Can T5 be Taught Twice for Data-to-Text Generation?

Chenhao Zhou
Yuanpei College, Peking University
2100017709@stu.pku.edu.cn

Abstract

In order to compare the efficiency of multiple fine-tuning iterations versus a singular fine-tuning process for pre-trained models, I framed the background within the context of a data-to-text generation tasks. By employing end-to-end full-parameter fine-tuning and PEFT (parameter-efficient fine-tuning) methods on T5 model, I ultimately arrived at a direct response to the posed inquiry.

1 Introduction

In industrial settings, it is customary to employ strategies involving fine-tuning pre-trained language models for downstream tasks to fulfill specific business requirements. When confronted with a novel downstream task, the following dilemma often arises: whether to employ a previously fine-tuned model tailored for prior requirements or to initiate training anew from an entirely non-fine-tuned pre-trained model.

To investigate this issue, I framed the context as follows: Given that a pre-trained model has undergone fine-tuning for a data-to-text generation task in the restaurant domain, can subsequent fine-tuning be conducted for a data-to-text generation task in the video game domain while ensuring the model's performance on both tasks? This report contains experiments description in Sec.2 and current limitations in Sec.**??**

2 Experiments

2.1 Datasets

I conduct experiments on 2 English datasets for data-to-text generation task.

E2E is a new dataset (Novikova et al., 2017) for training end-to-end, data-driven natural language generation systems in the restaurant domain, which is ten times bigger than existing, frequently used datasets in this area.

Flat MR	NL reference
name[Loch Fyne], eatType[restaurant], food[French], priceRange[less than £20], familyFriendly[yes]	Loch Fyne is a family-friendly restaurant providing wine and cheese at a low cost.
	Loch Fyne is a French family friendly restaurant catering to a budget of below £20.

Table 1: An example of a data instance from e2e dataset (Novikova et al., 2017).

give_opinion(NAME [**SpellForce 3**], RATING [**poor**], GENRES [**real-time strategy, role-playing**], PLAYER_PERSPECTIVE [**bird view**])
I think that **SpellForce 3** is **one of the worst games** I've ever played. Trying to combine the **real-time strategy** and **role-playing** genres just doesn't work, and the **bird's eye view** makes it near impossible to play.
verify_attribute(NAME [**Little Big Adventure**], RATING [**average**], HAS_MULTIPLAYER [**no**], PLATFORMS [**PlayStation**])
I recall that you were **not that fond** of **Little Big Adventure**. Does **single-player** gaming on the **PlayStation** quickly get boring for you?

Table 2: Examples of MRs and corresponding references utterances in the ViGGO dataset (Juraska et al., 2019).

ViGGO is an English data-to-text generation dataset (Juraska et al., 2019) in the video game domain, with target responses being more conversational than information-seeking, yet constrained to the information presented in a meaning representation.

2.2 Training

Data preprocessing In order to ensure the rationality of the experimental setup, efforts were made to align the two datasets as closely as possible in both format and scale. Specifically, with

Type	E2E			ViGGO		
	BLEU	ROUGE-L	METEOR	BLEU	ROUGE-L	METEOR
t5-base-E	23.2	0.44	0.24	-	-	-
t5-base-V	-	-	-	24.8	0.44	0.26
t5-base-E-V	18.8	0.40	0.18	24.9	0.44	0.25
t5-large-E	23.7	0.44	0.24	-	-	-
t5-large-V	-	-	-	23.8	0.42	0.22
t5-large-E-V	16.5	0.36	0.18	23.0	0.43	0.22
t5-3b-E	24.6	0.45	**0.25**	-	-	-
t5-3b-V	-	-	-	**26.4**	**0.46**	**0.27**
t5-3b-E-V	**24.8**	**0.46**	0.24	11.7	0.28	0.14
t5-3b-V-E	17.7	0.40	0.19	26.2	**0.46**	0.26

Table 3: Evaluation results of different types of fine-tuned model. Note that t5-xxx-A-B means fine-tuned on dataset A then dataset B.

reference to the 9 slots[1] of mean representations within the ViGGO dataset, I appended the prefix "inform" to all meaning representations in the E2E dataset. Simultaneously considering that the size of the ViGGO dataset is only 13% of the E2E dataset, I opted to sample 1/6 of the E2E training set.

Pre-trained model Drawing inspiration from prior work that transforms data-to-text tasks into text-to-text formulations (Kale and Rastogi, 2020), I opted for the utilization of the T5 pre-trained model (Raffel et al., 2020) with different size, including t5-base (220 million), t5-large (770 million) and t5-3b (3 billion).

Implementation details I decided to full-parameter fine-tune t5-base and t5-large, while using LoRA (Hu et al., 2022) to parameter-efficient fine-tune t5-3b. The specific procedure involved initial separate training on both datasets, followed by subsequent training using the ViGGO dataset to fine-tune the model that had been previously trained on the E2E dataset. To ensure uniformity, training was consistently conducted with a duration of 3 epochs and 8 batch size across all experiments.

2.3 Evaluation

2.3.1 Evaluation Metrics

Following the current general metrics for data-to-text generation tasks, three metrics are used to evaluate the generated sentences, *i.e.* BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005)

[1]Indicate 9 dialogue acts: inform, confirm, give_opinion, recommend, request, request_attribute, request_explanation, suggest, verify_attribute

2.3.2 Results Analysis

- The performance of fine-tuning on a pre-trained model is observed to be more pronounced with increasing scale, as evidenced in either of the two data-to-text datasets.
- Fine-tuning t5 model twice on the mentioned two data-to-text datasets could cause the loss, despite efforts to process the datasets' format to be as similar as possible.
- While ViGGO is only 13% the size of the E2E dataset, the lexical diversity of its utterances is 77% of that in the E2E dataset (Juraska et al., 2019). Therefore, transitioning from the ViGGO dataset to E2E dataset appears to be a more straightforward process than the reverse.

3 Conclusion and Limitations

This simple experiment suggests that, when confronted with two distinct downstream tasks, conducting separate fine-tuning operations may prove to be more efficient than performing dual fine-tuning iterations, given the specific conditions of the study.

Future work may dive into a broader spectrum of tasks to further investigate this issue. Additionally, it would be beneficial to specifically explore the impact of training order, an aspect not addressed in my current study.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Pro-*

ceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Juraj Juraska, Kevin Bowden, and Marilyn Walker. 2019. ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 164–172, Tokyo, Japan. Association for Computational Linguistics.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Saarbrücken, Germany. ArXiv:1706.09254.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.