On the Relationship between Skill Neurons and Robustness in Prompt Tuning

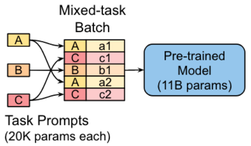
Leon Ackermann, Xenia Ohmer
Osnabrück University

Abstract

This paper investigates the robustness of RoBERTa and T5 after Prompt Tuning against Adversarial GLUE with the help of Skill Neurons.
The following novel contributions are made.
1. Prompt Tuning does not produce more robust models than model tuning
2. Skill Neurons can be found in other models than RoBERTa (like T5)
3. Skill Neurons suggest to be important for model robustness

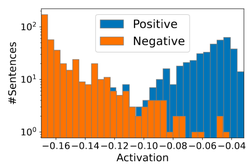
Background

Prompt Tuning
- PEFT method
- Prepend tunable parameters to prompt and train them



Prompt Tuning

Skill Neurons
- highly predictive, task specific and important neurons in FFN network
- Can be found with Prompt Tuning



Activation Histogram of one neuron for SST2

Adversarial Robustness
- Evaluation Performance on perturbed inputs

Methods

Repeat 5 times for each model-dataset combination
1. Prompt Tuning $(\mathbf{P}, \mathbf{X}) = [\mathbf{p}_1, \ldots, \mathbf{p}_p, \mathbf{x}_1, \ldots, \mathbf{x}_s]$, with $(\mathbf{P}, \mathbf{X}) \in \mathbb{R}^{(p+s)\times h}$
2. Skill Neurons
   1. Baseline activations $a_{\text{bsl}}(\mathcal{N}, \mathbf{p}_i) = \frac{1}{|D_{\text{train}}|} \sum_{\mathbf{X}_i \in D_{\text{train}}} a(\mathcal{N}, \mathbf{p}_i, (\mathbf{P}, \mathbf{X}_i))$
   2. Neuron accuracies $\text{Acc}(\mathcal{N}, \mathbf{p}_i) = \frac{\sum_{(\mathbf{X}_i, y_i) \in D_{\text{dev}}} \mathbf{1}_{[a(\mathcal{N}, \mathbf{p}_i, (\mathbf{P}, \mathbf{X}_i)) > a_{\text{bsl}}(\mathcal{N}, \mathbf{p}_i)] = y_i]}}{|D_{\text{dev}}|}$
   3. Neuron predictivities $\text{Pred}(\mathcal{N}, \mathbf{p}_i) = \max\left(\text{Acc}(\mathcal{N}, \mathbf{p}_i), 1 - \text{Acc}(\mathcal{N}, \mathbf{p}_i)\right)$

Maximum Aggregation of neuron predictivities $\text{Pred}(\mathcal{N}) = \frac{1}{k} \sum_{\mathbf{P}_i \in \mathcal{P}} \max_{\mathbf{p}_j \in \mathbf{P}_i} \text{Pred}(\mathcal{N}, \mathbf{p}_j)$

Analyses on Robustness, Transferability of Prompts and Predictivity, Task-specificity and model importance of Skill Neurons

Experiments

Models	Tasks	Datasets
RoBERTa	Ethical Judement	*EthicsDeontology, EthicsJustice*
	Paraphrase Identification	*MRPC, QQP, AdvQQP*
	Natural Language Inference	*QNLI, AdvQNLI*
T5	Sentiment Analysis	*IMDB, movierationales, SST2, AdvSST2*

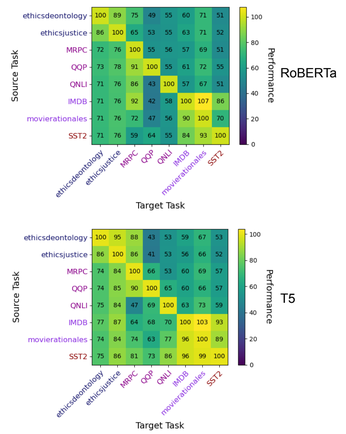
Results - Prompt Tuning

T5 is more robust than RoBERTa

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 ± 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

Mean and standard deviation of model's accuracy after Prompt Tuning across five different random seeds

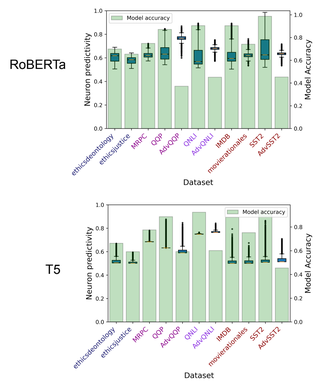
Prompt Tuning creates highly transferable prompts



Relative Transferability of Continuous Prompts

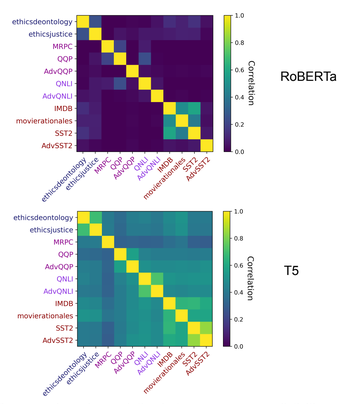
Results - Skill Neurons

Skill Neurons are highly predictive



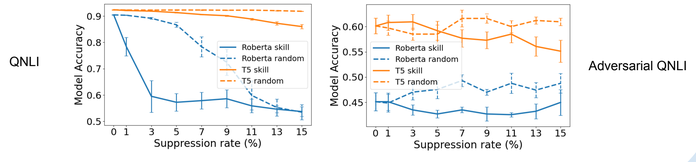
Model Performance (bar) and skill neuron predictivities (box)

Skill Neurons are task specific



Spearman's rank correlation between neuron predictivities

Skill Neuron are important for model performance

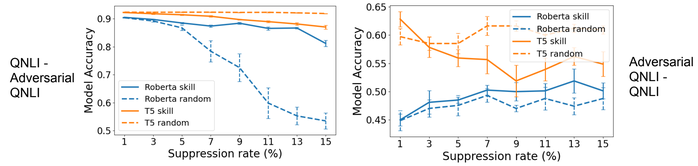


Suppressing 1-15% of activations of skill neurons and random neurons

Discussion

Results suggest connection between Skill Neurons and Robustness
1. T5 is more robust than RoBERTa
2. T5's skill neurons from non-adversarial datasets and adversarial datasets show a higher overlap than RoBERTa's

Robustness is higher if skill neurons of non-adversarial and corresponding adversarial dataset show high overlap



Model accuracies on the adversarial dataset when suppressing the skill neurons on the corresponding non-adversarial dataset and vice versa

Conclusion

1. Prompt Tuning is not inherintly more robust than model tuning
2. T5 is more robust than RoBERTa against Adversarial GLUE
3. Skill Neurons can be found in T5 (in addition to RoBERTa)
4. Activation of relevant skill neurons of non-adversarial datasets for adversarial datasets might increase robustness

Limitations

1. It cannot be excluded that some Skill Neurons resemble spurious correlations
2. Only the encoder of T5 was in scope of study
3. The investigated models are relatively small, larger ones might behave differently

Acknowledgements

We would further like to thank Elia Bruni and Michael Rau for helpful discussions. We would like to thank the Universitätsgesellschaft Osnabrueck for sponsoring our participation at the NeurIPS R0-FoMo Workshop.

References

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3045–3059

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. On transferability of prompt tuning for natural language processing. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3949–3969

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. Finding skill neurons in pre-trained transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11132–11152

Full Paper