efficient BERT pretraining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3774–3784, Dublin, Ireland. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models.

Peter Izsak, Moshe Berchansky, and Omer Levy. 2021. How to train BERT with an academic budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10644–10652, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2022. Non-linguistic supervision for contrastive learning of sentence embeddings. *arXiv preprint arXiv:2209.09433*.

Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. 2022. Promptbert: Improving bert sentence embeddings with prompts. *arXiv preprint arXiv:2201.04337*.

Yuxin Jiang and Wei Wang. 2022. Deep continuous prompt for contrastive learning of sentence embeddings. *arXiv preprint arXiv:2203.06875*.

Hunter Lang, Monica Agrawal, Yoon Kim, and David Sontag. 2022. Co-training improves prompt-based learning for large language models. *arXiv preprint arXiv:2202.00828*.

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.

Jason D Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. 2021. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems*, 34.

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.

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.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. 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 565–576.

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul Bennett, Jiawei Han, Xia Song, et al. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, 34.

Yu Meng, Chenyan Xiong, Payal Bajaj, Paul N Bennett, Jiawei Han, Xia Song, et al. 2022. Pretraining text encoders with adversarial mixture of training signal generators. In *International Conference on Learning Representations*.

Fei Mi, Wanhao Zhou, Lingjing Kong, Fengyu Cai, Minlie Huang, and Boi Faltings. 2021. Self-training improves pre-training for few-shot learning in task-oriented dialog systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1887–1898, Online and

Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575.

Jonathan Pilault, Christopher Pal, et al. 2020. Conditionally adaptive multi-task learning: Improving transfer learning in nlp using fewer parameters & less data. In *International Conference on Learning Representations*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247.

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, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization.

Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. 2020. A mathematical exploration of why language models help solve downstream tasks. In *International Conference on Learning Representations*.

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.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913.

Yixuan Su, Fangyu Liu, Zaiqiao Meng, Lei Shu, Ehsan Shareghi, and Nigel Collier. 2021. Tacl: Improving bert pre-training with token-aware contrastive learning. *arXiv preprint arXiv:2111.04198*.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2022. Scale efficiently: Insights from pre-training and fine-tuning transformers. In *International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. 2022. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, Dublin, Ireland. Association for Computational Linguistics.

Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. 2021. STraTA: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Alex Wang, Jan Hula, Patrick Xia, Raghavendra Pappagari, R Thomas McCoy, Roma Patel, Najoung Kim, Ian Tenney, Yinghui Huang, Katherin Yu, et al. 2019. Can you tell me how to get past sesame street? sentence-level pretraining beyond language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4465–4476.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. 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. 2022. Finetuned language models are zero-shot learners.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2022. Should you mask 15% in masked language modeling? *arXiv preprint arXiv:2202.08005*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. 2022. Zeroprompt: Scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization. *arXiv preprint arXiv:2201.06910*.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. 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 5065–5075.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*.

Miaoran Zhang, Marius Mosbach, David Ifeoluwa Adelani, Michael A Hedderich, and Dietrich Klakow. 2022. Mcse: Multimodal contrastive learning of sentence embeddings. *arXiv preprint arXiv:2204.10931*.

Tianyi Zhang and Tatsunori B Hashimoto. 2021. On the inductive bias of masked language modeling: From statistical to syntactic dependencies. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5131–5146.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? 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 1112–1125, Online. 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 *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Yanan Zheng, Jing Zhou, Yujie Qian, Ming Ding, Chonghua Liao, Li Jian, Ruslan Salakhutdinov, Jie Tang, Sebastian Ruder, and Zhilin Yang. 2022. FewNLU: Benchmarking state-of-the-art methods for few-shot natural language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 501–516, Dublin, Ireland. Association for Computational Linguistics.

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.