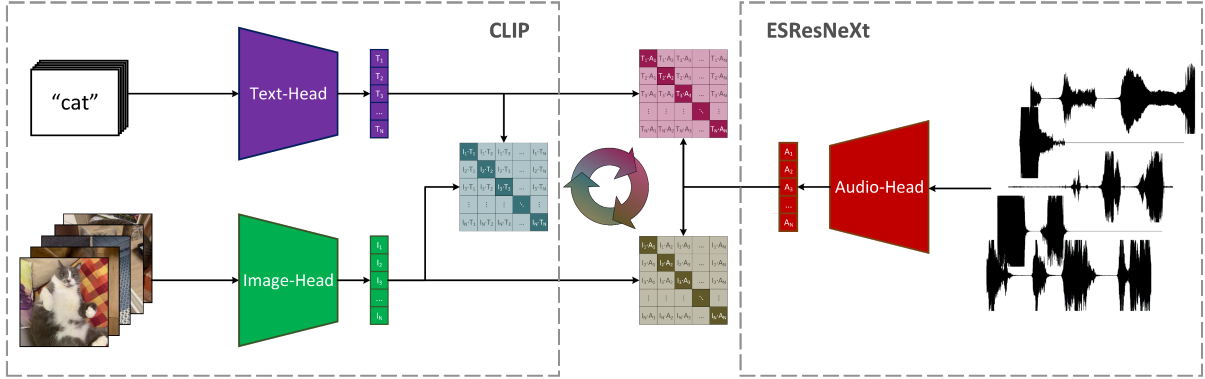

AudioCLIP

Extending CLIP to Image, Text and Audio



This repository contains implementation of the models described in the paper arXiv:2106.13043. This work is based on our previous works: * ESResNe(X)t-fbsp: Learning Robust Time-Frequency Transformation of Audio (2021). * ESResNet: Environmental Sound Classification Based on Visual Domain Models (2020).

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.

Downloading Pre-Trained Weights

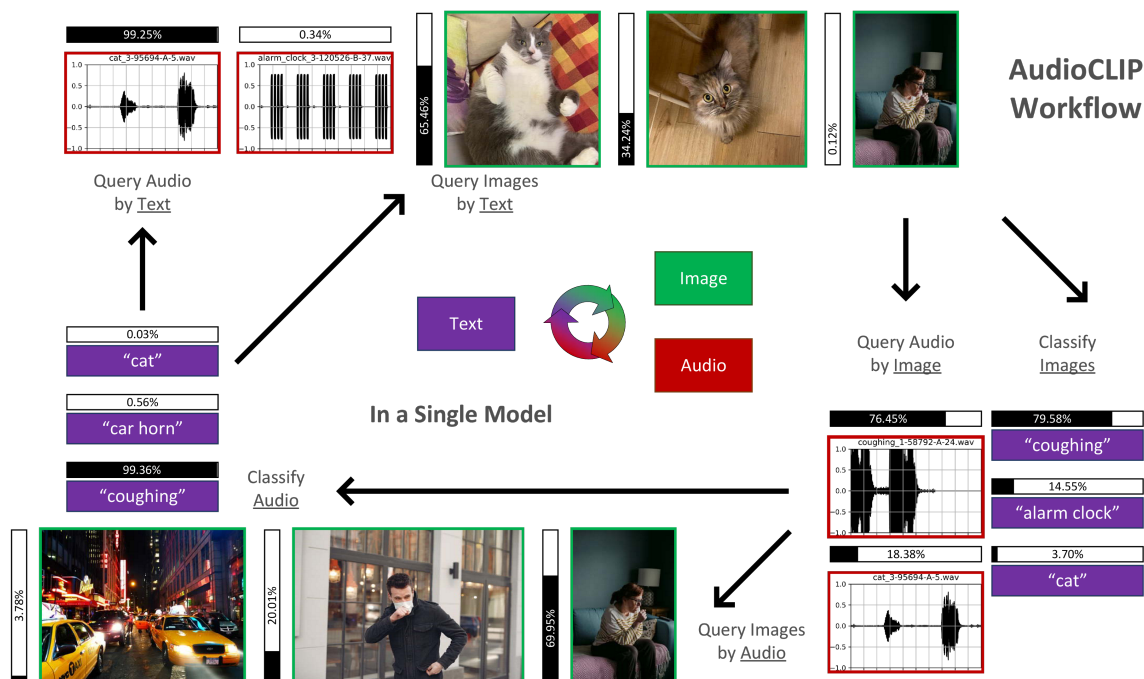
The pre-trained model can be downloaded from the releases.

```
1 # AudioCLIP trained on AudioSet (text-, image- and audio-head simultaneously)
2 wget https://github.com/AndreyGuzhov/AudioCLIP/releases/download/v0.1/AudioCLIP-Full-Training.pt
```

Important Note If you use AudioCLIP as a part of GAN-based image generation, please consider downloading the partially trained model, as its audio embeddings are compatible with the vanilla CLIP (based on ResNet-50).

Demo on Use Cases

Jupyter Notebook with sample use cases is available under the link.



How to Run the Model

The required Python version is ≥ 3.7 .

AudioCLIP

On the ESC-50 dataset

```
1 python main.py --config protocols/audioclip-esc50.json --Dataset.args.  
   root /path/to/ESC50
```

On the UrbanSound8K dataset

```
1 python main.py --config protocols/audioclip-us8k.json --Dataset.args.  
   root /path/to/UrbanSound8K
```

More About AudioCLIP

The AI Epiphany channel made a great video about AudioCLIP. [Learn more here.](#)

Cite Us

```
1 @misc{guzhov2021audioclip,  
2     title={AudioCLIP: Extending CLIP to Image, Text and Audio},  
3     author={Andrey Guzhov and Federico Raue and Jörn Hees and Andreas  
4         Dengel},  
5     year={2021},  
6     eprint={2106.13043},  
7     archivePrefix={arXiv},  
8     primaryClass={cs.SD}  
9 }
```