Enhancing Low-Resource LLMs Classification with PEFT and Synthetic Data

Parth Patwa, Simone Filice, Zhiyu Chen, Giuseppe Castellucci, Oleg Rokhlenko, Shervin Malmasi

Amazon, USA

Intro

- 0-shot classification is convenient and fast inference.
- Few-shot In Context Learning (ICL) is more accurate but the latency increases.
- Few-shot examples for parameter efficient fine-tuning (PEFT) is faster but as we show, it performs much worse than few-shot ICL.

Our Contributions

- Our framework for few-shot PEFT is faster, cheaper and more accurate.
- We use the same LLM - no external LLM or dataset.
- Our classifier has better accuracy on 3 text classification tasks while being a lot more efficient (~2x to 5x speedup).
- To the best of our knowledge, we are the first to improve few-shot PEFT without additional resources (external datasets or models).

Methodology

- We explore a low-resource setting: few (4) labeled examples per class.
- We augment the training data with synthetic data to ensure better training of PEFT.
- Our method has 3 steps: generate data, filter data, and train.

Method Overview Diagram

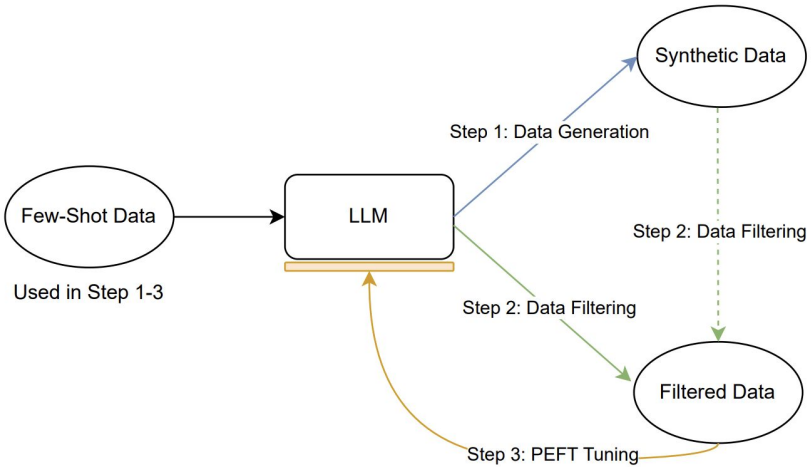


Figure 1: The overview of our method. First, very few real data points are used to generate synthetic data using ICL. Then, the synthetic data is filtered using ICL by LLM again. Finally, the filtered data and the real data are combined to train the LLM using LoRA.

Data Generation

- Similar to Chavan et al. (2023), we observe that in a few-shot setting, the performance of PEFT improves as the #training samples increases.
- Since ICL performs well, we hypothesize that the model has the inherent knowledge of the task but the low PEFT results are due to sub-optimal usage of the available resources.
- To fill this gap, we first use the LLM L in the ICL setting to generate class-wise synthetic data.

Data Generation Prompt

Few examples of movie reviews having positive sentiment are given. Generate more positive reviews
Text: [Positive review 1]
Label: Positive
...
Text: [Positive review 4]
Label: Positive
Text: [the model generates this]

Figure 2: An example of a prompt used for generating positive reviews for SST2 data. Four examples of the positive class are provided in the prompt.

Data Filtering

- We first apply a basic filtering to discard duplicates and too short/long texts.
- On manual inspection of the generated data, we found some instances that are not valid examples of the class they should represent due to hallucination.
- To remove these cases, we classify the generated data using ICL with L.

Classify the sentiment of the given movie review into Positive or Negative
Text: [review 1]
Label: [Label 1]
...
Text: [review 8]
Label: [Label 8]
Text: [generated review]
Label: [Predicted Label]

Training and Inference

- Finally, we use the filtered data along with the few (4 per class) real examples for the PEFT of the LLM L with LoRA.
- Note that L is used for all 3 steps, as we want to validate our hypothesis that L does not need additional knowledge to work in the PEFT setting.
- **Inference** - Given an unlabeled example, the trained LLM is asked to predict its label. Conversely to the ICL setting, the LLM does not use any example at inference time.

Experiments

- Models: Vicuna7b and Vicuna13b.
- Datasets: SST2 (sentiment analysis, binary), AG news (news classification, 4 classes) and TREC (question classification, 6 classes).
- No tuning LoRA hyper-parameters (rank=8, alpha=32,dropout=0.1).
- Similarly, no prompt engineering in ICL.
- Implementation: Hugging Face and PEFT library with torch backend.
- Compute: 4 v100 GPUs of 16GB each.

Results

- The few-shot ICL is more accurate than the 0-shot method but significantly slower.

Model	Method	# real	#syn	SST		Trec	
				acc.	inf.	acc.	inf.
Vicuna-7b	0-shot	0	0	0.55	0.27	0.16	0.28
Vicuna-7b	ICL	4	0	0.95	0.6	0.6	0.9

Results

- Few-shot LoRA - worse than few-shot ICL but as fast as 0-shot

Model	Method	# real	#syn	SST		Trec	
				acc.	inf.	acc.	inf.
Vicuna-7b	0-shot	0	0	0.55	0.27	0.16	0.28
Vicuna-7b	ICL	4	0	0.95	0.6	0.6	0.9
Vicuna-7b	LoRA	4	0	0.51	0.27	0.49	0.28

Results

- Our approach is comparable to or better than ICL

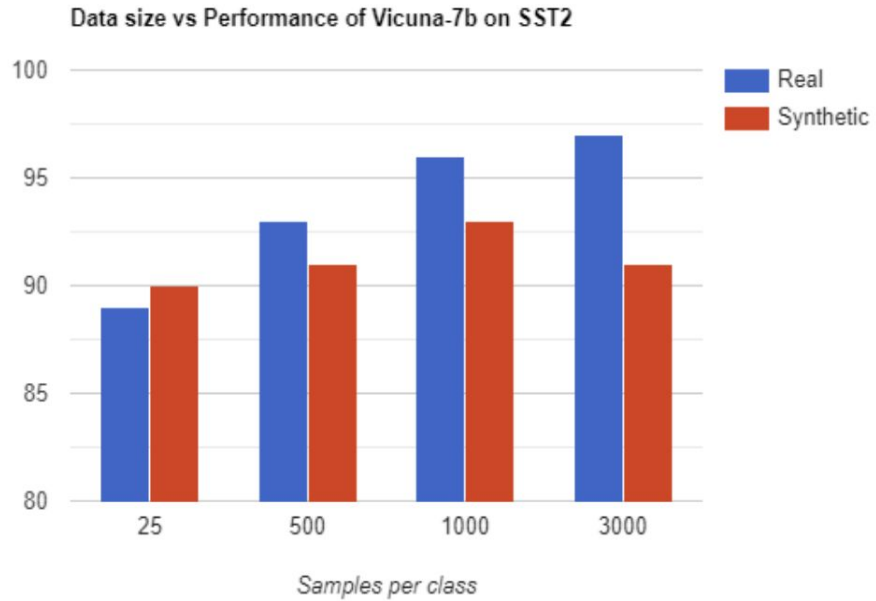
Model	Method	# real	#syn	SST		Trec	
				acc.	inf.	acc.	inf.
Vicuna-7b	0-shot	0	0	0.55	0.27	0.16	0.28
Vicuna-7b	ICL	4	0	0.95	0.6	0.6	0.9
Vicuna-7b	LoRA	4	0	0.51	0.27	0.49	0.28
Vicuna-7b	ours	4	21	0.9	0.27	0.79	0.28

Ablation Studies

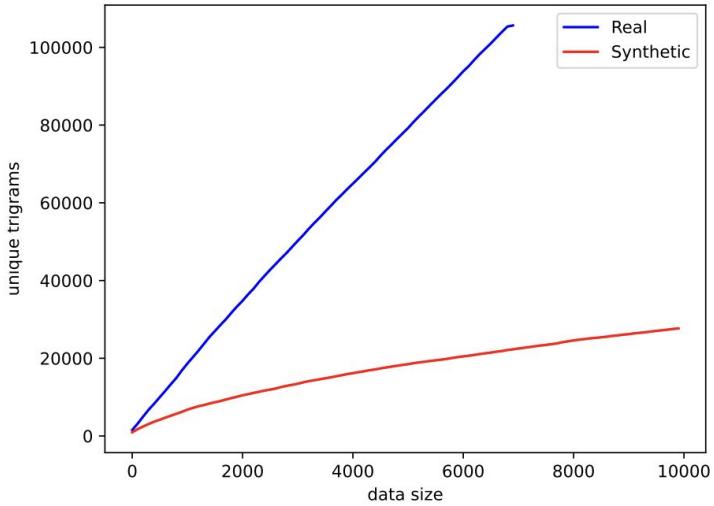
- Trec without filtering:
  - 33/125 incorrect examples, 0.68 accuracy overall
  - "Who is called the Father of Geometry” was incorrectly generated for the location class.
- Trec with filtering:
  - 5/125 incorrect examples, 0.79 accuracy overall

Effect of Data Size

- Increasing the size of real data is always beneficial
- adding more synthetic data does always not provide a clear benefit.



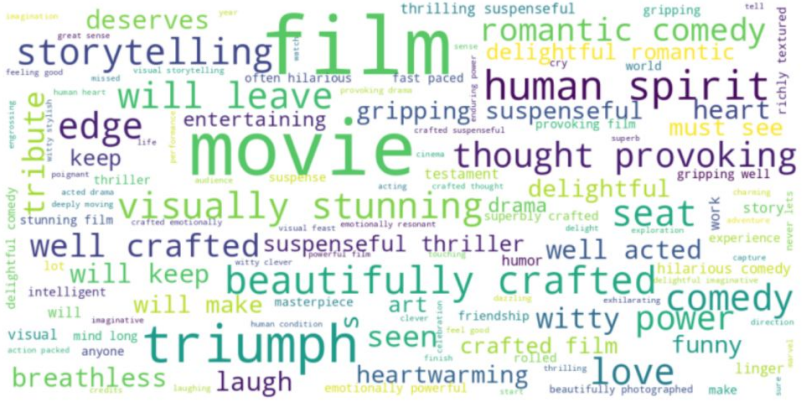
Data Diversity



- For smaller data sizes, the real data's diversity is comparable to that of synthetic data.
- However, as the data size increases, the real data diversity increases faster.
- Hence it is difficult to generate a large amount of diverse synthetic data with only 4 seed examples.

Qualitative Analysis: Positive Data Word Clouds

(a) Real data



(b) Synthetic data

- SST2 - movie review dataset. We can see that words like film, movie appear in both word clouds.
- Positive words like entertaining, beautiful, funny appear in both the word clouds.
- Other positive words like stunning, delightful only in the synthetic data word cloud whereas subtle positive words like compelling, solid are seen only in the real data word cloud.
- Hence, synthetic data has a slightly different distribution and can capture the meaning of the positive class.

Conclusion and Future Work

- Our framework makes LLMs more efficient and effective few-shot classifiers.
- First, the LLM is used to augment a very small training set with synthetic data.
- then, the LLM is used to filter the data (remove label inconsistent examples).
- finally, we use the resulting data to fine-tune the LLM using LoRA.
- Our approach+model surpasses several baselines in a few-shot setting on 3 datasets.
- Future work - improve the generation quality by promoting data diversity.

References

- Arnav Chavan, Zhuang Liu, Deepak Gupta, Eric Xing, and Zhiqiang Shen. 2023. One-for-all: Generalized lora for parameter-efficient fine-tuning.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mtbench and chatbot arena.

Related Work

- Classification can be performed in a generative way (0-shot or ICL). ChatGPT, Falcon etc show impressive results.
- He et al. (2023) propose a two-step approach where they first use ChatGPT to generate a few-shot Chain-of-Thought prompt, which they then use to annotate unlabeled data. Their procedure is relatively slow since the LLM is invoked twice with prompts that need to contain both examples and explanations. Conversely, we propose a solution whose computational complexity at inference time corresponds to the 0-shot setting case.
- Fine-tuning LLMs is expensive, but a viable solution is offered by Parameter Efficient Fine-Tuning (PEFT) techniques (Liu et al., 2022c; Liu et al., 2021), where only a small number of added or selected parameters are updated. These methods match the performance of full fine-tuning when large training datasets are available.
- On the contrary, there has been little focus on PEFT in low-resource settings. Our paper targets this scenario - we assume we can access only a few annotated examples (e.g., 4/ class) and no unlabeled data.
- A similar work is Liu et al. (2022b), which proposes a novel PEFT technique that works well in low resource settings when the PEFT weights are pre-trained and multiple tasks are trained in parallel. We differ from their work as we do not pre-train the PEFT weights and we target a single task at a time, without assuming (possibly related) data from other tasks is available.
- Relaxing this assumption is useful when dealing with peculiar tasks not sharing similarities with other available datasets.
- We are the first to improve few-shot PEFT without additional resources (external datasets, models).