Modular and Parameter-Efficient Fine-Tuning for NLP Models

Sebastian Ruder
Google Research
ruder@google.com

Jonas Pfeiffer
Google Research
jonaspfeiffer@google.com

Ivan Vulić
LTL, University of Cambridge
iv250@cam.ac.uk

Abstract

State-of-the-art language models in NLP perform best when fine-tuned even on small datasets, but due to their increasing size, fine-tuning and downstream usage have become extremely compute-intensive. Being able to efficiently and effectively fine-tune the largest pre-trained models is thus key in order to reap the benefits of the latest advances in NLP. In this tutorial, we provide a comprehensive overview of parameter-efficient fine-tuning methods. We highlight their similarities and differences by presenting them in a unified view. We explore the benefits and usage scenarios of a neglected property of such parameter-efficient models—modularity—such as composition of modules to deal with previously unseen data conditions. We finally highlight how both properties—parameter efficiency and modularity—can be useful in the real-world setting of adapting pre-trained models to under-represented languages and domains with scarce annotated data for several downstream applications.[1]

1 Motivation and Objectives

The emergence of large pre-trained language models (Devlin et al., 2019) has led to a watershed moment in NLP, accelerating progress and improving performance across a wide range of NLP benchmarks. These models have quickly superseded previous baseline models and are now a core part of every NLP researcher and practitioner's toolkit. While pre-training such models has always been prohibitively expensive, recent pre-trained models have been getting so large (Brown et al., 2020) that even their fine-tuning and downstream usage are extremely challenging. In practice, the largest models perform best, even when fine-tuned on small datasets (Li et al., 2020). Therefore, being able to *efficiently and effectively fine-tune* the largest pre-trained models is key in order to reap the benefits of the latest advances in NLP. This is a major challenge that threatens to further exacerbate the inequality between resource-rich and resource-constrained research and production environments.

Recent work has highlighted the benefit of parameter-efficient methods to fine-tune such large pre-trained models. These parameter-efficient fine-tuning methods include soft prompt methods that preprend a small set of trainable continuous parameters to the input or intermediate layers (Li and Liang, 2021; Lester et al., 2021; Mahabadi et al., 2022), low-rank methods that train a small number of parameters in a low-dimensional subspace using random projections (Li et al., 2018; Aghajanyan et al., 2021), and adapter methods that insert trainable transformations at different layers (Houlsby et al., 2019; Pfeiffer et al., 2020). Other methods only tune a subset of the model's parameters (Lee et al., 2019; Zaken et al., 2021). An alternative set of methods relies on identifying performant sparse subnetworks, which can be updated in isolation (Frankle and Carbin, 2019; Guo et al., 2021; Xu et al., 2021; Sung et al., 2021). These methods reduce not only the number of parameters during fine-tuning but also have been shown to be more robust than standard fine-tuning and to outperform it in low-resource conditions (He et al., 2021b; Han et al., 2021; Mahabadi et al., 2021).

In the *first part* of this tutorial, we will give a comprehensive overview of such parameter-efficient fine-tuning methods. We will highlight the similarities and differences of a wide array of these methods by presenting them in a unified view, which expands on recent work (He et al., 2021a; Mao et al., 2021) highlighting the connections between adapters and prefix tuning. Based on this common view, we will be able to clearly show the respective benefits and trade-offs of a diverse set of parameter-efficient fine-tuning methods.

A commonality of parameter-efficient methods—

[1] Slides are available at: https://tinyurl.com/modular-fine-tuning-tutorial

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, pages 23 - 29
December 7-8, 2022 ©2022 Association for Computational Linguistics

illustrated clearly in this framework—is that they learn a modification vector that is added to the pre-trained model parameters, which are kept fixed. This property opens the door to *modularity*, which we view as a neglected benefit of the parameter-efficient usage of pre-trained models.

In the *second part* of the tutorial, we will explore the benefits and usage scenarios of such modular approaches. We will demonstrate how modular 'expert' modules can be learned for specific data settings (Chen et al., 2019; Rücklé et al., 2020; Gururangan et al., 2022; Li et al., 2022). Moreover, they can provide further benefits when combined and adapting to previously unseen settings (Pfeiffer et al., 2021a). We will additionally discuss how modular approaches can be used to augment models with new capabilities or knowledge, such as memory for lifelong learning (Kaiser et al., 2017), numerical reasoning (Andor et al., 2019), and factual or linguistic knowledge (Wang et al., 2021a). A key benefit of modularity is that it enables the storage and composition of modules to deal with previously unseen data conditions (Ponti et al., 2021, 2022). We will highlight this benefit based on prior work (Wortsman et al., 2020; Ponti et al., 2021; Ansell et al., 2022) and explore applications that it may enable in the future. Finally, as an NLP 'history lesson', we will revisit modular approaches that preceded pre-trained models (Andreas et al., 2016) and highlight how they may be relevant for recent approaches. Overall, we will encourage attendees to think of pre-trained models not as monoliths but as building blocks than can be augmented for specific purposes and data settings.

Tying both previous parts together, the *third part* of the tutorial will focus on applications: we will demonstrate how the properties explored so far—parameter efficiency and modularity—can be useful in practical settings. Specifically, we will focus on the important real-world setting of adapting pre-trained models to under-represented languages and domains with scarce annotated data for several downstream applications, e.g., cross-lingual transfer (Pfeiffer et al., 2020, 2022) and NMT (Bapna and Firat, 2019; Philip et al., 2020; Le et al., 2021; Üstün et al., 2021). We will highlight approaches that enable learning language-specific components using previously presented techniques such as adapters (Üstün et al., 2020; Pfeiffer et al., 2020, 2021b; Parović et al., 2022) or sparse subnetworks (Lin et al., 2021; Ansell et al., 2022). We will specifically discuss challenges and possible solutions when using such methods to adapt pre-trained models to extremely low-resource scenarios, such as test time adaptation (Wang et al., 2021b), parameter generation (Platanios et al., 2018; Ansell et al., 2021; Üstün et al., 2022), domain adaptation (Chronopoulou et al., 2022), and usage of alternative data sources (Ebrahimi and Kann, 2021; Faisal and Anastasopoulos, 2022).

1.1 What This Tutorial Does NOT Cover

We focus on parameter-efficient methods for adaptation of pre-trained models and thus only briefly discuss methods to make pre-training itself more efficient via efficient neural network architectures (Tay et al., 2020), including mixture-of-experts layers (Shazeer et al., 2017; Fedus et al., 2021). We will only briefly mention the emerging but already extensive literature on prompting,[2] and discuss its connections to the main topic of this tutorial. While prompting is itself parameter-efficient (requiring zero parameters) and can be combined with the fine-tuning methods we discuss, an extensive discussion of prompting would require its own tutorial. For similar reasons, we will only briefly highlight the extensive literature on controllable text generation. We will also only briefly discuss other techniques to improve efficiency such as knowledge distillation as these have been covered by the recent High Performance Natural Language Processing tutorial at EMNLP 2020 (Ilharco et al., 2020).

1.2 Tutorial Specifications

Tutorial Type: Cutting-edge, 3 hours

Target Audience: The target audience are researchers and practitioners in NLP who are interested in 1) extending research on this topic as well as 2) using state-of-the-art pre-trained models efficiently. In addition, target audience members will become familiar with diverse ways to make use of pre-trained models, beyond the standard prompting or fine-tuning setup.

Prerequisites: The target audience should be familiar with common neural network architectures (e.g., attention, Transformers), and also have a basic understanding of contemporary approaches in NLP, such as standard pre-trained models.

[2] For a comprehensive survey discussing prompting methods, we refer to (Liu et al., 2021).

2 Tutorial Outline

In what follows, we provide finer-grained descriptions of the main topics covered in the tutorial, along with tentative time allocation:

2.1 Parameter-efficient Models *[1h 10 mins]*

1. **Overview of Parameter-efficient Models** *[35 mins]*: We will begin the tutorial by introducing our audience to the range of techniques and methods used to fine-tune NLP models in a parameter-efficient way, from prompt tuning and adapters to pruning-based approaches. We will motivate the necessity and importance of research on parameter efficiency, and the main benefits of these approaches. To highlight a more pragmatic motivation, a comprehensive list of current and potential applications will also be provided.

2. **A Unified View of Parameter Efficiency** *[35 mins]*: We will provide the audience with a unified view of the parameter-efficient methods presented thus far. We will employ this view to highlight the key dimensions along which existing approaches differ as well as detail the resulting trade-offs that different approaches make. As part of this section, we will also provide a systematic general overview of the performance and computational efficiency of representative methods on an array of diverse benchmarks. In general, we will aim to provide the audience with a sense of the 'design space' of parameter-efficient methods so that they will not only be able to employ current methods, but expand and build upon them in future research.

2.2 Coffee Break *[30 mins]*

2.3 Modular Models *[55 minutes]*

1. **Learning Modular Experts** *[25 minutes]*: We will first highlight how modular experts can be learned in different settings and how these experts can be used to adapt to novel data distributions. We will also discuss how experts can provide access to new capabilities or new types of knowledge, such as numerical reasoning or factual and linguistic knowledge.

2. **Storing and Composing Modules** *[15 minutes]*: Having described the general setting and scenarios where modularity can be useful, we will highlight how modularity can lead to extremely efficient storage as well as composition of modules to adapt to unseen data settings: in the long run, the modular design leads to (re)composable and more sustainable NLP methods.

3. **Modularity Before Pre-training** *[15 minutes]*: We will finally revisit classic modular approaches and describe how some of the techniques and lessons from prior work may be applicable to the current generation of models.

2.4 Application: Multilingual and Low-Resource NLP *[55 minutes]*

1. **Parameter-efficient Methods for Multilingual NLP** *[25 minutes]*: In the last part of the tutorial, we will describe how the previosly discussed methods can be used to adapt pre-trained models to low-resource scenarios, with a focus on adapting pre-trained multilingual models to under-represented languages and domains, and enhancing multilingual NMT models for such resource-poor languages. This part focuses mainly on how language-specific components can be learned effectively, and how they can be combined with domain-specific and task-specific components, reaping the benefits of the modular design (from the previous part). This section will also discuss very recent methods based on efficient multilingual and language-specific contextual parameter generation and learning language-specific sub-networks. We will also highlight connections to pre-neural research on parameter-efficient methods for multilingual NLP.

2. **Adapting to Extremely Low-resource Languages** *[15 minutes]*: In addition, we will discuss challenges when learning such modular components in the extremely low-resource settings that are common when dealing with under-represented languages. Going beyond data scarcity, we will highlight challenges when learning languages with a different script, word order, or rich morphology. We will then describe strategies that can be used to effectively adapt models to such languages, including the use of external information (e.g., linguistic typology) to condition and enrich the modular design.

3. **Open Research Directions** *[15 minutes]*: In the last section, we will provide the audience

with an overview of research directions in this area and key pointers that will help them to pursue their own research, and apply the current technology in downstream NLP applications. Some time will also be reserved for a short QA session with the presenters.

3 Diversity

The third part of the tutorial focuses on how the described methods can be applied to improve models especially for low-resource and under-represented languages. This aligns with a long-term aim and promise of multilingual NLP to bring language technology to virtually *any* language of the world. We aim to make scripts available that demonstrate how the discussed methods can be applied in this setting. We hope this will help to diversify the audience, especially in the emerging regions such as Africa and Central and South America, and make the tutorial accessible to both beginners and advanced researchers.

4 Ethics Statement

The methodology introduced in the tutorial potentially inherits standard undesirable biases stemming from pretraining language models on large (and unverified) multilingual text collections. During the tutorial, we will ensure to remind NLP researchers and practitioners to bear in mind these biases, and apply appropriate data filtering and debiasing techniques before deploying any text encoders and relevant methodology to real-world language technology applications.

5 Presenters

Name: Sebastian Ruder
Affiliation: Google Research
Email: ruder@google.com
Website: http://ruder.io
Sebastian is a research scientist at Google Research where he works on transfer and cross-lingual learning and on parameter-efficient models. He was the Program Co-Chair for EurNLP 2019 and has co-organized the 4th Workshop on Representation Learning for NLP at ACL 2019 and the First Workshop on Multilingual Representation Learning at EMNLP 2021 and 2022. He has taught tutorials on "Transfer learning in natural language processing", "Unsupervised Cross-lingual Representation Learning", and "Multi-domain Multilingual Question Answering" at NAACL 2019, ACL 2019, and EMNLP 2021 respectively. He has also co-organized and taught at the NLP Session at the Deep Learning Indaba 2018, 2019, and 2022.

Name: Jonas Pfeiffer
Affiliation: Google Research
Email: jonaspfeiffer@google.com
Website: https://pfeiffer.ai
Jonas is a research scientist at Google Research. He is interested in modular and compositional representation learning in multi-task, multilingual, and multi-modal contexts. Jonas has received the IBM PhD Research Fellowship award in 2020. He has given invited talks in academia (e.g. University of Cambridge, ETH, EPFL, NYU), industry (e.g. Facebook AI Research, IBM Research), as well as at Machine Learning Summer/Winter Schools (e.g. Lisbon ML Summer School (LxMLS) 2021, Advanced Language Processing Winter School (ALPS) 2022).

Name: Ivan Vulić
Affiliation: University of Cambridge & PolyAI
Email: iv250@cam.ac.uk
Website: https://sites.google.com/site/ivanvulic/
Ivan is a Principal Research Associate and a Royal Society University Research Fellow in the Language Technology Lab at the University of Cambridge, and a Senior Scientist at PolyAI. His research interests are in multilingual and multimodal representation learning, and transfer learning for low-resource languages and applications such as task-oriented dialogue systems. He has extensive experience giving invited and keynote talks, and co-organising tutorials (e.g., ECIR 2013, WSDM 2014, EMNLP 2017, NAACL-HLT 2018, ESSLLI 2018, ACL 2019, 2 tutorials at EMNLP 2019, AILC Lectures 2021, ACL 2022) and workshops in areas relevant to the tutorial proposal (e.g., VL'15, SIGTYP 2019-2021, DeeLIO 2020-2022, RepL4NLP 2021, MML 2022, publication chair of ACL 2019, program chair of *SEM 2021, tutorial co-chair of EMNLP 2021).

6 Acknowledgments

Ivan Vulić is supported by a personal Royal Society University Research Fellowship, and his work has also been supported by a Huawei research donation awarded to the Language Technology Lab.

References

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2021. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. In *Proceedings of ACL 2021*.

Daniel Andor, Luheng He, Kenton Lee, and Emily Pitler. 2019. Giving BERT a calculator: Finding operations and arguments with reading comprehension. In *Proceedings of the EMNLP 2019*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to Compose Neural Networks for Question Answering. In *Proceedings of NAACL 2016*.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of ACL 2022*.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of EMNLP 2021*.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. In *Proceedings of EMNLP 2019*.

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 NeurIPS 2020*.

Vincent S Chen, Sen Wu, Zhenzhen Weng, Alexander Ratner, and Christopher Ré. 2019. Slice-based learning: A programming model for residual learning in critical data slices. *Advances in NeurIPS 2019*.

Alexandra Chronopoulou, Matthew Peters, and Jesse Dodge. 2022. Efficient hierarchical domain adaptation for pretrained language models. In *Proceedings of NAACL-HLT 2022*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL 2019*.

Abteen Ebrahimi and Katharina Kann. 2021. How to Adapt Your Pretrained Multilingual Model to 1600 Languages. In *Proceedings of ACL 2021*.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. *CoRR*, abs/2205.09634.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv preprint*.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proceedings of ICLR 2019*.

Demi Guo, Alexander M. Rush, and Yoon Kim. 2021. Parameter-efficient transfer learning with diff pruning. In *Proceedings of ACL 2021*.

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A. Smith, and Luke Zettlemoyer. 2022. DEMix layers: Disentangling domains for modular language modeling. In *Proceedings of the NAACL 2022*.

Wenjuan Han, Bo Pang, and Yingnian Wu. 2021. Robust Transfer Learning with Pretrained Language Models through Adapters. In *Proceedings of ACL 2021*.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021a. Towards a Unified View of Parameter-Efficient Transfer Learning. *arXiv preprint*.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021b. On the Effectiveness of Adapter-based Tuning for Pretrained Language Model Adaptation. In *Proceedings of ACL 2021*.

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 *Proceedings of ICML 2019*.

Gabriel Ilharco, Cesar Ilharco, Iulia Turc, Tim Dettmers, Felipe Ferreira, and Kenton Lee. 2020. High performance natural language processing. In *Proceedings of EMNLP 2020: Tutorial Abstracts*.

Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. 2017. Learning to Remember Rare Events. In *Proceedings of ICLR 2017*.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of ACL-IJCNLP 2021*.

Jaejun Lee, Raphael Tang, and Jimmy Lin. 2019. What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning. *arXiv preprint*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of EMNLP 2021*.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. Measuring the Intrinsic Dimension of Objective Landscapes. In *Proceedings of ICLR 2018*.

Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. 2022. Branch-train-merge: Embarrassingly parallel training of expert language models. *arXiv preprint*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of ACL 2021*.

Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joseph E. Gonzalez. 2020. Train large, then compress: Rethinking model size for efficient training and inference of transformers. *arXiv preprint*.

Zehui Lin, Liwei Wu, Mingxuan Wang, and Lei Li. 2021. Learning language specific sub-network for multilingual machine translation. In *Proceedings of ACL 2021*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in Natural Language Processing. *arXiv preprint*.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient Low-Rank Hypercomplex Adapter Layers. In *Advances in NeurIPS 2021*.

Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Lambert Mathias, Marzieh Saeidi, Veselin Stoyanov, and Majid Yazdani. 2022. Prompt-free and efficient few-shot learning with language models. In *Proceedings of ACL 2022*.

Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Wen-tau Yih, and Madian Khabsa. 2021. UniPELT: A Unified Framework for Parameter-Efficient Language Model Tuning. *arXiv preprint*.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of NAACL-HLT 2022*.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the NAACL 2022*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rückĺe, Cho Kyunghyun, and Iryna Gurevych. 2021a. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of EACL 2021*.

Jonas Pfeiffer, Ivan Vulic, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: an adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of EMNLP 2020*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021b. UNKs Everywhere: Adapting Multilingual Language Models to New Scripts. In *Proceedings of EMNLP 2021*.

Jerin Philip, Alexandre Berard, Matthias Gallé, and Laurent Besacier. 2020. Monolingual adapters for zero-shot neural machine translation. In *Proceedings of EMNLP 2020*.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of EMNLP 2018*.

Edoardo M. Ponti, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen. 2021. Parameter Space Factorization for Zero-Shot Learning across Tasks and Languages. *Transactions of the ACL 2021*.

Edoardo Maria Ponti, Alessandro Sordoni, and Siva Reddy. 2022. Combining modular skills in multitask learning. *CoRR*, abs/2202.13914.

Andreas Rücklé, Jonas Pfeiffer, and Iryna Gurevych. 2020. MultiCQA : Zero-Shot Transfer of Self-Supervised Text Matching Models on a Massive Scale. In *Proceedings of EMNLP 2020*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of ICLR 2017*.

Yi-Lin Sung, Varun Nair, and Colin Raffel. 2021. Training neural networks with fixed sparse masks. In *Advances in NeurIPS 2021*.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient Transformers: A Survey. *arXiv preprint*.

Ahmet Üstün, Alexandre Berard, Laurent Besacier, and Matthias Gallé. 2021. Multilingual unsupervised neural machine translation with denoising adapters. In *Proceedings of EMNLP 2021*.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. In *Proceedings of EMNLP 2020*.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord, and Sebastian Ruder. 2022. Hyper-X: A unified hypernetwork for multi-task multilingual transfer. *CoRR*, abs/2205.12148.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021a. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of ACL 2021*.

Xinyi Wang, Yulia Tsvetkov, Sebastian Ruder, and Graham Neubig. 2021b. Efficient Test Time Adapter Ensembling for Low-resource Language Varieties. In *Findings of EMNLP 2021*.

Mitchell Wortsman, Vivek Ramanujan, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, and Ali Farhadi. 2020. Supermasks in Superposition. In *Advances in NeurIPS 2020*.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a Child in Large Language Model : Towards Effective and Generalizable Fine-tuning. In *Proceedings of EMNLP 2021*.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. *arXiv preprint*.