Multi-Task Zero-Shot modeling with Test Domain Shift: an exploration of sampling and fine-tuning techniques on DistilGPT-2 and BIG-bench

Lara Malinov
malinov@stanford.edu

Summary

Motivation: how to make small models learn better different skills in order to adapt to new situations

Method: Train DistilGPT-2 on a subset of BIG-bench tasks using sampling techniques and different fine-tune techniques and evaluate on out-of-domain tasks

Result:

- Training the linear layer with no sampling procedures achieves the best ROUGE-LSum scores
- Training with domain weighted sample performs better than task weighted samples on evaluation tasks
- Testing results show that the model with best OOS training scores are for logical reasoning and mathematics and performs best on evaluation tasks in emotional intelligence and emotional understanding

Background

- **MetaICL Research [1]:** Used 142 datasets for different experimental settings and 0-shot and in-context learning models. Result shows how models perform on unseen domains.
- **CS330 Multi-Task and Meta-Learning**
- **Surgical fine tuning [2]**

Method

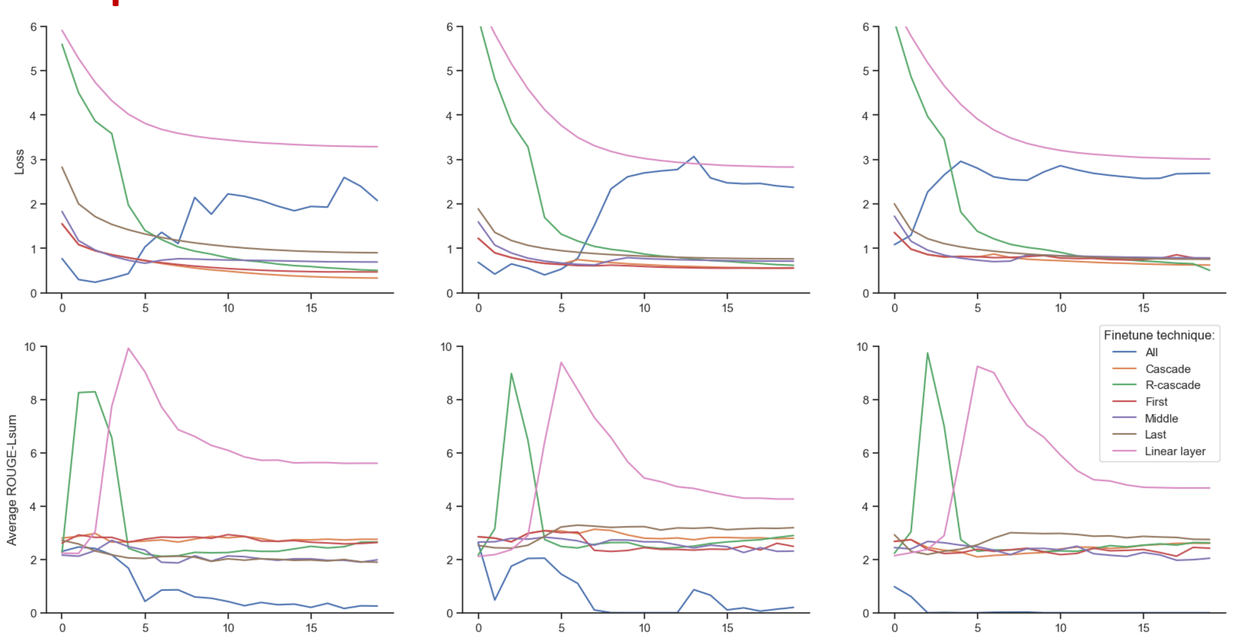
- Training using pretrained DistilGPT-2 (82M parameters, 6 Transformers Blocks)
- 3 training scenarios:
  - using all the training data
  - task weighted sample : up/down sampling each task; 500 observations per task
  - domain weighted sample: tasks are sampled s.t. all domains are equally represented
- 7 finetune approaches, training :
  - Linear layer (LL)
  - First block and LL
  - Third block and LL
  - Last block and LL
  - All parameters
  - Cascade: blocks are trained one after the other (Block 0 for 5 epochs, the following for 3 epochs) as well as LL
  - Reverse Cascade: linear layer trained for 5 epochs, Blocks 5, 4 and 3 for 3 epochs; Blocks 1 and 0 for 2 epochs) as well as LL

[1] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to Learn In Context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.
[2] Lee, Y., Chen, A.S., Tajwar, F., Kumar, A., Yao, H., Liang, P. and Finn, C., 2022. Surgical fine-tuning improves adaptation to distribution shifts. *arXiv preprint arXiv:2210.11466*.

Data

- BIG-bench tasks were first run on GPT-2 and filtered by selecting the tasks which had ROUGE-1 scores larger than 0
- Apriori itemset algorithm was applied on the keywords of the tasks to select the most common keywords/ domains
- **26 training tasks** with main keywords: *common sense, mathematics, numerical response, social reasoning, reading comprehension, contextual question-answering, logical reasoning, free response* (~10,000 obs)
  Of the training examples: 707 observations are held out for validation and testing
- **6 evaluation tasks** with no overlap with training tasks keywords: *analogical reasoning, emotional understanding, morphology, non-English, medicine, emotional intelligence, dialogue system, intent recognition* (~4,000 obs)
- Validation dataset: 25% of the held-out training observations, 25% of the evaluation task observations (~1,000 obs)
- Test dataset: remaining 75% of the held out training tasks and evaluation tasks (~3,000 obs)
- 80% of both validation and test dataset are evaluations task observations

Experiments



a) No sampling b) Task weighted sampling c) Domain weighted sampling

- As the training loss decreases the validation ROUGE-LSum scores are relatively stable across all training sample scenarios; lowest losses achieved by Cascade (no sampling, task weighted) and R-cascade (domain weighted)
- **Except:** Fine-tuning the last linear reaches the highest validation ROUGE-LSum in a couple of epochs even though the training loss is still the highest across fine-tuning approaches
- **Overall best model:** training the last linear layer indifferent of the sampling technique

- For testing, model checkpoints are selected with the highest validation ROUGE-LSum scores
- Results show that **training the linear layer** on the training tasks achieves the best scores across sampling procedure
- **Highest scores on test training task** observations is achieved by training the **last block** together with a **task weighted** training sample
- Domain weighted training sample is better than task weighting samples for evaluation tasks (with linear layer finetuning)
- **Best evaluation score** attained by training the linear layer with no sampling procedure.

	-			task weighted			domain weighted		
	All	Train	Evaluation	All	Train	Evaluation	All	Train	Evaluation
Raw DistilGPT-2	1.73	4.92	0.93	-	-	-	-	-	-
Middle	2.13	5.89	1.19	2.48	6.2	1.55	2.22	5.94	1.29
All	2.33	6.47	1.29	2.32	6.45	1.28	0.88	2.08	0.58
Last	2.53	5.40	1.82	3.25	**11.08**	1.29	2.97	**10.23**	1.15
Cascade	2.58	5.76	1.78	2.88	6.04	2.09	2.51	6.14	1.6
First	2.60	6.19	1.70	2.81	6.86	1.79	2.51	6.14	1.6
R-Cascade	7.97	6.11	8.44	8.14	6.12	8.65	8.75	6.31	9.36
Linear layer	**9.40**	**6.73**	**10.06**	**8.26**	6.11	**8.8**	**8.78**	6.3	**9.39**

*Note R-Cascade is selected at epoch 2, thus it is equivalent to Linear Layer fine-tuning. The only difference is that R-Cascade is trained in batches of 16 and Linear Layer in batches of 32.

Analysis

Tasks have multiple keywords, thus the visualization represents the average ROUGE-LSum testing tasks scores for pairs of keywords using the model trained without sampling procedures and fine-tuned on the linear layer

- Model performs best on OOS training observations with pairs of keywords: **logical reasoning, mathematics, contextual question-answering and numerical response**
- BUT on testing evaluation observations the highest score is for the keyword pair: **emotional intelligence – emotional understanding**