Improving Natural-Language-Based Audio Retrieval with Transfer Learning and Audio & Text Augmentations

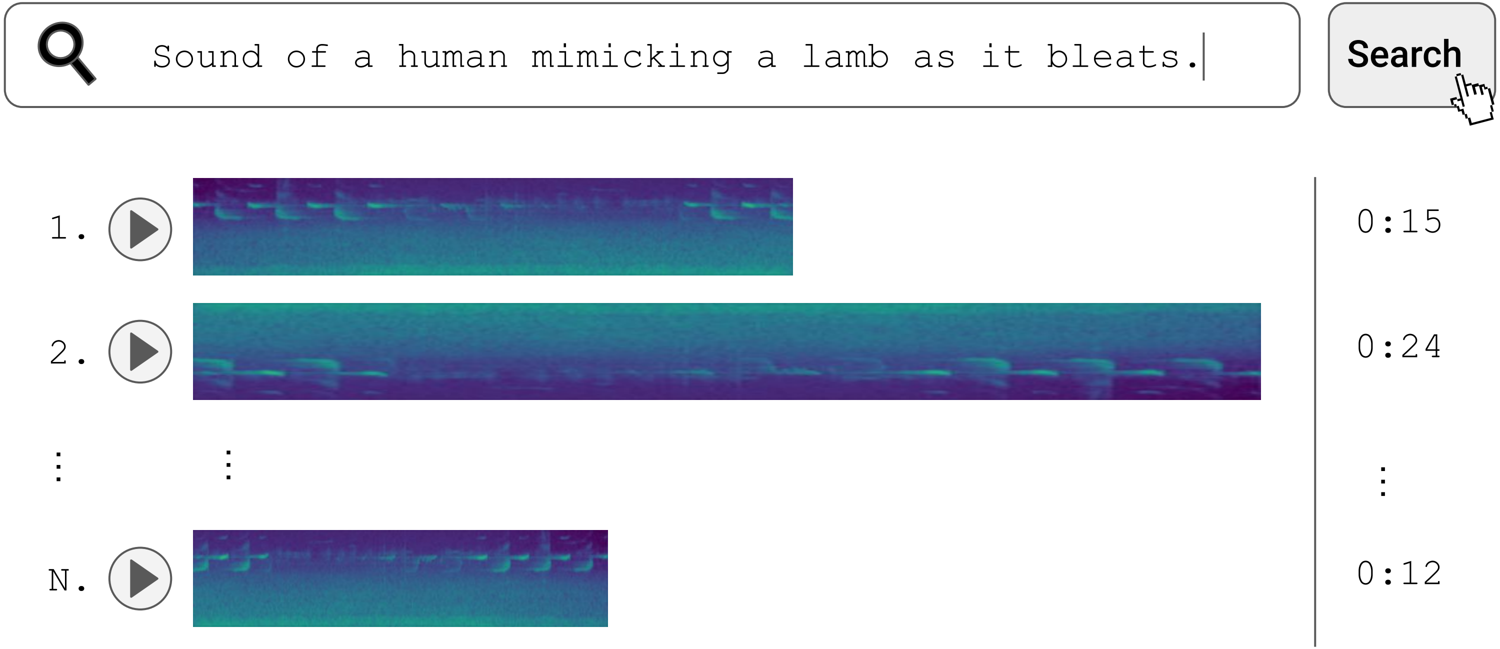
Paul Primus[1], and Gerhard Widmer[1,2]

[1]Institute of Computational Perception, [2]LIT Artificial Intelligence Lab

JOHANNES KEPLER UNIVERSITY LINZ

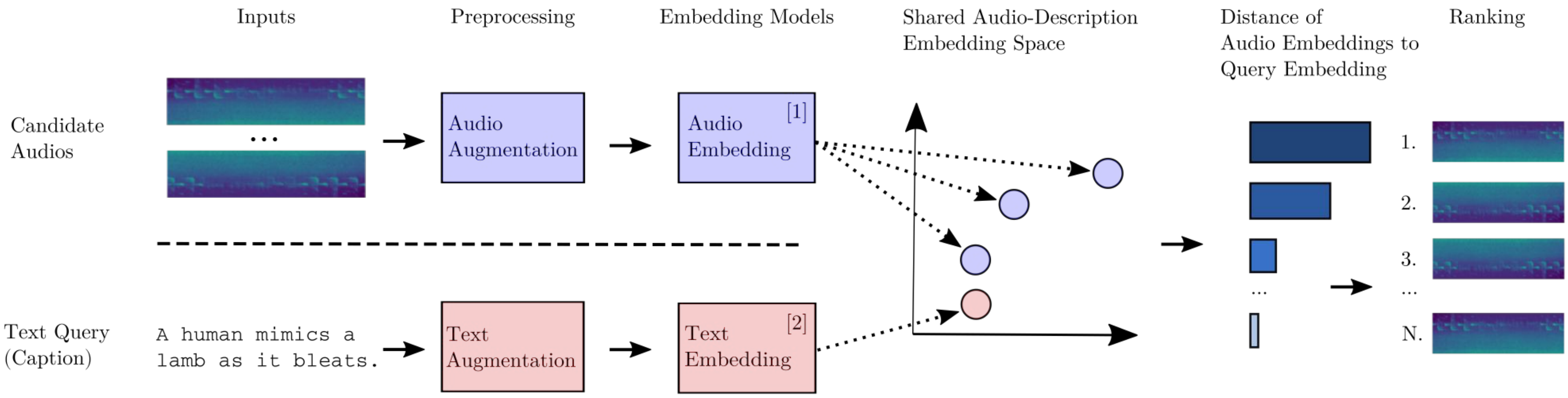
Institute of Computational Perception

Natural-Language-Based Audio Retrieval



JOHANNES KEPLER
UNIVERSITY LINZ

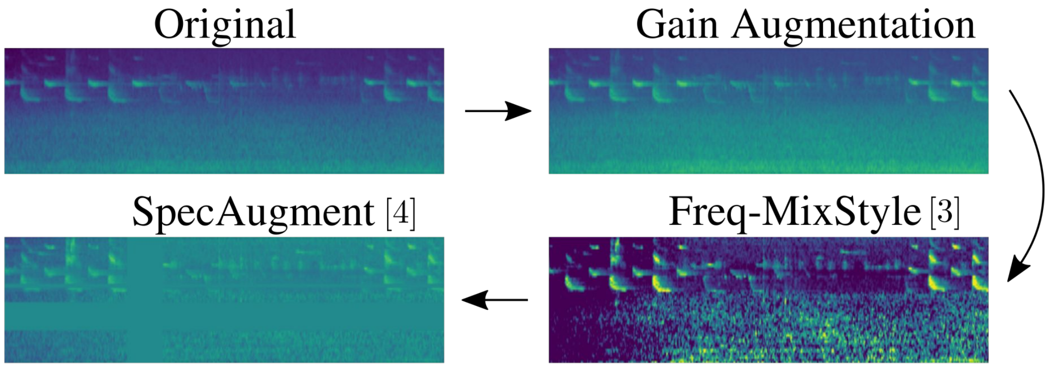
Our Retrieval Framework



[1] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE ACM Trans. Audio Speech Lang. Process., 2020.

[2] Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pretraining of deep bidirectional transformers for language understanding," in Proc. of the North American Ch. of the Ass. for Computational Linguistics: Human Language Technologies, NAACL-HLT, 2019.

Audio Augmentations



[3] B. Kim, S. Yang, J. Kim, H. Park, J. Lee, and S. Chang, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," in 23rd Annual Conference of the International Speech Communication Association, Interspeech, 2022.

[4] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in 20th Annual Conf. of the Int. Speech Communication Association, Interspeech, 2019.

Text Augmentations

Augmentation		Caption
Original		`The rain pours down.`
Back Translation	[5]	`It rains cats and dogs.`
Insert	[6]	`It tree rains cats and dogs.`
Delete	[6]	`It rains cats and ~~dogs~~.`
Swap	[6]	`It and cats rains dogs.`
Synonym	[6]	`It drizzles cats and dogs.`

[5] Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in Proc. of the 54th Annual Meeting of the Association for Computational Linguistics, ACL, 2016.

[6] W. Wei and K. Zou, "EDA: easy data augmentation techniques for boosting performance on text classification tasks," in Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing, EMNLP-IJCNLP, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., 2019

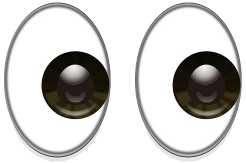
Results

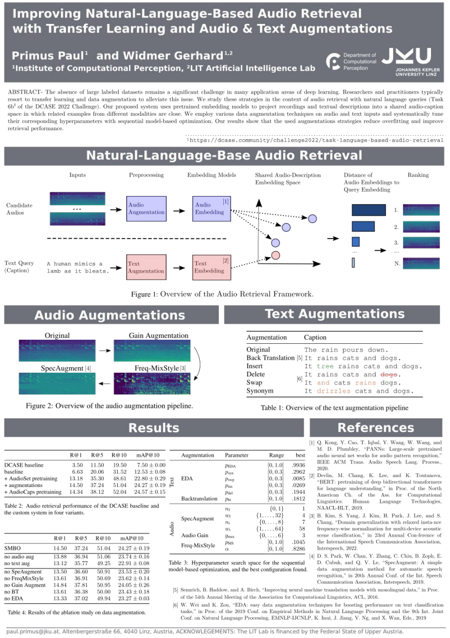
	R@1	R@5	R@10	mAP@10
DCASE baseline	3.50	11.50	19.50	7.50 ± 0.00
baseline	6.63	20.06	31.52	12.53 ± 0.08
+ AudioSet pretraining	13.18	35.30	48.61	22.80 ± 0.29
+ augmentations	14.50	37.24	51.04	24.27 ± 0.19
+ AudioCaps pretraining	14.34	38.12	52.04	24.57 ± 0.15

Ablation Study Results

	R@1	R@5	R@10	mAP@10
SMBO	14.50	37.24	51.04	24.27 ± 0.19
no audio aug	13.88	36.94	51.06	23.74 ± 0.16
no text aug	13.12	35.77	49.25	22.91 ± 0.08
no SpeAugment	13.50	36.60	50.91	23.53 ± 0.20
no FreqMixStyle	13.61	36.91	50.69	23.62 ± 0.14
no Gain Augment	14.84	37.81	50.95	24.05 ± 0.26
no BT	13.61	36.38	50.00	23.43 ± 0.18
no EDA	13.33	37.02	49.94	23.27 ± 0.03

Poster & Paper





Improving Natural-Language-Based Audio Retrieval with Transfer Learning and Audio & Text Augmentations

Primus Paul[1] and Widmer Gerhard[1,2]
[1]Institute of Computational Perception, [2]LIT Artificial Intelligence Lab
Department of Computational Perception
JKU JOHANNES KEPLER UNIVERSITY LINZ

ABSTRACT- The absence of large labeled datasets remains a significant challenge in many application areas of deep learning. Researchers and practitioners typically resort to transfer learning and data augmentation to alleviate this issue. We study these strategies in the context of audio retrieval with natural language queries (Task 6b[1] of the DCASE 2022 Challenge). Our proposed system uses pretrained embedding models to project recordings and textual descriptions into a shared audio-caption space in which related examples from different modalities are close. We employ various data augmentation techniques on audio and text inputs and systematically tune their corresponding hyperparameters with sequential model-based optimization. Our results show that the used augmentations strategies reduce overfitting and improve retrieval performance.

[1]https://dcase.community/challenge2022/task-language-based-audio-retrieval

Natural-Language-Base Audio Retrieval

Figure 1: Overview of the Audio Retrieval Framework.

Audio Augmentations

Figure 2: Overview of the audio augmentation pipeline.

Text Augmentations

Augmentation	Caption
Original	The rain pours down.
Back Translation [5]	It rains cats and dogs.
Insert	It tree rains cats and dogs.
Delete	It rains cats and ~~dogs~~.
Swap	It and cats rains dogs.
Synonym	It drizzles cats and dogs.

Table 1: Overview of the text augmentation pipeline

Results

	R@1	R@5	R@10	mAP@10
DCASE baseline	3.50	11.50	19.50	7.50 ± 0.00
baseline	6.63	20.06	31.52	12.53 ± 0.08
+ AudioSet pretraining	13.18	35.30	48.61	22.80 ± 0.29
+ augmentations	14.50	37.24	51.04	24.27 ± 0.19
+ AudioCaps pretraining	14.34	38.12	52.04	24.57 ± 0.15

Table 2: Audio retrieval performance of the DCASE baseline and the custom system in four variants.

	R@1	R@5	R@10	mAP@10
SMBO	14.50	37.24	51.04	24.27 ± 0.19
no audio aug	13.88	36.94	51.06	23.74 ± 0.16
no text aug	13.12	35.77	49.25	22.91 ± 0.08
no SpecAugment	13.50	36.60	50.91	23.53 ± 0.20
no FreqMixStyle	13.61	36.91	50.69	23.62 ± 0.14
no Gain Augment	14.84	37.81	50.95	24.05 ± 0.26
no BT	13.61	36.38	50.00	23.43 ± 0.18
no EDA	13.33	37.02	49.94	23.27 ± 0.03

Table 4: Results of the ablation study on data augmentation.

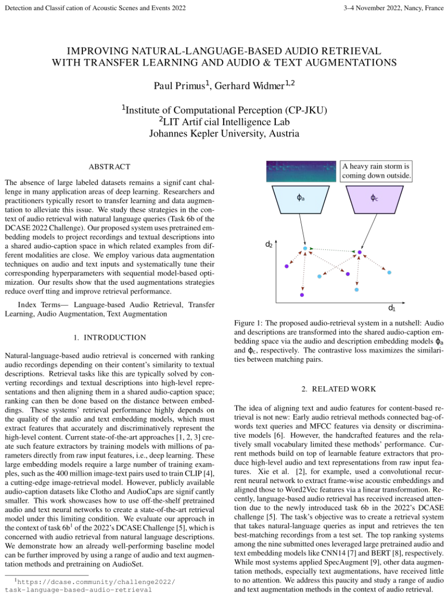
	Augmentation	Parameter	Range	best
Text	EDA	p_{EDA}	[0, 1.0]	.9936
		p_{syn}	[0, 0.3]	.2962
		p_{rep}	[0, 0.3]	.0085
		p_{ins}	[0, 0.3]	.0269
		p_{del}	[0, 0.3]	.1944
	Backtranslation	p_{bt}	[0, 1.0]	.1812
Audio	SpecAugment	m_f	{0, 1}	1
		w_f	{1, . . . , 32}	4
		m_t	{0, . . . , 8}	7
		w_t	{1, . . . , 64}	58
	Audio Gain	g_{max}	{0, . . . , 6}	3
	Freq-MixStyle	p_{MS}	[0, 1.0]	.1045
		α	[0, 1.0]	.8286

Table 3: Hyperparameter search space for the sequential model-based optimization, and the best configuration found.

References

[1] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural net works for audio pattern recognition," IEEE ACM Trans. Audio Speech Lang. Process., 2020.
[2] Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pretraining of deep bidirectional transformers for language understanding," in Proc. of the North American Ch. of the Ass. for Computational Linguistics: Human Language Technologies, NAACL-HLT, 2019.
[3] B. Kim, S. Yang, J. Kim, H. Park, J. Lee, and S. Chang, "Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification," in 23rd Annual Conference of the International Speech Communication Association, Interspeech, 2022.
[4] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in 20th Annual Conf. of the Int. Speech Communication Association, Interspeech, 2019.
[5] Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," in Proc. of the 54th Annual Meeting of the Association for Computational Linguistics, ACL, 2016.
[6] W. Wei and K. Zou, "EDA: easy data augmentation techniques for boosting performance on text classification tasks," in Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing, EMNLP-IJCNLP, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds., 2019

paul.primus@jku.at, Altenbergerstraße 66, 4040 Linz, Austria, ACKNOWLEDGEMENTS: The LIT Lab is financed by the Federal State of Upper Austria.



IMPROVING NATURAL-LANGUAGE-BASED AUDIO RETRIEVAL WITH TRANSFER LEARNING AND AUDIO & TEXT AUGMENTATIONS

Paul Primus[1], Gerhard Widmer[1,2]

[1]Institute of Computational Perception (CP-JKU)
[2]LIT Artificial Intelligence Lab
Johannes Kepler University, Austria

ABSTRACT

The absence of large labeled datasets remains a significant challenge in many application areas of deep learning. Researchers and practitioners typically resort to transfer learning and data augmentation to alleviate this issue. We study these strategies in the context of audio retrieval with natural language queries (Task 6b of the DCASE 2022 Challenge). Our proposed system uses pretrained embedding models to project recordings and textual descriptions into a shared audio-caption space in which related examples from different modalities are close. We employ various data augmentation techniques on audio and text inputs and systematically tune their corresponding hyperparameters with sequential model-based optimization. Our results show that the used augmentations strategies reduce overfitting and improve retrieval performance.

Index Terms— Language-based Audio Retrieval, Transfer Learning, Audio Augmentation, Text Augmentation

1. INTRODUCTION

Natural-language-based audio retrieval is concerned with ranking audio recordings depending on their content's similarity to textual descriptions. Retrieval tasks like this are typically solved by converting recordings and textual descriptions into high-level representations and then aligning them in a shared audio-caption space; ranking can then be done based on the distance between embeddings. These systems' retrieval performance highly depends on the quality of the audio and text embedding models, which must extract features that accurately and discriminatively represent the high-level content. Current state-of-the-art approaches [1, 2, 3] create such feature extractors by training models with millions of parameters directly from raw input features, i.e., deep learning. These large embedding models require a large number of training examples, such as the 400 million image-text pairs used to train CLIP [4], a cutting-edge image-retrieval model. However, publicly available audio-caption datasets like Clotho and AudioCaps are significantly smaller. This work showcases how to use off-the-shelf pretrained audio and text neural networks to create a state-of-the-art retrieval model under this limiting condition. We evaluate our approach in the context of task 6b[1] of the 2022's DCASE Challenge [5], which is concerned with audio retrieval from natural language descriptions. We demonstrate how an already well-performing baseline model can be further improved by using a range of audio and text augmentation methods and pretraining on AudioSet.

[1]https://dcase.community/challenge2022/task-language-based-audio-retrieval

Figure 1: The proposed audio-retrieval system in a nutshell: Audio and descriptions are transformed into the shared audio-caption embedding space via the audio and description embedding models ϕ_a and ϕ_c, respectively. The contrastive loss maximizes the similarities between matching pairs.

2. RELATED WORK

The idea of aligning text and audio features for content-based retrieval is not new: Early audio retrieval methods connected bag-of-words text queries and MFCC features via density or discriminative models [6]. However, the handcrafted features and the relatively small vocabulary limited these methods' performance. Current methods build on top of learnable feature extractors that produce high-level audio and text representations from raw input features. Xie et al. [2], for example, used a convolutional recurrent neural network to extract frame-wise acoustic embeddings and aligned those to Word2Vec features via a linear transformation. Recently, language-based audio retrieval has received increased attention due to the newly introduced task 6b in the 2022's DCASE challenge [5]. The task's objective was to create a retrieval system that takes natural-language queries and retrieves the ten best-matching recordings from a test set. The top ranking systems among the nine submitted ones leveraged large pretrained audio and text embedding models like CNN14 [7] and BERT [8], respectively. While most systems applied SpecAugment [9], other data augmentation methods, especially text augmentations, have received little to no attention. We address this paucity and study a range of audio and text augmentation methods in the context of audio retrieval.

Contact

email to paul.primus@jku.at