On the Relationship between Skill Neurons and Robustness in Prompt Tuning

LREC-COLING 2024

Leon Ackermann
lackermann@uni-osnabrueck.de
Osnabrück University, Germany

Xenia Ohmer
xenia.ohmer@uni-osnabrueck.de
Osnabrück University, Germany

Main idea

- **Prompt tuning** (Lester et al., 2021) is parameter-efficient finetuning method.
- Wang et al. (2022) used Prompt Tuning to identify **skill neurons** in the transformer feed-forward networks of RoBERTa.
- These skill neurons are highly predictive of the task labels, task-specific, and important for solving the task.

➔ In this paper, we **study the robustness of prompt tuning in relation to these skill neurons**, using RoBERTa and T5.

Funding

Universitätsgesellschaft OSNABRÜCK e.V.

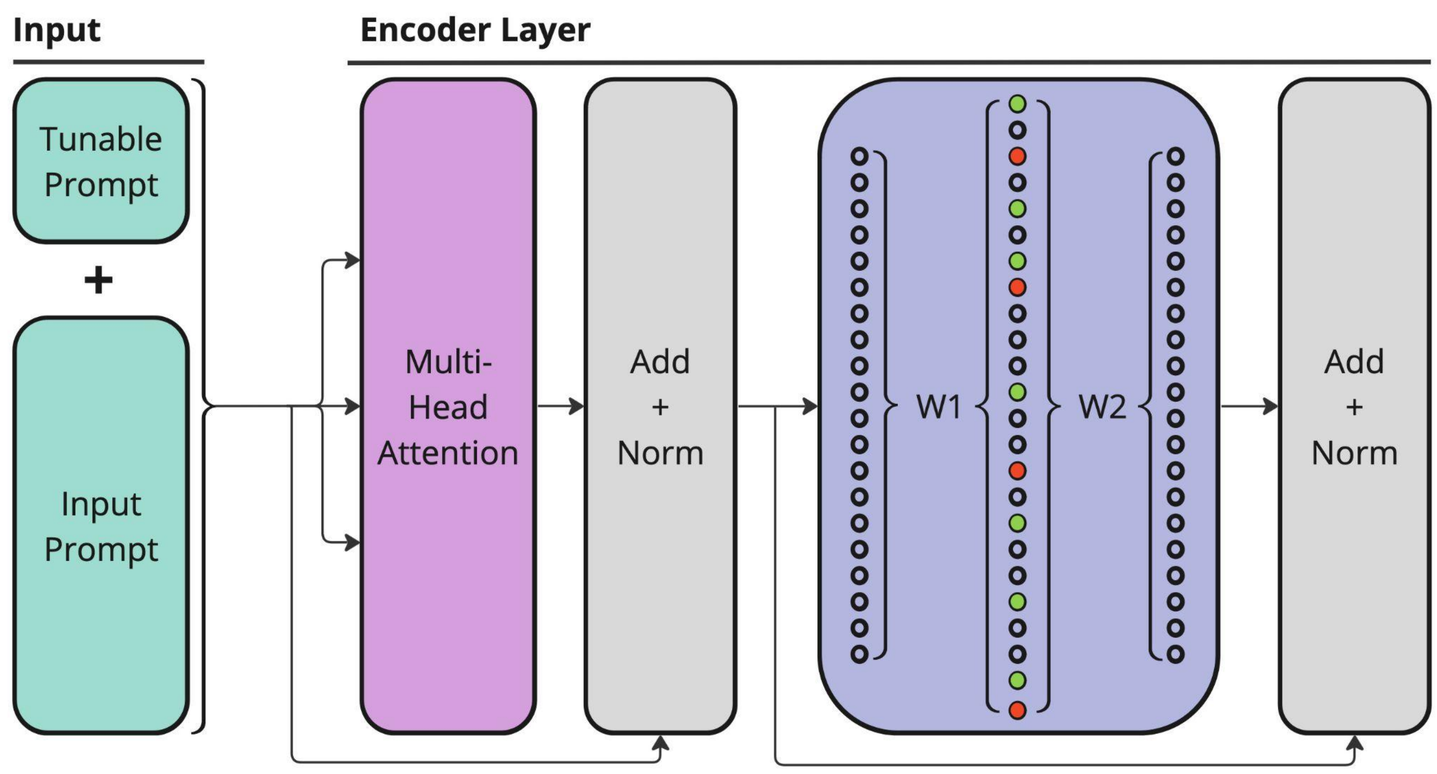
Methods

Prompt tuning

Prepend tunable parameters to the input (in the embedding space) and train them after model pretraining.

Skill neurons

1. Compute **baseline activations** of FFN neurons on training set.
2. Compute **neuron accuracies** on validation set with respect to baseline activation.
3. Compute **neuron predictivities** by aggregating the most excitatory and inhibitory neuron accuracies.
4. Aggregate the neuron predictivities across prompts tuned with different seeds (in our case 5).



Main results

Prompt tuning and robustness

Accuracy

Dataset	RoBERTa	T5
ethicsdeontology	69.9 ± 2.0	66.3 ± 1.6
ethicsjustice	65.4 ± 1.6	59.1 ± 2.9
MRPC	74.8 ± 5.9	77.5 ± 2.6
QQP	$87.1 \pm 0,2$	88.7 ± 1.1
AdvQQP	37.2 ± 4.1	59.2 ± 8.0
QNLI	90.4 ± 0.2	92.4 ± 0.2
AdvQNLI	45.1 ± 3.5	60.1 ± 3.1
IMDB	90.4 ± 0.3	88.2 ± 0.2
movierationales	74.1 ± 2.4	75.2 ± 1.4
SST2	98.7 ± 2.6	94.0 ± 0.4
AdvSST2	45.3 ± 4.5	45.4 ± 3.3

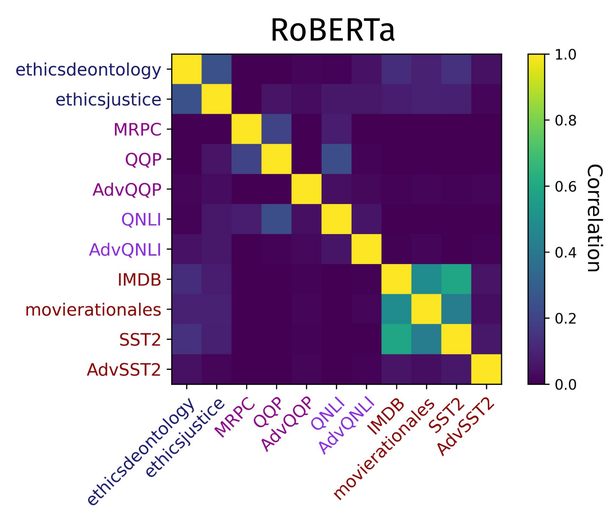
➔ Prompt tuning is relatively successful.
➔ Prompts are highly transferable within the same type of task (not shown here).
➔ Prompt tuning is not robust to adversarial data but T5 is more robust than RoBERTa.

Skill neurons

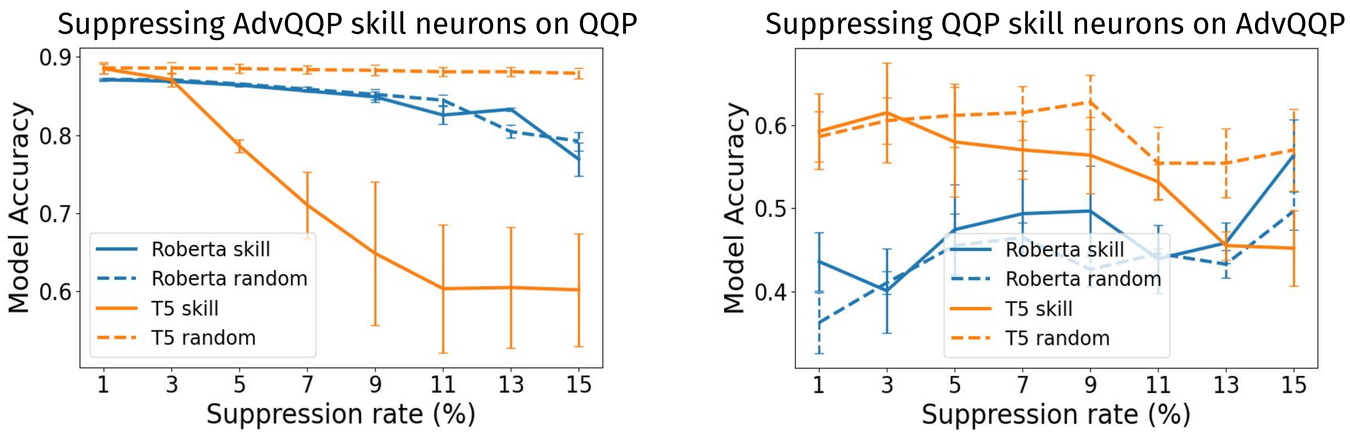
➔ We identify skill neurons in RoBERTa and T5.
➔ These neurons are highly predictive, task-specific (correlation analysis), and important (perturbation analysis).

Relation between skill neurons and adversarial robustness

Neuron predictivity correlation



Ablation: Perturbation Analysis



➔ There is a strong correlation between neuron predictivities on the adversarial datasets and their non-adversarial counterparts for T5 (ρ: 0.57–0.84), but not for RoBERTa (ρ: -0.01–0.07).
➔ An ablation analysis confirms that there is a stronger overlap in skill neurons between adversarial and non-adversarial datasets for T5 than RoBERTa.

Conclusion

- Both RoBERTA and T5 exhibit skill neurons. These neurons are highly predictive and task-specific. Suppressing these skill neurons significantly impacted task performance, highlighting their importance.
- Prompt tuning yields prompts that are transferable between similar tasks but not robust to adversarial attacks. T5 is more robust than RoBERTa.
- The skill neurons of T5 determined on non-adversarial data are also among the most predictive neurons on the adversarial data, which is not the case for RoBERTa.

➔ **Higher adversarial robustness may be related to a model's ability to consistently activate the relevant skill neurons on adversarial data.**

References & link

- Lester et al. (2021). [The Power of Scale Parameter-Efficient Prompt. EMNLP.
- Wang et al. (2022). Finding Skill Neurons in Pre-trained Transformer-based Language Models. EMNLP.