Paper Review - Synthio: Augmenting Small-Scale Audio Classification Datasets with Synthetic Data (ICLR 2025)

Summary:

This paper primarily focuses on a novel pipeline to generate synthetic data for augmenting small scale audio classification datasets by leveraging Large Language Models (LLM) and Text-to-Audio Diffusion Model (T2A). The authors propose a very simple yet effective pipeline. By leveraging mechanisms such as Direct preference optimization (DPO), MixCap and Self-Reflection, they are able to generate high quality and diverse synthetic samples which are well aligned with the original small-scale dataset. The experimental setup is exhaustive spanning across 10 datasets and a lot of analysis confirming the methods capability.

Figure:

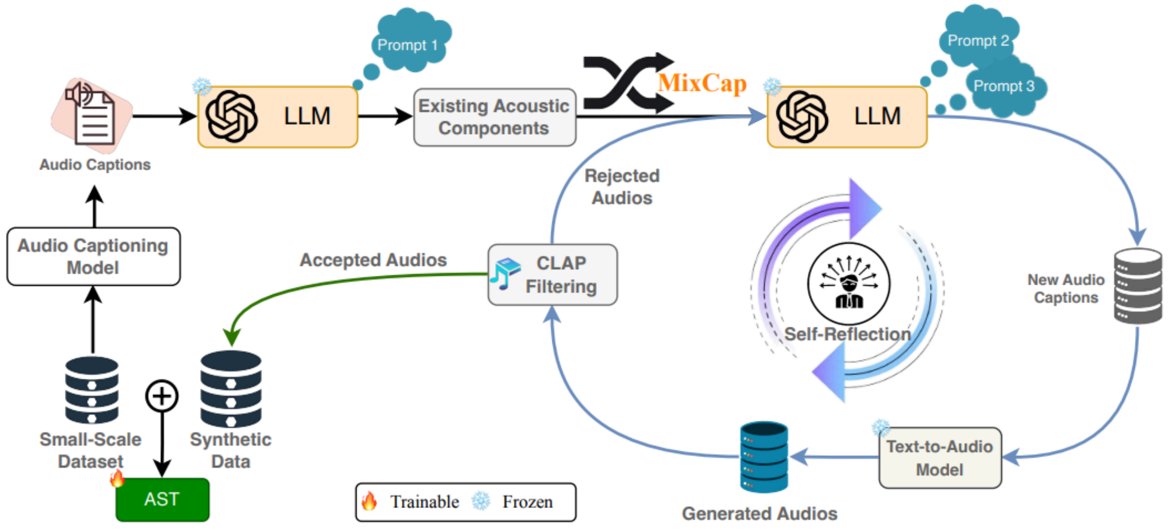

Figure 3: Overview of our proposed ***Language-Guided Audio Imagination*** for generating diverse synthetic augmentations. Starting with the small-scale dataset, we first generate audio captions and use an LLM to extract acoustic components (Prompt 1). Using these components and audio labels, we prompt the LLM to generate new and diverse captions (Prompt 2), which are then used to prompt the aligned T2A model for audio generation. The generated audios are filtered for label consistency using CLAP, with accepted audios added to the final synthetic dataset. Rejected audios undergo caption revision (Prompt 3) through a self-reflection process, and the revised captions are used to regenerate audios, iterating this process i times. Example captions are in Table 6.

Strengths:

1. The experiments are exhaustive with covering a wide range of datasets. I am happy to see at least 1 multi-label classification dataset.
2. The gains are significant especially when the dataset is very small (50,100). Also each component seems to have significant impact as indicated by the ablations in the results table.
3. The mathematical and concise explanation of concepts such as RLHF, DPO, DPO for diffusion are to the point.
4. As seen in Fig 1, increasing the number of augmentations is actually increasing the classification performance confirming the diversity of augmentation which doesn't seem to be the case for the competing method Vanilla Syn. Aug whose performance is dropping potentially indicating low diversity leading to overfitting.
5. Instead of just throwing away audios that do not pass the CLAP filtering, it regenerates the captions for the rejected audio to better align with the target label and the impact of this self-reflection model can be seen in the experiments.
6. In Table 3, it is interesting to see that only training on synthetic data leads to good performance, which indicates that the synthetic data is almost i.i.d with the original data.

Weaknesses:

1. It would be interesting if you had some experiments on noisy label learning to demonstrate the robustness of this method especially since your model is utilizing LLM and T2A diffusion models.
2. The only LLM that you use for this work is GPT-4-Turbo, making it difficult to gauge how much impact does the quality of the LLM have in the pipeline. Using at least one open source LLM would help make more sense of the numbers provided in the experimental section.
3. If you are using LLM and T2A which are already trained on large datasets, it could have been better to use contemporary audio self-supervised baselines which are self supervised on large datasets and fine-tuned on these small datasets to actually make the comparisons fair.

4. Zero shot performance of foundation models is not included, this makes one think do we need this method at all since there might be a zero shot method that already outperforms this for similar computation costs.

Minor Questions:

1. The details of the T2A adapter are not available. Is it a standard adapter with simple down and up projection? How many parameters?
2. Is it possible to use an ensemble of LLMs to generate captions?

Suggestions:

1. Trying out some recent variants of DPO to see which fine tuning methods works best for audio alignment would be very interesting and would certainly add an extra dimension which will only get better and better as the preference optimization space moves forward.
2. Since the task of supervised classification is considered somewhat easier, some experiments involving complex classification tasks (semi-supervised, domain shifts, open set etc.) would make the work more relevant.

Rating(out of 5):

4 (weak accept). The merit of this work lies in its simplicity and effectiveness of this pipeline as evident by the experiments. Some more experiments on tougher classification setups (eg. semi-supervised) could have made it a strong accept.