Mainstream machine listening models are trained to learn audio concepts under the paradigm of one class label to many recordings focusing on one task. Learning under such restricted supervision limits the flexibility of models because they require labeled audio for training and can only predict the predefined categories. Instead, we propose to learn audio concepts from natural language supervision. We call our approach Contrastive Language-Audio Pretraining (CLAP), which connects language and audio by using two encoders and a contrastive learning objective, bringing audio and text descriptions into a joint multimodal space. We trained CLAP with 128k audio and text pairs and evaluated it on 16 downstream tasks across 7 domains, such as classification of sound events, scenes, music, and speech. CLAP establishes state-of-the-art (SoTA) in Zero-Shot performance. Also, we evaluated CLAP’s audio encoder in a supervised learning setup and achieved SoTA in 5 tasks. The Zero-Shot capability removes the need of training with class labeled audio, enables flexible class prediction at inference time, and generalizes well in multiple downstream tasks. Code is available at: https://github.com/microsoft/CLAP.

Index Terms— contrastive learning, general purpose audio representation, zero-shot, sound event classification

1. INTRODUCTION

The human auditory system can hear sounds and extract the kind of decisions or meanings we need to interact with our surroundings [1]. For example, if we are in a soccer game and suddenly hear the crowd cheering joyfully, we can assume the local team scored! Machine Listening aims to understand audio cues by automatically processing sounds in audio signals and extract meaning [1]. Mainstream Machine Listening models break human hearing into tasks, such as the classification of sound events and acoustic scenes. Such models are trained by associating audio recordings to class labels of predefined categories for a specific task and can only predict specific categories [2]. Learning under such restricted supervision limits the flexibility to predict unseen classes.

Self-Supervised Learning (SSL) pretrains models with unlabeled audio, avoiding the limited supervision of learning from class labels. However, SSL excludes semantic knowledge from natural language. The pretrained model is then adapted to a downstream task in a supervised setup learning under the class label paradigm [3, 4, 5]. Mainstream and SSL models have static output layers that can only predict the predefined categories. On the other hand, models that enable Zero-Shot predictions can take an input audio and yield a prediction score for any class typed by the user. Zero-shot requires no training stage so there are no predefined categories. To enable such flexibility and generalization, models need to learn the relationships between the acoustic semantics and language semantics.

A middle path between both approaches is to learn audio concepts from natural language supervision, which is underexplored. Computer Vision has successfully developed models that learn image representations with natural language supervision, achieving high performance across different downstream tasks in Zero-Shot predictions and adapted (e.g. fine-tuned) to target datasets in a supervised setup. Examples are Open AI’s CLIP [6] and Florence [7]. In the audio domain, Wav2clip [8] and Audioclip [9] distill from CLIP and are trained with audio and class labels from AudioSet instead of audio and natural language. Although the performance was promising, it is yet to be explored how natural language can benefit flexibility and generalization to new classes and tasks.

We call our approach Contrastive Language-Audio Pretraining (CLAP), which connects natural language and audio into a joint multimodal space by using two encoders and contrastive learning. Our main contributions are first, to introduce our CLAP model trained with 128k audio and text pairs. Second, we show that our model enables Zero-Shot learning, thus removing the need of training with audio from predefined class labels. At inference time, enables prediction of classes from different downstream tasks. Third, CLAP generalizes well to 16 downstream tasks across 7 domains and establishes Zero-Shot SoTA performance. When CLAP’s audio encoder is finetuned it achieves SoTA in 5 tasks. We include insights about the model in Section 4.

2. METHOD

CLAP is illustrated in Fig 1. The input is audio and text pairs passed to an audio encoder and a text encoder. Both representations are connected in joint multimodal space with linear projections. The space is learned with the (dis)similarity of audio and text pairs in a batch using contrastive learning. The pretrained encoders with their projection layers can be used to compute audio and text embeddings and enable Zero-Shot Classification. Our method is inspired by the CLIP model [6].
2.1. Contrastive Language-Audio Pretraining

Let the processed audio be $X_a$ s.t. $X_a \in \mathbb{R}^{F \times T}$ where $F$ are the number of spectral components (e.g. Mel bins) and $T$ are the number of time bins. Let the text be represented by $X_t$. Each audio-text pair in a batch of $N$ is represented as $\{X_a, X_t\}_i$ where $i \in [0, N]$. For convenience, we drop the $i$ notation, and henceforth $\{X_a, X_t\}$ will denote a batch of $N$.

From the pairs, the audio and text are passed through an audio encoder and a text encoder respectively. Let $f_a(.)$ represent the audio encoder and $f_t(.)$ represent the text encoder. For a batch of $N$:

$$\hat{X}_a = f_a(X_a); \hat{X}_t = f_t(X_t)$$  \hspace{1cm} (1)

where $\hat{X}_a \in \mathbb{R}^{N \times V}$ are the audio representations of dimensionality $V$, and $\hat{X}_t \in \mathbb{R}^{N \times U}$ are the text representations of dimensionality $U$.

We brought audio and text representations, $\hat{X}_a$ and $\hat{X}_t$, into a joint multimodal space of dimension $d$ by using a learnable linear projection:

$$E_a = L_a(X_a); E_t = L_t(X_t)$$  \hspace{1cm} (2)

where $E_a \in \mathbb{R}^{N \times d}$, $E_t \in \mathbb{R}^{N \times d}$, $L_a$ and $L_t$ are the linear projections for audio and text respectively.

Now that the audio and text embeddings ($E_a$, $E_t$) are comparable, we can measure similarity:

$$C = \tau(E_t \cdot E_a^T)$$  \hspace{1cm} (3)

where $\tau$ is a temperature parameter to scale the range of logits. The similarity matrix $C \in \mathbb{R}^{N \times N}$ has $N$ correct pairs in the diagonal and $N^2 - N$ incorrect pairs in the off-diagonal.

$$L = 0.5(\ell_{text}(C) + \ell_{audio}(C))$$  \hspace{1cm} (4)

where $\ell_k = \frac{1}{N} \sum_{i=0}^{N} \log \text{diag}(\text{softmax}(C))$ along text and audio axis respectively. We used this symmetric cross-entropy loss ($L$) over the similarity matrix to jointly train the audio and text encoders along with their linear projections.

2.2. Zero-Shot Classification

For Zero-Shot classification, we used CLAP’s ability to determine the similarity between audio and text. Let’s consider a target dataset with $C$ class labels and $N$ test audios. First, we compute audio embeddings and text embeddings for $N$ audios and $C$ classes using the pretrained encoders and their projection layers. Second, because both the embeddings are in a common space, we compute the cosine similarity between each testing audio and all the class labels. Each audio will have as many logits as class labels. Third, logits are turned into a probability distribution by applying softmax for binary or multiclass and sigmoid for multilabel classification.

3. EXPERIMENTS

Training Datasets. We used 128,010 audio and text pairs from 4 datasets to construct the training dataset for CLAP. We extracted 36,796 pairs from FSD50k [17] (training and validation splits; title and description were concatenated); 29,646 pairs from ClothoV2 [18] (each audio has 5 captions so we created 5 pairs per clip); 44,292 from AudioCaps [19]; 17,276 pairs from MACS [20] (same process to create pairs as Clotho). An example of a Clotho caption: A camp fire crackles as the flames burn branches and leaves.

Downstream Tasks. We used 16 datasets from 7 different domains. Five tasks are Sound Event Classification: ESC50 [10] has 50 classes, FSD50K [10] has 200, US8K [21] has 10, DCASE17 Task4 [2] has 17, and AudioSet has 527. Five are music-related classification from HEAR [10]: GTZAN Music
Table 1. CLAP (ZS) Zero-Shot outperforms the literature. CLAP (Best) is the best performance among our supervised setups. CLAP (linear probe) uses CLAP as a feature extractor and a fully-connected layer. Higher is better for all numbers, DCASE17 employs F1 score, FSD50K and AudioSet employs mAP, everything else uses Accuracy.

| Model             | Sound Event Classification↑ | Music↑ |
|-------------------|----------------------------|--------|
|                   | Music | Acoustic Scene Classification↑ | Emotion Recognition↑ | Keyword Spotting↑ | Vocal Sound Classification↑ | Speaker Counting↑ |
| Benchmark (ZS)    |       | Beijing Opera                  | TUT2017 | CRE | MA-D | RAV | DESS | Speech Comm. | Vocal Sound | Libri Count |
| benchmark (ZS)    | 0.25  | 0.06                           | 0.1667  | 0.125 |       | 0.083 | 0.1667 | 0.090       |             |             |
| CLAP (ZS)         | 0.4746 | 0.2963                         | 0.1784  | 0.1599 |       | 0.1063 | 0.4945 | 0.1788      |             |             |
| Benchmark (Best)  | 0.9755 | 0.843 [13]                     | 0.752 [10] | 0.8182 [14] |       | 0.987 [15] | 0.905 [16] | 0.785 [10] |             |             |
| CLAP (linear probe) | 0.6399 | 0.5693                         | 0.2315  | 0.4044 |       | 0.3063 | 0.7732 | 0.5451      |             |             |
| CLAP (Best)       | 0.9026 | 0.7463                         | 0.6834  | 0.6436 |       | 0.9683 | 0.9130 | 0.7559      |             |             |

3.1. Experimental setup

**Pre-processing.** We used log Mel spectrogram representations of audio with a sampling rate of 44.1 KHz, hop size of 320 frames, window size 1024 frames, and 64 Mel bins in the range of 50-8000 Hz. During training, each audio clip is randomly truncated to a continuous segment of 5 secs, or padded if shorter. The captions were not altered. The batches with audio and text pairs are randomly sampled at training.

**Encoders.** We chose CNN14 [22] model as the audio encoder due to its SoTA performance. The model has 80M parameters, an embedding size of 2048, and was pretrained with 2M audio clips from AudioSet. The text encoder is the HuggingFace [23] implementation of BERT base uncased. The model has 110M parameters. We limited the max text sequence length to 100 tokens for computational efficiency. The [CLS] token from the final layer of BERT is used as the text embedding with a size of 768. Both, the audio and text embeddings are projected into a multimodal space with two learnable projection matrices resulting in an output dimension of 1024. The temperature parameter $\tau$ is learnable and initialised to 0.007. To prevent training instability, the logits scaled by $\tau$ are clipped to a maximum value of 100.

**Training.** We trained by unfreezing both encoders for 40 epochs. For Section 3.2 we used the CLAP model from the epoch that yielded the best performance in FSD50K. We use Adam Optimiser with an initial learning rate $10^{-3}$ and reduce the learning rate on plateau by $10^{-1}$ with a patience of 10. The models are implemented with PyTorch’s Distributed Data-Parallel and used 16GB V100 GPUs with scaling from 8 to 24 GPUs. Batch size was 128.

3.2. Evaluation setups for CLAP

**Zero-shot (ZS) Evaluation** studies the generalisation of CLAP to unseen classes and audios. The setup is explained in Section 2.2. Instead of using the class label, we constructed a natural language prompting, ‘this is a sound of [class label]’. The prompt was kept the same for all the domains except three. For Emotion Recognition we used ‘this person is feeling [class label]’, for Keyword Spotting we only use the keyword, and for Speaker Counting we used ‘[number between 0 - 10] persons speaking’.

**Supervised Feature Extraction Evaluation** studies the quality of audio representation learned by CLAP. Given a downstream task, we used CLAP as a feature extractor followed by training a classifier of 1 or 3 fully-connected layers (Freeze_L1 or linear probe and Freeze_L3, similar to [22]). We used a learning rate of $10^{-3}$ with Adam Optimizer for 30 epochs. We did not perform grid search for tuning hyperparameters due to computation constraints.

**Supervised Finetune Evaluation** benchmarks CLAP against the best performance for each task in the literature. Given a downstream task, we unfreeze and finetuned the audio encoder together with an attached 1 or 3 fully-connected layers. We used a learning rate of $10^{-4}$ with Adam Optimizer for 30 epochs. We did not perform grid search for tuning hyperparameters due to computation constraints.

---

1https://arxiv.org/pdf/2206.04769.pdf
4. RESULTS AND DISCUSSION

For baseline comparisons we considered the best performance in the literature for Zero-Shot Learning ‘Benchmark (ZS)’ and for Supervised Learning ‘Benchmark (Best)’ included in Table 1. We discuss the effect of freezing the encoders and the effect of prompts for Zero-Shot Classification.

4.1. Zero-Shot (ZS) results

CLAP (ZS) achieved SoTA on Sound Event Classification (SEC) for FSD50K, US8K and ESC50. For ESC50, CLAP achieved 82.6% accuracy (acc) beating human performance of 81% and AudioCLIP (69%) by an absolute 12%. On US8K, achieved 73% acc outperforming AudioCLIP (65%) by an absolute 8%. For the multi-label dataset FSD50K, CLAP beat Wav2CLIP (3%) by an absolute 27% mAP. On GTZAN’s Music vs Speech Classification, CLAP even beat supervised models achieving 100% acc. Results suggests the possibility of reliable audio models with no training involved.

CLAP (ZS) performed better than random on all downstream tasks, but slightly better than random on some music and speech-related tasks. CLAP achieved 47% acc in instrument classification (Beijing Opera), an absolute 22% higher than random. In the Vocal Sound dataset achieved 50% acc, an absolute 33% improvement over random. In Emotion Recognition (ER) and Keyword Spotting (KWS) CLAP outperformed random by up to an absolute 4% acc.

4.2. Supervised results

CLAP (Best) is the best performance among supervised setups and achieved SoTA on 5 datasets. CLAP achieved in GTZAN Music vs Speech Classification 100% acc, in GTZAN Music Genres Classification 91.3% acc, in Mr. Stroke Classification 97.94% acc, in Mr. Tonic Classification 95.34% acc, and in Vocal Sounds Classification 97.95% acc. In other tasks CLAP underperformed SoTA by at most 7%. The lowest performing task was Emotion Recognition’s RAVDESS with 64% acc vs a SoTA of 81%

CLAP performs better in domains like Sound Event Classification than in others like Emotion Recognition, which was more evident in ZeroShot than in Supervised. We hypothesize that SEC tasks perform better because CLAP’s training data consist of audio captioning datasets, which mainly include the description of sound events, acoustic scenes, actions, and objects. On the other hand, the training data is scarce on human speech and the captions do not describe aspects of its content or context. Therefore, CLAP underperforms on speech tasks like KWS and ER. We posit that as we increase training data and increase human speech-based captioning, CLAP’s performance on speech datasets will increase, as suggested in [24].

4.3. Effect of freezing CLAP encoders

We studied how freezing the audio and/or text encoder during training affected performance of the downstream tasks. The projection layers for both encoders are set to learnable. We computed the average of the CLAP (ZS) performance across the downstream tasks, where higher is better. The results are shown in Table 2. The best Avg. CLAP (ZS) score ~0.3265– is obtained by unfreezing both encoders and the worst score – 0.2809– by freezing both encoders. This is expected because unfreezing both encoders allow them to learn the multimodal information from the pairs. Surprisingly, unfreezing the text encoder performed better than unfreezing the audio encoder. Our intuition was that unfreezing the audio encoder would enable learning beyond sound events coming from the pre-trained AudioSet information. However, unfreezing the text encoder was better. A similar insight was found for CLIP models in Computer Vision [25]. This valuable finding suggests that, under the CLAP learning paradigm, it is possible to use an audio encoder of choice and turn it into a Zero-Shot classifier.

| Audio encoder frozen | Text encoder frozen | Avg. CLAP (ZS) score ↑ | ESC50 (acc)↑ |
|----------------------|---------------------|------------------------|--------------|
| ✓                    | ✓                   | 0.2809                 | 0.5555       |
| ✗                    | ✓                   | 0.2818                 | 0.6415       |
| ✓                    | ✗                   | 0.3109                 | 0.7631       |
| ✗                    | ✗                   | 0.3265                 | 0.826        |

Table 2. Effect of freezing text and/or audio encoders on CLAP (ZS) performance across all tasks and ESC50 only.

4.4. Changing prompts in Zero-Shot classification

The training data of CLAP consists of natural language captions in the form of sentences. However, the vast majority of datasets have class labels defined by a few words (i.e. ‘dog barking’ or ‘sneezing’). Using single words instead of language description at (ZS) inference affects performance. To overcome this difference in distributions, we used template prompts, such as ‘this is a sound of [class label]’ (see Section 3.2). Table 3, shows how different prompts for ESC50 can results in a range of 4% acc. Using the suggested prompt vs only the class label improved performance by 1.4%.

| Prompt                      | ESC50 (ZS) (acc)↑ |
|-----------------------------|------------------|
| ‘i can hear [class label]’  | 0.786            |
| ‘this is an audio of [class label]’ | 0.8005  |
| ‘[class label]’             | 0.812            |
| ‘this is [class label]’     | 0.8135           |
| ‘this is a sound of [class label]’ | 0.826 |

Table 3. Prompts used for ZS on ESC50 affects accuracy.

5. CONCLUSION

We introduced CLAP for learning audio concepts from natural language supervision. CLAP does not require class labels for training, enables flexible class prediction, and generalizes well to multiple downstream tasks. CLAP establishes SoTA in Zero-Shot performance and achieves strong supervised performance. CLAP shows potential for building a foundation model for audio that can generalize to a wide range of downstream tasks.
6. REFERENCES

[1] Richard F Lyon, “Machine hearing: An emerging field,” IEEE signal processing magazine, 2010.

[2] Annamaria Mesaros, Aleksandr Diment, Benjamin Elizalde, Toni Heittola, et al., “Sound event detection in the dcase 2017 challenge,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019.

[3] Yu Zhang, Daniel S Park, Wei Han, James Qin, et al., “Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition,” IEEE Journal of Selected Topics in Signal Proc, 2022.

[4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, et al., “Wavlm: Large-scale self-supervised pretraining for full stack speech processing,” IEEE Journal of Selected Topics in Signal Proc., 2022.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, 2020.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, et al., “Learning transferable visual models from natural language supervision,” in International Conference on Machine Learning. PMLR, 2021.

[7] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, et al., “Florence: A new foundation model for computer vision,” arXiv preprint arXiv:2111.11432, 2021.

[8] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, et al., “Wav2clip: Learning robust audio representations from clip,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[9] Andrey Guzhev, Federico Raue, Jörn Hees, and Andreas Dengel, “Audcioclip: Extending clip to image, text and audio,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[10] Joseph Turian, Jordie Shier, et al., “HEAR: Holistic Evaluation of Audio Representations,” in NeurIPS 2021 Competitions and Demonstrations Track, 2022.

[11] Qiuyi Kong, Yong Xu, et al., “Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization,” IEEE/ACM Transactions on Audio, Speech, and Language Proc., 2020.

[12] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, et al., “Efficient training of audio transformers with patchout.” arXiv preprint arXiv:2110.05069, 2021.

[13] Yoonchang Han, Jeongsoo Park, and Kyogu Lee, “Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification,” the Detection and Classification of Acoustic Scenes and Events (DCASE), 2017.

[14] Cristina Luna-Jiménez, Ricardo Kleinlein, David Griol, Zoraida Callejas, Montero, et al., “A proposal for multi-modal emotion recognition using aural transformers and action units on ravdess dataset,” Applied Sciences, 2022.

[15] Byeongjeun Kim, Simyung Chang, Jinkyu Lee, et al., “Broadcasted residual learning for efficient keyword spotting,” arXiv preprint arXiv:2106.04140, 2021.

[16] Yuan Gong, Jin Yu, and James Glass, “Vocalsound: A dataset for improving human vocal sounds recognition,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[17] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, “Fsd50k: An open dataset of human-labeled sound events,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022.

[18] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen, “Clotho: an audio captioning dataset,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

[19] Chris Dongjoon Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, “AudioCaps: Generating Captions for Audios in The Wild,” in NAACL-HLT, 2019.

[20] Irene Martín-Morató and Annamaria Mesaros, “What is the ground truth? reliability of multi-annotator data for audio tagging,” in 2021 29th European Signal Processing Conference (EUSIPCO), 2021.

[21] Justin Salamon, Christopher Jacoby, et al., “A dataset and taxonomy for urban sound research,” in 22nd ACM international conference on Multimedia, 2014.

[22] Qiuyi Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, et al., “Panns: Large-scale pretrained audio neural networks for audio pattern recognition,” IEEE/ACM Trans. Audio, Speech and Lang. Proc., 2020.

[23] Thomas Wolf, Lysandre Debut, et al., “Transformers: State-of-the-art natural language processing,” in Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020.

[24] Huang Xie and Tuomas Virtanen, “Zero-shot audio classification via semantic embeddings,” IEEE/ACM Transactions on Audio, Speech, and Language Proc., 2021.

[25] Xiaohua Zhai, Xiao Wang, et al., “Lit: Zero-shot transfer with locked-image text tuning,” in IEEE/CVF Conference on Computer Vision and Pattern Recog., 2022.