AudioCLIP: Extending CLIP to Image, Text and Audio[*]

Andrey Guzhov[1,2], Federico Raue[1], Jörn Hees[1], and Andreas Dengel[1,2]

[1] DFKI GmbH, Trippstadter Str. 122, 67663 Kaiserslautern, Germany
[2] TU Kaiserslautern, Kaiserslautern, Germany
firstname.lastname@dfki.de

Abstract In the past, the rapidly evolving field of sound classification greatly benefited from the application of methods from other domains. Today, we observe the trend to fuse domain-specific tasks and approaches together, which provides the community with new outstanding models. In this work, we present an extension of the CLIP model that handles audio in addition to text and images. Our proposed model incorporates the ESResNeXt audio-model into the CLIP framework using the AudioSet dataset. Such a combination enables the proposed model to perform bimodal and unimodal classification and querying, while keeping CLIP's ability to generalize to unseen datasets in a zero-shot inference fashion. AudioCLIP achieves new state-of-the-art results in the Environmental Sound Classification (ESC) task, out-performing other approaches by reaching accuracies of 90.07 % on the UrbanSound8K and 97.15 % on the ESC-50 datasets. Further it sets new baselines in the zero-shot ESC-task on the same datasets (68.78 % and 69.40 %, respectively).
Finally, we also assess the cross-modal querying performance of the proposed model as well as the influence of full and partial training on the results. For the sake of reproducibility, our code is published.

Keywords: Multimodal learning · Audio classification · Zero-shot inference.

1 Introduction

The latest advances of the sound classification community provided many powerful audio-domain models that demonstrated impressive results. Combination of widely known datasets – such as AudioSet [7], UrbanSound8K [25] and ESC-50 [19] – and domain-specific and inter-domain techniques conditioned the rapid development of audio-dedicated methods and approaches [15,10,30].

Previously, researchers were focusing mostly on the classification task using the audible modality exclusively. In the last years, however, popularity of multimodal approaches in application to audio-related tasks has been increasing [14,2,34]. Being applied to audio-specific tasks, this implied the use of either

---

textual or visual modalities in addition to sound. While the use of an additional modality together with audio is not rare, combination of more than two modalities is still uncommon in the audio domain. However, the restricted amount of qualitatively labeled data is constraining the development of the field in both, uni- and multimodal directions. Such a lack of data has challenged the research and sparked a cautious growth of interest for zero- and few-shot learning approaches based on contrastive learning methods that rely on textual descriptions [13,32,33].

In our work, we propose an approach to combine a high-performance audio model – ESResNeXt [10] – into a contrastive text-image model, namely CLIP [21], thus, obtaining a *tri-modal* hybrid architecture. The base CLIP model demonstrates impressive performance and strong domain adaptation capabilities that are referred as "zero-shot inference" in the original paper [21]. To keep consistency with the CLIP terminology, we use the term "zero-shot" in the sense defined in [21].

As we will see, the joint use of three modalities during the training results in out-performance of previous models in environmental sound classification task, extends zero-shot capabilities of the base architecture to the audio modality and introduces an ability to perform cross-modal querying using text, image and audio in any combination.

The remainder of this paper is organized as follows. In Section 2 we discuss the current approaches to handle audio in a standalone manner as well as jointly with additional modalities. Then, we describe models that serve as a base of our proposed hybrid architecture in Section 3, its training and evaluation in Section 4 and the obtained results in Section 5. Finally, we summarize our work and highlight follow-up research directions in Section 6.

2   Related Work

In this section, we provide an overview of the audio-related tasks and approaches that are intersecting in our work. Beginning with description of the environmental sound classification task, we connect it to the zero-shot classification through the description of existing methods to handle multiple modalities in a single model.

The environmental sound classification task implies an assignment of correct labels given samples belonging to sound classes that surround us in the everyday life (e.g., "alarm clock", "car horn", "jackhammer", "mouse clicking", "cat"). To successfully solve this task, different approaches were proposed that included the use of one- [27,28] or two-dimensional Convolutional Neural Networks (CNN) operating on static [18,24,32,9,15,17,33,8,30] or trainable [23,10] time-frequency transformation of raw audio. While the first approaches relied on the task-specific design of models, the latter results confirmed that the use of domain adaptation from visual domain is beneficial [9,17,10]. However, the visual modality was used in a sequential way, implying the processing of only one modality simultaneously.

The joint use of several modalities occurred first in video-related [14,34,6] tasks and was adapted to the sound classification task later [13,31]. However, despite the multimodal design, such approaches utilized two modalities simultaneously at most, while recent studies suggest that the use of more modalities is beneficial [2,1].

The multimodal approaches described above share a common key idea of contrastive learning. Such a technique belongs to the branch of self-supervised learning that, among other features, helps to overcome the lack of qualitatively labeled data. That makes it possible to apply contrastive learning-based training to the zero-shot classification tasks [13,32,33].

Summarizing, our proposed model employs contrastive learning to perform training on textual, visual and audible modalities, is able to perform modality-specific classification or, more general, querying and is implicitly enabled to generalize to previously unseen datasets in a zero-shot inference setup.

3   Model

In this section, we describe the key components that make up the proposed model and the way how it handles its input. On a high level, our hybrid architecture combines a ResNet-based CLIP model [21] for visual and textual modalities and an ESResNeXt model [10] for audible modality, as can be seen in Figure 1.
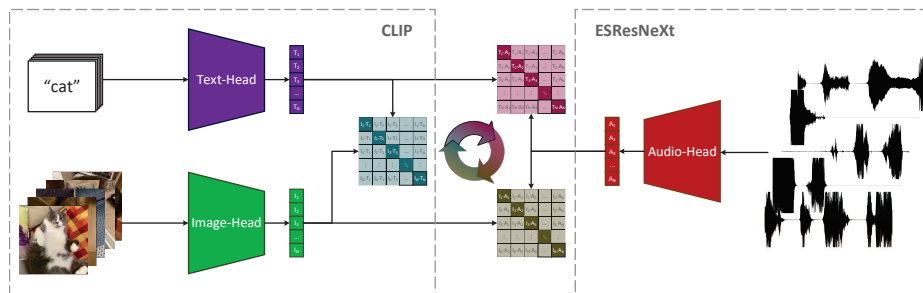

Figure 1. Overview of the proposed AudioCLIP model. On the left, the workflow of the text-image-model CLIP is shown. Performing joint training of the text- and image-heads, CLIP learns to align representations of the same concept in a shared multimodal embedding space. On the right, the audio-model ESResNeXT is shown. Here, the added audible modality interacts with two others, enabling the model to handle 3 modalities simultaneously.

3.1   CLIP

Conceptually, the CLIP model consists of two subnetworks: text and image encoding heads. Both parts of the CLIP model were pre-trained jointly under

natural language supervision [21]. Such a training setup enabled the model to generalize the classification ability to image samples that belonged to previously unseen datasets according to the provided labels without any additional fine-tuning.

For the text encoding part, a slightly modified [22] Transformer [29] architecture was used [21]. For the chosen 12-layer model, the input text was represented by a lower-case byte pair encoding with a vocabulary of size 49 408 [21]. Due to computational constraints, the maximum sequence length was clipped at 76 [21].

For the image encoding part of the CLIP model, two different architectures were considered. One was a Vision Transformer (ViT) [21,5], whose architecture made it similar to the text-head. Another option was represented by a modified ResNet-50 [11], whose global average pooling layer was replaced by a QKV-attention layer [21]. As we mentioned in Section 3.1, for the proposed hybrid model we chose the ResNet-based variant of the CLIP model because of its lower computational complexity, in comparison to the ViT-based one.

Given an input batch (text-image pairs) of size N, both CLIP-subnetworks produce the corresponding embeddings that are mapped linearly into a multi-modal embedding space of size 1 024 [21]. In a such setup, CLIP learns to maximize the cosine similarity between matching textual and visual representations, while minimizing it between incorrect ones, which is achieved using symmetric cross entropy loss over similarity measures [21].

3.2   ESResNeXt

For the audio encoding part, we decided to apply ESResNeXt model [10] that is based on the ResNeXt-50 [3] architecture and consists of a trainable time-frequency transformation based on complex frequency B-spline wavelets [26]. The chosen model contains moderate number of parameters to learn (~ 30 M), while performing competitive on a large-scale audio dataset, namely AudioSet [7], and providing state-of-the-art-level classification results on the UrbanSound8K [25] and ESC-50 [19] datasets. Additionally, the ESResNeXt model supports an implicit processing of a multi-channel audio input and provides improved robustness against additive white Gaussian noise and sample rate decrease [10].

3.3   Hybrid Model – AudioCLIP

In this work, we introduce an additional – audible – modality into the novel CLIP framework, which is naturally extending the existing model. We consider the newly added modality as equally important as the originally present. Such a modification became possible through the use of the AudioSet [7] dataset that we found suitable for this, as described in Section 4.1.

Thus, the proposed AudioCLIP model incorporates three subnetworks: text-, image- and audio-heads. In addition to the existing text-to-image-similarity loss term, there are two new ones introduced: text-to-audio and image-to-audio. The proposed model is able to process all three modalities simultaneously, as well as any pair of them.

4   Experimental Setup

In this section, we describe datasets that were used, data augmentation methods we applied, the training process and its corresponding hyper-parameters, finalizing with the performance evaluation methods.

4.1   Datasets

In this work, five image, audio and mixed datasets were used directly and indirectly. Here, we describe these datasets and define their roles in the training and evaluation processes.

Composite CLIP Dataset: In order to train CLIP, a new dataset was constructed by its authors. It consisted of roughly $400\,\mathrm{M}$ text-image pairs based on a set of $\sim 500\,\mathrm{k}$ text-based queries, and each query covered at least $\sim 20\,\mathrm{k}$ pairs [21]. In this work, the CLIP dataset was used indirectly as a weight initializer of the text- and image-heads (CLIP model).

ImageNet: ImageNet is a large-scale visual datasets described in [4] that contains more than $1\,\mathrm{M}$ images across $1\,000$ classes. For the purposes of this work, the ImageNet dataset served as a weight initializer of the ESResNeXt model and as a target for the zero-shot inference task.

AudioSet: Being proposed in [7], the AudioSet dataset provides a large-scale collection ($\sim 1.8\,\mathrm{M}$ & $\sim 20\,\mathrm{k}$ evaluation set) of audible data organized into 527 classes in a non-exclusive way. Each sample is a snippet up to $10\,\mathrm{s}$ long from a YouTube-video, defined by the corresponding ID and timings.

For this work, we acquired video frames in addition to audio tracks. Thus, the AudioSet dataset became the glue between the vanilla CLIP framework and our tri-modal extension on top of it. In particular, audio tracks and the respective class labels were used to perform image-to-audio transfer learning for the ESResNeXt model, and then, the extracted frames in addition to audio and class names served as an input for the hybrid AudioCLIP model.

During the training part, ten equally distant frames were extracted from a video recording, and one of them was picked randomly ($\sim \mathcal{U}$) and passed through the AudioCLIP model. In the evaluation phase, the same extraction procedure was performed, with the difference that only a central frame was presented to the model. Performance metrics are reported based on the evaluation set of the AudioSet dataset.

UrbanSound8K: The UrbanSound8K dataset provides $8\,732$ mono- and binaural audio tracks sampled at frequencies in the range $16 - 48\,\mathrm{kHz}$, each track is not longer than 4s. The audio recordings are organized into ten classes: "air conditioner", "car horn", "children playing", "dog bark", "drilling", "engine idling", "gun shot", "jackhammer", "siren", and "street music". To ensure correctness during the evaluation phase, the UrbanSound8K dataset was split by its authors into 10 non-overlapping folds [25] that we used in this work.

On this dataset, we performed zero-shot inference using the AudioCLIP model trained on AudioSet. Also, the audio encoding head was fine-tuned to

the UrbanSound8K dataset in both, standalone and cooperative fashion, and the classification performance in both setups was assessed.

ESC-50: The ESC-50 dataset provides 2 000 single-channel 5 s long audio tracks sampled at 44.1 kHz. As the name suggests, the dataset consists of 50 classes that can be divided into 5 major groups: *animal*, *natural and water*, *non-speech human*, *interior*, and *exterior* sounds. To ensure correctness during the evaluation phase, the ESC-50 dataset was split by its author into 5 non-overlapping folds [19] that we used in this work.

On this dataset, we performed zero-shot inference using the AudioCLIP model trained on AudioSet. Also, the audio encoding head was fine-tuned to the ESC-50 dataset in both, standalone and cooperative fashion, and the classification performance in both setups was assessed.

4.2   Data Augmentation

In comparison to the composite CLIP dataset (Section 4.1), the audio datasets provide two orders of magnitude less training samples, which makes overfitting an issue, especially for the UrbanSound8K and ESC-50 datasets. To address this challenge, several data augmentation techniques were applied that we describe in this section.

Time Scaling: Simultaneous change of track duration and its pitch is achieved using random scaling along the time axis. This kind of augmentation combines two computationally expensive ones, namely time stretching and pitch shift. Being a faster alternative to the combination of the aforementioned techniques, the time scaling in the range of random factors $[-1.5, 1.5]$, $\sim \mathcal{U}$ provides a lightweight though powerful method to fight overfitting [9].

Time Inversion: Inversion of a track along its time axis relates to the random flip of an image, which is an augmentation technique that is widely used in the visual domain. In this work, random time inversion with the probability of 0.5 was applied to the training samples similarly to [10].

Random Crop and Padding: Due to the requirement to align track duration before the processing through the model we applied random cropping or padding to the samples that were longer or shorter than the longest track in a non-augmented dataset, respectively. During the evaluation phase, the random operation was replaced by the center one.

Random Noise: The addition of random noise was shown to be helpful to overcome overfitting in visual-realted tasks [12]. Also, the robustness evaluation of the ESResNeXt model suggested the improved sustainability of the chosen audio encoding model against the additive white Gaussian noise (AWGN) [10]. In this work, we extended the set of data augmentation techniques using AWGN, whose sound-to-noise ratio varied randomly ($\sim \mathcal{U}$) from 10.0 dB to 120 dB. The probability of the presense of the noise was set to 0.25.

4.3   Training

The entire training process was divided into subsequent steps, which made acquisition of the final AudioCLIP model reliable and assured its high performance. As described in Section 3.1, we took a ResNet-based CLIP text-image-model pre-trained on its own dataset (Section 4.1) [21] and combined it with the ESResNeXt audio-model initialized using ImageNet weights and then pre-trained on the AudioSet dataset [10].

While the CLIP model was already pre-trained on text-image pairs, we decided to perform an extended AudioSet pre-training of the audio-head first, as it improved performance of the base ESResNeXt model (Table 1), and then to continue training in a tri-modal setting combining it with two other heads. Here, the whole AudioCLIP model was trained jointly on the AudioSet dataset using audio snippets, the corresponding video frames and the assigned textual labels.

Finally, audio-head of the trained AudioCLIP model was fine-tuned on the UrbanSound8K and ESC-50 datasets in a bimodal manner (audio and text) using sound recordings and the corresponding textual labels.

The trained AudioCLIP model and its audio encoding head were evaluated on the ImageNet dataset as well as on the three audio-datasets: AudioSet, UrbanSound8K, and ESC-50.

Audio-Head Pre-Training The initialization of the audio-head's parameters was split into two steps. First, the ImageNet-initialized ESResNeXt model was trained on the AudioSet dataset in a standalone fashion. Then, the pre-trained audio-head was incorporated into the AudioCLIP model and trained further under the cooperative supervision of the text- and image-heads.

Standalone: The first pre-training step implied the use of the AudioSet dataset as a weight initializer. Here, the ESResNeXt model was trained using the same setup as described in [10], with the difference in the number of training epochs. In this work, we increased the training time, which turned out into better evaluation performance on the AudioSet dataset and the subsequent downstream tasks, as described in Section 5.1 and independently quantified.

Cooperative: The further pre-training of the audio-head was done jointly with the text- and image-heads. Here, the pre-trained (in a standalone manner) audio-head was modified slightly through the replacement of its classification layer with a randomly initialized one, whose number of output neurons was the same as the size of CLIP's embedding space.

In this setup, the audio-head was trained as a part of the AudioCLIP model, which made its outputs compatible with the embeddings of the vanilla CLIP model. Parameters of the two other subnetworks, namely text- and image-head, were frozen during the cooperative pre-training of the audio encoding head, thus, these heads served as teachers in a multi-modal knowledge distillation setup.

The performance of the AudioCLIP model trained in such a fashion was assessed and is described in Section 5.

AudioCLIP Training The joint training of the audio-head made it compatible with the vanilla CLIP model, however, the distribution of images and textual descriptions in the AudioSet dataset does not follow the one from the CLIP dataset. This could lead to suboptimal performance of the resulting AudioCLIP model on the target dataset as well as on the downstream tasks.

To address this issue, we decided to perform the training of the whole tri-modal model on the AudioSet dataset. Here, all three modality-dedicated heads were tuned together, making the resulting model take into account the distributions of images and textual descriptions (video frames and names of the assigned AudioSet classes, respectively), in addition to the distribution of audio samples. The influence of the whole model training on the network's performance in comparison to the audio-head-only training is described in Section 5.

Audio-Head Fine-Tuning The trained AudioCLIP model provides general-purpose multimodal classification, or more general, querying abilities. However, under some conditions, it is required to acquire a more domain-specific model, which is able to distinguish concepts that differ just slightly.

To address this need, we performed experiments on tuning of the audio encoding head to two target datasets: UrbanSound8K and ESC-50.

Standalone: The ESResNeXt model that served as the audio-head demonstrated strong classification abilities on the chosen downstream tasks [10]. As we performed the AudioSet pre-training step instead of using a pre-trained ESResNeXt model, we fine-tuned it to the UrbanSound8K and ESC-50 datasets as well, in order to assess the change of the classification accuracy.

The fine-tuning step was done the same way as in [10], which implied the replacement of the classification layer with a randomly initialized one, whose number of outputs was defined by the number of targets in the downstream task. We report the performance of the fine-tuned ESResNeXt model in Section 5.1.

Cooperative: During the fine-tuning of the AudioCLIP model to the downstream tasks, only the parameters of the audio-head were being updated, so the text- and image-heads were frozen at this step. In comparison to the AudioSet training, which implied a multi-label setup, the corresponding textual class labels from the UrbanSound8K and ESC-50 datasets were represented by one class per audio sample.

For the fine-tuned AudioCLIP model, we assess the downstream classification performance as well as the querying performance, as described in Section 5.2.

4.4   Hyper-Parameters

In this work, we trained our model on the AudioSet, UrbanSound8K and ESC-50 datasets. The required hyper-parameters are reported in the current section.

In all training phases, the model parameters were optimized using Stochastic Gradient Descent [20] optimizer with Nesterov's momentum [16] of 0.9, weight decay of $5 \cdot 10^{-4}$ and batch size of 64.

Table 1. Evaluation results of the ESResNeXt model trained on the AudioSet dataset for more epochs. In comparison to the original training, performance improves.

Dataset	Score (%)	ESResNeXt Training	
		[10] (5 epochs)	Our (30 epochs)
AudioSet	mAP	28.17	34.14
UrbanSound8K	accuracy	89.14	89.49
ESC-50		95.20	95.90

The learning rate value decreased exponentially, varying its value η and the decrease factor γ from 10^{-4} and 0.95, respectively, during the standalone pre-training of the audio-head to $5 \cdot 10^{-5}$ and 0.98 during the fine-tuning of the AudioCLIP model to the downstream tasks.

The number of epochs was set to 30 for the AudioSet-based training, and to 50 for the fine-tuning to the downstream tasks.

4.5 Performance Evaluation

The model performance was assessed based on two tasks: classification and querying. While the evaluation of the first was possible for both, audio-head itself and the full AudioCLIP model, the performance on the latter task was assessed for the multimodal network only.

Classification The evaluation of the classification performance was done using the AudioCLIP model as well as its audio-head, namely ESResNeXt. The latter predicted the class labels directly, as the number of its outputs was equal to the number of targets in the datasets. For the AudioCLIP model, the classification task implied an intermediate step, which included construction of a target from textual labels [21].

In this work, the performance of the proposed model was evaluated after the training on the AudioSet dataset given audio and/or image as an input. For the UrbanSound8K and ESC-50 datasets, two downstream tasks were evaluated: classification after the training on the target dataset and without the training. The corresponding accuracies are reported in Section 5.1.

Querying The multimodal nature and symmetry of AudioCLIP allowed to perform querying of samples represented by another modality. Here, classification can be considered as a sub-task of querying, whose query consists of image and/or audio while the result is represented by text.

In this work, we assessed the querying performance of the trained AudioCLIP model on the ImageNet, AudioSet, UrbanSound8K and ESC-50 datasets. The results include Top-1 Precision/Recall (P@1/R@1) and Mean Average Precision (mAP), and presented in Section 5.2.

Table 2. Evaluation of AudioCLIP after partial (audio-head) and full training on AudioSet. The latter improves, in general, the results on the downstream tasks.

Dataset	Modality	Score (%)	Training On Target	Training	
				Audio-Head	Full Model
ImageNet	image	accuracy		40.51	21.79
AudioSet	image	mAP	✓	8.93	14.82
	audio		✓	25.85	28.36
	both		✓	25.11	32.38
UrbanSound8K				65.31	68.78
	audio	accuracy	✓	89.95	**90.07**
ESC-50				69.40	68.60
			✓	96.65	**97.15**

5   Results

5.1   Classification

Audio-Head Only The extended pre-training (30 epochs instead of 5) on the AudioSet dataset provided an audio encoding head that had increased performance, in comparison to the original training (from 28.17 % to 34.14 %, mAP). Such an improvement was also beneficial for the downstream tasks, making the newly trained audio-head to out-perform its base variant on the UrbanSound8K and ESC-50 datasets and achieving accuracy of 89.49 % and 95.90 %, respectively (Table 1).

AudioCLIP Our tri-modal training setup – through the use of video frames – introduced more diversity into audio-head's target distribution, thus fighting the overfitting issue and further increasing performance in audio classification tasks, in comparison to the audio-only ESResNeXt model. Also, the joint training of all three heads provided an additional performance boost and the ability to use multiple modalities to perform classification, as well as the zero-shot inference capabilities on previously unseen datasets (Table 2).

Partial Training: The training of the audio-head under the supervision of the text- and image-heads already allows to out-perform current state-of-the-art results on the UrbanSound8K and ESC-50 datasets by achieving accuracy of 89.95 % and 96.65 %, respectively.

Moreover, even the partial training of the AudioCLIP model sets new highest zero-shot classification accuracy on the ESC-50 dataset (69.40 %, Table 3) as well as out-performs performance of the commonly trained baseline CNN (64.50 %, Table 3).

Full Training: The joint training of the AudioCLIP model provides further performance improvements in comparison to the partial one. Such a trained AudioCLIP model sets the new state-of-the-art classification accuracy on the

Table 3. Evaluation results on UrbanSound8K (US8K) and ESC-50, accuracy (%).

	Model	Source	Training On Target	Target Dataset	
				US8K	ESC-50
Others	Human (2015)	[19]	–	–	81.30
	Piczak-CNN (2015)	[18]	✓	73.70	64.50
	SB-CNN (2017)	[24]	✓	79.00	–
	VGGish + Word2Vec (2019)	[32]		–	26.00
	ESResNet (2020)	[9]	✓	85.42	91.50
	WEANET N^4 (2020)	[15]	✓	–	94.10
	DenseNet-201 × 5, ensemble (2020)	[17]	✓	87.42	92.89
	VGGish + Word2Vec + GloVe (2021)	[33]		–	33.00
	ESResNeXt (2021)	[10]	✓	89.14	95.20
	AST (2021)	[8]	✓	–	95.60
	ERANN (2021)	[30]	✓	–	96.10
	Audio-Head (ESResNeXt, our training)		✓	89.49	95.90
Ours	AudioCLIP (partial training)			65.31	**69.40**
			✓	89.95	96.65
	AudioCLIP (full training)			**68.78**	68.60
			✓	**90.07**	**97.15**

UrbanSound8K and ESC-50 datasets (90.07 % and 97.15 %, respectively). Also, given the full model training setup, a new zero-shot classification baseline was set for the UrbanSound8K dataset (68.78 %, Table 3).

5.2 Querying

The original CLIP model introduced the ability to perform querying using both supported modalities – text and image – in any direction. Given a query (e.g., text), model provided similarity scores for the samples represented by another (visual) modality. Thus, given a dataset and a modality, the set of queries was defined by the unique samples of the chosen modality. In this work, we added the support of audio, enabling the model to query between text, images and audio in any combination. We evaluated the querying performance on the ImageNet, AudioSet, UrbanSound8K and ESC-50 datasets and summarized it in Table 4.

Image by Text: In this setup, all unique sets of class names assigned to the samples from a target dataset were collected and served as textual queries while the results were represented by images (ImageNet, AudioSet). Thus, only the visual samples possessing the same set of labels were considered as relevant results.

For the AudioSet dataset, the full training contributed to the increase of the performance score measured by mAP. However, such a training led to the decrease of the querying performance on the ImageNet dataset, as its distribution is likely different from the AudioSet one.

Table 4. Querying scores of AudioCLIP after partial and full training on AudioSet. The latter in general improves results on AudioSet and the downstream tasks.

Dataset	Modality		Audio-Head			Full Model		
	Query	Result	P@1	R@1	mAP	P@1	R@1	mAP
ImageNet	text	image	5.42	84.15	**52.91**	1.61	89.00	33.13
AudioSet	text	image	0.81	46.79	9.51	1.31	74.95	**17.22**
	text	audio	2.51	84.38	23.54	5.33	76.13	**30.79**
	audio	image	0.62	56.39	5.45	1.03	52.12	**7.22**
	image	audio	0.61	61.94	4.86	1.20	53.15	**6.86**
UrbanSound8K	text	audio	40.81	47.69	77.43	42.28	48.18	**80.04**
ESC-50			51.25	85.20	**77.20**	51.78	80.40	76.97

Audio by Text: Having the same type of query as in the previous setup, here, the result was represented by an audio recording and considered correct if the labels matched the query.

On the AudioSet and UrbanSound8K datasets, the full training increases the querying performance. For the ESC-50 dataset it is not the case, however, the gap is not large and is close to be marginal.

Audio by Image and Vice Versa: For both types of queries – audio by image and image by audio – the full training of the AudioCLIP model was beneficial in terms of querying performance (mAP).

6   Conclusion

In this work, we extended the CLIP model [21] from textual and visual modalities to audio using an effective sound classification model [10].

The proposed AudioCLIP model achieves new state-of-the-art classification results on two datasets: UrbanSound8K (90.07 %) and ESC-50 (97.15 %). To ease reproducibility, the details on hyper-parameters and implementation as well as weights of the trained models are made available for the community[1].

Additionally, for the zero-shot inference, our model out-performs previous approaches on the ESC-50 dataset with a large gap (69.40 %) and sets a baseline for the UrbanSound8K dataset (68.78 %).

We also evaluated the performance of our model on cross-modal querying tasks as well as the influence of the partial and full training on the results in classification and querying tasks.

In the future, we would like to further investigate the performance of the proposed AudioCLIP model on a wider variety of datasets and tasks. Also, changing the backbones of image- and audio-heads to more powerful networks could further improve the model performance.

---

[1] https://github.com/AndreyGuzhov/AudioCLIP

References

1. Akbari, H., Yuan, L., Qian, R., Chuang, W.H., Chang, S.F., Cui, Y., Gong, B.: Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. arXiv preprint arXiv:2104.11178 (2021)
2. Alayrac, J.B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., Zisserman, A.: Self-supervised multimodal versatile networks. arXiv preprint arXiv:2006.16228 (2020)
3. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Dzabraev, M., Kalashnikov, M., Komkov, S., Petiushko, A.: Mdmmt: Multidomain multimodal transformer for video retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3354–3363 (2021)
7. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017)
8. Gong, Y., Chung, Y.A., Glass, J.: Ast: Audio spectrogram transformer (2021)
9. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Esresnet: Environmental sound classification based on visual domain models. In: 25th International Conference on Pattern Recognition (ICPR). pp. 4933–4940 (January 2021)
10. Guzhov, A., Raue, F., Hees, J., Dengel, A.: Esresne(x)t-fbsp: Learning robust time-frequency transformation of audio. In: 2021 International Joint Conference on Neural Networks (IJCNN) (2021)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hussain, Z., Gimenez, F., Yi, D., Rubin, D.: Differential data augmentation techniques for medical imaging classification tasks. In: AMIA Annual Symposium Proceedings. vol. 2017, p. 979. American Medical Informatics Association (2017)
13. Islam, M.T., Nirjon, S.: Soundsemantics: exploiting semantic knowledge in text for embedded acoustic event classification. In: Proceedings of the 18th International Conference on Information Processing in Sensor Networks. pp. 217–228 (2019)
14. Kim, C.D., Kim, B., Lee, H., Kim, G.: Audiocaps: Generating captions for audios in the wild. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 119–132 (2019)
15. Kumar, A., Ithapu, V.: A sequential self teaching approach for improving generalization in sound event recognition. In: International Conference on Machine Learning. pp. 5447–5457. PMLR (2020)
16. Nesterov, Y.: A method of solving a convex programming problem with convergence rate o $(1/k^2)$ o $(1/k2)$. In: Sov. Math. Dokl. vol. 27 (1983)

17. Palanisamy, K., Singhania, D., Yao, A.: Rethinking cnn models for audio classification (2020)
18. Piczak, K.J.: Environmental sound classification with convolutional neural networks. In: 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2015)
19. Piczak, K.J.: Esc: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 1015–1018 (2015)
20. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM journal on control and optimization **30**(4), 838–855 (1992)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
22. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
23. Sailor, H.B., Agrawal, D.M., Patil, H.A.: Unsupervised filterbank learning using convolutional restricted boltzmann machine for environmental sound classification. In: INTERSPEECH. pp. 3107–3111 (2017)
24. Salamon, J., Bello, J.P.: Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters **24**(3), 279–283 (2017)
25. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 1041–1044 (2014)
26. Teolis, A., Benedetto, J.J.: Computational signal processing with wavelets, vol. 182. Springer (1998)
27. Tokozume, Y., Harada, T.: Learning environmental sounds with end-to-end convolutional neural network. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2721–2725 (March 2017). https://doi.org/10.1109/ICASSP.2017.7952651
28. Tokozume, Y., Ushiku, Y., Harada, T.: Learning from between-class examples for deep sound recognition (2017), https://arxiv.org/abs/1711.10282
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
30. Verbitskiy, S., Vyshegorodtsev, V.: Eranns: Efficient residual audio neural networks for audio pattern recognition (2021)
31. Wang, L., Luc, P., Recasens, A., Alayrac, J.B., Oord, A.v.d.: Multimodal self-supervised learning of general audio representations. arXiv preprint arXiv:2104.12807 (2021)
32. Xie, H., Virtanen, T.: Zero-shot audio classification based on class label embeddings. In: 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). pp. 264–267. IEEE (2019)
33. Xie, H., Virtanen, T.: Zero-shot audio classification via semantic embeddings. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 1233–1242 (2021)
34. Zhang, J., Xu, X., Shen, F., Lu, H., Liu, X., Shen, H.T.: Enhancing audio-visual association with self-supervised curriculum learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3351–3359 (2021)