ASSIGNMENT 3: RESEARCH ON SFT METHOD

PENG Qiheng
225040065
Data Science
Shenzhen, China

1 RESEARCH TOPIC

Research topic Research Topic: Fine-tuning Method on Pre-trained Models

Importance of the topic The goal of this experiment was to explore how supervised fine-tuning (SFT) can improve the model's instruction-following ability in this domain.

Task to do In this assignment, I fine-tuned a pre-trained large language model (LLM) for Chinese medical dialogue using Huatuo26M-Lite dataset.

2 EXPERIMENT DESIGN

Dataset I used the Huatuo26M-Lite from Huggingface.

Base Model I selected Qwen3-4B as the backbone model due to its strong performance in Chinese tasks.

Training Method I employed SFT with QLoRA for parameter-efficient fine-tuning. The training was conducted with a RTX 5090 GPU.

Evaluation I used ChatGPT-3 as an evaluator, following the prompt template from the document. I used 20 test questions and compared my model's outputs with a baseline.

3 CODE IMPLEMENTATION

3.1 EXPERIMENT SETTING

```
# CUDA=12.8, python==3.12
torch==2.8.0+cu128
```

3.2 MODEL LOADING

```
bnb_4bit_quant_type = nf4
use_nested_quant = True
model_id = Qwen/Qwen3-14B
bnb_config = BitsAndBytesConfig(
    load_in_4bit=True,
    bnb_4bit_use_double_quant=use_nested_quant,
    bnb_4bit_quant_type=bnb_4bit_quant_type,
    bnb_4bit_compute_dtype=torch.bfloat16
)
tokenizer = AutoTokenizer.from_pretrained(model_id)
model = AutoModelForCausalLM.from_pretrained(
    model_id,
    quantization_config=bnb_config,
    device_map={ :0}
)
```

3.3 EXPERIMENT

Setting 1: **Baseline model**

Setting 2: **Model scaling up (Qwen3-4B → Qwen3-14B)** The parameter count has increased from 4 billion to 14 billion, significantly increasing the model capacity. Keep other training parameters unchanged for fair comparison.

Setting 3: **Data Augmentation** Randomly select 30% of the samples from the training set for enhancement, generate synonym replacement versions for each selected question, keep the original answer unchanged, and only enhance the part of the question.

Setting 4: **Training Tricks** Modify training hyperparameters:

```
training_arguments = transformers.TrainingArguments(
    output_dir="./checkpoint",
    num_train_epochs=1,
    per_device_train_batch_size=10,
    per_device_eval_batch_size=10,
    gradient_accumulation_steps=1,
    optim="paged_adamw_32bit",
    save_steps=0,
    logging_steps=1,
    learning_rate=1e-4,
    weight_decay=0.001,
    max_steps=-1,
    warmup_ratio=0.03,
    group_by_length=True,
    lr_scheduler_type="cosine",
    gradient_checkpointing=False,
    report_to="none",
    eval_strategy="steps",
    eval_steps=1,
)
```

ChatGPT	winning rate
Ours (*setting 1*)	75%
Ours (*setting 2*)	90%
Ours (*setting 3*)	80%
Ours (*setting 4*)	80%

Table 1: Experimental Results

3.4 EVALUATION

We use ChatGPT for evaluation with the following prompt:

We would like to request your feedback on the two AI assistants in response to the user question displayed above.
Please evaluate the helpfulness, relevance, accuracy, level of details of their responses. You should tell me whether Assistant 1 is 'better than', 'worse than', or 'equal to' Assistant 2.
Please first compare their responses and analyze which one is more in line with the given requirements.
In the last line, please output a single line containing only a single label selecting from 'Assistant 1 is better than Assistant 2', 'Assistant 1 is worse than Assistant 2', and 'Assistant 1 is equal to Assistant 2', avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

4 RESULT ANALYSIS

4.1 QUANTITATIVE EVALUATIONS

Based on the experimental results, we can draw the following quantitative insights:

Model size scaling has a significant impact on performance. The winning rate increased by 15 percentage points when scaling from Qwen3-4B to Qwen3-14B. This demonstrates that larger model capacity substantially enhances the model's ability to understand and generate professional medical dialogue, consistent with the scaling laws observed in large language models. Data augmentation (Setting 3) yielded a 5 percentage point improvement, indicating that increasing training data diversity through synonym replacement effectively improves the model's adaptability to different user question formulations. This is particularly valuable in medical dialogue where patients may express the same concern in various ways. Training technique optimization (Setting 4) also achieved a 5 percentage point improvement, highlighting the importance of proper hyperparameter configuration. Adjustments to learning rate, batch size, and optimizer selection significantly impacted fine-tuning effectiveness. Comparative impact: The performance gain from model size scaling was substantially higher than from data augmentation or training technique optimization. This suggests that when resources are limited, prioritizing model size expansion may yield better returns than other techniques. Consistent improvement: All four settings significantly outperformed the baseline model, confirming that supervised fine-tuning is an effective approach for improving instruction-following ability in medical dialogue tasks.

4.2 CASE STUDY

User question: ”请描述一下病毒性感染的传播途径以及如何防止传播？”

Baseline (Setting 1): ”病毒性感染是指由病毒感染引起的疾病，其传播方式和预防措施因不同的病原体而异。以下是对常见病毒性感染的传播途径及防范方法的概述：一、常见的病毒性感染及其传播途径1. 流感（如甲型H1N1）传播途径：空气飞沫传播：咳嗽、打喷嚏时产生的飞沫可被他人吸入。接触传播：手部接触污染表面后触摸...”

Qwen3-14B (Setting 2): ”病毒性感染是指由病毒感染引起的疾病，这类疾病的传播方式多种多样。了解其主要传播途径有助于采取有效的预防措施以减少传染风险。一、常见的病毒性感染传播途径：1. 呼吸道飞沫传播：当感染者咳嗽、打喷嚏或说话时会释放含有病毒的飞沫。健康人吸入这些带有病毒的小液滴后可能被感染（如流感、新冠等）。密切接触传播...”

Analysis: The answer of the 14B model first lists the transmission channels, and then answers the measures to prevent transmission based on the channels, with a chain thinking logic, while the baseline model answers directly.

5 CONCLUSION

This research explored the effectiveness of supervised fine-tuning (SFT) methods for enhancing Chinese medical dialogue systems. Our key findings include:

Model size is a critical factor: Scaling from 4B to 14B parameters yielded the most significant performance improvement, confirming that larger models better capture the complex medical knowledge required for accurate dialogue.

Data augmentation provides meaningful enhancement: Increasing training data diversity improves the model's ability to handle varied patient inquiries, a crucial aspect of medical dialogue systems.

Training methodology matters: Proper fine-tuning configuration is essential for maximizing model performance in specialized domains.

ACKNOWLEDGMENT

This is the Assignment3 for DDA6307 / CSC6052 / MDS6002, see details in https://nlp-course-cuhksz.github.io/.

REFERENCES