IMPROVING LANGUAGE-BASED AUDIO RETRIEVAL USING LLM AUGMENTATIONS

Bartłomiej Zgórzyński, Jan Kulik, Juliusz Kruk, Mateusz Matuszewski

Samsung R&D Institute Poland, Warsaw, Poland
{b.zgorzynski, j.kulik, j.kruk, m.matuszews2}@samsung.com

ABSTRACT

This work explores the integration of large language models (LLMs) in multimodal machine learning, focusing on their usefulness in augmenting and generating audio-caption datasets. The study is structured around three primary objectives. The first objective is to evaluate the capability of LLMs to enhance existing audio-caption datasets by generating augmented and improved captions. The second objective explores the potential of LLMs to create new audio-caption datasets by extracting relevant text and audio from video-caption datasets. Various LLMs and hyperparameter configurations are tested to determine their effectiveness in these two tasks. The final objective is to evaluate the impact of these augmented and newly created datasets on training outcomes, providing insights into their potential contributions to audio related machine learning tasks. The results demonstrate the potential of LLMs to significantly advance the field by improving data quality and availability, in result enhancing model training and performance.

Index Terms— Language-Based Audio Retrieval, DCASE 2024, Large Language Models, Caption augmentation, Bi-encoder architecture, Multimodal learning

1. INTRODUCTION

The rapid advancement of multimodal machine learning has led to an increasing demand for high-quality and diverse audio-caption datasets. However, collecting such datasets by gathering audio samples and captioning them manually can be a time-consuming and resource-intensive task, often resulting in limited availability and quality issues. Recent developments in large language models have shown promising capabilities in natural language processing tasks, sparking interest in their potential applications in multimodal learning.

The effectiveness of utilizing text augmentations has been demonstrated by Primus et al. [1] through the application of various tools. However, with the advent of GPT-3 and subsequent advancements in large language models, it has become evident that these models alone can perform a range of complex augmentations, notably improving results of language-audio retrieval. The objective of this work is to evaluate and benchmark the effectiveness of augmentations performed by LLMs. We mainly focus on exploring two potential applications of large language models in language-audio retrieval task. The first is to augment captions using back-translation and mixing described in subsection 2.2. Primus et al. [2] showed that paraphrasing captions to Clotho v2.1 dataset using GPT-3.5 Turbo can successfully enrich the training data. Xu [3] introduced AudioSetMix, which employed LLM-assisted transformations for audio captions, demonstrating how LLMs can dynamically augment audio-caption datasets to improve both the diversity and quality of the data. In alignment with this, Wu et al. [4, 5] have proven that mixing captions improves the performance of audio captioning. Audio retrieval and captioning are in principle closely related [6] and therefore we further explore the effect of mixing captions on the former. The second LLM implementation is to generate a new audio-caption dataset from existing video-caption datasets by extracting audio descriptions from captions using LLMs as presented in subsection 2.3). The WavCaps [7] dataset utilizes LLMs to refine raw audio descriptions into more structured, caption-like sentences, demonstrating an effective use of LLMs in creating cleaner, more useful datasets for multimodal learning.

We further explore this concept by experimenting with different LLMs, hyperparameters and systematically measuring the results. The model we are using for comparison to verify and measure the results of proposed methods is a custom bi-encoder architecture inspired by the work of Primus et al. [2], and is described in subsection 2.1. In accordance with the objectives of this work, experiments are conducted exclusively using LLMs. Simpler methods of data augmentation are not tested, as LLMs are expected to yield more sophisticated and effective results. The conducted experiments and the final results are analyzed in section 3.

The findings presented in this work demonstrate promising results, warranting further investigation to fully explore the vast potential of LLMs in multimodal learning.

2. METHOD

2.1. Model training

To evaluate selected augmentations and data generation methods, we use a two-phase approach with an audio retrieval model: pre-training and fine-tuning. For pre-training, we employ datasets Clotho v2.1 [8], AudioCaps [9], and WavCaps [7]. We then fine-tune the model using only the Clotho and AudioCaps datasets. The effectiveness of the model is assessed by comparing the mean average precision at rank 10 (mAP@10) across test splits of these two datasets. The model consists of a bi-encoder architecture designed to estimate similarity between audio and text data. Input audio and text are mapped to a 1,024-dimensional latent space, where pairs with similar meanings are positioned close to each other, while pairs with different meanings are positioned further apart. The similarity between embeddings is determined using cosine similarity. For textual embeddings we utilize the RoBERTa-large model [10], and to encode audio we use the PaSST-S [11] encoder. We train the entire model simultaneously without freezing any layers.

To train our systems, we employ the InfoNCE loss with a trainable temperature. After calculating embeddings of all n audios and texts from a given batch, we compute the similarity matrix S, where S_{ij} denotes the similarity between text i and audio j. The diagonal of the matrix represents matching pairs, while all other elements are considered non-matching. We then calculate the mean

Original caption	Back-translated caption
Brakes squeak, and a quiet engine idles nicely	The brakes squeal, and a quiet engine slows down smoothly
Loud deep tone cascading through a large room	Deep and loud tone resonating in a large room
A siren wails into the open air while waves lap the shore	A siren sounds in the air while the waves hit the shore

Table 1: Examples of back-translation

cross-entropy loss on each row (text-to-audio loss) and each column (audio-to-text loss) after applying the softmax function. The final loss is the mean of the audio-to-text and text-to-audio components.

We analyze 30-second audio segments based on Clotho's maximal audio length. Since the audio encoder processes 10-second segments, we split the input audio into 10-second windows with a 10-second hop size, without any additional overlap between windows. Subsequently, we average all embeddings from a given audio to obtain final representation.

2.2. Augmentations

In this section, we explore two augmentation methods: back-translation and mixing. The back-translation method subtly modifies captions by translating them into a random language and then back to English, leveraging linguistic nuances to introduce minor yet impactful changes. Meanwhile, mixing involves combining audio samples and captions to generate new, coherent captions that expand our dataset significantly.

2.2.1. Back-translation

Each caption is translated to a random language and then back to English using a large language model. The following prompt was used for this task:

> *You will be given audio captions. The captions are going to be used for training of an audio captioning model. Translate every caption to a random language and then translate it back to English. When translating, feel free to make proper adjustments to ensure the phrase is natural and coherent. Do not comment on translations.*

At first glance, this method appears similar to simple paraphrasing; however, it offers two significant advantages. First, it introduces subtle yet noticeable modifications to the original text, leveraging the inherent differences between languages. Second, by operating within the constraints of translation, the LLM preserves the core meaning of the caption. Consequently, this approach produces slightly modified captions, often with a different word order, while minimizing the risk of altering the fundamental message. We present some examples of this augmentation on captions from Clotho v2.1 dataset using GPT-4o in Table 1.

2.2.2. Mixing

Audio samples from the Clotho and AudioCaps datasets are mixed with each other and the LLM is prompted to combine the corresponding captions in a sensible manner. This process results in the creation of 50,000 new audio-caption pairs. The following prompt was used as input for this task:

> *You will be given a list of audio captions. Your task is to mix them together to generate a new caption. The caption that you generate should be a mix of all the input captions. Keep the generated caption under 15 words. Do not write introductions or explanations. The caption should be a natural and coherent sentence in the style of the input captions. The captions are not chronological, so don't refer to time dependencies between them.*

2.3. VideoCaps

2.3.1. Dataset generation

In order to create a new high-quality dataset, we collected commonly used video-caption datasets: Activity-Net [12], Charades-Ego [13], MSRVTT [14], MSVD [15], VATEX [16], VIOLIN [17] and WebVid [18]. This resulted in obtaining around 10.8 million samples. Then, we extracted samples that contained valid audio, which narrowed the dataset down to around 770,000 audio-caption pairs. The main challenge was that many of the captions were primarily video-focused and did not contain any meaningful information about the audio content.

Therefore, in order to filter out such cases, we employ the fine-tuned model described in subsection 2.1 to obtain audio and text embeddings for each sample. Then, cosine similarity between embeddings of each ground-truth pair is computed. This approach leverages the model's capability to represent complex semantic relationships and can be used to estimate the quality of ground-truth pairs in an arbitrary dataset. This method was applied to select top 100,000 samples for further processing.

To further process the selected samples, the LLM was used to rephrase original captions and remove any visual context that would be irrelevant during audio retrieval training. The following prompt was used as input to the LLM:

> *You will be given video captions. Rephrase them and remove parts that couldn't possibly be inferred from audio events. Remove any details from the captions that refer to visual or spoken events. Focus on the audio content only. Remove dates, time and names of places and persons. Do not write introductions or explanations. Each audio caption should be one sentence with less than 15 words. Use grammatical sentences.*

Finally, since rephrasing is prone to outliers and low-quality results as the LLM may deem the input captions as inadequate or simply fail to perform the task properly, we perform final filtration on the rephrased captions and extract the top 70,000.

2.3.2. Selecting LLM temperature

The aforementioned method of evaluating dataset quality can also serve as a benchmark to evaluate performance of various LLMs in processing captions and aid in selecting hyperparameters. The latter is especially important, since temperature can have significant impact on the quality of LLM output [19] and its optimal value can only be determined empirically. Therefore, we conducted a grid search and used various commercial and open-source LLMs with different temperature settings to rephrase 1,000 captions that were randomly sampled from the top 100,000 pairs obtained earlier. Then, cosine similarity between each rephrased caption and the corresponding audio clip was computed.

Experiment	LLM used	AudioCaps mAP@10	Clotho mAP@10
Pre-training	-	56.70	37.36
Pre-training + VideoCaps	GPT-4o	56.73	38.12
Pre-training + VideoCaps (without WavCaps)	GPT-4o	54.77	34.21
Base fine-tuning	-	59.43	38.68
Back-translation	Llama 3 8B	59.10	38.95
Back-translation	GPT-3.5 Turbo	59.76	39.14
Back-translation	GPT-4o	59.71	39.11
Mixing	Llama 3 8B	**60.61**	39.17
Mixing	GPT-3.5 Turbo	59.23	39.18
Mixing	GPT-4o	59.81	**39.24**
VideoCaps	Llama 3 8B	58.57	38.70
VideoCaps	GPT-3.5 Turbo	58.57	38.56
VideoCaps	GPT-4o	58.87	38.44
VideoCaps + Mixing	GPT-4o	59.08	38.82
VideoCaps + Back-translation	GPT-4o	59.38	39.02
Back-translation + Mixing	GPT-4o	59.44	38.94
VideoCaps + Mixing + Back-translation	GPT-4o	58.87	38.44

Table 2: Performance of text-to-audio retrieval on the AudioCaps and Clotho v2.1 test sets was evaluated. Each model was trained three times, with the values reported in the tables representing the average performance on each dataset.

3. EXPERIMENTS AND RESULTS

In this section, we describe all conducted experiments. Table 2 presents the performance of all models on the Clotho v2.1 and AudioCaps datasets, including both pre-training and fine-tuning results. For all experiments, we utilize the InfoNCE loss function. We update the model parameters using the AdamW optimizer with a batch size of 128. Additionally, we employ a cosine decay learning rate scheduler with warmup. During training, we select the best model checkpoints based on the mAP@10 value evaluated on the validation set, which is assessed twice within each training epoch.

3.1. Pre-training

Initially, we aimed to develop an audio-retrieval model for subsequent fine-tunings and data filtering. The training phase utilized Clotho-training, AudioCaps-training, and WavCaps datasets, with Clotho-validation and AudioCaps-validation employed for validation purposes. The training consists of 16 epochs, with a learning rate schedule from 1×10^{-5} to 5×10^{-7}. We utilize structured patchout of 15 and 2 for time and frequency dimensions, respectively. Additionally, random deletion and synonym replacement are applied with a probability of 0.8.

3.2. Fine-tuning

The next step involves further fine-tuning the model. For this purpose, the same datasets were used for both training and validation, excluding WavCaps. The number of epochs has been reduced to 6, and the learning rate has been decreased to range from 3×10^{-6} to 6×10^{-8}. To increase model regularization, we changed the optimizer weight decay parameter from 0.0 to 0.1. The results indicate that additional fine-tuning without WavCaps significantly increases the mAP@10 on both the Clotho and AudioCaps datasets.

3.3. Back-translation and mixing

A certain degree of randomness in generation is particularly desirable, especially during back-translation, to avoid literal translation. We decided to set the temperature parameter to 0.7 for each of the LLMs. Then, we have prepared augmented datasets: for each caption in training splits of AudioCaps and Clotho v2.1 we have generated exacly one back-translated caption. For mixing, we randomly selected 50,000 data pairs, equalized their audio energies, and used the LLMs to combine their captions.

To evaluate the effectiveness of augmented datasets in the development of audio retrieval systems, we conducted additional fine-tuning experiments. The effectiveness of these augmentations was assessed by comparing the mAP@10 value on the AudioCaps and Clotho v2.1 datasets, based on the training data used. The results, presented in Table 2, demonstrate that both back-translation and mixing significantly enhance the model's performance. For back-translation, utilizing more advanced language models leads to improved results, while for mixing, the best results were obtained with the smallest tested model.

Table 3 demonstrates that our best model, enhanced through data augmentation using Llama 3 8B, outperforms most of current state-of-the-art solutions. The single models by Primus et al. and Chen et al., submitted to the DCASE 2023 and 2024 Challenges, were trained on the full AudioCaps dataset and validated exclusively on the Clotho v2.1, which naturally resulted in improved performance on this dataset.

3.4. VideoCaps

3.4.1. Temperature selection

To select the optimal temperature settings for each large language model, we conducted experiments across a spectrum of temperature settings, ranging from 0.0 to 1.5. These settings adjust the randomness in the model's output: lower temperatures result in more deterministic outputs, while higher temperatures allow for greater

Model	AudioCaps mAP@10	Clotho mAP@10
CLAP [20]	51.3	20.4
Chen et al. [21]	-	37.00
Primus et al. [2]	-	38.56
Primus et al. [22]	-	**39.77**
Ours	**60.61**	39.17

Table 3: Comparison of our solution with other state-of-the-art text-to-audio retrieval systems.

diversity but potentially less coherence and relevance to the original audio content. We measured the median cosine similarity of 1,000 selected samples after rephrasing.

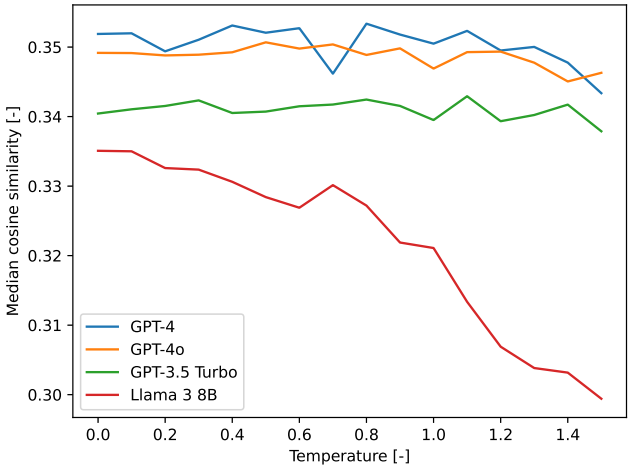

Figure 1: Median cosine similarity of rephrased datasets obtained using various LLMs across a spectrum of temperature settings

The results, shown in Figure 1, indicate that all models demonstrate higher median cosine similarity at lower temperatures. This finding suggests that more deterministic settings produce captions that are more closely aligned with the reference captions, highlighting the trade-off between creativity and accuracy in model-generated captions. Additionally, GPT-4 and GPT-4o consistently outperform GPT-3.5 Turbo and Llama 3 8B across most temperature settings, suggesting that newer and more sophisticated model architectures may better maintain semantic accuracy even as the output becomes more diverse at higher temperatures. For further experiments, we selected a temperature of 0.0 for Llama 3 8B, 1.1 for GPT-3.5 Turbo, and 0.7 for GPT-4o.

3.4.2. Trainings

After filtering, VideoCaps contains approximately 70,000 new audio-caption pairs. To evaluate the dataset's quality, we conducted additional pre-training to measure the influence of VideoCaps on model performance, keeping pre-training settings the same. The results of this experiment are shown in Table 2. Substituting WavCaps with VideoCaps alone resulted in a significant performance decrease on both datasets. However, integrating both WavCaps and VideoCaps notably enhanced performance on Clotho v2.1 while maintaining comparable results on AudioCaps.

We also conducted fine-tuning after base pre-training. The results showed slightly lower performance compared to fine-tuning without VideoCaps. This difference may be attributed to the larger volume of VideoCaps data compared to AudioCaps and Clotho v2.1. Additionally, variations in performance could stem from the fine-tuning process adapting to specific styles of descriptions and audio content present in the evaluation sets.

3.5. Joint Trainings

We also conducted experiments using different augmentations during fine-tuning, exclusively utilizing data generated by GPT-4o. The results are shown in Table 2, revealing that none of the combinations surpassed the performance achieved with Mixing and Llama 3 8B.

4. DISCUSSION AND CONCLUSION

In this study we explored various approaches to augmenting and generating new datasets for the training of text-to-audio retrieval systems using large language models. The results demonstrate that the utilization of presented techniques can significantly enhance retrieval model performance.

The novel approach introduced in this paper by creating VideoCaps dataset, shows promising results for generating large-scale text-audio datasets, thereby improving model pre-training. We also showed that large language model choice and the value of the temperature parameter can significantly impact the quality of the generated dataset. Additionally, our experiments indicate that mixing audio and captions, especially when augmented using Llama 3 8B, yields the best results for our system. This method produced a new state-of-the-art model in text-to-audio retrieval, achieving a mAP@10 score of 60.61 on the AudioCaps dataset. Additionally, it achieved a comparable performance to the current state-of-the-art models on the Clotho v2.1 dataset, with a score of 39.17.

However, fine-tunings with multiple generated data resulted in lower performance. One possible explanation for this is that there is a larger quantity of artificially generated data compared to the original data, which are likely of better quality.

5. FUTURE RESEARCH

In our research, we introduced three distinct ways to create and enhance datasets using large language models. There is still a room for experiments and improvements.

In addition to our methods, there are alternative approaches to generating synthetic audio-caption datasets. One such approach involves using the outputs of an audio captioning model as part of the prompt, along with other relevant metadata. Another method is to mix audio clips sequentially without significant overlap and prompt the LLM to generate corresponding captions, taking into account the temporal aspects of the audio.

Further research should investigate the outcomes of utilizing different LLMs and fine-tuning their hyperparameters, such as temperature, top-p, and top-k. In addition, experimenting with prompt engineering seems be an essential approach. It is reasonable to assume that as the quality of available large language models improves over time, the quality of synthetically generated datasets will also enhance.

6. REFERENCES

[1] P. Primus and G. Widmer, "Improving natural-language-based audio retrieval with transfer learning and audio & text augmentations," 2022. [Online]. Available: https://arxiv.org/abs/2208.11460

[2] P. Primus, K. Koutini, and G. Widmer, "Advancing natural-language based audio retrieval with passt and large audio-caption data sets," in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, Tampere, Finland, September 2023, pp. 151–155.

[3] D. Xu, "Audiosetmix: Enhancing audio-language datasets with llm-assisted augmentations," 2024. [Online]. Available: https://arxiv.org/abs/2405.11093

[4] S.-L. Wu, X. Chang, G. Wichern, J.-w. Jung, F. Germain, J. L. Roux, and S. Watanabe, "Beats-based audio captioning model with instructor embedding supervision and chatgpt mix-up," DCASE2023 Challenge, Tech. Rep., May 2023.

[5] S.-L. Wu, X. Chang, G. Wichern, J. weon Jung, F. Germain, J. L. Roux, and S. Watanabe, "Improving audio captioning models with fine-grained audio features, text embedding supervision, and llm mix-up augmentation," 2024. [Online]. Available: https://arxiv.org/abs/2309.17352

[6] E. Labbé, T. Pellegrini, and J. Pinquier, "Killing two birds with one stone: Can an audio captioning system also be used for audio-text retrieval?" in *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, Tampere, Finland, September 2023, pp. 86–90.

[7] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," 2023.

[8] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," 2019.

[9] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *NAACL-HLT*, 2019.

[10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[11] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Interspeech 2022*. ISCA, 2022. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2022-227

[12] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 961–970.

[13] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Charades-ego: A large-scale dataset of paired third and first person videos," 2018.

[14] G. Tan, D. Liu, M. Wang, and Z.-J. Zha, "Learning to discretely compose reasoning module networks for video captioning," 2020.

[15] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, Portland, OR, 6 2011.

[16] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," 2020.

[17] J. Liu, W. Chen, Y. Cheng, Z. Gan, L. Yu, Y. Yang, and J. Liu, "Violin: A large-scale dataset for video-and-language inference," 2020.

[18] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *IEEE International Conference on Computer Vision*, 2021.

[19] M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous, "Is temperature the creativity parameter of large language models?" 2024.

[20] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," 2024. [Online]. Available: https://arxiv.org/abs/2211.06687

[21] M. Chen, Y. Liu, B. Peng, and J. Chen, "Dcase 2024 challenge task 8 technical report," DCASE2024 Challenge, Tech. Rep., June 2024.

[22] P. Primus and G. Widmer, "A knowledge distillation approach to improving language-based audio retrieval models," DCASE2024 Challenge, Tech. Rep., June 2024.