Dual Context-Guided Continuous Prompt Tuning for Few-Shot Learning

Jie Zhou, Lei Tian, Houjin Yu, Xiao Zhou, Hui Su, Jie Zhou
Pattern Recognition Center, WeChat AI, Tencent Inc, Beijing, China
sannyzhou@tencent.com

Abstract

Prompt-based paradigm has shown its competitive performance in many NLP tasks. However, its success heavily depends on prompt design, and the effectiveness varies upon the model and training data. In this paper, we propose a novel dual context-guided continuous prompt (DCCP) tuning method. To explore the rich contextual information in language structure and close the gap between discrete prompt tuning and continuous prompt tuning, DCCP introduces two auxiliary training objectives and constructs input in a pair-wise fashion. Experimental results demonstrate that our method is applicable to many NLP tasks, and can often outperform existing prompt tuning methods by a large margin in the few-shot setting.

1 Introduction

With the rise of pretrained language models(PLMs), natural language processing(NLP) shifted from the fully-supervised paradigm to pretrain and fine-tune paradigms (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019). To further utilize the large capacity of PLMs, a prompt-based paradigm is proposed to reformulate downstream tasks into an LM-like task upon the context and task-specific prompt.

There are some issues with the prompt-based paradigm, especially prompt engineering. Discrete prompts (a.k.a hard prompts) (Petroni et al., 2019; Wang et al., 2021) need expert-level experience to manually discover templates. To address this problem, automatic prompt design is conducted on gradient-based search (Shin et al., 2020), generation (Ben-David et al., 2021), ensembles (Schick and Schütze, 2021) and scoring (Davison et al., 2019). ADAPET (Tam et al., 2021) provide a denser supervision during fine-tuning based on the label-conditioned language modeling task. However, these methods might get sub-optimal templates and require adequate validation data (Zhao et al., 2021; Perez et al., 2021).

What's more, it is unnecessary to limit prompts to hard-crafting text. Continuous prompts (a.k.a soft prompts) (Liu et al., 2021b; Li and Liang, 2021) take templates as additional trainable parameters. Thus, prompt search can be simplified as optimizing parameters based on downstream task. Recent works add layer-wise adaptive prompt parameters (Qin and Eisner, 2021; Liu et al., 2021a), data-dependent mixture (Qin and Eisner, 2021) and hard-soft hybrid prompt (Han et al., 2021) based on adequate training data. When it comes to the few-shot learning scenario, it remains unclear how to effectively learn continuous prompts. Previous works mainly improve continuous prompts by additional prompt and target encoder (Gao et al., 2021; Zhang et al., 2021; Liu et al., 2021a).

This paper presents a new model-agnostic perspective of further utilizing deep LM features. We propose a novel **D**ual **C**ontext-guided **C**ontinuous **P**rompt (DCCP) tuning approach that makes PLMs better few-shot learners. Our main concern is how to learn better continuous prompts with only a few samples, averting dependency on hand-craft engineering and large validation samples.

Considering that prompt-based models predict based on both prompt and context, the vanilla models learn about $P(Y|X_{context}, H_{Prompt})$. Notably, additional prompt embeddings H_{prompt} are optimized based on the given context $X_{context}$ with LM decoding task on the downstream target Y in previous works. We give an insight into better continuous prompt tuning throughout the dual view of context-aware prompt and label-aware context representations. Technically, we introduce a new label-aware context-masked input aligning with the vanilla context-aware prompt-masked input. We add two auxiliary training objectives for coupling layer-wise linguistic features. The dual view makes the model learn $P(X_{context}|Y, H_{prompt})$ and $P(H_{context}|Y, H_{prompt})$ throughout LM inner features, further optimizing the prompt embed-

Findings of the Association for Computational Linguistics: ACL 2022, pages 79 - 84
May 22-27, 2022 ©2022 Association for Computational Linguistics

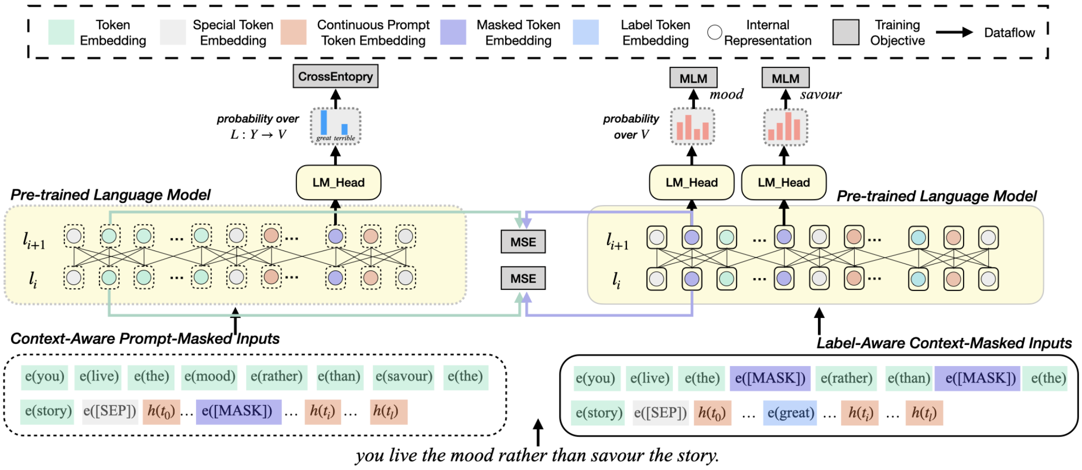
Token Embedding
Special Token Embedding
Continuous Prompt Token Embedding
Masked Token Embedding
Label Token Embedding
Internal Representation
Training Objective
Dataflow
CrossEntopry
probability over $L : Y \to V$
great terrible
LM_Head
Pre-trained Language Model
l_{i+1}
l_i
MSE
MSE
MLM
mood
MLM
savour
probability over V
LM_Head
LM_Head
Pre-trained Language Model
Context-Aware Prompt-Masked Inputs
e(you) e(live) e(the) e(mood) e(rather) e(than) e(savour) e(the)
e(story) e([SEP]) $h(t_0)$ … e([MASK]) … $h(t_i)$ … $h(t_l)$
Label-Aware Context-Masked Inputs
e(you) e(live) e(the) e([MASK]) e(rather) e(than) e([MASK]) e(the)
e(story) e([SEP]) $h(t_0)$ … e(great) … $h(t_i)$ … $h(t_l)$
you live the mood rather than savour the story.

Figure 1: The Architecture of Dual Context-Guided Continuous Prompt Tuning.

dings. In a nutshell, DCCP goes deep into PLMs representations for better continuous prompt tuning.

We conduct experiments on 10 NLP datasets in the few-shot learning setting. DCCP significantly outperforms conventional fine-tuning, discrete prompts, and previous works on continuous prompts. DCCP achieves 89.6% (on average) of the full-supervised fine-tuning performance across all datasets with only 16 training samples. It obtains gain 11.8%, 2.5%, and 1.6% absolute improvement on average compared to conventional fine-tuning, vanilla continuous prompts (Gao et al., 2021), and state-of-the-art continuous prompts (Zhang et al., 2021). We empirically demonstrate that DCCP makes LM a better few-shot learner.

2 Methodology

In this section, we first introduce the vanilla continuous prompt tuning model and then clarify our dual context-guided prompt tuning method.

2.1 Vanilla Continuous Prompt Tuning

Given a pretrained language model, a context input sequence $X_{context} = (x_0, \dots, x_n)$ is tokenized as $[\text{CLS}]X_{context}[\text{SEP}]$. The conventional fine-tuning model predicts based on [CLS] output. For prompt-based methods, a task-specific prompt $\widetilde{X}_{prompt} = (t_0, \dots, [\text{MASK}], \dots, t_l)$ is added into the input as $X_{in} = [\text{CLS}]X_{context}[\text{SEP}]\widetilde{X}_{prompt}[\text{SEP}]$. t_i is represented by a trainable pseudo token embedding h_i. It takes downstream tasks as a masked language modeling(MLM) task. Assume that verbalizer $L : Y \to V$ maps the class set Y to vocabulary set V, the probability of predicting $y_j \in Y$ is:

$$p(y_j|X_{in}) = p([\text{MASK}] = V_j|\widetilde{X}_{prompt}, X_{content}) \quad (1)$$

where X_{prompt} is represented by the additional trainable embedding parameters. Here we discard the prompt and target encoders used in (Liu et al., 2021b,a), retaining origin LM architectures. Based on downstream tasks, the vanilla model is optimized based on cross-entropy loss:

$$L_c = -\frac{1}{N}\Sigma_i^N \Sigma_j^{|Y|} y_{ij} \log p_{ij}. \quad (2)$$

2.2 Dual Context-Guided Prompt Tuning

Continuous prompt tuning simply introduces a trainable pseudo template for automatic prompt searching. It faces optimization challenges of word embedding discreteness and prompt embedding association (Liu et al., 2021b), which makes it hard tune continuous prompt with only a few samples.

Although continuous prompts are pseudo tokens and not referred to any real word, we propose that continuous prompt tuning should be consistent with natural language modeling. It could take more language modeling constraints into account, thus further reducing the gap between pretraining and fine-tuning. Vanilla continuous prompt tuning has considered an MLM-like objective L_c for matching target verbalizer and masked token output, focusing on prompt and downstream tasks. Furthermore, we propose two auxiliary language modeling tasks for both pluggable prompt and origin context.

We aim to further leverage context information for better guiding prompt tuning. The auxiliary

	SST-2 (acc)	**MR** (acc)	**CR** (acc)	**SUBJ** (acc)	**TREC** (acc)
Majority[†]	50.9	50.0	50.0	50.0	18.8
Prompt-based zero-shot[‡]	83.6	80.8	79.5	51.4	32
"GPT-3" in-context learning	84.8 (1.3)	80.5 (1.7)	87.4 (0.8)	53.6 (1.0)	26.2 (2.4)
Fine-tuning	81.4 (3.8)	76.9 (5.9)	75.8 (3.2)	90.8 (1.8)	88.8 (2.1)
LMBFF (Gao et al., 2021)	92.3 (1.0)	85.5 (2.8)	89.0 (1.4)	91.2 (1.1)	88.2 (2.0)
PTuning (Liu et al., 2021b)	92.4 (0.6)	86.4 (1.5)	91.1 (0.6)	91.8 (0.8)	90.5 (1.6)
DART (Zhang et al., 2021)	93.5 (0.5)	88.2 (1.0)	91.8 (0.5)	90.7 (1.4)	87.1(3.8)
DCCP	**94.1** (0.6)	**89.2** (0.7)	**92.6** (0.6)	**92.8** (1.0)	**92.1** (2.3)
Fine-tuning (full)[†]	95.0	90.8	89.4	97.0	97.4
	MNLI (acc)	**SNLI** (acc)	**QNLI** (acc)	**MRPC** (F1)	**QQP** (F1)
Majority[†]	32.7	33.8	49.5	81.2	0.0
Prompt-based zero-shot[‡]	50.8	49.5	50.8	61.9	49.7
"GPT-3" in-context learning	52.0 (0.7)	47.1 (0.6)	53.8 (0.4)	45.7 (6.0)	36.1 (5.2)
Fine-tuning	45.8 (6.4)	48.4 (4.8)	60.2 (6.5)	76.6 (2.5)	60.7 (4.3)
LMBFF (Gao et al., 2021)	68.3 (2.5)	**77.1** (2.1)	68.3 (7.4)	76.2 (2.3)	67.0 (3.0)
PTuning (Liu et al., 2021b)	65.7 (4.0)	68.3 (7.3)	67.6 (7.3)	78.6 (1.1)	65.8 (3.9)
DART (Zhang et al., 2021)	67.5 (2.6)	75.8 (1.6)	66.7 (3.7)	78.3 (4.5)	67.8 (3.2)
DCCP	**68.6** (2.6)	74.1 (3.9)	**71.3** (3.2)	**80.3** (1.3)	**67.9** (3.5)
Fine-tuning (full)[†]	89.8	92.6	93.3	91.4	81.7

Table 1: Main results using RoBERTa-large. † refers to using the full training set while ‡ refers to using no training samples. The others involve $K = 16$ (per class) for few-shot experiments. Note that the mean (and standard deviation) performances are reported over 5 different splits. "GPT-3" in-context learning: using the in-context learning proposed in (Brown et al., 2020) with RoBERTa-large (no parameter updates).

tasks are constructed for context language modeling. Technically, a label-aware context-masked input $\widetilde{X}_{in}$ is fed as another model input aligning with origin context-aware prompt-masked X_{in}. Given the ground-truth label y, we obtain a semantically intact prompt X_{prompt}. A masked context input $\widetilde{X}_{context}$ is generated by randomly masking context tokens at position of $z_i \in Z$ in the same manner as the pretrained MLM task. The new input is:

$$\begin{aligned}&\widetilde{X}_{in} = \text{[CLS]}\widetilde{X}_{context}\text{[SEP]}X_{prompt}\text{[SEP]},\\&\widetilde{X}_{context} = (x_0, \ldots, \text{[MASK]}, \ldots, \text{[MASK]}, \ldots, x_n),\\&X_{prompt} = (t_0, \ldots, L(y), \ldots, t_l).\end{aligned} \quad (3)$$

As depicted in Fig 1, we obtain a couple of model inputs. The origin context-aware prompt-masked input has intact context information but lacks downstream label information. On the contrary, the label-aware context-masked input is aware of the ground-truth label but misses partial context features. Although these two dual inputs separately lack partial semantic information, they should be semantically paraphrased.

Specifically, the first auxiliary constraint L_{mlm} is for the masked language modeling task of label-aware masked context tokens Z. It is calculated as:

$$\begin{aligned}&\ell^{i,j}_{mlm} = -\log(p(\widetilde{x}^{i,j}_c = x^{i,j}_c|y_i, X_p, \widetilde{X}_c, j \in Z)),\\&L_{mlm} = \frac{1}{N}\Sigma^N_i \frac{1}{|Z|}\Sigma^{|Z|}_j \ell^{i,j}_{mlm},\end{aligned} \quad (4)$$

where c and p refer to the context and prompt. The origin text token $x^{i,j}_c$ serves as the hard label of this label-aware context cloze task.

In addition, paraphrased texts can be closely related to each other throughout the language structure. We further exploit different-level linguistic features for aligning the dual context input as a paraphrased couple. According to the large capacity of PLMs, the LM encoder could be directly utilized as a linguistic feature encoder. We add a metric constraint on internal representations of the pairwise masked context tokens Z. It aligns the label-aware masked context $\widetilde{X}_{context}$ with the origin context $X_{context}$ upon linguistic features across LM layers. The training objective is calculated based on internal representations as a mean square error loss L_{mse}.

$$\begin{aligned}&\ell_{mse}(\widetilde{x}^{i,j}_c, x^{i,j}_c) = \frac{1}{S}\Sigma^S_k||\widetilde{h}^{i,j,k}_c - h^{i,j,k}_c||^2_2,\\&L_{mse} = \frac{1}{N}\Sigma^N_i \frac{1}{|Z|}\Sigma^{|Z|}_j \ell_{mse}(\widetilde{x}^{i,j}_c, x^{i,j}_c)\end{aligned} \quad (5)$$

where h is the hidden state, S indicates the depth of the LM model, j refers to the masked context

Dataset	Verbalizer	Prompt
SST-2	terrible/great	[unused1] [unused2] [unused3] <mask>[unused4][unused5].
MR		
Cr		
SUBJ	subjective/objective	
TREC	Description/Entity/Expression/Human/Location/Number	
MNLI	No/Yes/Maybe	[unused1] <mask>[unused2]
SNLI		
QNLI	No/Yes	
MRPC		
QQP		

Table 2: Verbalizer and Pseudo prompt templates for continuous prompt tuning experiments.

tokens, and i means the i-th sample. The overall training objective is $L = L_c + L_{mlm} + L_{mse}$. These two auxiliary constraints are trained in a self-supervised learning manner, which leverages more information than the vanilla prompt tuning within the same dataset size. In the other words, this model-agnostic training method makes full use of the current training data from the view of going deep into the internal representations.

All in all, the vanilla model predicts downstream task via filling the blank of prompts based on the context information. Our proposed auxiliary tasks reconstruct the masked context based on the ground-truth label and prompt semantics. Therefore, our dual context-guided continuous prompt (DCCP) tuning method would advance few-shot learning based on the dual implementation of prompt and context features.

3 Experiments

We experiment our proposed architecture on 10 NLP tasks in the few-shot setting (k=16) according to LMBFF (Gao et al., 2021). The datasets involve sentiment analysis (SST-2, MR, CR), subjective analysis (SUBJ), question type (TREC), natural language inference (MNLI, SNLI, QNLI), paraphrase detection (MRPC, QQP).

3.1 Experimental Settings

The experiment is conducted in the same setting as (Gao et al., 2021; Zhang et al., 2021), which is based on RoBERTa-large (Liu et al., 2019). We conduct a grid search on multiple hyper-parameters for each set, and choose the best setting on the development subset. We use AdamW (Loshchilov and Hutter, 2019) as the optimizer. We average the performance on the test set with five fixed random few-shot training datasets for each task. The verbalizer and pseudo prompt template can be referred to Tab 2.

3.2 Results and Analysis

Main Result In Table 1, we compare the performance of our DCCP to the state-of-the-art prompt-based methods and conventional fine-tuning method. Our model achieves great performance gain compared to the conventional fine-tuning and vanilla continuous prompt tuning model over all 10 tasks. DCCP outperforms the SOTA prompt-based methods (Gao et al., 2021; Zhang et al., 2021) across 9 datasets, which indicates the great advancement of our DCCP on few-shot learning. Especially, in the condition of only 16 training and development samples, DCCP could obtain a competitive result compared to the full training set in SST-2, MR and CR dataset. Our results obtain up to 5% and 4% absolutely improvement when compared to DART (Zhang et al., 2021) and LMBFF (Gao et al., 2021).

Ablation Study According to Table 3, both auxiliary tasks outperform the vanilla model and previous works. It denotes that the context-view language modeling tasks are beneficial for the continuous prompt tuning approach in the few-shot learning scenario. The results reveal that the metric constraint on internal representation is complementary to the masked language modeling.

Our overall methodology achieves 2.5% perfor-

Method	Avg. Performance
Fine-tuning	74.51
LMBFF (Gao et al., 2021)	80.31
PTuning (Liu et al., 2021b)	79.79
DART (Zhang et al., 2021)	80.74
DCCP	**82.3**
w/o MLM	81.02
w/o MSE	80.97

Table 3: Ablation Study of DCCP. The score refers to the average performance across all datasets.

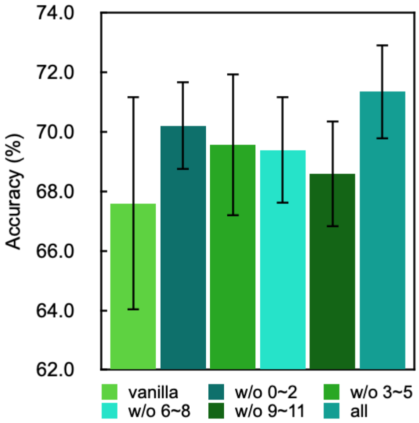

Figure 2: Average Performance on QNLI of arious supervised layers for dual context constraint. Note that "w/o k-i" denotes "not supervise on k-th to i-th LM layer". "w/o 9-11" additionally averts the masked language modeling.

mance gain upon the vanilla model without modifying the model architecture or leveraging more external data.

Will the layers of metric constraint affects performances? Referring to Fig 2, it is necessary to consider internal representations at different LM layers as we couple the dual context linguistic features. It could get more stable and better results by comparing all linguistic features of the label-aware masked context and origin context.

Performance on various training dataset size. Fig 3 illustrates our stable improvement compared to conventional fine-tuning and vanilla prompt tuning as the number K of training samples increases. Even though it converges with vanilla prompt tuning around $K = 256$, it retains better stability and performance.

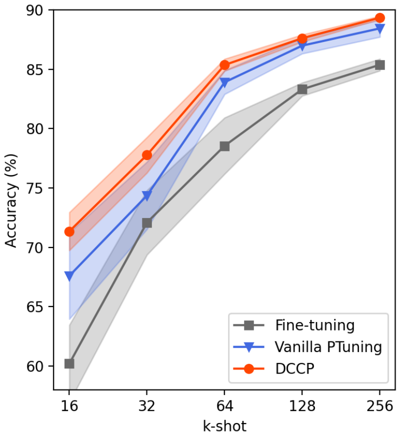

Figure 3: Conventional Fine-tuning vs Vanilla PTuning vs our DCCP across various K-shot (i.e. # instances per class) settings on QNLI.

4 Conclusion

In this paper, we present a model-agnostic approach for advancing continuous prompt. Specifically, we propose a novel dual context-guided continuous prompt tuning method for few-shot learning. Our approach constructs a couple of dual inputs, including the origin context-aware prompt-masked input and label-aware context-masked one. Then, we go deep into the language model to leverage linguistic features for two auxiliary constraints on the pairwise context inputs. The empirical results show that continuous prompts can be further revised during the procedure of reconstructing context.

References

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2021. PADA: A prompt-based autoregressive approach for adaptation to unseen domains. *CoRR*, abs/2102.12206.

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*.

Joe Davison, Joshua Feldman, and Alexander M. Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1173–1178. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3816–3830. Association for Computational Linguistics.

Xu Han, Weilin Zhao, Ning Ding, Zhiyuan Liu, and Maosong Sun. 2021. PTR: prompt tuning with rules for text classification. *CoRR*, abs/2105.11259.

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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4582–4597. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021a. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. GPT understands, too. *CoRR*, abs/2103.10385.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *CoRR*, abs/2105.11447.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

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, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5203–5212. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Timo Schick and Hinrich Schütze. 2021. 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, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269. 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 2020, Online, November 16-20, 2020*, pages 4222–4235. 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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4980–4991. Association for Computational Linguistics.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *CoRR*, abs/2104.14690.

Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Differentiable prompt makes pre-trained language models better few-shot learners. *CoRR*, abs/2108.13161.

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.