MULTI-MODAL CONTINUAL PRE-TRAINING FOR AUDIO ENCODERS

Gyuhak Kim[1]*, *Ho-Hsiang Wu*[2], *Luca Bondi*[2], *Bing Liu*[1]†,

[1] University of Illinois at Chicago
[2] Bosch Research, USA - Bosch Center for Artificial Intelligence

ABSTRACT

Several approaches have been proposed to pre-train an audio encoder to learn fundamental audio knowledge. These training frameworks range from supervised learning to self-supervised learning with a contrastive objective under multi-modal supervision. However, these approaches are constrained to a single pretext task, preventing their adaptability to multi-modal interactions beyond the modalities provided in training data. Continual learning (CL), in the meantime, allows machine learning systems to incrementally learn a new task while preserving the previously acquired knowledge, making the system more knowledgeable over time. The existing CL approaches are limited to learning downstream tasks such as classification. In this work, we propose to combine CL methods with several audio encoder pre-training methods. The audio encoders, when pre-trained continually over a sequence of multi-modal tasks, namely audio-visual and audio-text, exhibit improved performance across various downstream tasks compared to their non-continual learning counterparts, due to knowledge accumulation. The audio encoders are also capable of performing cross-modal tasks of all learned modalities.

***Index Terms*—** Continual Learning, Multi-Modal Learning, Audio Representation Learning, Audio Classification, Cross-Modal Retrieval

1. INTRODUCTION

Audio encoders have been pre-trained for various pretext tasks. As a supervised pre-training method, PANNs [1] create an audio encoder by employing a sequence of convolutional neural network (CNN) and pre-train the encoder on AudioSet [2], an annotated large scale audio data. As self-supervised learning methods gain popularity, several audio pre-training techniques have leveraged multi-modal signals through contrastive loss inspired by contrastive language-image pre-training (CLIP) [3]. For instance, Wav2CLIP [4] builds an audio encoder capable of handling audio-visual interactions through video data while CLAP [5] uses audio caption data to pre-train audio encoders with audio-text interactions.

These audio encoder pre-training methods have limitations as they are pre-trained with a single task, making them difficult to perform downstream tasks that encompass modalities not involved in the original pre-training. The major challenge in learning many multi-modal (e.g., tri-modal) interactions is the complexity of acquiring training data that spans more modalities. AudioCLIP [6] attempts to learn tri-modalities, audio-visual-text, but it requires the training data, where all modalities need to be jointly aligned by each sample. This prevents the method from broad applications.

Continual learning (CL), meanwhile, is a learning paradigm in which a system learns a sequence of tasks to become more capable of performing various tasks by knowledge accumulation in the process. The major challenge in CL is catastrophic forgetting [7], a phenomenon where the system forgets the previous knowledge after learning a new task. The existing approaches to CL mainly focus on downstream tasks such as classification [8]. Recently, [9] proposes a continual pre-training method, but it is limited to unimodal tasks in natural language processing. Although [10] proposed an evaluation benchmark for CL systems in pre-training multi-modal tasks, it is limited to bimodal interactions, namely, vision-language.

In this work, we propose to combine CL methods with several audio encoder pre-training methods and show the efficacy of CL techniques in constructing general audio encoders for various multi-modal tasks in the audio domain. This work makes the following contributions. 1) Our work provides the first study on continual pre-training of audio encoders for a sequence of multi-modal tasks. Namely, using an audio encoder pre-trained with audio data, we continually pre-train it on two sequences of multi-modal tasks: i) from audio to audio-text task and ii) from audio to audio-vision task, followed by audio-text task. 2) We adapt existing CL methods and provide a training framework for continual pre-training of audio encoders. 3) We systematically evaluate the continually pre-trained audio encoders for various downstream tasks including linear probing, zero-shot audio classification, and cross-modal retrieval of all the learned modalities. We found that the audio encoders continually pre-trained with CL methods show improvements in most downstream tasks due to knowledge accumulation and they are also capable of performing cross-modal retrieval tasks of all modalities.

2. METHOD

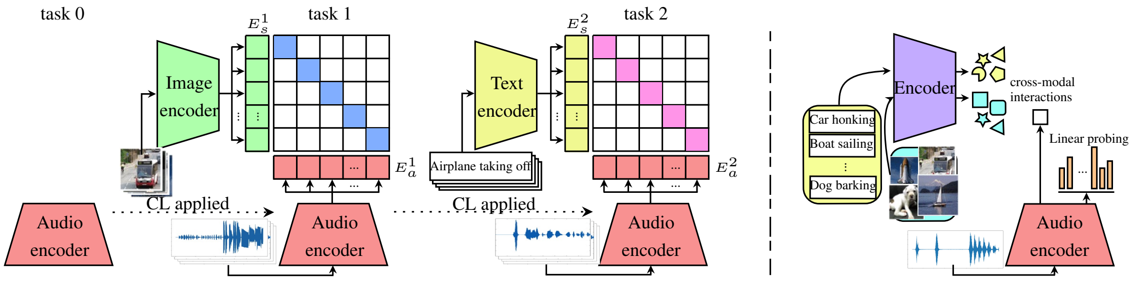
We pre-train an audio encoder on a sequence of multi-modal data involving audio signals to construct an audio encoder capable of performing various downstream tasks in the audio domain. The overview of the training process is presented in Fig. 1.

For training an audio encoder, we follow the training framework of CLIP [3], where an image encoder is trained with contrastive loss on visual data under the supervision of the corresponding text samples. Suppose an audio encoder f_a is initialized from a pre-trained parameters and it is trained on pairs of audio samples X_a^k and the corresponding supervision data X_s^k for task k. An example of the pair in our case is an audio-text or audio-image samples. However, simply training the encoder without considering the general knowledge in the pre-trained initial encoder leads to the loss of the general knowledge. This is due to the fact that training on task-specific data $\{X_a^k, X_s^k\}$ substantially alters the model's weights, subsequently erasing the foundational knowledge. This phenomenon is called catastrophic forgetting [7].

In order to mitigate the forgetting problem and promote knowledge accumulation, we utilize techniques from continual learning.

[*] The author performed the work during an internship at Bosch.

[†] The work of the author was supported in part by a research contract from KDDI and three NSF grants (IIS-1910424, IIS-1838770, and CNS-2225427)



(a) Pre-training

(b) Downstream tasks

Fig. 1. The overview of the training framework. (a) Initialized with a pre-trained audio encoder f (i.e., task 0), the encoder is pre-trained for a sequence of multi-modal tasks using the symmetric cross-entropy loss and a continual learning objective. The big square boxes in task 1 and 2 indicate the pairwise cosine similarities between the audio and the supervision embeddings, which are denoted as E_a^k and E_s^k for task $k \in \{1, 2\}$, in the contrastive loss. The audio encoder is transferred to the next task and is trained for learning a new task. (b) After pre-training, the audio encoder is used for downstream tasks of unimodal classification as well as all the learned multi-modal interactions. The tasks include audio classification via linear probing and cross-modal interactions including zero-shot classification and cross-modal retrieval.

In the following subsections, we present the objective functions designed to facilitate the learning of multi-modal interactions in the new task while preserving the accumulated knowledge obtained throughout the sequence of tasks.

2.1. Continual Pre-Training for Multi-Modal Tasks

Let f_a^k be the audio encoder, g^k the projection function, and f_s^k a pre-trained encoder (e.g., image or text encoder) employed to generate embeddings for the data X_s^k paired with the audio data X_a^k for supervision in task k. Denote the parameters of the functions by θ_a^k, ϕ_a^k, and θ_s^k, respectively. For N pairs of data $\{X_a^k, X_s^k\}$, obtain the audio and supervision embeddings as

$$E_a^k = g^k(f_a^k(X_a^k; \theta_a^k); \phi_a^k); E_s^k = f_s^k(X_s^k; \theta_s^k),$$

where both E_a^k and E_s^k are in $\mathcal{R}^{N \times D}$. Using the symmetric cross-entropy loss l introduced in [3], we minimize the contrastive loss

$$\mathcal{L}_c(\theta_a^k, \phi_a^k) = l(E_a^k, E_s^k). \tag{1}$$

We do not train the supervision model f_s^k as it is already pre-trained and its purpose is to simply generate embeddings for the audio encoder.

The contrastive loss computes pairwise cosine similarity between E_a^k and E_s^k, and trains the model by aligning the audio embeddings with their corresponding embeddings correctly paired by supervision while also pushing apart the audio embeddings from other incorrectly paired embeddings.

2.2. Continual Learning Methods

Continual learning leverages knowledge distillation (KD) [11]. We will briefly introduce KD and discuss how to apply it for continual pre-training of audio encoders. KD is originally proposed to transfer the knowledge of a large pre-trained teacher network h^{tchr} into a compact student network h^{std}. The knowledge of a model is characterized by the acquired mapping from the input X of the current task to the output vectors $h^{std}(X)$ and $h^{tchr}(X)$. While minimizing the main objective function $\mathcal{L}$ (e.g., cross-entropy loss), the student network is instructed to mimic the behavior of the teacher network by minimizing the KL-Divergence (KLD) loss $\text{KLD}(h^{std}(X), h^{tchr}(X))$. Here, only h^{std} is trained while h^{tchr} is fixed, and and h^{tchr} is discarded after training.

Continual learning adapts KD to alleviate catastrophic forgetting. At task k, a new network is initialized using the parameters from task $k-1$. The current network for the new task is seen as the student network and the previous network containing the previous knowledge is regarded as the teacher network. However, the existing CL methods cannot be directly applicable to our problem as they are designed for a sequence of unimodal tasks [12, 13] or bimodal tasks of the same modality across tasks [14]. We pre-train audio encoders for a sequence of tasks, where different tasks may involve different modalities.

We implement three representative CL methods in knowledge distillation: LwF [12], CaSSLe [13], and Mod-X [14]. LwF was proposed for continual learning of classification tasks. As our goal is to build an audio encoder instead of a classifier, we distill the features for task k as

$$\mathcal{L}_{kd}(\theta_a^k) = \text{KLD}(g^{k-1}(f_a^k(X_a)), g^{k-1}(f_a^{k-1}(X_a))).$$

As for the self-supervised continual learning method of unimodal tasks, CaSSLe introduces an adapter p with parameters γ^k. We train the audio encoder so that the embeddings are easily adaptable by a simple linear function p with loss

$$\mathcal{L}_{kd}(\theta_a^k, \phi_a^k, \gamma^k) = \text{KLD}(p(g^k(f_a^k(X_a))), g^{k-1}(f_a^{k-1}(X_a))).$$

The final method Mod-X is designed for bimodal tasks of the same modality, language-vision. Mod-X defines knowledge through the cosine similarities between embeddings from audio and supervision encoders, and instructs the current encoder to follow the previous knowledge. For KD from an audio encoder trained for a multimodal task, we directly apply the Mod-X technique. For KD from an audio encoder trained for a unimodal task, since there is no encoder used for supervision (i.e., f_s), we utilize the supervision encoder of the current task and distill the cosine similarities between embeddings from the previous audio encoder and the current supervision encoder.

The final objective function we minimize is formulated as

$$\mathcal{L} = (1-\lambda)\mathcal{L}_c + \lambda\mathcal{L}_{kd}, \tag{2}$$

where $\mathcal{L}_{kd}$ is one of the KD losses defined above and λ is a hyperparameter that controls the importance of knowledge distillation. We set λ to 0.3 in our experiments.

3. EXPERIMENT DESIGN

Pre-training task sequences. We are interested in building an audio encoder for a sequence of multi-modal tasks including audio-visual

(AV) and audio-text (AT) data. In order to study the effect of continual pre-training of audio encoder on multi-modal data and its generalization ability to novel downstream tasks in the audio domain, we conduct our pre-training across two distinct sequences: A-AT and A-AV-AT. In these sequences, the initial task "A" refers to the pre-training step on audio-only data. We focus on the AT interactions for our final tasks with potential increasing applications.

Datasets for pre-training. We have three pre-training tasks: audio-only (A), audio-visual (AV), and audio-text (AT). The audio-only (A) pre-training uses supervised learning on the unimodal audio data AudioSet [2]. For AV pre-training, we use the audio-visual dataset VGGSound [15] collected from YouTube videos by following the training protocol in [4]. For AT pre-training, we use the same collection of datasets employed in the audio-text pre-training method [5]. The collection comprises of FSD50k [16], ClothoV2 [17], AudioCaps [18], and MACS [19] datasets.

Datasets for downstream tasks. After continually pre-training the audio encoders in each sequence, we evaluate the audio encoders on various downstream tasks. For audio-only tasks, we adopt the HEAR Benchmark [20]. We also add UrbanSound8K (US8k) [21] and AudioSet (balanced segment) [2]. We adopt the standard classification metrics provided in HEAR Benchmark and report the accuracy and mAP for the three additional tasks.

To evaluate cross-modal interactions, we use audio-visual event (AVE) [22] and VGGSound [15] test set for audio-visual tasks, and Clotho [17] for language-based audio retrieval, along with zero-shot classification from US8k [21] and ESC-50 [23] for audio-text tasks. For zero-shot classification, we report the accuracy, and for retrieval tasks, we report Recall@10.

Implemented methods. We implement the audio encoder pre-training methods CLAP [5] and Wav2CLIP [4] and the representative CL methods LwF [12], CaSSLe [13], and Mod-X [14]. CLAP was designed for learning AT task, while Wav2CLIP was for learning AV task. For details about the CL methods, consult Sec. 2.2. For comparison, we also implement a multi-task pre-training with WavCaps [24] (MT-WC), a curated subset from AudioSet augmented with captioning, resulting in a joint audio-visual-text data unlocking contrastive with both image and text modalities concurrently.

Pre-training details. At the initial task A, instead of training an audio encoder on audio data from scratch, we simply use the popular pre-trained audio encoder PANNs [1]. We use ResNet-38. In the subsequent multi-modal pre-training tasks (i.e., AV and AT), we integrate the audio encoder pre-training methods with the CL methods. For all tasks, we use Adam [25] optimization with the standard early stopping criteria.

For task AV, we implement Wav2CLIP both without and with the CL methods as Eq. 2. We use the CLIP image encoder for supervision, train with VGGSound with learning rate as 0.001, and utilize ReduceLROnPlateau scheduler.

For task AT, we train the audio encoder using the AT pre-training method CLAP both without and with the CL methods. We follow closely the training recipes from [26], with CLIP text encoder instead. Since CLIP already learned the joint embedding space of both image and text, this should help our model to generalize to unseen modalities. We train the audio encoder with learning rate of 0.0001 and utilize StepLR scheduler.

4. RESULTS AND DISCUSSION

After pre-training on the last task in each pre-training sequences A-AT and A-AV-AT, we evaluate the audio encoders on the downstream tasks. In the following discussion, for both sequences, we use **Wav2CLIP** and **CLAP** to indicate the audio encoders trained without any CL methods for task AV and AT, respectively. We use **LwF**, **CaSSLe**, and **Mod-X** to denote the audio encoders pre-trained with the three CL methods throughout the tasks, respectively.

Table 1. Model performances on audio-text downstream tasks after continual pre-training on the task sequence A-AT. The column T2A indicates text-based audio retrieval while the column A2T represents audio-based text retrieval.

	ECS-50	US8k	Clotho	
	Zero-shot	Zero-shot	T2A	A2T
CLAP	75.63	77.46	40.75	45.01
LwF	**78.17**	**78.95**	40.04	**47.21**
CaSSLe	74.70	75.36	**41.17**	43.70
Mod-X	63.18	70.82	34.16	32.63

After continual pre-training on the sequence A-AT. Tab. 1 gives the results of the audio encoders evaluated on the downstream tasks including zero-shot audio classification and audio-text retrieval. For zero-shot evaluation, we follow the protocol proposed in [3]. We consider CLAP as the target baseline as it is originally proposed for learning audio-text (AT) interactions. CLAP achieves 75.63 and 77.46 on the zero-shot classification data ESC-50 and US8k, respectively. In contrary, the audio encoder trained with the continual learning method LwF achieves 78.17 and 78.95 on the same data. The performance differences between LwF and CLAP in the two zero-shot evaluations are 2.54 and 1.49, respectively. The positive *forward transfer*, which measures the rate of performance improvement in the new task by continual learning methods compared to non-continual learning methods, in both data indicates successful knowledge accumulation through LwF. Similar observations can be made for audio-text retrieval tasks when comparing LwF to CLAP

The two CL methods, CaSSLe and Mod-X, are not as effective as LwF. This difference can be attributed to the fact that LwF's original application aligns more closely with the task sequence A-AT. LwF is designed for transferring knowledge of a classification model. The initial model which is transferred to task AT is pre-trained with the audio classification data AudioSet. However, CaSSLe is designed for self-supervised CL for unimodal tasks and Mod-X is designed for bimodal learning.

After continual pre-training on the sequence A-AV-AT. We evaluate the performances of the audio encoders for the downstream tasks involving all the modalities learned throughout the pre-training tasks. For all the methods except Wav2CLIP, we use the final audio encoders after learning the last task AT. Since Wav2CLIP is originally designed for learning AV interactions, the model cannot be trained for AT as training with AT implies changing the original method. Therefore, we fix the audio encoder after training the task AV and fine-tune only the adapter when learning AT.

Tab. 2 presents the evaluation results. First, consider the audio-text interactions: zero-shot classification and audio-text retrieval using Clotho. The audio encoders pre-trained using the CL methods CaSSLe and Mod-X significantly outperform both Wav2CLIP and CLAP. The improvements (i.e., forward transfers) by Mod-X from CLAP are 7.11 in ESC-50 and 2.96 in US8k. For the audio-text retrieval tasks, both CaSSLe and Mod-X outperforms the non-continual learning methods. Although Mod-X performs slightly lower than CaSSLe in A2T, on average over the two audio-text retrieval tasks, Mod-X achieves 46.28 while CaSSLe achieves 44.69. This demonstrates the effectiveness of Mod-X in distilling the knowledge of encoders trained with multi-modal tasks of different modalities despite the fact that it is originally designed for contin-

Table 2. The model performances on various downstream tasks including zero-shot, audio-text, and audio-visual interactions in A-AV-AT. Wav2CLIP is evaluated using the audio encoder after training the task AV while all the other methods are based on the audio encoder after the last task AT. The column I2A indicates image-based audio retrieval while the column A2I represents audio-based image retrieval.

	ESC-50	US8k	Clotho		AVE		VGGSound	
	Zero-shot	Zero-shot	T2A	A2T	I2A	A2I	I2A	A2I
MT-WC	0.48	7.53	0.75	1.05	43.28	47.76	3.37	4.97
Wav2CLIP	50.04	41.66	11.93	10.91	**65.42**	**66.42**	**11.53**	**13.52**
CLAP	72.58	75.06	40.69	42.87	37.44	33.46	3.16	2.61
LwF	73.50	74.84	41.77	44.40	39.55	38.18	3.00	2.75
CaSSLe	77.47	77.00	41.10	**48.28**	35.32	36.94	2.64	2.79
Mod-X	**79.69**	**78.02**	**45.80**	46.75	42.04	43.78	3.81	3.97

Table 3. Results after linear probing. Here, number1/number2 in the table correspond to the results of audio encoders after pre-training on the last task in the sequences A-AT and A-AV-AT, respectively. The column GZTAN M/S represents GZTAN Music/Speech data in HEAR Benchmark. The second last column Avg. indicates the average result across the seven experiments while the last column Full HEAR indicates the average result across 16 full HEAR Benchmark tasks, including the 7 tasks in the table. For PANNs, MT-WC, and Wav2CLIP, the results for A-AT and A-AV-AT are identical. This is because, in both sequences, PANNs is based on the initial pre-trained audio encoder, MT-WC is a multi-task method, and Wav2CLIP is based only on the sequence A-AV-AT.

	ESC-50	FSD50K	Gunshot	US8k	AudioSet	GZTAN M/S	CREMA-D	Avg.	Full HEAR
				A-AT/A-AV-AT					
PANNs	92.85	**60.89**	77.08	83.35	**47.84**	**100.0**	53.32	73.62	59.65
MT-WC	91.10	57.03	**89.58**	81.50	46.89	98.46	52.94	73.93	59.72
Wav2CLIP	91.82	59.66	82.29	81.98	46.83	98.08	51.68	73.19	59.74
CLAP	92.87/91.00	59.78/58.20	77.38/85.86	83.72/81.69	46.20/45.82	97.41/96.44	48.43/49.78	72.26/72.68	59.07/56.06
LwF	93.18/92.35	60.10/59.92	82.74/92.41	83.64/**84.27**	46.34/45.27	96.88/97.66	**54.43**/50.62	73.90/**74.64**	**63.73**/59.61
CaSSLe	92.42/92.80	58.34/58.93	82.84/82.14	82.16/84.00	45.87/46.48	98.46/96.92	52.70/51.92	73.26/73.31	62.98/61.56
Mod-X	90.20/**93.35**	55.01/59.87	77.18/82.14	82.56/84.04	45.14/46.09	96.11/97.30	46.42/49.68	70.37/73.21	53.95/60.44

ual learning of multi-modal tasks of the same modality. Note that Wav2CLIP performs very poorly since its audio encoder cannot be trained for AT. Surprisingly, MT-WC trained with three co-existing modalities concurrently performs even worse than Wav2CLIP, given that we use both CLIP image and text encoders, which are already projected in the joint space. We think that this might be due to that the supervisions provided from two modalities needs to be weighted, as currently one loss term is dominating.

For the downstream tasks of audio-visual interactions, all the methods experience performance drops compared to Wav2CLIP. We term this performance drops of the audio encoders in audio-visual downstream task after learning the pre-training task AT as *forgetting* or negative *backward transfer*. Notably, the audio encoder of Wav2CLIP has preserved the knowledge as it is fixed after pre-trained on the task AV. Only an adapter is trained to AT to perform audio-text downstream tasks. Mod-X shows relatively mild negative backward transfer compared to the other CL methods due to the positive transfer from AV to AT. All the CL methods achieve less than 40 in the audio-visual retrieval tasks of AVE while Mod-X achieves at least 42 from both tasks. We also observe that LwF is stronger than CaSSLe. This implies that the indirect distillation via additional projection function p in CaSSLe is good for forward transfer, but is in fact less effective in knowledge protection. Finally, although MT-WC shows strong performances in the audio-visual retrieval tasks, it requires the three modalities in the training data, which is expensive to acquire and less flexible in reality.

Audio-only downstream task with linear probing. We evaluate the audio encoders continually pre-trained on both A-AT and A-AV-AT on the downstream tasks involving only audio. Following the evaluation protocol in [3], we fix the audio encoder and fine-tune only a classifier for the audio classification data. The results are presented in Tab. 3. We report only partial and average of (unreported) full results from HEAR Benchmark due to constraint on space.

The audio encoders by CL methods except Mod-X in the sequence A-AT exhibit comparative performances to the average 73.62 of the initial audio encoder PANNs despite their additional capacity in handling multi-modal interactions. The non-CL methods Wav2CLIP and CLAP show slightly lower performances, which are 73.19 and 72.68, respectively, than PANNs. In contrast, the audio encoders by the CL methods, when pre-trained on the longer sequence A-AV-AT, achieve higher performances than the non-CL counterparts. This demonstrates the efficacy of CL methods in knowledge accumulation for general audio encoder.

Compare the average performances of CL methods between the two task sequences. Although Mod-X shows a great improvement in performance when it is pre-trained for more diverse tasks (i.e., A-AV-AT), the other CL methods do not experience much improvements. However, as we have seen previously in Tab. 2, the performance gains in audio-text interactions and the abilities in performing audio-visual downstream tasks facilitated through A-AV-AT demonstrates the advantage of continual pre-training over diverse tasks. Finally, although MT-WC is trained on dataset jointly satisfying the three modalities (audio-visual-text), it only achieves 73.93 on average, which is slightly lower than 94.64 of LwF.

5. CONCLUSION

This paper studied continual pre-training of audio encoders for multimodal tasks and evaluated the models using the standard benchmark data within the audio domain. The audio encoders, trained with continual learning techniques, are able to accumulate the knowledge across a series of multimodal interaction tasks. The models are evaluated on various downstream tasks covering all the learned modalities and exhibit superior performances over models trained for a single task or without any continual learning technique.

6. REFERENCES

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

[2] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*. IEEE, 2017, pp. 776–780.

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

[4] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello, "Wav2clip: Learning robust audio representations from clip," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4563–4567.

[5] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[6] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.

[7] Michael McCloskey and Neal J Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*, vol. 24, pp. 109–165. Elsevier, 1989.

[8] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars, "A continual learning survey: Defying forgetting in classification tasks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 7, pp. 3366–3385, 2021.

[9] Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu, "Continual pre-training of language models," in *The Eleventh International Conference on Learning Representations*, 2022.

[10] Tejas Srinivasan, Ting-Yun Chang, Leticia Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason, "Climb: A continual learning benchmark for vision-and-language tasks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 29440–29453, 2022.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[12] Zhizhong Li and Derek Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[13] Enrico Fini, Victor G Turrisi Da Costa, Xavier Alameda-Pineda, Elisa Ricci, Karteek Alahari, and Julien Mairal, "Self-supervised models are continual learners," in *CVPR*, 2022, pp. 9621–9630.

[14] Zixuan Ni, Longhui Wei, Siliang Tang, Yueting Zhuang, and Qi Tian, "Continual vision-language representaion learning with off-diagonal information," *arXiv preprint arXiv:2305.07437*, 2023.

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

[16] 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*, vol. 30, pp. 829–852, 2021.

[17] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[18] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "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)*, 2019, pp. 119–132.

[19] 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)*. IEEE, 2021, pp. 76–80.

[20] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al., "Hear: Holistic evaluation of audio representations," in *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 125–145.

[21] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[22] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu, "Audio-visual event localization in unconstrained videos," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 247–263.

[23] Karol J Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[24] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *arXiv preprint arXiv:2303.17395*, 2023.

[25] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] Ho-Hsiang Wu, Oriol Nieto, Juan Pablo Bello, and Justin Salomon, "Audio-text models do not yet leverage natural language," in *ICASSP*. IEEE, 2023, pp. 1–5.