Meta-Training Methods in Natural Language Processing

Tianjian Li

Johns Hopkins University

June 30, 2022

Abstract

Meta learning(i.e. learning to learn) is a learning paradigm where models are not directly trained on a specific task but rather trained on instructions on how to perform these tasks. There have been an emergent trend of applying meta learning methods to Natural Language Processing. This note will cover some of the canonical meta learning methods including T0[8], where a pretrained T5[4] model is fine-tuned on various tasks; FLAN[9] where a very large pretrained language model(137B parameters) is fine-tuned on various tasks with instructions; MetaICL[6], where a pretrained language model is finetuned to perform in-context learning; Few shot meta learning methods such as STILT:Intermediate Pre-Training on selected tasks for low resource downstream task [1]; Multi-Task Fine-Tuning [5] which co-finetunes a low resource task with a additional self-supervised objectives;

1 Zero-Shot

1.1 T0

T0[8] hypothesizes that the reason why large language model have reasonable zero-shot performance is that they are training implicitly under a **multitask** setting. To verify their hypothesis, the authors of T0 conducted a set of experiments by explicitly multitask training a pretrained language model. The authors mainly studies two questions

1. Can multitask training improve performance on unseen tasks?

   The answer is yes. T0[9] outperforms T5+LM[4] on unseen tasks significantly. Moreover, T0 outperforms GPT-3[3] on 9/11 unseen tasks. The author also finds out that scaling the number of tasks during such a intermediate multitask fine-tuning phase leads to better performance.

2. Can we vary the choice of prompts to get a a model that is robust to prompt wording?

   The answer is yes. Increasing the number of prompts does yield better performance and the variance is generally lower(more robust) if we increase the number of prompts. Note that the variance is sometimes the lowest when using no prompts. The author concludes by saying that "Adding more datasets consistently leads to higher median performance but does not always reduce interquartile range for held-out tasks"

1.2 FLAN: Instruction Tuning

FLAN[9] is very similar to T0[8], in that they both pre-finetune a large pretrained langauge model on a collection of tasks. On a higher level intuition, this work(FLAN) is actually telling the model to understand the "instructions", so that given a new task, we can leverage that the model understands instructions to give it new instructions for the new task to achieve better performance. Compared with T0, FLAN uses a much larger model(137B vs 11B), and uses a decoder-only Transformer model as the base model as opposed

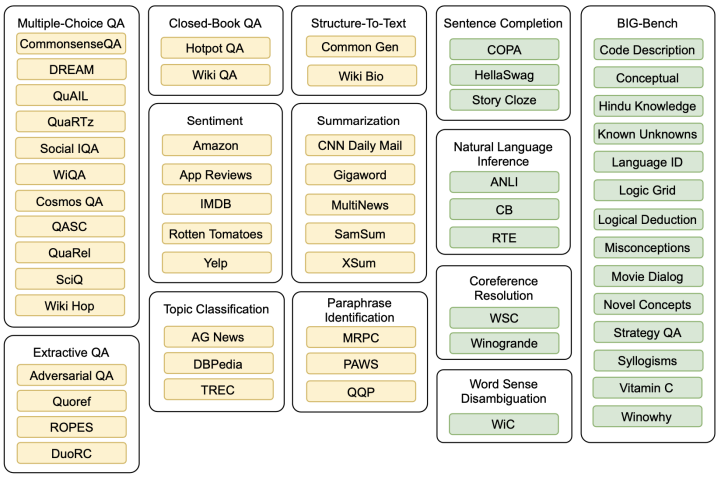

Figure 1: detailed datasets of T0, figure pasted from [9]

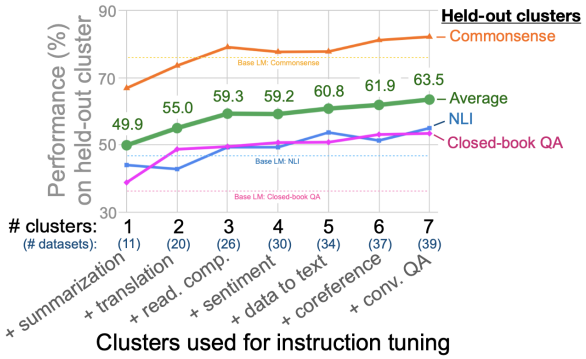

Figure 2: detailed datasets of T0, figure pasted from [9]

to the T5 model with adopts a encoder-decoder Transformer as their base model.

The experimental results of FLAN echoes with the results of T0. Notably the number of tasks and the size of the model contribute most to the zero-shot performance. From the following figure we can see that adding more datasets consistently improve the zero-shot performance on held-out tasks.

Some major drawbacks of FLAN multitasking is that it is not useful in everytask, especially when the held-out task is similar to the mask language modeling objective used during pretraining. From figure 3 we can see that FLAN instruction tuning only works when the model size is large than 8B. The authors extrapolate that "potentially all model capacity is used to learn the mixture of instruction tuning tasks" (instead of learning the semantic meaning of instructions).

1.3 MetaICL

In Context Learning(ICL) have been popularized since the release of the famous GPT-3[3] model, where a pretrained language model(usually very large, GPT-3 uses 175 billion parameters) is given a few examples, concatenated with the actual input at the end to generate an answer. The model learns structured and semantic knowledge of the task from the given examples. MetaICL[6] aims to let the pretrained language model(in this case, a GPT-2[2] model with 770M parameters) learn how to perform ICL through feeding it ICL examples of different tasks.

	Meta-training	Inference
Task	*C meta-training* tasks	An unseen *target* task
Data given	Training examples $\mathcal{T}_i = \{(x_j^i, y_j^i)\}_{j=1}^{N_i}$, $\forall i \in [1, C]$ $(N_i \gg k)$	Training examples $(x_1, y_1), \cdots, (x_k, y_k)$, Test input x
Objective	For each iteration, 1. Sample task $i \in [1, C]$ 2. Sample $k+1$ examples from $\mathcal{T}_i$: $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ 3. Maximize $P(y_{k+1}\|x_1, y_1, \cdots, x_k, y_k, x_{k+1})$	$\text{argmax}_{c \in \mathcal{C}} P(c\|x_1, y_1, \cdots, x_k, y_k, x)$

Figure 3: overview of metaICL, figure pasted from [6]

One major edge of MetaICL over other zero-shot frameworks is that MetaICL does not rely on task reformatting, which results in high variance in template engineering[7] but rather use all datasets in their as-is condition. The major results are:

1. MetaICL outperforms GPT-3[3], T0[8] and FLAN[9] under zero-shot settings.
2. The gains over zero-shot transfer are particularly significant with meta-training tasks and target tasks are **dissimilar**.

Experimental results show that:

- MetaICL significantly outperforms baselines in most settings, it only marginally outperforms Multi-task 0-shot in the QA→QA setting, as an exception. This is likely because the meta-training and target tasks are relatively similar
- Gains over Multi-task 0-shot are more significant on target tasks in **unseen domains**.
- Fine-tuning with meta-train yields the best results with the only exception with the downstream task is QA(both non-QA to QA and QA to QA).
- Number of tasks and task diversity does a lot of the heavy lifting, which echos the finding of T0 and FLAN.

Again I would like to reiterate that all of the downstream tasks falls into the NLI category. Further investigations are required to prove that meta-training method works for **zero-shot natural language generation**.

2 Few-Shot

References

[1] Jason Phang, Thibault Févry, and Samuel R. Bowman. "Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks". In: *CoRR* abs/1811.01088 (2018). arXiv: 1811.01088. URL: http://arxiv.org/abs/1811.01088.

[2] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019).

[3] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[4] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[5] Ahmed Magooda, Diane Litman, and Mohamed Elaraby. "Exploring Multitask Learning for Low-Resource Abstractive Summarization". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1652–1661. DOI: 10.18653/v1/2021.findings-emnlp.142. URL: https://aclanthology.org/2021.findings-emnlp.142.

[6] Sewon Min et al. "MetaICL: Learning to Learn In Context". In: *NAACL-HLT*. 2022.

[7] Swaroop Mishra et al. "Reframing Instructional Prompts to GPTk's Language". In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 589–612. DOI: 10.18653/v1/2022.findings-acl.50. URL: https://aclanthology.org/2022.findings-acl.50.

[8] Victor Sanh et al. "Multitask Prompted Training Enables Zero-Shot Task Generalization". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=9Vrb9D0WI4.

[9] Jason Wei et al. "Finetuned Language Models are Zero-Shot Learners". In: *International Conference on Learning Representations*. 2022. URL: https://openreview.net/forum?id=gEZrGCozdqR.