Learning How to Ask: Querying LMs with Mixture of Soft Prompts

Guanghui Qin and Jason Eisner
Johns Hopkins University

Abstract
Task: Extract factual knowledge from cloze language models.
Basic method: Run cloze LM on prompts (sentence with blanks)
This work: Replace prompts with vectors that can be optimized with back-propagation in a continuous space.

Soft Prompts
Words in prompts are discrete, but transformers are concerned with vectors. Instead of feeding LMs with words, we can feed vectors (layer-0 in diagram) that can be optimized with back-propagation.

Why Soft?
Easy to optimize: Back-propagation.
Larger space of prompts: Vectors can be more expressive. = {performed, played, painted} can be profession-neutral; = {his, her, their} can be gender-neutral.
Focusness: Certain keywords can be emphasized by adjusting vector length.

Deep Perturbation
Transformers encode the prompts with many layers of sequences of vectors. Extending the idea of soft prompts to all the layers, we add “perturbation biases” to each layer of transformers.

Ensembling
We exploit a mixture of prompts, whose weights can be trained with EM algorithm.

Experiments
Experiments conducted on T-Rex datasets show:
1. Performance improved by 12.2% if trained from other people’s prompts.
2. Training from random initialization is almost as good as from other prompts.
Others: Extensive experiments on many datasets with many language models all show the effectiveness of soft prompts & deep perturbation. Please look at our paper for more details.