ADSFT: Adaptive and Domain-Specific Fine-Tuning Paradigms for Generative Models

Tao Liu **Yuanpei Sui** **Junyi Wang**

1 Background

In the e-commerce industry, generating high-quality advertisements tailored to specific products is crucial for enhancing product visibility and improving sales conversion rates. The Amazon Product dataset offers a rich set of product information, including keywords and key-value attribute lists (kv-list), that supports the creation of precise advertisement generation tasks. However, the generated advertisements need to ensure both high generation speed and factual consistency with the input product information. Therefore, the goal of this research is to fine-tune ChatGLM3-6B to generate e-commerce advertisements more quickly and accurately, improving the overall quality and efficiency of the advertisement generation process.

2 Problem Formulation

Given a product's keyword and attribute list (kv-list), the goal is to generate an advertisement (adv) that fits the product and ensures consistency with the product information. Formally, this can be defined as:

$$adv = M(K, V) \tag{1}$$

where $K = \{k_1, k_2, ..., k_n\}$ represents the set of keywords, and $V = \{v_1, v_2, ..., v_m\}$ represents the set of product attributes. M is the model, which needs to be fine-tuned based on ChatGLM3-6B to generate the advertisement text. The final output advertisement (adv) needs to maintain high consistency with the input information while meeting the speed requirements for real-time advertisement generation.

3 Related Work

Generative models have been applied to advertisement generation tasks, with transformer-based models [4] like GPT-3 and BERT [1] showing promising performance. However, these models still face challenges in terms of improving generation speed, especially in large-scale advertisement tasks, and ensuring factual consistency with the input information. To address these issues, techniques such as LoRA [2] and Prompt Tuning [3] have been introduced, which can effectively reduce the cost of fine-tuning and improve generation efficiency. LoRA (Low-Rank Adaptation) reduces the computational complexity by introducing trainable low-rank matrices to the pre-trained model, making it suitable for efficient model adaptation in resource-limited settings. Prompt Tuning, on the other hand, modifies the prompt to guide the model's generation process. Combining these methods can further enhance the precision and flexibility of advertisement generation.

4 Proposed Methodology

To optimize the speed and quality of advertisement generation, we propose the following three innovative methods for fine-tuning the ChatGLM3-6B model:

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1. Subsetting the Dataset: We propose using a subset selection method inspired by the B-S problem to select a representative subset from the records in the Amazon Product dataset. This will reduce the computational overhead during model training and inference, while maintaining the diversity and representativeness of the training data.

2. Dynamic Low-Rank Adjustment (LoRA): We will start the training with a higher rank value to allow the model to learn sufficient features from the data. As the model stabilizes, we will gradually lower the rank to reduce the complexity of parameter convergence, thereby speeding up the training process and optimizing resource utilization.

3. Combining Prompt Tuning and LoRA: By combining Prompt Tuning with LoRA, we can guide the advertisement generation process more precisely while minimizing parameter adjustments. This combination will improve the accuracy of generated advertisements and enhance model performance while maintaining a high generation speed and low resource consumption.

5 Datasets and Evaluation

We will use the Amazon Product dataset, which consists of e-commerce product descriptions and their corresponding keyword-attribute pairs. The dataset contains millions of records, and we plan to optimize the dataset using a subset selection approach to maintain diversity while reducing computational complexity. Performance will be evaluated by accuracy along with efficiency, which represented by Bleu, Meteor, Rouge, Cider and inference speed separately.

References

[1] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[3] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[4] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.