CLAP: Contrastive Language-Audio Pre-training based on Large-Scale Synthetic Parallel Audio-Text Data

Anonymous Author(s)

ABSTRACT

Compared with ample visual-text pre-training research, few works explore audio-text pre-training, which is mostly due to the lack of sufficient parallel audio-text data. In this work, we utilize an audio-captioning based approach to expand parallel audio-text data with the large-scale audio event dataset AudioSet. With the expanded synthetic 1.22M parallel audio-text pairs, we proposed Contrastive Language-Audio Pre-training (CLAP) where contrastive learning is used to pre-train an audio-text bi-encoder. For the first time, we comprehensively demonstrate the performance of such a pre-trained bi-encoder audio-text model on a series of downstream audio-related tasks, including single modality tasks like audio classification and tagging, as well as cross-modal tasks consisting of audio-text retrieval and audio-based text generation. Experimental results indicate that our CLAP pre-trained on enlarged audio-text data achieves the state-of-the-art zero-shot classification performance on most datasets, which demonstrates the benefits of pre-training from a large amount of high-quality synthetic data. Kenny: Perhaps no deed to mention this: The audio encoder also serves as an efficient pattern recognition model by fine-tuning it on audio-related tasks.

CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence*; **Natural language processing**.

KEYWORDS

multi-modal learning, contrastive learning, audio captioning, audio classification, zero-shot inference

ACM Reference Format:

Anonymous Author(s). 2018. CLAP: Contrastive Language-Audio Pre-training based on Large-Scale Synthetic Parallel Audio-Text Data. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Multi-modal machine learning has become increasingly popular since it mimics our experience of learning: we accept and handle information from different modalities. With the success of deep neural networks and large-scale datasets, we have witnessed rapid development of multi-modal learning in recent years. Vision-language pre-training [7, 24, 27, 34] using Transformer has pushed the state of the art (SOTA) on a wide range of cross-modal tasks, such as visual question answering (VQA) [3], Image-Text Retrieval [25], visual commonsense reasoning (VCR) [41], etc. In these works, a joint representation of vision and language modalities is learnt through pre-training on large-scale image-text datasets and then fine-tuned on specific downstream vision-language tasks.

Contrary to substantial efforts on vision-language pre-training, audio-related multi-modal learning, is still in its infancy. Although audio is an important modality, few works explore pre-training involving both audio and language. The bottleneck of audio-language cross-modal learning lies in the scarcity of audio-text parallel data. Compared with large-scale parallel image-text datasets such as COCO [25] (~1.64M pairs), Visual Genome [23] (~5.06M pairs), and Conceptual Captions [33] (~ 12M pairs), current parallel audio-text datasets contain only about 100K pairs (see Section 3.1). The lack of large-scale parallel audio-text datasets may be attributed to the fact that not only the annotation of audio is much more costlier than that of image [42] since it takes longer to listen to the audio samples, but even the loosely coupled audio-text pairs are rare on the web [43].

To alleviate the above problem of data scarcity, previous works on audio-text cross-modal learning mostly incorporate CLIP [32], a powerful model enabling image-text alignment, to facilitate audio-language representation learning. The visual modality works as a pivot to connect audio and text since aligned video-audio data is abundant from most video clips. However, mismatch between audio and visual modalities is commonly observed when detecting objects and events via sound and image. For example, visible objects in videos do not necessarily make sounds while sounds may be produced by objects off-the-screen. Such mismatch leads to noise in audio-visual and audio-text alignment based on visual pivoting, indicated by the limited improvement achieved by these studies [17, 40, 43].

To overcome the data scarcity bottleneck while circumventing the data noise sourcing from other modalities, we propose an audio-captioning based approach to expand parallel audio-text data using AudioSet [14], the most large-scale audio event dataset. AudioSet contains only audio clips and corresponding audio event tags in the original dataset. Based on the provided event tags, we train an AudioSet tag-guided audio captioning model. With the guidance of human-annotated event tags, the generated captions are expected to be more related to the audio content. We thus create a large-scale synthetic parallel audio-text corpus containing 1.22M pairs. Similar to CLIP, we perform contrastive language-audio pre-training (CLAP) on an audio-text bi-encoder, which consists of separate audio and text encoders. The pre-training is comprised

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

of two steps: 1) pre-training on the large-scale synthetic data; 2) further pre-training on real data to adapt to the real distribution.

After pre-training, CLAP can be further transferred to cross-modal and single-modality tasks. For the first time, we present a comprehensive performance analysis of such a pure audio-text bi-encoder on a series of downstream tasks, including audio-text retrieval, audio captioning and audio classification. Results show that significant achievements can be achieved by fine-tuning CLAP on a panel of tasks. With the supervision of natural language, CLAP achieves SOTA zero-shot classification performance on most datasets. By linear probing and fine-tuning the audio encoder of CLAP, we show that CLAP works as an efficient audio feature extractor which can be conveniently transferred to audio classification tasks of different domains.

The main contribution of this paper can be summarized as follows:

- We propose an AudioSet tag-guided audio captioning based approach to generate large-scale synthetic audio-text data to facilitate audio-text pre-training.
- We perform the first pure audio-text contrastive pre-training paradigm based on the synthetic parallel data. The exclusion of CLIP from audio-text pre-training helps eliminate the noise induced by the visual modality.
- We validate the effect of pre-training by fine-tuning CLAP on cross-modal and single-modality tasks including classification, retrieval and generation. CLAP fine-tuning consistently outperforms the baseline trained from scratch.
- We achieve SOTA zero-shot classification performance on several datasets with the supervision of natural language.

2 RELATED WORK

2.1 Vision-Language Pre-training

The research on multi-modal pre-trained models initially thrives in the intersection of vision and language modality. Vision-language pre-trained models generally handles three groups of tasks: understanding tasks like Classification, VQA and Visual Entailment, generation tasks like Image Captioning, and Image-Text Retrieval/Matching tasks. Researchers have proposed different model structures that are specifically suitable for certain group(s) of tasks. Cross-Encoder models process multi-modal inputs in the same encoder to allow full interaction of the two modalities and thus are generally performing well on understanding tasks [7, 24]. Bi-Encoder models encode the visual and textual inputs with different encoders to get separate embeddings [19, 32]. Since the embeddings can be pre-computed and stored for query, they are favorable for efficient retrieval. Encoder-Decoder models encode single or both modalities in the encoder, and use a decoder for generation, which provides the capability for generation tasks [38, 39]. Our model mainly adopts the Bi-Encoder paradigm. We exhibit that it can achieve competitive performance across all three groups of tasks.

For pre-training models, the data size has been shown to be vital for performance regardless of model structure and data quality. Experimental results from the bi-encoder model CLIP shows that its zero-shot image classification performance steadily increase with the number of images involved in pre-training. Another bi-encoder ALIGN [19] further scales up the pre-training data with noisy images from the web and shows that the models pre-trained on noisy data can still outperform those trained on higher-quality data given larger data size. SimVLM [39], an encoder-decoder model, also achieves great success in both understanding and generation tasks with the large pre-training data ALIGN. Inspired by their findings, we propose to synthesize parallel audio-text data for audio-language pre-training, despite the potential noise in the synthetic data.

2.2 Audio-Language Pre-training

With the success of visual-language pre-training, a few recent works start to incorporate audio into multi-modal pre-training. For instance, an audio encoder is added into CLIP with the contrastive learning paradigm. Large-scale video-text datasets are often utilized since the dataset provides visual-text alignment while audio-visual alignment is naturally available from the video data. VATT [1] and MMV [2] uses HowTo100M [30] and AudioSet for pre-training. The audio-text alignment is learnt implicitly through the pivot of visual modality. They validate the effect of pre-training mainly on video action recognition and image classification tasks. AudioCLIP [17] performs the tri-modal contrastive learning explicitly by using AudioSet event tags as the corresponding text. It focuses on transferring to image and audio classification tasks. Wav2CLIP [40], in contrast, does not incorporate text into pre-training. It distills CLIP by pre-training on audio-visual alignment in VGGSound [6]. The learnt audio representation is evaluated on several downstream tasks. Following these works, we adopt contrastive pre-training to learn audio-text cross-modal representation.

Compared with either textual AudioSet tags or video description, VIP∼A$_{\text{N}}$T [43] is proposed recently to curating audio-text pairs by providing natural language audio-focused description for audio clips. CLIP with the prompt "the sound of" is utilized to retrieve captions from the current audio captioning corpus for AudioSet audio clips. A frame of the corresponding video is used as the query. In this way, large-scale parallel audio-text pairs are automatically curated using the visual pivot. Audio-language pre-training without explicitly incorporating the visual modality is conducted on the curated parallel audio-text data. They achieve competitive zero-shot audio-text retrieval and audio classification performance. Inspired by VIP∼A$_{\text{N}}$T, we generate large-scale parallel audio-text data based on AudioSet and audio captioning.

2.3 Audio Event Recognition

Audio event recognition is an emerging field which attracts increasing amount of attention recently. It requires recognizing the rich information in the sounds surrounding us, including the acoustic scenes where we are and what events are present. Audio event recognition contains various tasks like acoustic scene classification [29], audio tagging [14] and sound event detection [5]. In recent years, the release of Detection and Classification of Acoustic Scenes and Events (DCASE) challenges encourages the development of novel datasets, tasks and approaches. The release of AudioSet is also a milestone for audio event recognition. It contains 2.08M 10-second audio clips[1] with 527 annotated sound events. Robust audio representations can be learnt by pre-training a deep neural

[1] Only 1.95M clips are available in this work since some videos are removed.

network on AudioSet. Outstanding performance can be achieved by fine-tuning the pre-trained network on downstream classification tasks. Besides AudioSet, datasets like VGGSound and FSD50K [12] are also released recently to facilitate further research.

More recently, audio captioning [9] is proposed. Beyond audio event tags, a caption provides unconstrained natural language description of an audio clip. Several datasets (see Section 3.1) are proposed to enable audio captioning research. Audio-text retrieval [31] is also proposed recently which requires retrieving audio signals using their textual descriptions and vice versa. The audio-language pre-training in this work is conducted based on these audio-text datasets. We evaluate the benefit of pre-training by fine-tuning CLAP on these audio event recognition and audio-text tasks.

3 DATA PREPARATION

In this work, we use both currently available audio-text datasets and synthetic parallel audio-text data for pre-training. We describe these datasets and synthetic audio-text data generation approach in this section.

3.1 Existing Audio-Text Datasets

Dataset	# Audio-text pairs			Avg # words	Duration /h
	train	val	test		
AudioCaps	49501	2475	4820	8.80	127
Clotho	19195	5225	5225	11.33	44
MACS	17275			9.25	11
Total	85971	7700	10045	9.60	182

Table 1: Statistics of current English audio-text datasets.

Current parallel audio-text datasets are from audio captioning, including AudioCaps [20], Clotho [10] and MACS [28]. AudioCaps is a subset of AudioSet, containing about 50K audio clips. Each audio clip in the training set has one caption annotation while five annotations are provided for audio clips in the validation and test set. Clotho contains 5,929 audio clips with five caption annotations provided for each clip. The audio data are collected from Freesound [13] platform. MACS is a recently released dataset built on TAU Urban Acoustic Scenes 2019 dataset containing 3,930 audio clips. Each audio clip is annotated by several captions, ranging from two to five. The dataset does not provide splits of training, validation or test. A summary of these datasets is in Table 1.

3.2 Data Synthesis by Audio Captioning

Although about 100K audio-text pairs are available in current audio-text datasets, the dataset size is much smaller than image-text datasets (see Section 1). However, large-scale audio event data are available from AudioSet. To leverage the large-scale audio-only data without caption description, we aim to generate captions for audio clips in AudioSet. Since AudioCaps is a subset of AudioSet, we first train a captioning model on AudioCaps and then use it to generate parallel audio-text data from AudioSet. Recent works tend to make use of supplementary information to guide captioning such as keyword [11] and similar captions [21]). However, these systems often suffer from poor prediction accuracy of supplementary information since the guidance can only be inferred from the input audio during inference. In AudioSet, the label, consisting of audio event tags presented in the audio clip, serve as an effective guidance since it is available for all clips. Therefore, to enhance the quality of generated captions, we incorporate the event tags into caption generation. The model generates a caption conditioned on both the input audio and the hint from AudioSet tags. The architecture is shown in Figure 1. It contains an audio encoder and a text decoder. A sequence of audio features $\mathbf{x}$ is fed to the encoder and transformed into a sequence of high-level representations $\mathbf{e}^a$.

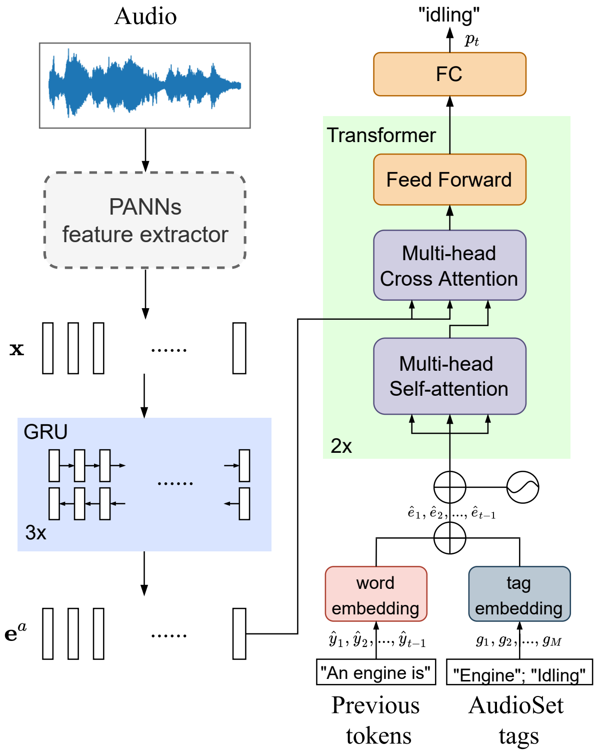


Figure 1: The proposed audio captioning system with AudioSet tag guidance. The system generates caption based on both the input audio clip and the provided AudioSet tags.

$$\mathbf{e}^a = \text{Encoder}(\mathbf{x})$$

The decoder predicts the probability of each token at the time-step t conditioned on partly decoded tokens $\{\hat{y}_n\}_{n=1}^{t-1}$, the provided AudioSet tags $\{g_m\}_{m=1}^{M}$ (M is the number of tags) and $\mathbf{e}^a$:

$$p_t = \text{Decoder}(\mathbf{e}^a, \{\hat{e}_n\}_{n=1}^{t-1})$$

$$\hat{e}_n = e_n^w + e^g$$

$$e_n^w = \text{WE}(\hat{y}_n), \quad e^g = \frac{1}{M}\sum_{m=1}^{M}\text{TE}(g_m)$$

where WE and TE denote word embedding and tag embedding layer which transform $\hat{y}_n$ and g_m into fixed-dimensional vectors. Starting from the special "<BOS>" token, the decoder auto-regressively predicts the next token until "<EOS>" is reached.

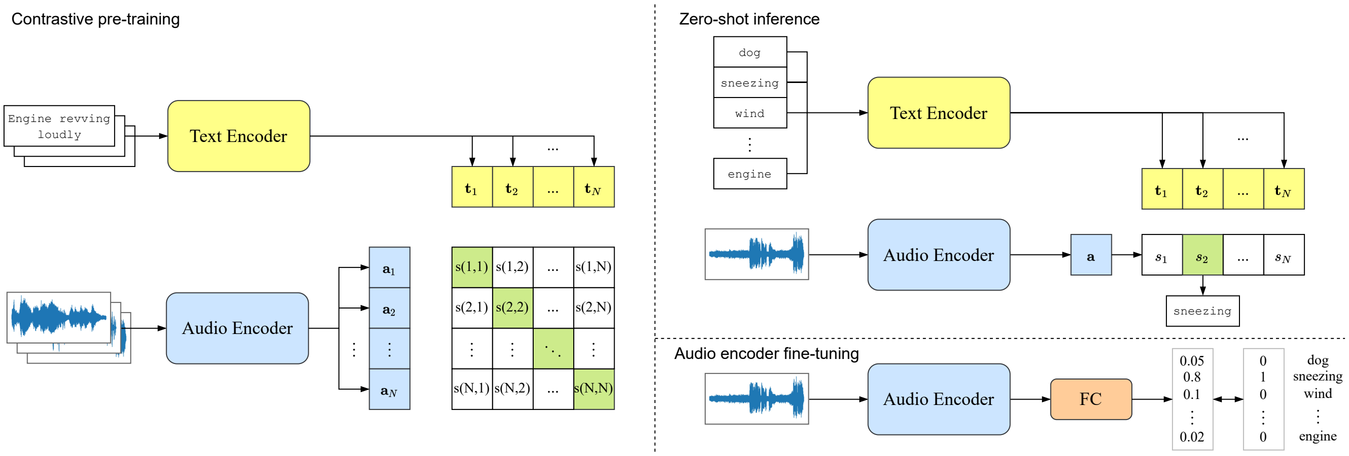


Figure 2: An overview of our proposed language-audio pre-training approach. Similar to CLIP, we first train an audio encoder and a text encoder jointly by contrastive learning. Then the pre-trained bi-encoder can be transferred to zero-shot classification. The pre-trained audio encoder can also be treated as a better feature extractor and be further fine-tuned on downstream tasks.

In this work, we utilize deep embeddings from PANNs, specifically the *CNN14* variant, as the input audio feature $\mathbf{x}$. The encoder is a three-layer bidirectional gated recurrent unit (GRU) following [11] while the decoder is a two-layer Transformer with the final fully connected (FC) layer. The captioning system is trained by word-level cross entropy (CE) loss:

$$\mathcal{L} = \sum_{t=1}^{T} -\log\left(p_t(y_t)\right)$$

where y_t is the ground truth token at the time-step t.

After training the AudioSet tag-guided captioning model, we use it to generate captions for large-scale AudioSet audio clips. However, the data distribution of AudioCaps is different from AudioSet since audio clips with specific event tags are excluded during the construction process of AudioCaps [20]. To circumvent the distribution bias problem, we exclude audio clips with tags that never appear in AudioCaps, with about 1.22M audio clips left. One caption is generated for each audio clip using the enhanced captioning model, resulting in about 1.22M audio-text pairs. We use this large-scale synthetic parallel audio-text data for pre-training.

4 AUDIO-TEXT PRETRAINING FRAMEWORK

In this section, we describe the proposed framework of contrastive audio-text bi-modal representation learning from audio-text pairs. The framework consists of an audio encoder and a text encoder for the two modalities. As Figure 2 shows, the model is first pre-trained by contrastive learning. Then the pre-trained model is used for zero-shot inference by calculating the similarity scores between the audio clip and all textual labels. The audio encoder can be further fine-tuned on downstream tasks to boost performance.

We first illustrate the contrastive pre-training approach. Then the architectures of the two encoders are introduced respectively.

4.1 Contrastive Language-Audio Pre-training

Similar to CLIP, the proposed contrastive learning approach learns the correspondence between the text content and the audio events in an arbitrary audio-text pair. For an audio clip $\mathcal{A}$ and a sentence $\mathcal{T}$, the audio and text encoders Enc_A and Enc_T transform them into two embeddings $\mathbf{a}$ and $\mathbf{t}$ respectively. A multi-modal embedding space is learned by maximizing the similarity between $\mathbf{a}$ and $\mathbf{t}$ of matched audio-text pairs and minimizing that of mismatched pairs. Following CLIP, the training objective is to minimize the InfoNCE loss [37]. Given a minibatch of N audio-text pairs $(\mathcal{A}_1, \mathcal{T}_1), (\mathcal{A}_2, \mathcal{T}_2), \ldots, (\mathcal{A}_N, \mathcal{T}_N)$, their embeddings are calculated:

$$\mathbf{a}_i = \text{Enc}_A(\mathcal{A}_i)$$
$$\mathbf{t}_i = \text{Enc}_T(\mathcal{T}_i)$$

The training loss is a symmetric cross entropy loss between the predicted cosine similarity scores and the ground truth pairing labels:

$$s(i,j) = \frac{\mathbf{a}_i \cdot \mathbf{t}_j^{\mathrm{T}}}{\|\mathbf{a}_i\| \cdot \|\mathbf{t}_j\|}$$
$$\mathcal{L}_i^{A\to T} = -\log \frac{exp\left(s(i,i)/\tau\right)}{\sum_{j=1}^{N} exp(s(i,j)/\tau)}$$
$$\mathcal{L}_i^{T\to A} = -\log \frac{exp\left(s(i,i)/\tau\right)}{\sum_{j=1}^{N} exp(s(j,i)/\tau)}$$
$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{L}_i^{A\to T} + \mathcal{L}_i^{T\to A})$$

where τ is the temperature optimized jointly with Enc_A and Enc_T.

4.2 Audio Encoder

Similar to the feature extractor in Section 3.2, we use the pre-trained CNN14 from PANNs [22] as Enc_A instead of training the model from scratch. Time-frequency representation Log Mel Spectrogram

(LMS) is extracted from the input audio and fed to 12 convolution blocks. 2×2 max pooling is done between every two blocks. After the convolution blocks, the audio embedding $\mathbf{a}$ is obtained by a global pooling on the feature map and transformation through a fully-connected layer. Although Transformer-based models are applied for audio classification recently [15] and achieves better performance than convolutional neural networks (CNN), it works on a sequence of patch embeddings without sub-sampling, resulting in high memory demand. Therefore, we adopt the pre-trained CNN14 to enable a larger minibatch size for training.

4.3 Text Encoder

For the text encoding part, we utilize BERT to transform $\mathcal{T}$ into $\mathbf{t}$. It is a deep Transformer pre-trained on large-scale corpus, including BooksCorpus and English Wikipedia, by self-supervised learning. Due to its powerful capability to extract representations with contextual semantics, BERT has exhibited superior performance on a series of language understanding tasks [8]. In this work, we employ BERT$_{\text{MEDIUM}}$ [36] as Enc$_{\mathcal{T}}$ for better computation efficiency and lower memory requirements. It consists of eight Transformer layers with a hidden embedding size of 512.

5 EXPERIMENTAL SETUP

In this section, we first present our experimental setup in training the audio captioning model to synthetic parallel audio-text data generation. Then the CLAP pre-training, fine-tuning and evaluation procedures are described in detail.

5.1 AudioSet Tag-Guided Audio Captioning Model Training

The AudioSet tag-guided captioning model is trained on AudioCaps for 25 epochs with a batch size of 64. The learning rate linearly warms up to 5×10^{-4} and then exponentially decays to 5×10^{-7} until the end of training. Scheduled sampling [4] and label smoothing [35] are used for regularization. We use stochastic weight average [18] by averaging the last five checkpoints as the final model. For inference on AudioSet audio clips, we use beam search with a beam size of three.

5.2 CLAP Pre-training

The contrastive pre-training consists two steps. First, the model is pre-trained on large-scale synthetic parallel audio-text data mentioned in Section 5.1. We use a batch size of 128 and train the model for 200K iterations. About 1,200 audio-text pairs are randomly selected from the synthetic data to form a separate validation set. The model is validated every 500 iterations on the validation set. We use the Adam optimizer with the maximum learning rate of 1×10^{-4}. The learning rate is decayed by a cosine scheduler [26] with linear warm up in the first 10k iterations.

After pre-training on the synthetic data, we further pre-train the model on the real data. The model with the best performance on the synthetic validation set is used to initialize parameters for this step. Since there is a gap between the quality of real and synthetic data, the second pre-training step is adopted to alleviate the bias caused by synthetic data. We use the combination of all training sets of real audio-text data introduced in Section 3.1 for training. The training setup is similar to the first step with several modifications on hyper-parameters. The total training iterations and warm up iterations are 15000 and 750 while the model is validated every 750 iterations. The bi-encoder trained after this step is referred to as CLAP.

5.3 Downstream Tasks

Task	Dataset	# Audio clips	Metric
Audio-text Retrieval	AudioCaps	50K	R@K
	Clotho	6K	
Audio Captioning	AudioCaps	50K	COCO & FENSE
	Clotho	6K	
Audio Classification	ESC50	2K	Accuracy
	UrbanSound8K	8K	
	TAU2019	14K	
	VGGSound	192K	
Audio Tagging	FSD50K	50K	mAP
	AudioSet	1.93M	

Table 2: A summary of downstream cross-modal and single-modality tasks. TAU2019 is the acoustic scene dataset from DCASE2019 challenge task1A.

The pre-trained CLAP can be transferred to a series of downstream tasks, which is summarized in Table 2, including both cross-modal tasks and single-modality tasks.

Cross-modal audio-text tasks include audio-text retrieval and audio captioning. For audio-text retrieval, we use recall at K (R@K) as the evaluation metric. Standard COCO evaluation metrics from image captioning are used to evaluate audio captioning performance. Besides, we also incorporate recently proposed FENSE [44] into evaluation for its higher correlation with human judgments.

Single-modality tasks include single-label (classification) and multi-label (tagging) audio classification. Accuracy and mean average precision (mAP) are used for evaluation. We include several datasets with the size ranging from 2K to 1.93M for comparison with previous works.

5.4 Zero-shot Classification

With the pre-trained CLAP, we can perform zero-shot classification. If a textual label contains "_", we replace "_" with a blank. CLAP calculates the similarity scores between a given audio clip and all these textual labels. These scores are treated as the predicted probability of each audio event for evaluation.

5.5 Fine-tuning

5.5.1 Audio-text Retrieval. The fine-tuning on audio-text retrieval tasks uses almost the same configuration as the pre-training step. For both AudioCaps and Clotho, we fine-tune the pre-trained bi-encoder model for 20 epochs using the InfoNCE loss with a batch size of 128. The learning rate linearly warms up to the maximum value in the first epoch. The maximum learning rate for AudioCaps and Clotho are 5×10^{-5} and 2×10^{-6}, respectively.

5.5.2 *Audio Captioning.*

The audio captioning system is similar to the model in Section 3.2 except 1) the audio feature is extracted by CLAP instead of PANNs; 2) the system does not receive guidance from AudioSet tags. For both AudioCaps and Clotho, the training and inference configuration follows Section 5.1.

5.5.3 *Audio Classification and Tagging.*

For single-modality tasks, we further fine-tune the pre-trained audio encoder Enc_A. An extra fully-connected (FC) layer is added to Enc_A for classification. We perform two types of fine-tuning: linear probing and fine-tuning the whole Enc_A. For linear probing, Enc_A is used as a feature extractor and only the final FC layer is trained while no parameters are frozen in the second setting. Cross entropy loss and binary cross entropy loss are used for classification and tagging training respectively.

6 RESULTS

In this section, the performance of CLAP is presented comprehensively. We first evaluate the quality of synthetic parallel audio-text data. Then for both cross-modal and single-modality downstream tasks, we reveal the influence of pre-training. For single-modality tasks, transferring is done by zero-shot classification and fine-tuning Enc_A.

6.1 Benefits of Synthetic Parallel Audio-text Data

	B_4	R	M	C	S	F
System	26.4	49.0	24.5	80.4	21.0	62.5
Human	29.0	49.5	28.8	90.8	28.8	68.0

Table 3: The comparison of synthetic parallel audio-text data and real data in terms of audio captioning performance. Metrics include BLEU_4 (B_4), ROUGE_L (R), METEOR (M), CIDEr (C), SPICE (S) and FENSE (F).

The quality of synthetic data is first evaluated in terms of captioning performance. We compare the performance of synthetic captions and human-annotated captions on AudioCaps test set. Since human annotations are used both as the candidate to be evaluated and the reference, we use a round-robin evaluation schedule. Specifically, in each round we exclude one reference annotation and evaluate the caption based on the left four annotations. The five scores are averaged as the performance indicator. Section 6.1 shows the comparison. Metrics reveal that the synthetic data performance is close to human annotations. Kenny: Any explanation why the synthetic data does better with B/R/M than C/S/F? With the guidance of AudioSet tags, the captioning model is capable of generating high-quality parallel audio-text data based on AudioSet.

Then we conduct the pre-training of the bi-encoder model on synthetic parallel audio-text data. The results are shown in the upper half of Table 4. For comparison, we include the curated audio-focused captions (AC) from VIP∼A_NT [43], which uses CLIP and the prompt "the sound of" to retrieve captions from AudioCaps and Clotho training corpus. The two synthetic datasets share a similar size (1.22M and 1.08M). Results indicate that the model trained on our synthetic data significantly outperforms VIP∼A_NT except for text to audio retrieval on Clotho. The inferior performance of pre-training on curated AC indicates that using the visual modality as a pivot between audio and text leads to noisy data. The noise may come from the the mismatch between audio and visual modalities, explained in Section 1. Although the comparison is not fair since curated AC do not use audio-text alignments which we use to train the captioning model, our focus is that generating captions directly from audio and AudioSet tags eliminate the noise induced by the visual modality. Note that Clotho captions are also used to curate audio-text data in VIP∼A_NT while we only use AudioCaps to train the captioning model. The model trained on our synthetic data may suffer from the distribution difference between the two datasets when evaluated on Clotho.

After the second pre-training stage, the model learns the distribution of real data and achieves good performance on both datasets, especially on Clotho. To validate the effectiveness of synthetic parallel audio-text data, we also train an bi-encoder on real data from scratch. Kenny: What is "real" in the table? is it the same as the curated AC in the upper half? The comparison is given in the lower half of Table 4. Pre-training on synthetic data achieves significant performance improvement on AudioCaps while the performance on Clotho is comparable with the model trained from scratch, indicating the benefit of large-scale pre-training. Kenny: It seems that the results on Clotho in the lower half is not that good, explain?

6.2 Cross-modal Audio-and-Language Tasks

Table 5 and Table 6 show the performance achieved by transferring CLAP to cross-modal audio-text tasks, including audio-text retrieval and audio caption generation.

Audio-text Retrieval. For audio-text retrieval, we compare the model fine-tuning from CLAP with the model trained from scratch with the same architecture. As the size of Clotho is small, the model trained from scratch performs poorly. With the initialization from large-scale pre-trained CLAP, significant improvement can be witnessed on both AudioCaps and Clotho.

Automated Audio Captioning. For audio captioning, CLAP is taken as the audio feature extractor. The captioning model takes the feature extracted from CLAP to generate captions. We compare CLAP with PANNs by training the model with the same architecture but taking PANNs CNN14 feature as the input. CLAP feature outperforms PANNs on metrics regarding the semantic content like CIDEr, SPICE and FENSE while no improvement is observed on metrics measuring N-gram overlaps, including BLEU_4 and ROUGE_L. This means CLAP feature helps the model generate more content-related audio descriptions while the N-grams may be different from the references.

6.3 Single-modality Audio Classification

6.3.1 *Zero-shot Transfer.*

We first perform zero-shot inference on several datasets to reveal the transferring ability of CLAP. Previous works enabling zero-shot inference, including AudioCLIP, Wav2CLIP and VIP∼A_NT, are incorporated for comparison. In these works, CLIP are incorporated for synthetic audio-text data generation or pre-training. AudioCLIP is trained on AudioSet with audio

Training Data	AudioCaps						Clotho					
	Audio ⇒ Text			Text ⇒ Audio			Audio ⇒ Text			Text ⇒ Audio		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
synthetic (1.22M)	32.6	62.9	76.7	23.5	54.3	68.4	7.6	22.0	31.5	5.6	15.8	23.8
Curated AC (1.08M) [43]	15.2	-	52.9	9.9	-	45.6	7.1	-	30.7	6.7	-	29.1
real	40.0	72.9	84.9	33.7	69.3	82.1	18.5	42.2	54.4	13.3	34.4	48.0
synthetic → real (CLAP)	44.2	77.5	87.4	35.2	71.3	84.3	18.1	40.1	52.5	12.4	34.0	47.3

Table 4: Audio-text retrieval performance of pre-trained models.

Model	AudioCaps						Clotho					
	Audio ⇒ Text			Text ⇒ Audio			Audio ⇒ Text			Text ⇒ Audio		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
from scratch	38.6	74.3	86.0	33.4	68.8	82.2	13.6	34.4	46.1	11.5	32.2	45.0
fine-tune from CLAP	47.4	78.1	87.6	38.0	72.9	84.5	18.4	38.5	53.4	14.2	36.6	49.6

Table 5: A comparison of audio-text retrieval performance between the model trained from scratch and that fine-tuned with CLAP initialization.

Audio Feature	AudioCaps						Clotho					
	B_4	R	M	C	S	F	B_4	R	M	C	S	F
PANNs	27.1	49.7	24.4	71.6	18.1	60.1	16.2	37.9	17.6	40.4	12.2	44.1
CLAP	27.1	49.6	24.8	73.2	18.6	61.1	16.2	37.6	17.8	41.1	12.8	44.9

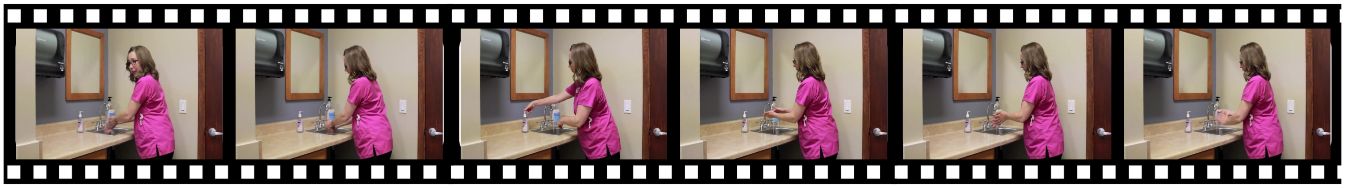
Table 6: A comparison of audio captioning performance between using audio features extracted from PANNs and CLAP.

	Model	ESC50	UrbanSound8K	TAU2019	VGGSound (mAP)	FSD50K	AudioSet
	SOTA	95.9	89.5	85.1	52.5	56.7	45.9
Zero-shot	AudioCLIP	69.4	68.8	-	-	-	-
	Wav2CLIP	41.4	40.4	-	- (10.0)	3.0	-
	VIP$\sim$A$_N$T	69.2	71.7	-	-	-	13.3
	CLAP	80.6	77.3	32.2	14.9 (13.5)	31.3	10.5
Linear probing	PANNs	90.8	82.5	58.9	41.1	29.8	-
	CLAP	94.4	86.1	63.2	43.6	33.3	38.8
Fine-tuning	PANNs	94.7	87.7	76.4	54.4	57.3	-
	CLAP	95.7	88.8	77.1	53.8	59.7	43.9

Table 7: Audio classification and tagging performance in different settings: 1) zero-shot transfer 2) linear probing a pre-trained model 3) fine-tuning a pre-trained model.

event labels while Wav2CLIP is trained on VGGSound without using labels. VIP$\sim$A$_N$T takes a similar pre-training scheme as CLAP: training on synthetic data first and then fine-tuning on real data. We also include current SOTA results as topline for reference. Results are shown in the upper half of Table 7. On both ESC50 and UrbanSound8K, under the setting that AudioSet or AudioCaps are incorporated into pre-training, CLAP significantly outperforms AudioCLIP and VIP$\sim$A$_N$T. On VGGSound, we list mAP in the parentheses to compare with Wav2CLIP. CLAP outperforms Wav2CLIP even though the latter is pre-trained on the same dataset, indicating the effective transfer ability of CLAP. However, CLAP achieves a very low mAP on AudioSet even though Enc$_A$ is trained on AudioSet before the contrastive pre-training. Apart from the data distribution bias caused by the creation of AudioCaps (see Section 3.2), we observe that the noise in AudioSet labels exacerbates the problem. Previous works find that the annotation errors that not annotating events present in an audio clip are common in AudioSet [16]. An example is shown in Figure 3. Speech from a woman can be clearly heard in the audio clip and CLAP assigns high probability to the event "Female speech, woman speaking". However, the event does

Filename: -1Hub6Ps_cc_10.000_20.000.wav



Event labels: Sink (filling or washing) | Water tap, faucet | Speech | Hands | Inside, small room

Top-5 CLAP prediction: Frying (food) (0.35) | Female speech, woman speaking (0.34) | Sizzle (0.34) | Water tap, faucet (0.32) | Boiling (0.32)

Figure 3: An example of annotation errors in AudioSet. A woman is speaking in the audio clip while the corresponding event "Female speech, woman speaking" is not annotated.

not occur in the AudioSet annotation. We assume such annotation errors make the results of certain audio event classes not reliable. On FSD50K where annotations are more reliable, CLAP achieves a much higher mAP.

6.3.2 Audio Encoder Fine-tuning. Besides zero-shot inference, we also evaluate the transferring ability of CLAP by fine-tuning Enc_A on these audio classification tasks. We compare it with PANNs since they share the same CNN14 architecture. The lower half of Table 7 shows the results. In both linear probing and fine-tuning settings, CLAP outperforms PANNs on ESC50 and TAU2019. With only one FC layer as the classifier, the performance of linear probing CLAP on ESC50 and UrbanSound8K is even close to current SOTA results, indicating that CLAP serves as a powerful audio feature extractor. Especially for small audio event datasets, CLAP is able to extract highly discriminative features for classification.

7 CONCLUSION

In this work, we propose an AudioSet tag-guided audio captioning model to generate large-scale parallel audio-text data on AudioSet. The audio-text data generation approach does not incorporate CLIP to eliminate the noise induced from the visual modality. Based on the large-scale synthetic parallel audio-text data, we pre-train a bi-encoder audio-text model using contrastive learning. After pre-training the model on the synthetic data and the real data successively, we obtain CLAP which can be transferred to a series of downstream tasks. Experimental results on both cross-modal and single-modality tasks, including retrieval, generation and classification, validate the effectiveness of CLAP. Under the difficult zero-shot condition that no training data are available, CLAP exhibits the state-of-the-art performance on most datasets.

8 ACKNOWLEDGEMENTS

This work has been supported by National Natural Science Foundation of China (No.61901265), State Key Laboratory of Media Convergence Production Technology and Systems Project (No.SKLMCPTS 2020003) and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102). Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

REFERENCES

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Proceedings of Conference on Neural Information Processing Systems (NIPS)* 34 (2021).

[2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. *Proceedings of Conference on Neural Information Processing Systems (NIPS)* 33 (2020), 25–37.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2425–2433.

[4] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent Neural networks. In *Proceedings of Conference on Neural Information Processing Systems (NIPS)*. 1171–1179.

[5] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen. 2015. Polyphonic sound event detection using multi label deep neural networks. In *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–7.

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 721–725.

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Proceedings of European Conference on Computer Vision (ECCV)*. Springer, 104–120.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 4171–4186.

[9] Konstantinos Drossos, Sharath Adavanne, and Tuomas Virtanen. 2017. Automated audio captioning with recurrent neural networks. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 374–378.

[10] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 736–740.

[11] Ayşegül Özkaya Eren and Mustafa Sert. 2020. Audio Captioning Based on Combined Audio and Semantic Embeddings. In *Proceedings of IEEE International Symposium on Multimedia (ISM)*. https://doi.org/10.1109/ISM.2020.00014

[12] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 30 (2022), 829–852.

[13] Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of ACM International Conference on Multimedia*. 411–412.

[14] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

[15] Yuan Gong, Yu An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proceedings of Conference of the International Speech Communication Association*. ISCA, 56–60.

[16] Yuan Gong, Yu-An Chung, and James Glass. 2021. PSLA: Improving Audio Tagging With Pretraining, Sampling, Labeling, and Aggregation. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 29 (2021), 3292–3306.

[17] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2021. Audioclip: Extending clip to image, text and audio. *arXiv preprint arXiv:2106.13043* (2021).

[18] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407* (2018).

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. PMLR, 4904–4916.

[20] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating Captions for Audios in The Wild. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 119–132.

[21] Yuma Koizumi, Yasunori Ohishi, Daisuke Niizumi, Daiki Takeuchi, and Masahiro Yasuda. 2020. Audio Captioning using Pre-Trained Large-Scale Language Model Guided by Audio-based Similar Caption Retrieval. *arXiv preprint arXiv:2012.07331* (2020).

[22] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing* 28 (2020), 2880–2894.

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123, 1 (2017), 32–73.

[24] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of European Conference on Computer Vision (ECCV)*. Springer, 121–137.

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision (ECCV)*. Springer, 740–755.

[26] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[28] Irene Martin and Annamaria Mesaros. 2021. Diversity and Bias in Audio Captioning Datasets. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*. Barcelona, Spain, 90–94.

[29] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2018. A multi-device dataset for urban acoustic scene classification. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*. 9–13.

[30] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[31] Andreea-Maria Oncescu, A Koepke, João F Henriques, Zeynep Akata, and Samuel Albanie. 2021. Audio Retrieval with Natural Language Queries. In *Proceedings of Conference of the International Speech Communication Association*.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[33] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*. 2556–2565.

[34] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826.

[36] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962* (2019).

[37] Aaron Van den Oord, Yazhe Li, Oriol Vinyals, et al. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* 2, 3 (2018), 4.

[38] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. *arXiv preprint arXiv:2202.03052* (2022).

[39] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904* (2021).

[40] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2021. Wav2CLIP: Learning Robust Audio Representations From CLIP. *arXiv preprint arXiv:2110.11499* (2021).

[41] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. 6720–6731.

[42] Zhiling Zhang, Zelin Zhou, Haifeng Tang, Guangwei Li, Mengyue Wu, and Kenny Q Zhu. 2021. Enriching Ontology with Temporal Commonsense for Low-Resource Audio Tagging. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 3652–3656.

[43] Yanpeng Zhao, Jack Hessel, Youngjae Yu, Ximing Lu, Rowan Zellers, and Yejin Choi. 2021. Connecting the Dots between Audio and Text without Parallel Data through Visual Knowledge Transfer. *arXiv preprint arXiv:2112.08995* (2021).

[44] Zelin Zhou, Zhiling Zhang, Xuenan Xu, Zeyu Xie, Mengyue Wu, and Kenny Q Zhu. 2021. Can Audio Captions Be Evaluated with Image Caption Metrics? *arXiv preprint arXiv:2110.04684* (2021).