ing (Cf. Section 4.2).

Our main contribution in this paper is a novel empirical finding that properly optimized prompt tuning can be comparable to fine-tuning universally across various model scales and NLU tasks. In contrast to observations in prior work, our discovery reveals the universality and potential of prompt tuning for NLU.

Technically, our approach P-tuning v2 is not conceptually novel. It can be viewed as an optimized and adapted implementation of **Deep Prompt Tuning** (Li and Liang, 2021; Qin and Eisner, 2021) designed for generation and knowledge probing. The most significant improvement originates from appling continuous prompts for every layer of the pretrained model, instead of the mere input layer. Deep prompt tuning increases the capacity of continuous prompts and closes the gap to fine-tuning across various settings, especially for small models and hard tasks. Moreover, we present a series of critical details of optimization and implementation to ensure finetuning-comparable performance.

Experimental results show that P-tuning v2 matches the performance of fine-tuning at different model scales ranging from 300M to 10B parameters and on various hard sequence tagging tasks such as extractive question answering and named entity recognition. P-tuning v2 has 0.1% to 3% trainable parameters per task compared to fine-tuning, which substantially reduces training time memory cost and per-task storage cost.

2 Preliminaries

NLU Tasks. In this work, we categorize NLU challenges into two families: *simple classification tasks* and *hard sequence labeling tasks*.[3] Simple classification tasks involve classification over a label space. Most datasets from GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) are in this category. Hard sequence labeling tasks involve classification over a sequence of tokens, such as named entity recognition and extractive question answering.

Prompt Tuning. Let $\mathcal{V}$ be the vocabulary of a language model $\mathcal{M}$ and let $\mathbf{e}$ be the embedding layer of $\mathcal{M}$. In the case of discrete prompting (Schick and Schütze, 2020), prompt tokens {"It", "is", "[MASK]"} $\subset$ $\mathcal{V}$ can be used to classify a movie review. For example, given the input text $\mathbf{x}$ ="Amazing movie!", the input embedding sequence is formulated as $[\mathbf{e}(\mathbf{x}), \mathbf{e}(\text{"It"}), \mathbf{e}(\text{"is"}), \mathbf{e}(\text{"[MASK]"})]$.

Lester et al. (2021) and Liu et al. (2021) introduce trainable continuous prompts as a substitution to natural language prompts for NLU with the parameters of pretrained language models frozen. Given the trainable continuous embeddings $[h_0, ..., h_i]$, the input embedding sequence is written as $[\mathbf{e}(\mathbf{x}), h_0, ..., h_i, \mathbf{e}(\text{"[MASK]"})]$, as illustrated in Figure 2. Prompt tuning has been proved to be comparable to fine-tuning on 10-billion-parameter models on simple classification tasks (Lester et al., 2021; Kim et al., 2021; Liu et al., 2021).

3 P-Tuning v2

3.1 Lack of Universality

Lester et al. (2021); Liu et al. (2021) have been proved quite effective in many NLP applications (Wang et al., 2021a,b; Chen et al., 2021; Zheng et al., 2021; Min et al., 2021), but still fall short at replacing fine-tuning due to lack of universality, as discussed below.

Lack of universality across scales. Lester et al. (2021) shows that prompt tuning can be comparable to fine-tuning when the model scales to over 10 billion parameters. However, for medium-sized models (from 100M to 1B) that are widely used, prompt tuning performs much worse than fine-tuning.

Lack of universality across tasks. Though Lester et al. (2021); Liu et al. (2021) have shown superiority on some of the NLU benchmarks, the effectiveness of prompt tuning on hard sequence tagging tasks is not verified. Sequence tagging predicts a sequence of labels for each input token, which can be harder and incompatible with verbalizers (Schick and Schütze, 2020). In our experiments (Cf. Section 4.2 and Table 3), we show that Lester et al. (2021); Liu et al. (2021) perform poorly on typical sequence tagging tasks compared to fine-tuning.

Considering these challenges, we propose P-tuning v2, which adapts deep prompt tuning (Li and Liang, 2021; Qin and Eisner, 2021) as a universal solution across scales and NLU tasks.

3.2 Deep Prompt Tuning

In (Lester et al., 2021) and (Liu et al., 2021), continuous prompts are only inserted into the input

[3] Note that the notions of "simple" and "hard" are specific to prompt tuning, because we find sequence labeling tasks are more challenging for prompt tuning.