MULTITASK PROMPTED TRAINING ENABLES ZERO-SHOT TASK GENERALIZATION

BigScience

https://github.com/bigscience-workshop/t-zero

https://huggingface.co/bigscience/T0pp

Large language models have recently been shown to attain reasonable zero-shot generatlization on a diverse set of tasks. This is supposedly a consequence of implicit multitask learning. **Can zero-shot generalization instead be directly induced by explicit multitask learning?**

Our contributions

- We converted a large number of supervised datasets into a unified human-readable prompted format, each with multiple prompts.
- We fine-tuned a pre-trained language model (T5 LM Adapted [2]) on these prompted datasets in a supervised and massively multitask fashion.
- On 9 out of 11 of unseen tasks, our model **T0** matches or outperforms models 16x bigger (GPT-3), while being more robust to the prompt formulation.

BigScience

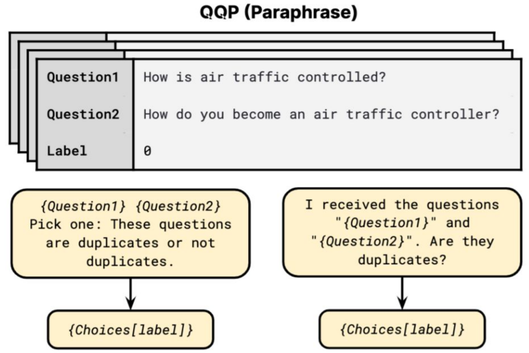
- BigScience is a year-long research workshop on large multilingual models and datasets, gathering 900+ researchers from 60 countries and more than 250 institutions.
- We are currently training a 176B parameters model which is expected to end early July.

References

[1] PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. Bach et al. ACL2022
[2] The Power of Scale for Parameter-Efficient Prompt Tuning. Lester et al. EMNLP2021

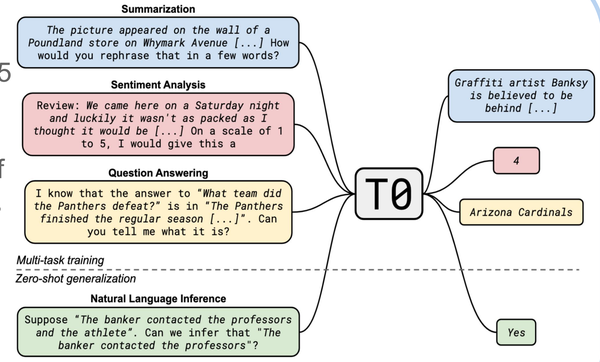
Prompting datasets

- We developed PromptSource [1], an interactive application with a templating language that make it easy to convert diverse datasets into prompts.
- As of March 2022, PromptSource contains 2'000+ prompts for 170 datasets.



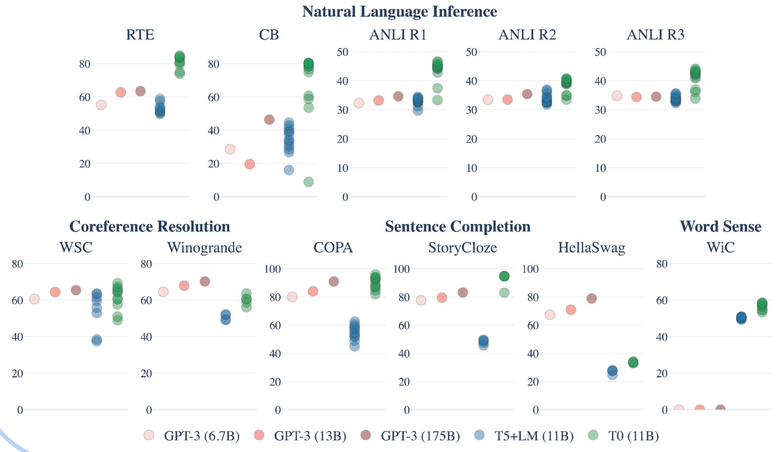
Training setup

- We fine-tuned a pre-trained language model (T5 LM Adapted [2]) on a massively multitask mixture of prompted datasets.
- We selected tasks in the training mixture and evaluated the model on held-out tasks.



Results

- Multitask prompted training outperforms language modeling training while matching or exceeding 16x bigger models.



- Adding more training prompts per dataset consistently leads to higher median performance and lower spread between best and worst prompt.