Learning Audio-Video Language Representations

by

Andrew Rouditchenko

S.B. Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2019

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
Department of Electrical Engineering and Computer Science
May 19, 2021

Certified by .....
James Glass
Senior Research Scientist
Thesis Supervisor

Certified by .....
David Harwath
Assistant Professor
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Learning Audio-Video Language Representations

by

Andrew Rouditchenko

Submitted to the Department of Electrical Engineering and Computer Science
on May 19, 2021, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Automatic speech recognition has seen recent advancements powered by machine learning, but it is still only available for a small fraction of the more than 7,000 languages spoken worldwide due to the reliance on manually annotated speech data. Unlabeled multi-modal data, such as videos, are now increasingly available in many different languages and provide opportunities to scale speech technologies. In this thesis, we introduce models and datasets for learning visually grounded spoken language from raw audio in videos. We propose a self-supervised audio-video model that learns from the English narration naturally present in instructional videos to relate spoken words and sounds to visual content. Our model can recognize spoken words and natural sounds in audio queries to retrieve relevant visual clips, supporting its application to video search directly using audio and spoken queries, without needing to transcribe speech to text. We further demonstrate that our model can learn multilingual audio-video representations and can successfully perform retrieval on Japanese videos. Since our approach only requires audio-visual data without transcripts, we believe it is a promising direction to enable novel speech processing tools.

Thesis Supervisor: James Glass
Title: Senior Research Scientist

Thesis Supervisor: David Harwath
Title: Assistant Professor

Acknowledgments

First and foremost, I would like to thank my advisors Dr. Jim Glass and Professor Dave Harwath. I am grateful to Jim for providing me the opportunity to join the Spoken Language Systems group. I thank Dave and Jim for their support, mentorship, advice, and patience. They inspired me to pursue exciting research directions while encouraging me to grow as an independent researcher.

I would like to thank the talented students and researchers that I worked with. I thank Angie Boggust for being a dedicated and inspiring collaborator. I thank Ian Palmer for allowing me to participate in his research project. I thank the IBM researchers that I worked closely with, including Dr. Sam Thomas, Dr. Dhiraj Joshi, Dr. Hilde Kuehne, Dr. Kartik Audhkhasi, Brian Chen, and Kevin Duarte. I also thank the other researchers, advisors, and technical staff at MIT and IBM that supported my research.

Being a member of the Spoken Language Systems group has been an incredible experience. I am grateful to all students, postdoctoral associates, research scientists, and lab members for creating a supportive and collaborative research environment. I would especially like to thank my fellow MEng students Moin Nadeem, Kunal Tangri, Rami Manna, Dotun Oseni-Adegbite, and Emmanuel Azuh, as well as undergraduate students Assel Ismoldayeva, Michael Wu, and Chris Song, whom I spent time with in the lab discussing experiments and in classes like NLP. I also thank administrative staff Marcia Davidson.

I am thankful that my research funding was supported by the MIT-IBM Watson AI lab and by an EECS Teaching Assistanship. I thank Professor Stefanie Jegelka for selecting me as a teaching assistant for an exciting machine learning course.

I would also like to thank the faculty and researchers that I worked with as an undergraduate. I thank Professor Josh McDermott and Professor Antonio Torralba for providing me opportunities to work with them on several projects and to establish my research foundation. I was fortunate to work with their students, postdoctoral associates, and other lab members, including Dr. Hang Zhao, Maddie Cusimano, Dr.

James Traer, and Dr. Chuang Gan. I also thank Dr. Tyler Lee for advising me during an internship at Intel.

I am blessed to have a loving family, and I thank them for their support and sacrifice. I am grateful for all of the friends that I have made throughout the years and for their support during challenging times. Finally, I give thanks and praise to God.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

While technologies like Automatic Speech Recognition (ASR) and Machine Translation (MT) enable us to interact better with computers and each other, they are currently only available for less than 2% of the world's more than 7,000 spoken languages, in part due to the large amount of manually labelled data required for each language [63]. Recently, researchers have proposed models that can instead learn to recognize words from raw audio by associating them to semantically related images [15,28–30,37,46,74]. By training models to retrieve images from associated spoken captions, they learn to identify words in speech and objects in images without supervised speech recognition or object detection. These models are appealing because they are trained only on images and spoken audio captions of images, without requiring conventional text transcripts. However, these methods require the collection of recorded spoken captions, limiting their scalability to other languages and visual contexts. Further, use of still images precludes application to real-world scenarios where multiple speakers and visual actions often occur.

Videos instead provide a natural source of paired visual and audio data that does not require manual annotation and exists publicly in large quantities. Thus, self-supervised audio-video models [6–8, 42, 60, 66, 86] have been applied to cross-modal tasks focused on identifying non-speech sounds and localizing the objects that produced them. In our work, we instead focus on relating spoken words to visual entities in videos such as objects and actions, which is a more challenging task than

sound localization since human speech is more semantically complex and the objects of interest do not produce the sound. Towards this goal, we leverage instructional videos which provide opportunity to learn semantic relationships between raw speech and visual content given the rich narration naturally present in them. While audio-video models typically assume sound-making objects are visible on screen, in instructional video there may be misalignment between when a speaker describes an object and when it is on screen. Thus, while instructional videos provide opportunity to learn semantic relationships between raw speech and visual content, they are a noisy and challenging source of data.

In this thesis, we explore several approaches for learning visually grounded spoken language from the raw audio in instructional videos. We propose a self-supervised audio-video model that learns from the narration naturally present in instructional videos to relate spoken words and sounds to visual content. Our model only requires audio-visual data without transcripts, and can perform video retrieval directly from input audio queries containing spoken words and natural sounds. We further propose a tri-modal that jointly processes raw audio, video, and text captions from videos to learn a multi-modal semantic embedding space. To understand our model's multilingual capabilities, we collected a new dataset of Japanese cooking videos. We demonstrate that our audio-video model can learn multilingual audio-video representations and can successfully perform retrieval on the Japanese videos.

1.1 Related Work

In this section, we review the related work relevant to the entire thesis.

1.1.1 Learning Visually-Grounded Speech

The task of matching spoken audio captions to semantically relevant images was introduced in the effort to build models that learn language from raw audio and visual semantic supervision [25, 30, 74]. Models are typically trained to learn an audio-visual embedding space where true image-caption pairs are similar to each other,

while non-matching pairs are far apart. Over the years, researchers have proposed modeling improvements with more complex image encoders, audio encoders, and loss functions [15, 27–29, 37, 46, 55, 68, 73]. In terms of training data, Harwath et al. [28–30] collected 400k spoken audio captions of images in the Places205 [87] dataset in English from 2,683 speakers, which is one of the largest spoken caption datasets. Other work has proposed synthetic speech captions as training data, which are less natural [15, 31, 37]. The models have been explored for other tasks such as speech retrieval given spoken queries or text captions [38, 39], discovering word-like speech units [26, 79, 80], and for other data such as handwritten digits and spoken captions [19, 34, 44]. For a recent survey of visually grounded models of spoken language, see Chrupała [14]. We instead use videos naturally present on the internet as the primary source of training data, which are available in English and in other languages. While we focus on the spoken narration naturally present in instructional videos, researchers have collected spoken captions for videos [53, 58] in concurrent work.

1.1.2 Self-Supervised Audio-Video Learning

Self-supervised audio-video learning has been explored in recent years to learn representations of objects and sounds without manually labelled data. Some works propose proxy tasks to learn representations for downstream tasks such as classification [4, 6, 8, 35, 42, 60, 61]. Other approaches use self-supervised learning for audio-video applications, such as audio-visual source separation [20, 66, 86] and spatial audio generation [21, 54, 84]. The most relevant works are those that apply audio-video models for cross-modal retrieval tasks. Tian et al. [77] proposed an audio-video model for localizing segments of audio within video clips. They train it on a video dataset of audio-visual events, ensuring that sound-making object are visible for at least 2 seconds in each clip. Surís et al. [72] train an audio-video model for retrieval on general YouTube videos, and incorporate a classification loss on the video class label. We build upon these works by learning from unlabeled instructional videos. Arandjelović and Zisserman [7] employ self-supervision between the audio and visual streams in video to relate objects with the sounds they make. They train their model

for binary classification of true audio and video pairs versus mismatched ones and apply their model for audio-video retrieval. In our work, we instead use audio-video self-supervision to relate objects to the speech that describes them and directly train our model for audio-video retrieval.

1.1.3 Multi-Modal Learning from Instructional Videos

The recent influx of instructional video datasets such as How2 [67], Inria Instructional Videos [2], COIN [75], CrossTask [89], YouCook2 [88], Mining YouTube [43], and HowTo100M [50] has inspired a variety of methods for semi-supervised text-video modelling. These works focus on learning a joint multi-modal embedding space between text and video, and typically do not incorporate the audio signal. Methods that do incorporate audio [3, 33, 45, 49, 52, 82, 85] still require text captions and do not learn from the raw videos alone. To create text captions, some methods rely on humans to generate textual descriptions of the visual scene [88]. Unlike raw audio which can be noisy and nondescript, human-generated textual captions provide a clean signal that is visually salient; however, collecting these descriptions is time-consuming, making it infeasible for large datasets. To reduce the need for annotation, other methods rely on ASR transcripts to provide text representative of the speech in videos [47, 50, 67, 70, 71]. ASR transcripts provide a cleaned version of the audio that no longer contains salient non-speech sounds. We build upon these works by learning a joint embedding space directly between video and the audio naturally present in videos, and showing that our method can also incorporate ASR text and annotated text captions when available.

1.2 Thesis Outline

This thesis is organized as follows. In Chapter 2, we describe all of the video datasets used in this thesis. In Chapter 3, we introduce our self-supervised audio-video model. We study the model through several experiments, ablations studies, and qualitative results. In Chapter 4, we introduce our tri-modal model that learns from audio, video, and text, and show quantitative results. In Chapter 5, we explore the multilingual

capabilities of our audio-video model and apply it on videos in Japanese and images and spoken captions in Japanese and Hindi. We also introduce a new dataset of Japanese cooking videos. Finally, in Chapter 6, we provide a conclusion, ideas for future work, and our parting thoughts.

1.3 Bibliographic Note

Content primarily in Chapters 3 and 4 has appeared previously in Rouditchenko et al. [65]. The work in this thesis was performed in collaboration with Angie Boggust, Sam Thomas, Dhiraj Joshi, and Brian Chen.

Chapter 2

Video Datasets

In this chapter, we provide an overview of the video datasets used in this thesis. We primarily use the HowTo100M dataset [50], which contains over 1 million videos, to train our models. We use the YouCook2 [88], CrossTask [89], MSR-VTT [83], and LSMDC [64] datasets to fine-tune and evaluate our models. These datasets are smaller than HowTo100M and typically contain only several thousand videos. HowTo100M, YouCook2, and CrossTask are instructional video datasets, while MSR-VTT contains general videos and LSMDC contains movies. We also propose the YouCook-Japanese video dataset, which contains instructional cooking videos in Japanese.

2.1 Video Downloading

Many of the experiments in this thesis are conducted on video datasets consisting of videos from YouTube. These video datasets are typically distributed via lists of URLs to comply with YouTube's terms of service, and each research group must scrape the videos independently. Over time, the number of videos available can shrink, making it challenging to reproduce and compare the results from different researchers. Therefore, we provide details about the number of videos we were able to download, and which experimental splits we used.

2.2 Dataset Details

2.2.1 HowTo100M

The HowTo100M dataset [50] contains instructional YouTube videos from domains such as *home and garden*, *computers and electronics* and *food and entertaining*. We downloaded the HowTo100M dataset from YouTube between Dec. 2019 - Mar. 2020. The original dataset contains 1,238,792 videos. At the time of download 1,166,089 videos were available on YouTube (72,703 less than the original dataset), which we used as our training set.

2.2.2 YouCook2

The YouCook2 dataset [88] consists of 2,000 instructional cooking videos from YouTube. The videos were separated into a 67-23-10 training-validation-testing split and categorized by humans into one of 89 recipe types (eg., *spaghetti and meatballs*). Videos were segmented by human annotators into clips representing recipe steps, and each clip was annotated with a text summary of the recipe step. Following Miech et al. [50], we use 9,586 training clips and 3,350 validation clips due to the unavailability of some videos on YouTube.

2.2.3 CrossTask

The CrossTask dataset [89] consists of 2,750 instructional videos from YouTube with 18 primary tasks and 65 related tasks. Each task is defined as list of steps, such as "*remove cap*" and "*spread mixture*". Each video is associated with one task and contains a subset of steps from the task. 20 videos from each of the 18 primary tasks are designated as the validation set (360 videos total), and the remaining videos are designated as the training set. The videos in the validation set were segmented into clips for each step by human annotators, while the videos in the training set were segmented into clips for each step automatically based on the ASR transcripts. The training set contains 17,840 clips while the validation set contains 2,819 clips.

2.2.4 MSR-VTT

The MSR-VTT [83] dataset consists of YouTube videos from categories such as *music* and *sports* that are not necessarily instructional. Videos were segmented into video clips by human annotators and annotated with 20 natural language sentences each. At the time of download, 5,722 videos with audio were available. We train our model on 6,783 training clips and evaluate on 968 audio containing test clips of the 1,000 test clips used in prior work [50, 85]. For a fair comparison, we count the 32 missing test clips without audio as mistakes in our retrieval calculations. We note that there are several other experimental splits [45].

2.2.5 LSDMC

The LSMDC dataset [64] consists of movies with audio description (AD) — audio descriptions of movie scenes for viewers with visual impairments. The movies were split into video clips corresponding to scenes with AD narration, and each clip is annotated with the text transcript of the AD narration. Following Miech et al. [50], we use 101,079 training clips and 1,000 testing clips. We use the audio from the original movie clips; however, the audio is often silent because AD narration is inserted at breaks in dialogue. The recorded AD narrations were not available.

2.2.6 YouCook-Japanese

The YouCook-Japanese dataset contains 1,174 Japanese cooking videos. We propose the dataset in Chapter 5 and describe the dataset collection details. The training set contains 737 videos, the validation set contains 224 videos, and the evaluation set contains 213 videos. We segmented the videos into clips containing speech, resulting in 10k clips for training, 3k clips for validation, and 3k clips for evaluation. We report results on the evaluation set and encourage other researchers to tune parameters only on the validation set.

2.3 Chapter Summary

In this chapter, we described the video datasets used throughout this thesis. We explained the number of videos that we were able to download and which experimental splits we used to make it easier to compare with our experiments.

Chapter 3

AVLnet: Audio-Video Language Network

The predominant approaches for learning from instructional videos rely on text annotations. In manually supervised cases, instructional video clips have been segmented and captioned by annotators [88]. To reduce the amount of supervision, some methods employ Automatic Speech Recognition (ASR) systems to generate text transcripts of the speech within the videos [50]. ASR transcripts bypass many of the challenges of learning from raw speech, which may contain background noise, variation across speakers, or multiple speakers. While speech is a continuous signal, ASR processes the speech into discrete words limited to a certain vocabulary. However, ASR transcripts are only available for less than 2% of the world's more than 7,000 spoken languages, so while models trained using ASR transcripts perform well in common languages such as English, they are inapplicable to many of the languages spoken across the globe. Further, ASR can also be errorful, especially when confronted with background sounds, reverberation, accents, and new vocabulary. For these reasons, we need models that can learn from the raw audio and visual channels in videos without any additional annotation or ASR transcription.

In response, we propose the Audio-Video Language Network (AVLnet) and a self-supervised framework to learn visually grounded language from raw video input. We circumvent the need for spoken or textual annotations by learning directly from the

raw audio channel in video clips. Our model consists of audio and video branches that extract local video clip features and pool them into single feature vectors representing the content in each modality. We apply non-linear feature gating [48] enabling our model to re-calibrate the feature activations before the final output embeddings. To train our model on the noisy audio signal in instructional videos, we utilize the Masked Margin Softmax (MMS) loss [37] to simulate audio and visual retrieval and robustly train against a large number of negative samples. This results in an audio-video embedding space that colocates semantically similar audio and visual inputs and can successfully be used for downstream retrieval tasks.

We train AVLnet on HowTo100M [50], a large-scale instructional video dataset. Instead of defining video clips at ASR boundaries, we train our model on randomly segmented clips, reducing the need for supervision. Despite training on unlabeled videos, our model achieves state-of-the-art retrieval results on speech-image pairs in the Places Audio Caption dataset [28]. We propose video retrieval tasks on three video datasets, YouCook2 [88], CrossTask [89], and MSR-VTT [83], and achieve state-of-the-art results over previous models. We further show how our model leverages audio cues from both speech and natural sounds for retrieval and semantically relates the audio and visual modalities to learn audio-visual concepts. Some of the results in this chapter were presented in Rouditchenko et al. [65].

3.1 Technical Approach

3.1.1 Audio-Video Models

The AVLnet architecture, shown in Figure 3-1, consists of parallel visual and audio branches that extract features at a local level and then pool them into visual and audio feature vectors representing the overall content within each modality. This procedure provides flexibility by allowing the model to handle variable length video clips, which is especially useful during inference where clip boundaries are determined by human annotators and can vary drastically in length. The visual branch consists

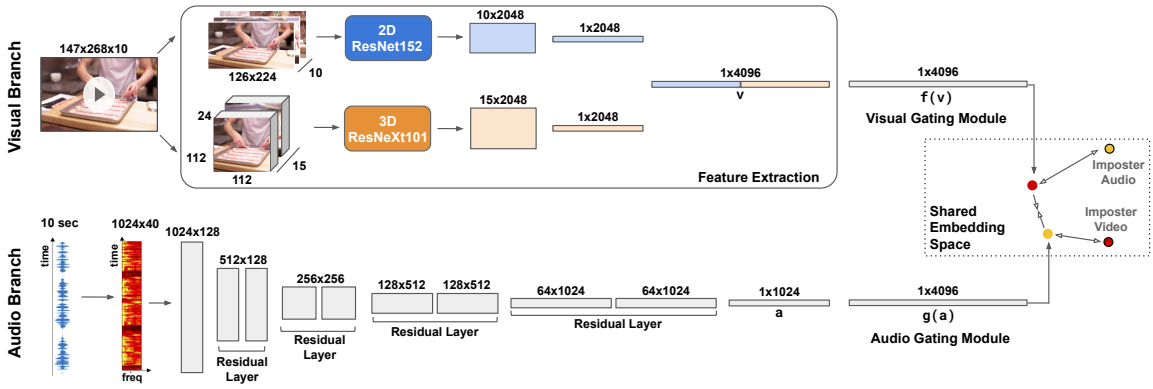


Figure 3-1: The Audio-Video Language Network (AVLnet) model consists of video and audio branches, non-linear feature gating, and an audio-video embedding space. The model is trained through self-supervision and applied to image and video retrieval tasks.

of a 2D and 3D CNN feature extraction pipeline. From each video clip, we compute 2D image features to obtain 1 feature per second using a ResNet-152 model [32] pretrained on ImageNet [17] and 3D video features to obtain 1.5 features per second using a ResNeXt-101 model [23] pretrained on Kinetics [11]. Each of the CNN outputs are temporally max-pooled to produce two 2048-dimensional feature vectors, which are then concatenated into a 4096-dimensional feature vector $\mathbf{v}$. The audio branch consists of a trainable CNN with residual layers [28] to process the raw audio in videos. The model takes in audio spectrograms and outputs a temporal feature map, which is temporally mean-pooled to obtain a 1024-dimensional feature vector $\mathbf{a}$. In contrast to text-video models that require pretrained word embeddings to process speech transcripts [47, 50], our audio model is not pretrained, so it can be applied to videos in any language, including those for which ASR is not available.

3.1.2 Audio-Video Gated Embeddings

After the visual feature vector $\mathbf{v}$ and audio feature vector $\mathbf{a}$ are extracted, we learn a projection of both vectors into a shared embedding space. While this could be achieved with a linear projection, we apply non-linear feature gating [48] which allows the model to re-calibrate each dimension based on its learned importance and encourages the model to activate dimensions in unison across both modalities. Non-linear gating

is defined as:

$$f(\mathbf{v}) = (W_1^v \mathbf{v} + b_1^v) \circ \sigma(W_2^v(W_1^v \mathbf{v} + b_1^v) + b_2^v) \tag{3.1}$$

$$g(\mathbf{a}) = (W_1^a \mathbf{a} + b_1^a) \circ \sigma(W_2^a(W_1^a \mathbf{a} + b_1^a) + b_2^a) \tag{3.2}$$

where $f(\mathbf{v})$ and $g(\mathbf{a})$ are the output 4096-dimensional embedding vectors, $W_1^a, W_2^a, W_1^v, W_2^v$ matrices and $b_1^a, b_2^a, b_1^v, b_2^v$ vectors are learnable parameters, $\circ$ denotes element-wise multiplication, and σ is an element-wise sigmoid activation.

3.1.3 Contrastive Loss for Audio-Video Retrieval

Due to the self-supervised nature of AVLnet, we use the Masked Margin Softmax (MMS) loss [37], a contrastive loss function that simulates retrieval within each batch. The MMS loss trains the model to discriminate between the true audio-visual embedding pairs ($\mathbf{a}_i$, $\mathbf{v}_i$), and imposter pairs ($\mathbf{a}_i$, $\mathbf{v}_j^{\text{imp}}$) and ($\mathbf{a}_j^{\text{imp}}$, $\mathbf{v}_i$). The indices (i, j) indicate the index of the video clip in the batch. Unlike the triplet loss used in prior unsupervised audio-image modeling [28] that samples imposter pairs randomly or via negative mining, the MMS loss enables comparisons of positives with a wider range of negatives. While the original MMS loss includes a masking component to handle multiple ground truth audio captions paired with each visual sample, we exclude the masking since it is inapplicable to our scenario where each visual clip contains only one ground truth audio pair. The loss $\mathcal{L}_{MMS}$ is defined as follows:

$$\mathcal{L}_{MMS}(f(\mathbf{v}), g(\mathbf{a})) = L(f(\mathbf{v}), g(\mathbf{a})) + L(g(\mathbf{a}), f(\mathbf{v})) \tag{3.3}$$

Where $f(\mathbf{v})$ and $g(\mathbf{a})$ are the gated embeddings, and the function L defined as:

$$L(\mathbf{x}, \mathbf{y}) = -\frac{1}{B} \sum_{i=1}^{B} \left(\log \frac{e^{\mathbf{x}_i \cdot \mathbf{y}_i - \delta}}{e^{\mathbf{x}_i \cdot \mathbf{y}_i - \delta} + \sum\limits_{\substack{j=1 \\ j \neq i}}^{B} e^{\mathbf{x}_i \cdot \mathbf{y}_j^{\text{imp}}}} \right) \tag{3.4}$$

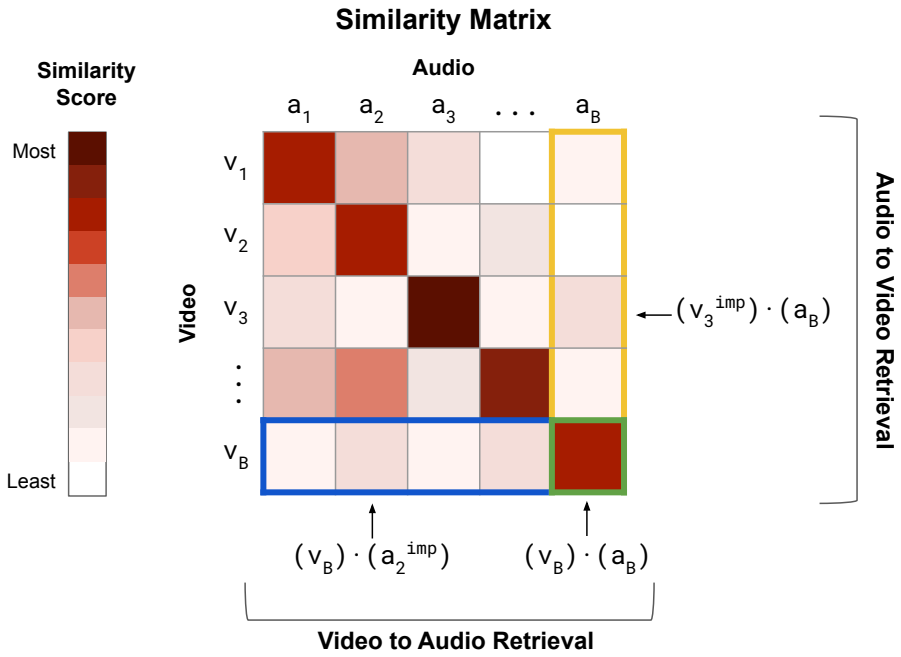


Figure 3-2: The MMS loss maximizes the similarity of the true audio-visual pair $(\mathbf{a}_i, \mathbf{v}_i)$ shown in green. It also minimizes the similarity of $\mathbf{a}_i$ paired with imposter videos $\mathbf{v}_j^{\text{imp}}$ (in yellow) and $\mathbf{v}_i$ paired with imposter audios $\mathbf{a}_j^{\text{imp}}$ (in blue).

The MMS loss $\mathcal{L}_{MMS}$ can be seen as the sum of two applications of InfoNCE [59] (with a margin), the first where the visual input is fixed and audio samples are retrieved, and the second where the audio input is fixed and visual samples are retrieved. However, whereas negatives are sampled from within the same audio sample for InfoNCE [59], we use audio and video samples from both within the same video and from others as negatives as this has been empirically shown to improve performance for text-video approaches [50]. During training, we use a batch of N videos and sample M clips per video, resulting in effective batch of $B = NM$ video clips, where $B - 1$ samples are used as negative for each ground truth pair. An illustration of the loss is provided in Figure 3-2.

3.2 System Description

3.2.1 Video Clip Sampling

Given a corpus of unlabeled instructional videos, we generate training samples without supervision by randomly segmenting each video into M clips of length t (which may overlap) to obtain a corpus of clips. The number of clips per video, M, is the same for all videos irregardless of video length. This procedure allows us to sample clips without supervised annotation (i.e., segmenting based on ASR transcripts.) As a result, it is applicable to instructional videos in languages not supported by ASR, and it enables greater flexibility to vary the number and length of clips in the resulting dataset. Although unsupervised clip selection may result in silent or non-salient clips, our experimental results in Section 3.3.4 show our model performs comparably whether trained on randomly sampled clips or on clips determined by ASR boundaries.

3.2.2 Training

We train AVLnet on the instructional YouTube videos from the HowTo100M [50] dataset. The HowTo100M dataset provides video clip segmentations according to time intervals of each video's ASR transcript and captions each clip with the text from its transcript. However, to reduce the amount of supervision in our method, we train AVLnet on the video and audio from randomly segmented clips.

3.2.3 Implementation Details

In the AVLnet audio branch, the audio input is represented as a log Mel filterbank spectrogram. We use a 16 kHz sampling rate, 25 ms Hamming window, 10 ms window stride, and 40 Mel filter bands. For the 2D and 3D visual feature extractors, we use the pretrained models from PyTorch [62] and feature extraction implementation provided by Miech et al. [50]. When training AVLnet, we do not update the weights of the 2D and 3D feature extractors due to GPU memory limitations. We use a batch of $N = 128$ videos, and sample $M = 32$ clips per video, each $t = 10$ seconds long.

We minimize the MMS loss with Adam [41] using a learning rate of 1e−3 and fix the margin hyperparameter $\delta = 0.001$. We train each model on 2 V100 GPUs for 30 epochs, which takes approximately 2 days. For fine-tuning on the variable length video clips in the YouCook2, CrossTask, and MSR-VTT datasets, we crop or pad the audio up to 50s in YouCook2 and CrossTask, and 30s for audio in MSR-VTT.

3.3 Experiments

3.3.1 Downstream Tasks

Image Retrieval. AVLnet is designed to learn from freely-available, uncurated, and noisy instructional videos that exist in the real world, as opposed to manually collected and annotated spoken caption datasets. Nonetheless, both data sources are focused on descriptive speech of visual scenes, so it could be expected that learning from instructional videos would provide a relevant initialization for learning from images and spoken captions. Therefore, we train AVLnet on HowTo100M videos and fine-tune it on images and spoken captions. We evaluate the performance on audio to image and image to audio retrieval tasks. We compute the similarity between a spoken caption and image as the dot product of their embedding vectors.

Video Retrieval. We evaluate our model on video clip retrieval (audio to video) and language retrieval (video to audio) tasks, which measure how well the model can retrieve content in one modality based on a query in the other modality. This follows prior work on audio to video retrieval on YouCook2 [10]. This procedure tests our model's capability for video search directly using audio and spoken queries, without needing to transcribe speech in the query to text. We report results in the zero-shot, fine-tuned, and no-pretraining settings. We compute the similarity between an audio sample and visual sample as the dot product of their embedding vectors. The ground truth pairing between the visual sample and audio sample of a video clip are used as the true labels. In other words, for audio to video retrieval, the result for a query audio sample is correct when its corresponding visual sample is within the top N most

Table 3.1: Image retrieval on the Places Audio Caption dataset.

Method	Audio to Image			Image to Audio			
	R@1	R@5	R@10	R@1	R@5	R@10	Avg R@10
Random	0.1	0.5	1.0	0.1	0.5	1.0	1.0
Harwath et al. [29]	27.6	58.4	71.6	21.8	55.1	69.0	75.3
Harwath et al. [27]	-	-	-	-	-	-	79.4
Ours, AVLnet	**44.8**	**79.9**	**86.4**	**42.8**	**76.2**	**84.8**	**85.6**

similar visual samples.

3.3.2 Datasets

Images and Spoken Captions: We fine-tune and evaluate our model on the Places Audio Caption dataset [28]. The dataset contains 400k images from the Places205 dataset [87] paired with 1,000 hours of unscripted spoken captions. Following prior work [27, 29], we evaluate performance on the validation set of 1,000 image and spoken caption pairs. To pre-process the data, we follow the procedure in Harwath et al. [28]. Images are cropped to 224 by 224 pixels, while audio samples are sampled at 16 kHz, padded or cropped to 20s, and processed into spectrograms using a 25 ms Hamming window, 10 ms window stride, and 40 Mel filter bands.

Videos: We fine-tune and evaluate our model on two instructional video datasets: YouCook2 [88] and CrossTask [89]. While YouCook2 contains cooking videos, CrossTask contains a wider range of instructional videos. We also fine-tune and evaluate on MSR-VTT [83] which contains general YouTube videos. We use the human-annotated clips defined in each dataset: 9,586 train clips and 3,350 validation clips for YouCook2, 17,840 train clips and 2,819 validation clips for CrossTask, and 6,783 train clips and 968 test clips for MSR-VTT. The full dataset details are in Chapter 2.

3.3.3 Comparison to State-of-the-Art

Image Retrieval. In this experiment, we train AVLnet on HowTo100M using the 2D CNN features, so the model can be fine-tuned on the downstream images without any modifications. During fine-tuning on Places, we update the weights of the visual encoder instead of keeping it frozen as in training on HowTo100M. In Table 3.1, we

Table 3.2: Video retrieval results. Models trained on: 1. target dataset only; 2. HowTo100M only; 3. HowTo100M and target dataset. A→V = Video Clip Retrieval; V→A = Language Retrieval.

Method	YouCook2						CrossTask						MSR-VTT					
	A→V			V→A			A→V			V→A			A→V			V→A		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.03	0.15	0.3	0.03	0.15	0.3	0.04	0.18	0.35	0.04	0.18	0.35	0.1	0.5	1.0	0.1	0.5	1.0
1. Boggust et al. [10]	0.5	2.1	3.4	0.6	2.2	3.7	0.4	1.9	3.7	0.6	2.8	5.7	1.0	3.8	7.1	1.8	4.5	8.1
1. Arandjelović et al. [7]	0.3	1.9	3.3	0.5	2.0	3.7	0.4	**2.5**	4.1	**0.7**	4.5	9.8	**1.3**	4.3	8.2	0.3	2.5	6.6
1. **Ours, AVLnet**	**0.7**	**2.3**	**3.9**	**0.8**	**3.0**	**4.9**	**0.7**	2.4	**4.6**	0.5	**5.2**	**11.0**	0.9	**5.0**	**9.0**	**0.8**	**4.6**	**8.1**
2. Boggust et al. [10]	6.8	22.4	31.8	7.9	23.8	32.3	5.5	18.7	28.3	5.2	18.2	27.6	7.6	21.1	28.3	9.3	20.7	28.8
2. Arandjelović et al. [7]	13.6	31.7	41.8	12.9	33.0	42.4	7.3	19.5	27.2	7.5	19.4	27.2	12.6	26.3	33.7	11.9	25.9	34.7
2. **Ours, AVLnet**	**27.4**	**51.6**	**61.5**	**27.3**	**51.2**	**60.8**	**11.9**	**29.4**	**37.9**	**10.8**	**27.3**	**35.7**	**17.8**	**35.5**	**43.6**	**17.2**	**26.6**	**46.6**
3. Boggust et al. [10]	8.5	26.9	38.5	9.9	30.0	41.1	6.6	20.8	31.2	6.0	21.5	31.4	10.3	27.6	35.9	11.8	29.0	38.6
3. Arandjelović et al. [7]	17.4	39.7	51.5	19.0	43.4	53.9	9.5	25.8	36.6	11.1	28.9	40.7	16.2	32.2	42.9	15.4	34.9	45.0
3. **Ours, AVLnet**	**30.7**	**57.7**	**67.4**	**33.0**	**58.9**	**68.4**	**13.8**	**34.5**	**44.8**	**15.5**	**37.0**	**52.9**	**20.1**	**40.0**	**49.6**	**22.0**	**41.4**	**50.3**

compare prior state-of-the-art models trained only on Places [27,29] to AVLnet trained on HowTo100M and fine-tuned on Places. Our method achieves large gains over prior results, showing AVLnet learns a relevant initialization that transfers to the images and captions in Places.

Video Retrieval. We compare AVLnet to prior audio-video models proposed for video clip retrieval in non-instructional contexts. The model from Boggust et al. [10] only uses the center image frame from each video clip during training and inference. The model from Arandjelović et al. [7] is trained with a binary cross-entropy loss. Compared with AVLnet, it does not use non-linear gating and uses an embedding dimension of 128 instead of 4096. For fair comparison, we train all models on HowTo100M, and, since the prior models each use different visual and audio pipelines, we change them to work with our 2D/3D visual features and deep audio network.

Table 3.2 shows the results for audio to video retrieval and video to audio retrieval on YouCook2, CrossTask, and MSR-VTT in the zero-shot, fine-tuned, and no-pretraining settings. Our method outperforms the baseline models, especially in the zero-shot and fine-tuned settings. We also note that training on HowTo100M significantly improves the performance compared with training only on the target dataset, including on MSR-VTT which contains general YouTube videos.

3.3.4 Ablation Studies

We evaluate our design choices via ablation studies comparing each model's video clip retrieval on YouCook2 and CrossTask (Table 3.3). Given the computational requirements of HowTo100M, we train for 15 epochs with a batch size of 64. First, we compare projections and find non-linear feature gating outperforms both linear and non-linear projection heads [12]. Next, we evaluate loss functions. MMS [37] outperforms MIL-NCE [47], Binary Cross Entropy [6], Max-Margin Ranking [50], and InfoNCE [59]. For MIL-NCE, we defined neighbors as the nearest non-overlapping 10s clips. For InfoNCE, we used negative samples from both within the same video and others. MIL-NCE, initially proposed for text-video models, performs the worst, suggesting loss functions designed for text may not transfer well to audio.

We also find AVLnet performs better when trained on both 2D and 3D visual features. AVLnet performs similarly when trained on random vs. ASR-defined clips, indicating our approach reduces supervision while maintaining performance.

Finally, we assess HowTo100M clip length and find it has a large effect on retrieval performance. While we propose 10s, speech-image models [28, 29] use spoken captions that are typically 20s, and text-video models [47] use ASR-defined clips that average 4s. We find 10s outperforms 2.5, 5, and 20s, suggesting short clips may not contain speech relevant to the visuals, whereas long clips may contain too many audio-visual concepts.

3.3.5 Retrieving Speech versus Non-Speech Sounds

To identify the audio cues AVLnet uses for retrieval, we investigate performance in the absence and presence of speech. We create two distinct evaluation sets: one containing videos without speech and one with speech. To assign videos to each set, we identify the number of words in each YouCook2 validation video clip via ASR [1]. We create a new evaluation set, Sounds-241, containing the 241 clips without a detected word. We randomly sample 241 clips with at least one word detected to create another evaluation set: Speech-241. AVLnet achieves higher retrieval performance on Speech-

Table 3.3: AVLnet ablation study video clip retrieval (R@10). YC=YouCook2; CT=CrossTask; ZS=zero-shot; FT=fine-tune.

Study	Configuration	YC-ZS	YC-FT	CT-ZS	CT-FT
Projection Heads	Linear	44.2	53.0	28.4	35.7
	Non-Linear	47.8	57.6	30.6	38.4
	Gating	**54.3**	**63.0**	**33.0**	**43.6**
Loss Function	MIL-NCE	24.8	29.6	15.2	22.1
	Max-Margin	27.4	39.1	18.7	30.1
	Binary Cross Entropy	46.2	54.6	28.4	41.3
	InfoNCE	51.6	60.5	31.9	41.9
	MMS	**54.3**	**63.0**	**33.0**	**43.6**
Clip Sampling / Visual Features	I. 2D features only	51.6	57.9	32.6	37.9
	II. ASR clips	**57.6**	62.8	**34.6**	**44.5**
	AVLnet	54.3	**63.0**	33.0	43.6
Clip Duration	2.5s	23.1	46.1	20.6	36.4
	5s	41.2	55.2	30.2	41.4
	10s	**54.3**	**63.0**	**33.0**	**43.6**
	20s	40.9	52.6	24.5	35.3

Table 3.4: Speech vs. non-speech retrieval results (R@10).

Method	Speech-241		Sounds-241	
	A→V	V→A	A→V	V→A
AVLnet zero-shot	88.0	88.0	32.4	33.6
AVLnet fine-tuned	**92.5**	**91.7**	44.0	46.8

241 (Table 3.4), suggesting our model is particularly effective when speech is present and supporting its application to speech to video search. The performance on Sounds-241 is far above chance, where chance performance is 4.1% R@10, demonstrating AVLnet also detects relevant cues in natural sounds.

3.3.6 Qualitative Retrieval Results

To better understand the performance gains AVLnet achieves over baseline methods, we analyze retrieval examples from our AVLnet model fine-tuned on YouCook2. We show retrieval examples from the YouCook2 validation set in Figure 3-3. We find the retrieved results display high semantic similarity to salient content in the query. For example, in the top row of Figure 3-3, the query audio contains speech instructing viewers to mix together flour and other dry ingredients, and all the retrieved videos show bowls of flour mixtures. The same is true for audio retrieval where, in the third row of Figure 3-3, the query video clip shows oil spread on bread and the retrieved

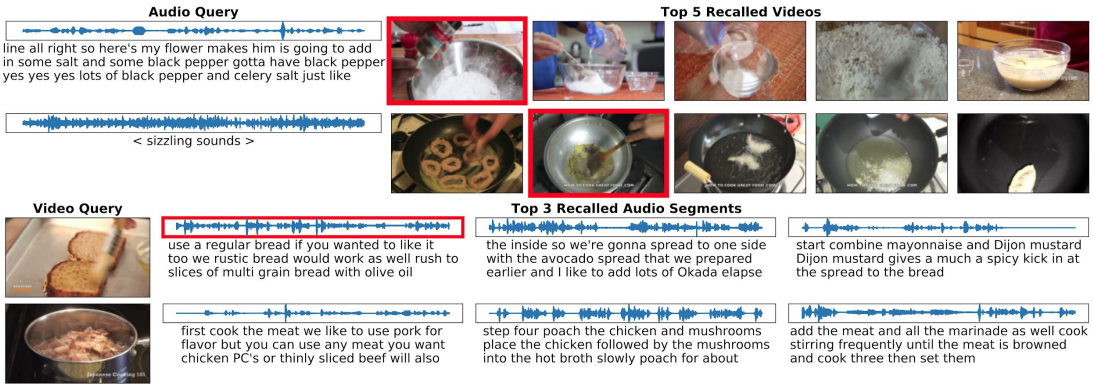


Figure 3-3: Video (top) and audio retrieval (bottom) results from AVLnet fine-tuned on YouCook2. Video clips are represented as their center frame, and audio clips are represented as their waveform and ASR transcript. The correct match is highlighted.

audio contains the words 'bread' and 'spread'. This semantic relationship persists even when the correct clip is not the top result. In the bottom row of Figure 3-3, the correct clip is not recalled in the top five results, yet the video and retrieved audio are both related to cooking meat. Further, we find AVLnet has learned to relate natural sounds to salient video clips. The second row of Figure 3-3 shows an audio query containing only sizzling sounds. Since there was no speech, the ASR system fails, but our model retrieves video clips of frying oil. These results suggest our model has learned the semantic relationships between speech, natural sounds, and visual content, and support its application to video search directly using audio without transcribing speech.

We show additional video and audio retrieval examples from AVLnet fine-tuned on YouCook2 (Figures 3-4 and 3-5) and from AVLnet fine-tuned on CrossTask (Figures 3-6 and 3-7). Consistent with our previous findings, AVLnet retrieves clips that are semantically similar to the query clip, regardless of dataset. In the YouCook2 examples, given an audio query instructing viewers to add ingredients to the blender (Figure 3-4a) AVLnet recalls video clips of blenders, and given a video clip making hamburger patties (Figure 3-5b) AVLnet recalls audio segments discussing burgers. We find similar results on the CrossTask dataset where, given an audio query *"lightly tighten the lug nuts clockwise"* (Figure 3-6b), AVLnet retrieves video clips tightening lug nuts on tires, and given a video query displaying cut lemons AVLnet retrieves audio segments about

lemons (Figure 3-7a). The similarity between queries and retrieved clips persists even when the correct result is not in AVLnet's top 5 results (Figures 3-4c, 3-5c, 3-6c, and 3-7c). For instance, in Figure 3-4c, given an audio query about chopping green onions, AVLnet does not recall the correct clip in the top 5 results, but recalls other highly related clips of chopping green onions. Overall, these results suggest AVLnet has learned to relate semantically similar audio and video channels of videos.

3.4 Chapter Summary

In this chapter, we present a self-supervised method for learning audio-video representations from instructional videos. Whereas prior audio-video work mainly focuses on sound localization, our goal is to relate spoken words to visual entities. We introduce the AVLnet model that learns directly from raw video, reducing the need for spoken or text annotations. We establish baselines on video retrieval tasks on YouCook2, CrossTask, and MSR-VTT and achieve state-of-the-art performance on image retrieval tasks on the Places Spoken Caption dataset. Finally, we show AVLnet learns audio-visual concepts by relating speech and sound to visual objects. In the following chapter, we explore methods for incorporating text into the model.

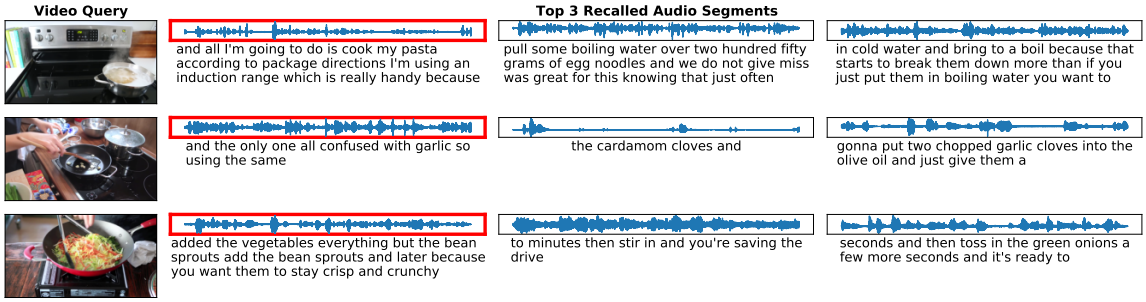
(a) Video clip retrieval examples for clips retrieved correctly (R@1).

(b) Video clip retrieval examples for clips retrieved in the top 5 results (R@5).

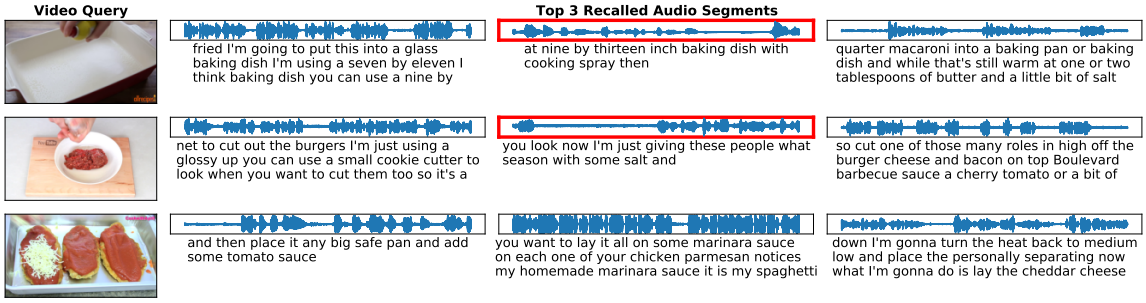
(c) Video clip retrieval examples for clips not retrieved in the top 5 results ($R > 5$).

Figure 3-4: Additional video clip retrieval examples from the YouCook2 validation set. Each row displays the top recalled video clips (shown as each clip's center frame) to the given audio (shown as its waveform and ASR transcript). The ASR transcripts contain mistakes, but are only used for visualization given AVLnet operates on raw audio. The correct match is highlighted.

(a) Language retrieval examples for clips retrieved correctly (R@1).



(b) Language retrieval examples for clips retrieved in the top 5 results (R@5).



(c) Language retrieval examples for clips not retrieved in the top 5 results ($R > 5$).

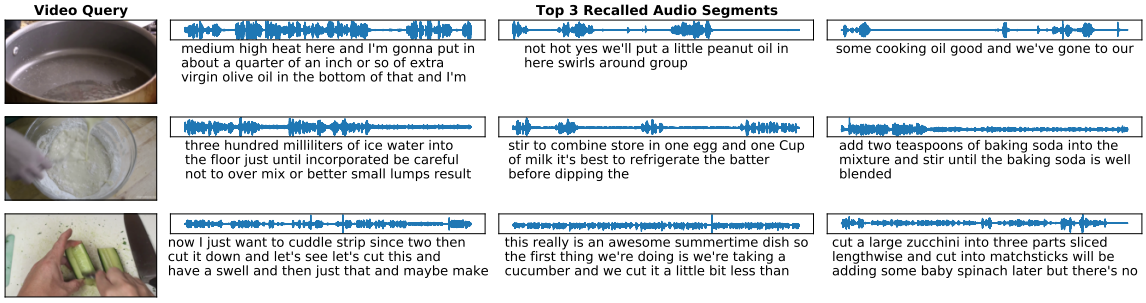
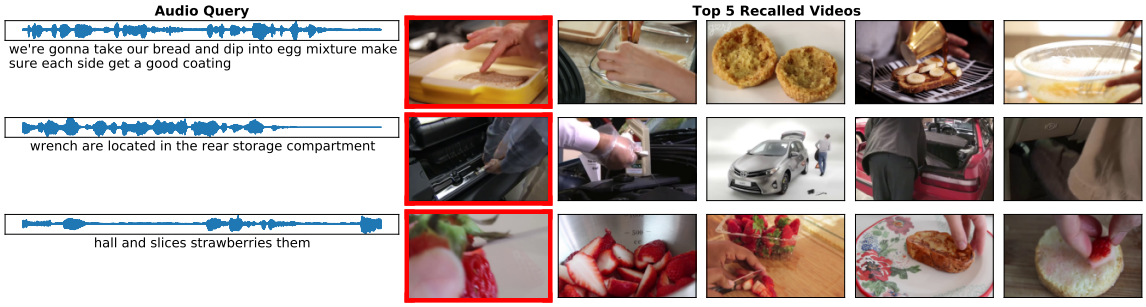


Figure 3-5: Additional audio retrieval examples from the YouCook2 validation set. Each row displays the top recalled audio segments (shown as each segment's waveform and ASR transcript) to the given video (shown as its center frame). The ASR transcripts contain mistakes, but are only used for visualization given AVLnet operates on raw audio. The correct match is highlighted.

(a) Video clip retrieval examples for clips retrieved correctly (R@1).



(b) Video clip retrieval examples for clips retrieved in the top 5 results (R@5).

(c) Video clip retrieval examples for clips not retrieved in the top 5 results ($R > 5$).

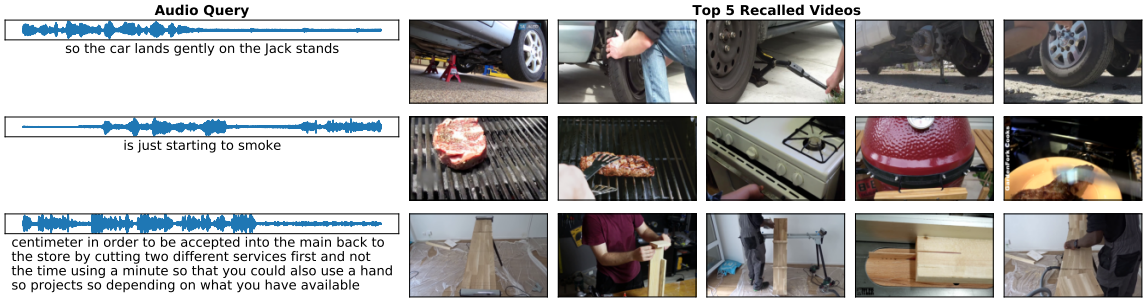
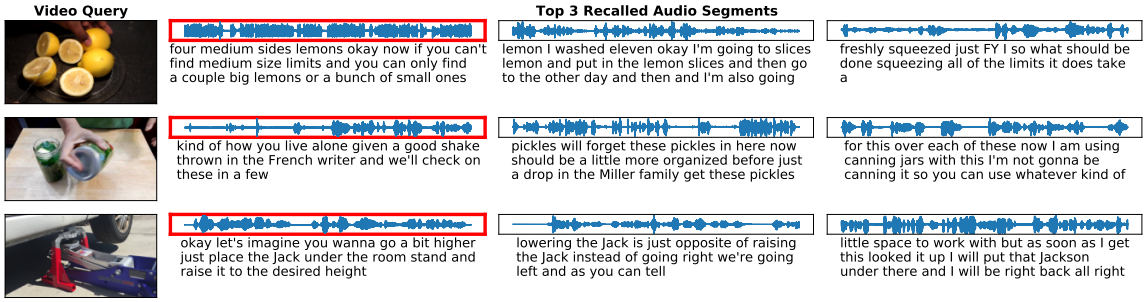
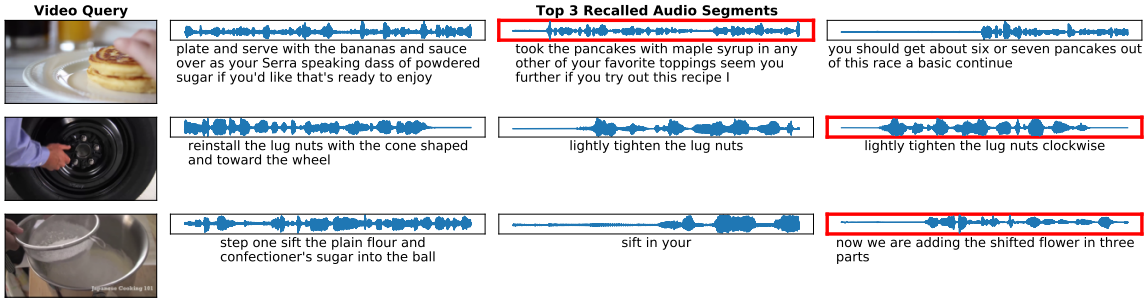


Figure 3-6: Additional video clip retrieval examples from the CrossTask validation set. Each row displays the top recalled video clips (shown as each clip's center frame) to the given audio (shown as its waveform and ASR transcript). The ASR transcripts contain mistakes, but are only used for visualization given AVLnet operates on raw audio. The correct match is highlighted.

(a) Audio retrieval examples for clips retrieved correctly (R@1).



(b) Audio retrieval examples for clips retrieved in the top 5 results (R@5).



(c) Audio retrieval examples for clips not retrieved in the top 5 results ($R > 5$).

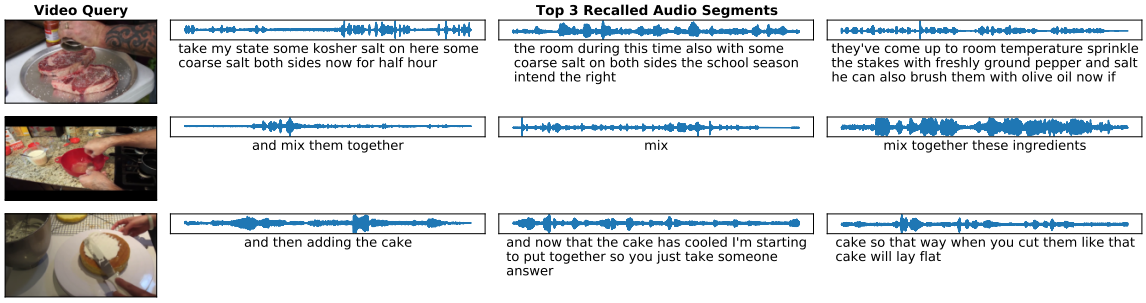


Figure 3-7: Additional audio retrieval examples from the CrossTask validation set. Each row displays the top recalled audio segments (shown as each segment's waveform and ASR transcript) to the given video (shown as its center frame). The ASR transcripts contain mistakes, but are only used for visualization given AVLnet operates on raw audio. The correct match is highlighted.

Chapter 4

AVLnet-Text: Learning from Audio, Video, and Text

In the previous chapter, we showed that the audio-video AVLnet model is able to learn visually grounded language without text captions. Learning without text captions is desirable since ASR is only supported for less than 2% of the world's spoken languages and manually annotating videos with captions is expensive and time-consuming. However, many existing video datasets already have text captions. Therefore, in this chapter we introduce a text branch into the AVLnet model to process text. We refer to the resulting class of models as AVLnet-Text. We propose two ways to incorporate the text branch into the AVLnet model with two corresponding training losses. We compare our approach with previous text-video models on several standard video and language datasets: YouCook2 [88], MSR-VTT [83], and LSMDC [64]. Finally, we show that AVLnet trained without text captions on HowTo100M can perform retrieval with text on the downstream datasets in both the zero-shot and fine-tuned settings, which suggests that the audio representations can be adapted with text representations from only a small amount of text captions. Some of the results in this chapter were presented in Rouditchenko et al. [65].

4.1 Technical Approach

4.1.1 Text Processing

To incorporate text into AVLnet, we add a third branch that processes the text caption from each video clip. We first extract word embeddings using a GoogleNews pretrained Word2Vec model [51] from a text feature extraction pipeline [50]. Following the design of the AVLnet audio and video branches, the word embeddings are max-pooled over the words in each clip's text caption. We integrate the resulting text embedding vector into AVLnet in two different ways, as discussed in the following sections. The differences are illustrated in Figure 4-1. Although our text model is shallower than recent transformer architectures, a study of deeper text models for learning a text-video embedding found little improvement over this simple text model [47].

4.1.2 Independent Tri-Modal Branch Architecture

In this architecture, which we denote as AVLnet-Text-Tri, we keep the text, audio, and video branches separate and apply gating to each branch independently. The motivation for this architecture is to learn a shared embedding space where any two modalities can be compared. For a given clip, we apply non-linear gating to the max-pooled word embedding vector $\mathbf{t}$ as follows:

$$h(\mathbf{t}) = (W_1^a \mathbf{t} + b_1^t) \circ \sigma(W_2^t(W_1^t \mathbf{t} + b_1^t) + b_2^t) \tag{4.1}$$

Where $h(\mathbf{t})$ is the output 4096-dimensional embedding vector, W_1^t, W_2^t matrices and b_1^t, b_2^t vectors are learnable parameters, $\circ$ denotes element-wise multiplication, and σ is the element-wise sigmoid activation. We apply the MMS loss over each of the modality pairs (audio-video, audio-text, and video-text), and the branches are jointly optimized through the sum of these three losses, as follows:

$$\mathcal{L}_{TRI}(f(\mathbf{v}), g(\mathbf{a}), h(\mathbf{t})) = \mathcal{L}_{MMS}(f(\mathbf{v}), g(\mathbf{a})) + \mathcal{L}_{MMS}(g(\mathbf{a}), h(\mathbf{t})) + \mathcal{L}_{MMS}(f(\mathbf{v}), h(\mathbf{t})) \tag{4.2}$$

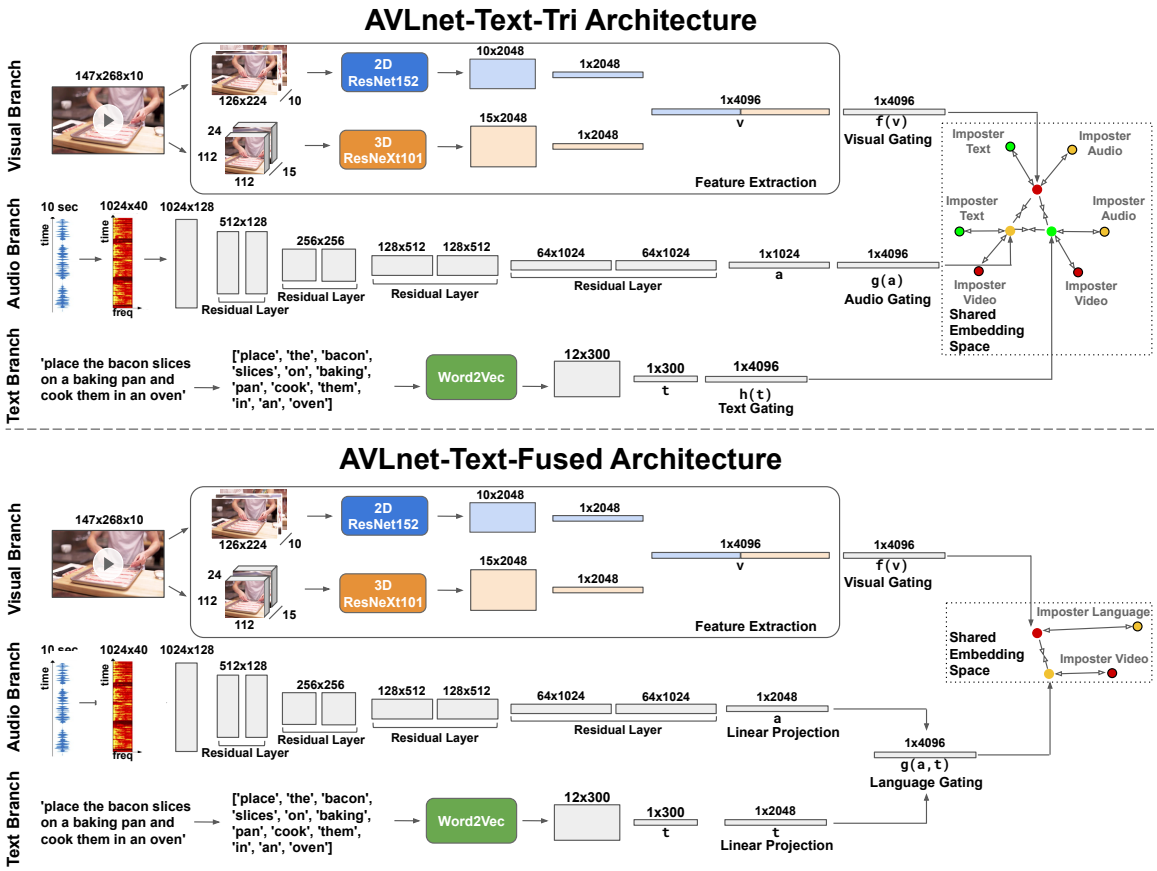


Figure 4-1: We integrate text into the AVLnet model in two different ways. The AVLnet-Text-Tri architecture keeps the text branch separate and projects all three modalities into a shared embedding space. The AVLnet-Text-Fused architecture fuses the audio and text branches into a language branch to learn a shared embedding space between the visual and language (audio-text) modalities.

where $\mathcal{L}_{MMS}$ is defined in Equation 3.3.

4.1.3 Audio-Text Fused Architecture

In this architecture, which we denote as AVLnet-Text-Fused, we fuse the outputs of the audio and text branches before non-linear gating due to the complementary language information in the raw audio and text. Specifically, instead of applying the non-linear gating solely to the audio embedding vector (as in Equation 3.2), we apply

the gating to both the audio and text embedding vectors as follows:

$$g(\mathbf{a}, \mathbf{t}) = (W_1^a \mathbf{a} + W_1^t \mathbf{t} + b_1^{a+t}) \circ \sigma(W_1^{a+t}(W_1^a \mathbf{a} + W_1^t \mathbf{t} + b_1^{a+t}) + b_2^{a+t}) \tag{4.3}$$

where $g(\mathbf{a}, \mathbf{t})$ represents the output language embedding vector combining speech and text information, W_1^a, W_1^t, W_1^{a+t} matrices and b_1^{a+t}, b_2^{a+t} vectors are learnable parameters, $\circ$ denotes element-wise multiplication, and σ is the element-wise sigmoid activation. To train this model, we optimize the following loss:

$$\mathcal{L}_{FUSED}(f(\mathbf{v}), g(\mathbf{a}, \mathbf{t})) = \mathcal{L}_{MMS}(f(\mathbf{v}), g(\mathbf{a}, \mathbf{t})) \tag{4.4}$$

where $\mathcal{L}_{MMS}$ is defined in Equation 3.3. The audio sample and text caption from each video clip are treated as inseparable and are sampled together.

4.2 Experiments

We train AVLnet-Text-Tri and AVLnet-Text-Fused on the instructional YouTube videos from the HowTo100M [50] dataset. For these experiments, we used the video clips defined by the time intervals of each video's ASR transcript, and we use the ASR text as the caption. We evaluate and fine-tune our models on the YouCook2 [88], MSR-VTT [83], and LSMDC [64] datasets. Each dataset provides human-annotated video clip boundaries and text summaries of the clips (full dataset details are in Chapter 2). We evaluate our models on the video clip and language retrieval tasks, in which a language query (text or text and audio) is used to retrieve video and vice versa. The previous results [5, 47, 50] on these datasets mainly focus on text to video retrieval (denoted by T→V). Some models [45, 85] also incorporate audio into the retrieval task, where the audio is considered jointly with the video (denoted by T→A+V). To compare with the prior work in this setting, we use the AVLnet-Text-Tri model. Since the model is trained with a loss that encourages all three modalities to project into a shared embedding space, we use the sum of the text-video and text-audio similarities to retrieve the most similar videos to a given text caption. We also consider the

setting where audio is integrated with text and both are used to retrieve visual clips (denoted by T+A→V). For this evaluation, we use the AVLnet-Text-Fused model. It is also possible to use the AVLnet-Text-Tri model for this evaluation, however, we found that that it typically performed worse than AVLnet-Text-Fused in this setting. We use the standard recall metrics R@1, R@5, R@10, and the median rank (Md. R). For AVLnet-Text-Tri, the hyperparameters are the same as AVLnet, except we increased the batch size to 256, increased the learning rate to 2.5e−4, used a larger embedding size of 6144, and used a clip length of 8 seconds instead of 10 seconds. For AVLnet-Text-Fused, the hyperparameters are also the same as AVLnet, except we used a smaller batch size of 64 and smaller learning rate of 1e−4. We trained both models for 15 epochs, using 4 V100 GPUs for AVLnet-Text-Tri and 2 V100 GPUs for AVLnet-Text-Fused.

4.2.1 Video Retrieval Results

The retrieval results on YouCook2, MSR-VTT, and LSMDC are shown in Table 4.1. In general, the models that incorporate audio typically perform better than those that do not. The improvement in performance when incorporating audio is more significant on YouCook2 and MSR-VTT than LSMDC, since the audio and visual channels in movies often have little salient alignment. AVLnet-Text-Fused typically outperforms AVLnet-Text-Tri in terms of recall metrics on all datasets, but the retrieval setups differ (T+A→V versus T→A+V). On YouCook2, both AVLnet-Text models outperform the previous state-of-the-art models, however, none of the previous models incorporated audio. On MSR-VTT, AVLnet-Text-Tri outperforms the previous state-of-the-art that incorporated audio [45]. On LSMDC, AVLnet-Text-Tri is on-par with the previous state-of-the-art model, achieving a higher R@1 result. We note that the results were current as of the original development of this work (June 2020), since then, there has been much progress on these datasets (especially MSR-VTT).

4.2.2 Training with Text in a Low-Resource Scenario

In this experiment, we explore a scenario where obtaining text annotations during training is expensive, but text exists or can be obtained for smaller evaluative datasets or real world applications. We train the audio-video AVLnet model on HowTo100M without text, and fine-tune/evaluate it with the audio, video, and text from YouCook2, MSR-VTT, and LSMDC. We integrate text into the model following the AVLnet-Text-Fused architecture design, and evaluate the model in the T+A→V setting. The results are shown in Table 3.3, where we compare the zero-shot and fine-tuned results with AVLnet-Text-Fused. Despite being trained on HowTo100M without any text and only fine-tuned with a small amount of text captions on the downstream datasets, the model can perform retrieval with text surprisingly well in both the zero-shot and fine-tuned conditions. AVLnet-Text-Fused still achieves higher results, indicating that using ASR text captions during training on HowTo100M is beneficial. Nonetheless, these results suggest that AVLnet learns language representations from speech, not just natural sounds or voice characteristics, and that audio representations can be adapted with text representations with only a small amount of text captions.

4.3 Chapter Summary

In this chapter, we incorporated a text branch into AVLnet to leverage the existing text captions that already exist in many video datasets. The text branch produces a text embedding vector by max-pooling over the word embedding representations of the text in each video clip. The text branch is then incorporated into the AVLnet model in two different ways, which offers flexibility in the retrieval capabilities. AVLnet-Text achieves strong results on existing retrieval tasks on the YouCook2, MSR-VTT, and LSMDC datasets. Further, we find that AVLnet trained without text on HowTo100M can be adapted with text from the evaluation datasets, suggesting that AVLnet learns audio representations that are complementary to text. This result encouraged us to investigate the model's abilities to learn representations from videos in another language, which we describe in the next chapter.

Table 4.1: Video clip and language retrieval results on YouCook2, MSR-VTT, and LSMDC. The best bi-modal and tri-modal results are bolded. Mod=Modalities.

(a) YouCook2

Method	Training Set	Video Clip Retrieval - YouCook2					Language Retrieval - YouCook2				
		Mod.	R@1	R@5	R@10	Md. R	Mod.	R@1	R@5	R@10	Md. R
Random	—	→V	0.03	0.15	0.3	1675	V→	0.03	0.15	0.3	1675
Miech et al. [50]	HT100M	T→V	6.1	17.3	24.8	46	V→T	5.3	16.5	25.2	42
Miech et al. [47]	HT100M	T→V	**15.1**	**38.0**	**51.2**	**10**	—	—	—	—	—
Miech et al. [50]	HT100M + YC2	T→V	8.2	24.5	35.3	24	V→T	**7.2**	**22.8**	**34.3**	**24**
AVLnet-Text-Tri	HT100M	T→A+V	19.9	36.1	44.3	16.0	V+A→T	28.5	53.7	65.3	6
AVLnet-Text-Tri	HT100M + YC2	T→A+V	**30.2**	**55.5**	**66.5**	**4**	V+A→T	**35.4**	**63.3**	**74.2**	**4**
AVLnet-Text-Fused	HT100M	T+A→V	25.6	52.7	64.4	5	V→T+A	29.3	55.3	65.5	4
AVLnet-Text-Fused	HT100M + YC2	T+A→V	**33.2**	**61.0**	**71.5**	**3**	V→T+A	**34.0**	**62.4**	**72.5**	**3**

(b) MSR-VTT

Method	Training Set	Video Clip Retrieval - MSR-VTT					Language Retrieval - MSR-VTT				
		Mod.	R@1	R@5	R@10	Md. R	Mod.	R@1	R@5	R@10	Md. R
Random	—	→V	0.1	0.5	1.0	500	V→	0.1	0.5	1.0	500
Miech et al. [50]	HT100M	T→V	7.5	21.2	29.6	38	V→T	8.4	21.3	28.9	42
Amrani et al. [5]	HT100M	T→V	8.0	21.3	29.3	33	—	—	—	—	—
Miech et al. [47]	HT100M	T→V	9.9	24.0	32.4	29.5	—	—	—	—	—
Miech et al. [50]	HT100M + MSR-VTT	T→V	14.9	40.2	52.8	9	V→T	**16.8**	**41.7**	**55.1**	**8**
Amrani et al. [5]	HT100M + MSR-VTT	T→V	**17.4**	**41.6**	**53.6**	**8**	—	—	—	—	—
JSFusion [85]	MSR-VTT	T→A+V	10.2	31.2	43.2	13	—	—	—	—	—
CE [45]	MSR-VTT	T→A+V	20.9	48.8	62.4	6	V+A→T	20.6	50.3	**64.0**	5.3
AVLnet-Text-Tri	HT100M	T→A+V	8.3	19.2	27.4	47.5	V+A→T	8.7	19.6	25.1	45
AVLnet-Text-Tri	HT100M + MSR-VTT	T→A+V	**22.5**	**50.5**	**64.1**	**5**	V+A→T	**22.5**	**50.8**	63.9	**5**
AVLnet-Text-Fused	HT100M	T+A→V	19.6	40.8	50.7	9	V→T+A	19.7	43.0	54.9	8
AVLnet-Text-Fused	HT100M + MSR-VTT	T+A→V	**27.1**	**55.6**	**66.6**	**4**	V→T+A	**28.5**	**54.6**	**65.2**	**4**

(c) LSMDC

Method	Training Set	Video Clip Retrieval - LSMDC					Language Retrieval - LSMDC				
		Mod.	R@1	R@5	R@10	Md. R	Mod.	R@1	R@5	R@10	Md. R
Random	—	→V	0.1	0.5	1.0	500	V→	0.1	0.5	1.0	500
Miech et al. [50]	HT100M	T→V	4.0	9.8	14.0	137	V→T	2.4	8.1	11.8	154
Amrani et al. [5]	HT100M	T→V	4.2	11.6	17.1	119	—	—	—	—	—
Miech et al. [50]	HT100M + LSMDC	T→V	7.1	19.6	27.9	40	V→T	**6.6**	**17.8**	**25.9**	**50**
Amrani et al. [5]	HT100M + LSMDC	T→V	**6.4**	**19.8**	**28.4**	**39**	—	—	—	—	—
JSFusion [85]	LSMDC	T→A+V	9.1	21.2	34.1	36	—	—	—	—	—
CE [45]	LSDMC	T→A+V	11.2	**26.9**	**34.8**	**25.3**	—	—	—	—	—
AVLnet-Text-Tri	HT100M	T→A+V	1.4	5.9	9.4	273.5	V+A→T	1.6	4.4	7.5	245.5
AVLnet-Text-Tri	HT100M + LSDMC	T→A+V	**11.4**	26.0	34.6	30	V+A→T	**12.1**	**25.5**	**32.9**	**34**
AVLnet-Text-Fused	HT100M	T+A→V	4.4	10.6	15.3	105.5	V→T+A	3.8	11.3	15.9	109
AVLnet-Text-Fused	HT100M + LSMDC	T+A→V	**17.0**	**38.0**	**48.6**	**11**	V→T+A	**16.5**	**37.6**	**47.6**	**13**

Table 4.2: Results on training with text in a low-resource scenario (R@10). Mod=Modalities, Eval=Evaluation, ZT=Zero-shot, FT=Fine-Tune.

HowTo100M Mod.	Eval. & FT Mod.	YouCook2		MSR-VTT		LSMDC	
		ZT	FT	ZT	FT	ZT	FT
A, V	T, A, V	49.3	66.3	37.0	59.7	10.4	44.4
T, A, V	T, A, V	**64.4**	**71.5**	**50.7**	**66.6**	**15.3**	**48.6**

Chapter 5

Cascaded Multilingual Audio-Video Learning

5.1 Introduction

In Chapter 3, we introduced the audio-video AVLnet model and demonstrated its ability to learn the relationships between raw speech audio and visual content in videos. However, the model was only trained and evaluated on videos in English. In this chapter, we demonstrate that the AVLnet model can learn multilingual audio-video representations. It would be challenging to collect large-scale instructional video datasets in other languages to train AVLnet given the significant engineering effort required to download and process video datasets as large as HowTo100M. Furthermore, there are currently fewer instructional videos available for other languages, especially low-resource languages. To address these limitations, we propose a cascaded approach that applies the AVLnet model trained on English videos to videos in Japanese. While spoken audio captions of images already exist for Japanese [57] and Hindi [24], there are no instructional video datasets similar in size to YouCook2 [88], a standard evaluation dataset of instructional videos in English, in other languages. Therefore, we introduce the YouCook-Japanese instructional video dataset. Applying our cascaded approach, we show an improvement in retrieval performance of nearly 10x on YouCook-Japanese compared to training on the Japanese videos solely. We also show that our cascaded

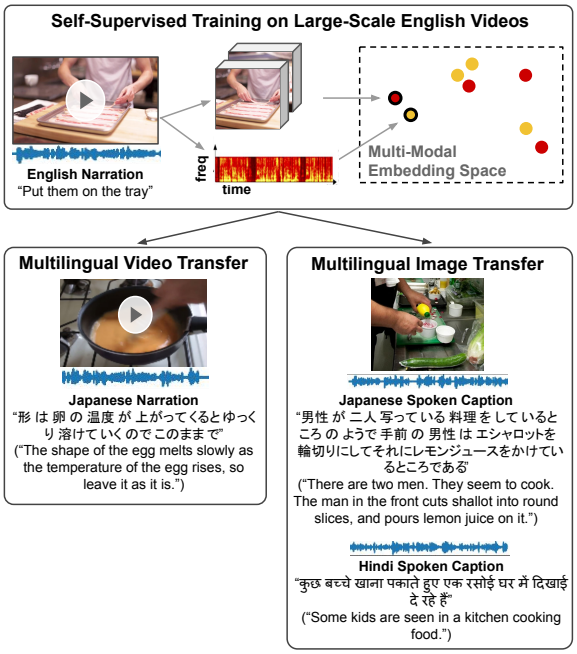


Figure 5-1: Given an audio-video model (AVLnet) trained on videos in English, we transfer the representations to videos in Japanese. We also transfer the representations to images and spoken captions in Japanese and Hindi.

approach can work as a bridge between English instructional videos and spoken audio captions of images in Japanese and Hindi. Given the AVLnet model trained on English videos, we fine-tune it on Japanese and Hindi spoken captions of images, achieving state-of-the-art performance. Finally, we provide an analysis of the impact of the amount of English training videos on downstream performance on both English and Japanese videos.

5.2 Related Work

5.2.1 Multilingual Speech and Video Processing

Image and spoken caption models have been explored in the multilingual setting. Harwath et al. [24] collected 100k Hindi captions of Places images and proposed a bilingual audio-visual model. Building from this, Ohishi et al. [57] collected 100k Japanese captions and proposed a trilingual model. Other work has proposed bilingual models with synthetic spoken captions [31] and image text taggers [39], clustering

for bilingual image-audio dictionaries [9], and pair expansion methods for learning from multilingual captions of disjoint images [56]. Instead of learning from multiple languages simultaneously, our approach is to learn from them one at a time in a cascade.

Several multilingual video datasets have been introduced, such as How2 [67] and VATEX [81] which contain parallel translations of English video captions in Portuguese and Chinese. Instead of collecting parallel translations, Sigurdsson et al. proposed versions of HowTo100M in Japanese, French, and Korean [69]. Thus far, all of the methods proposed on these datasets rely on text captions. Instead, we use AVLnet to learn from videos using speech audio and without requiring text.

Finally, multilingual ASR is a well-established research area. Methods include simultaneous training on multiple languages [13, 16, 36, 40] and cascaded approaches in which representations learned from one language are used as initialization to learn from other languages [22, 76]. Our approach is similar in spirit to the cascaded methods, but it only requires audio-visual data without transcripts.

5.3 Technical Approach

5.3.1 Videos

Our goal is to learn audio-visual representations for videos in languages other than English using AVLnet. AVLnet is trained through a contrastive loss to discriminate between temporally aligned audio-video pairs and temporally mismatched pairs from both within the same video and from other videos. This results in an audio-video embedding space which colocates semantically similar audio and visual inputs. Since AVLnet does not require any annotations besides the raw video data, we only assume that a set of videos in the target language is given, but without any additional annotation. One approach is to simply train AVLnet only on the target videos in the new language. However, we find that a large number of videos, typically hundreds of thousands, is necessary to learn strong representations from scratch, and there is

simply not enough videos in downstream datasets such as YouCook2 to train the model from scratch. Therefore, our proposed approach is simple: given the AVLnet model trained on English HowTo100M videos, we apply it to videos in Japanese by directly fine-tuning it on the Japanese videos. This represents a cascade since the model only learns from videos in one language at a time (ie. first English, then Japanese).

YouCook-Japanese. There are currently no other instructional video datasets in other languages similar in size to YouCook2. Therefore, we collected a dataset of Japanese cooking videos, and call it YouCook-Japanese to indicate the similarity in content and size to YouCook2. As a starting point, Sigurdsson et al. [69] proposed a version of HowTo100M in Japanese with approximately 300k videos. We followed the steps to download Japanese instructional videos from YouTube, except we limited the search to cooking videos only. We used a CNN-based audio segmentation toolkit [18] to segment the videos into clips containing speech, and then filtered the clips to be at least 5s and at most 50s. To make the dataset similar in size to YouCook2, we selected 10k random clips for training, 3k clips for validation, and 3k clips for evaluation, with the constraint that each video can only appear in one set. The training set contains 737 videos, the validation set contains 224 videos, and the evaluation set contains 213 videos.

5.3.2 Images and Spoken Captions

Since instructional videos and spoken captions of images both contain descriptive audio of visual scenes, our cascaded approach is also applicable to images and spoken captions. Specifically, we use the AVLnet model trained on HowTo100M videos and fine-tune it on the spoken captions and images in the Places Audio Caption Dataset in Japanese and Hindi. For these experiments, we train AVLnet using only the 2D features in the visual branch so that the model can work on both videos and images.

Table 5.1: Video retrieval on YouCook2 Videos (YC-EN) and YouCook-Japanese videos (YC-JP). HT100M=HowTo100M.

(a) English YouCook2 Videos (YC-EN)

AVLnet Train Data	**Video Clip (A→V)**			**Language (V→A)**		
	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.03	0.15	0.3	0.03	0.15	0.3
YC-EN	0.7	2.3	3.9	0.8	3.0	4.9
HT100M	27.4	51.6	61.5	27.3	51.2	60.8
HT100M + YC-EN	**30.7**	**57.7**	**67.4**	**33.0**	**58.9**	**68.4**
HT100M + YC-JP	19.4	40.4	51.3	19.8	43.5	53.7

(b) YouCook-Japanese Videos (YC-JP)

AVLnet Train Data	**Video Clip (A→V)**			**Language (V→A)**		
	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.03	0.17	0.33	0.03	0.17	0.33
YC-JP	0.7	2.4	3.8	0.5	1.8	3.0
HT100M	4.6	12.1	18.2	5.6	14.6	21.3
HT100M + YC-EN	5.1	13.2	18.9	5.6	14.5	20.7
HT100M + YC-JP	**7.0**	**20.4**	**29.3**	**7.6**	**20.9**	**29.7**

5.4 Experiments

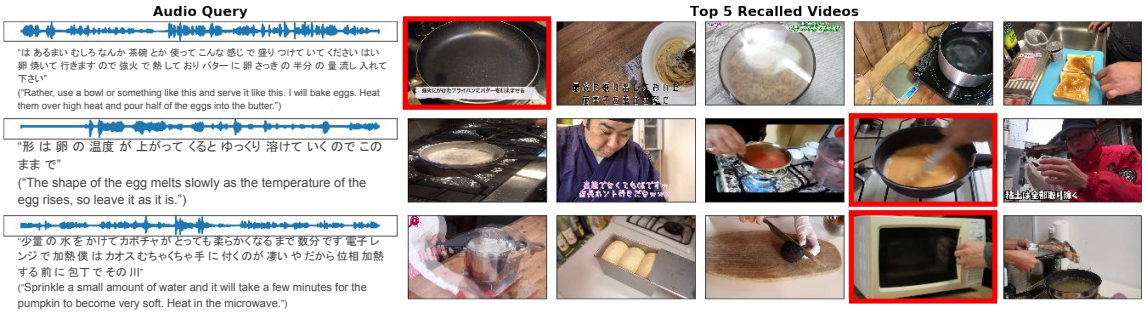
5.4.1 Datasets

Videos. We use the following instructional video datasets: HowTo100M [50], YouCook2 [88], and YouCook-Japanese. For YouCook2, we use 9,586 train clips and 3,350 validation clips. We evaluate performance on audio to video clip retrieval and video clip to audio retrieval. The full dataset details are in Chapter 2.

Images and Spoken Captions. We fine-tune and evaluate our model on the Places Audio Caption dataset [28], which contains 100k images from the Places205 dataset [87] each with a spoken caption in Japanese [57] and Hindi [24]. We evaluate the performance on audio to image and image to audio retrieval using the associated evaluation set of 1k images and spoken captions.

5.4.2 Implementation Details

For training AVLnet on HowTo100M, we follow the details in Section 3.2.3. For fine-tuning on video clips from YouCook2 and YouCook-Japanese, we use a batch size



(a) YouCook-Japanese video retrieval after training on English HowTo100M videos and without fine-tuning on YouCook-Japanese videos.

(b) YouCook-Japanese video retrieval after training on English HowTo100M videos and fine-tuning on YouCook-Japanese videos.

Figure 5-2: YouCook-Japanese video retrieval results with AVLnet - (a) zero-shot and (b) after fine-tuning. Japanese ASR transcripts and English translations are shown, but AVLnet only uses audio as input. Center frames of clips are shown, and the correct match is in red.

of 256 clips and a learning rate of 1e−4. We pad the audio or crop it up to 50 seconds in length. For fine-tuning AVLnet on images and spoken captions in Places, we either keep the ResNet-152 model frozen or fine-tune it. We use a learning rate of 1e−3 for the frozen setting and a learning rate of 1e−4 for the trainable setting.

5.4.3 Video Retrieval

YouCook2. Table 5.1a shows the video retrieval results on English YouCook2 videos. We note that some of the YouCook2 results have already been presented in Table 3.2, and we re-print them here for comparison with the results on YouCook-Japanese. Training on HowTo100M significantly improves performance compared with training only on YouCook2. In the zero-shot setting, ie. without fine-tuning on any YouCook2 videos, the model achieves strong retrieval performance, likely due to the similar

instructional domain of HowTo100M and YouCook2 and shared language (English). Performance further improves after fine-tuning on YouCook2 videos. The final row of Table 5.1a shows that fine-tuning the model on YouCook-Japanese videos reduces the performance on YouCook2, indicating that the model is sensitive to the language present in the videos.

YouCook-Japanese. Table 5.1b shows the video retrieval results on YouCook-Japanese videos. AVLnet's performance when trained only on YouCook-Japanese is similar to AVLnet's performance on YouCook2 when trained only on YouCook2 videos, indicating that the two datasets are similar in difficulty. Using our cascaded approach, we apply the AVLnet model trained on HowTo100M to the Japanese videos which significantly improves performance. In the zero-shot setting, ie. without fine-tuning, the retrieval performance is nearly 5x the performance compared with training on YouCook2-Japanese only. This is surprising considering that the model has only been trained on English videos. Fine-tuning the model on the Japanese videos further increases the performance to nearly 10x the performance compared with training on YouCook2-Japanese only. We also note that fine-tuning the model on English YouCook2 videos instead of Japanese videos is comparable to the zero-shot performance, further indicating that the model is actually sensitive to the language present in the videos.

Qualitative results. Figure 5-2 shows qualitative YouCook-Japanese video retrieval results. In the zero-shot setting, without fine-tuning on Japanese videos, the model seems to perform retrieval using salient natural sounds, for example, sizzling sounds or microwave beeps. After fine-tuning the model on YouCook-Japanese, the model can handle more complex queries and retrieve video clips with specific ingredients mentioned in the audio queries.

Varying the % of HowTo100M videos. In Figure 5-3, we show the video retrieval performance when training AVLnet model with a smaller percentage of HowTo100M videos. In Figure 5-3a, we use 10% or less of the HowTo100M videos, and in Figure 5-3b we use between 10% and 100% of the videos. The plots show that performance generally increases with the number of HowTo100M videos. However, the gap be-

Table 5.2: Image retrieval on the Places Audio Caption dataset. No HT100M = No training on HowTo100M (Model was trained on Places only).

(a) Places Audio Captions - Japanese

Method	**Audio to Image**			**Image to Audio**		
	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.1	0.5	1.0	0.1	0.5	1.0
Havard et al. [31]	18.2	48.5	62.2	15.3	41.4	57.6
Ohishi et al. [56]	20.1	49.7	63.9	16.7	44.3	57.8
Ohishi et al. [57]	20.3	52.0	66.7	20.0	46.8	62.3
Ours, No HT100M	20.7	48.8	63.6	16.8	44.9	58.8
Ours, AVLnet	**23.5**	**57.3**	**70.4**	**24.3**	**56.6**	**70.0**

(b) Places Audio Captions - Hindi

Method	**Audio to Image**			**Image to Audio**		
	R@1	R@5	R@10	R@1	R@5	R@10
Random	0.1	0.5	1.0	0.1	0.5	1.0
Harwath et al. [24]	8.0	25.0	35.6	7.4	23.5	35.4
Havard et al. [31]	9.6	28.2	40.7	8.0	27.6	37.1
Ohishi et al. [56]	9.4	29.8	41.8	9.3	29.5	38.2
Ohishi et al. [57]	11.2	31.5	44.5	10.8	31.3	41.9
Ours, No HT100M	9.2	26.0	35.7	8.7	23.6	33.5
Ours, AVLnet	**15.2**	**38.9**	**51.1**	**17.0**	**39.8**	**51.5**

tween performance on English and Japanese videos is lower when 10% or less of the HowTo100M videos are used.

5.4.4 Image Retrieval

Table 5.2 shows the retrieval results on the Places Audio Caption dataset in Japanese and Hindi. For our cascaded approach, we fine-tune AVLnet trained on HowTo100M videos to each language in Places independently. We compare our approach to the state-of-the-art models for each dataset. While previous models are not trained on HowTo100M videos, some of them [24, 57] are trained on images with parallel spoken captions in multiple languages. Our cascaded approach involves training on one language at a time, achieving large gains over prior baselines. We also show the results of training our model only on Places, without training on HowTo100M videos. The results are significantly lower but comparable to the previous baselines, indicating

Table 5.3: Comparison of frozen versus trainable image encoder for fine-tuning on the Places Audio Caption dataset.

Language	Frozen Img. CNN	Audio to Image			Image to Audio		
		R@1	R@5	R@10	R@1	R@5	R@10
English	Yes	37.3	71.6	82.3	36.5	71.3	82.8
	No	**44.8**	**79.9**	**86.4**	**42.8**	**76.2**	**84.8**
Japanese	**Yes**	**23.5**	**57.3**	**70.4**	**24.3**	**56.6**	**70.0**
	No	20.8	50.9	64.9	20.9	49.5	63.5
Hindi	**Yes**	**15.2**	**38.9**	**51.1**	**17.0**	**39.8**	**51.5**
	No	12.1	30.9	44.1	11.9	30.8	41.7

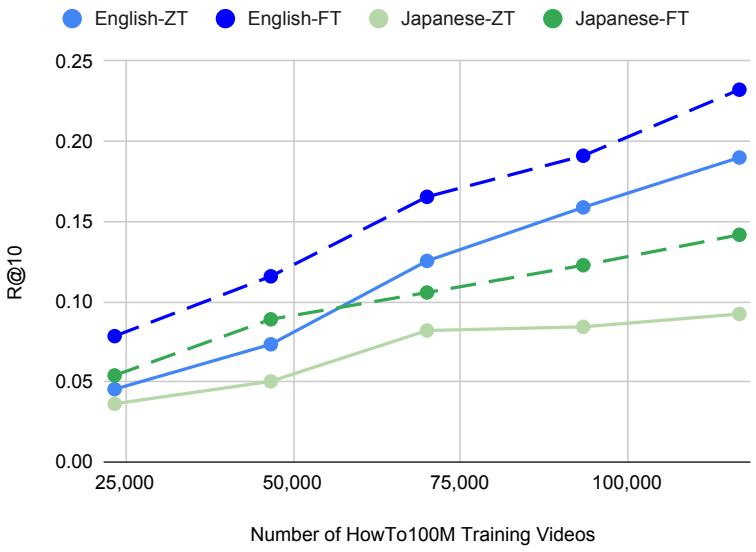
that training on HowTo100M videos is beneficial.

In Table 5.3, we show that the results on each language are sensitive to whether the visual encoder (ResNet-152) was made trainable or kept frozen during fine-tuning. The results were higher for English with a trainable encoder, while the results were higher for Japanese and Hindi with a frozen encoder. We hypothesize that the 400k images in the English set is enough data to train the ResNet-152 model, while the 100k images in the Japanese and Hindi set is not enough, and therefore it is better to leave it frozen for Japanese and Hindi. Furthermore, given that the visual encoders are also frozen during fine-tuning on videos, these results suggest that the visual branch is more language independent than the audio branch, and that the audio branch needs to be adapted to handle unseen languages.

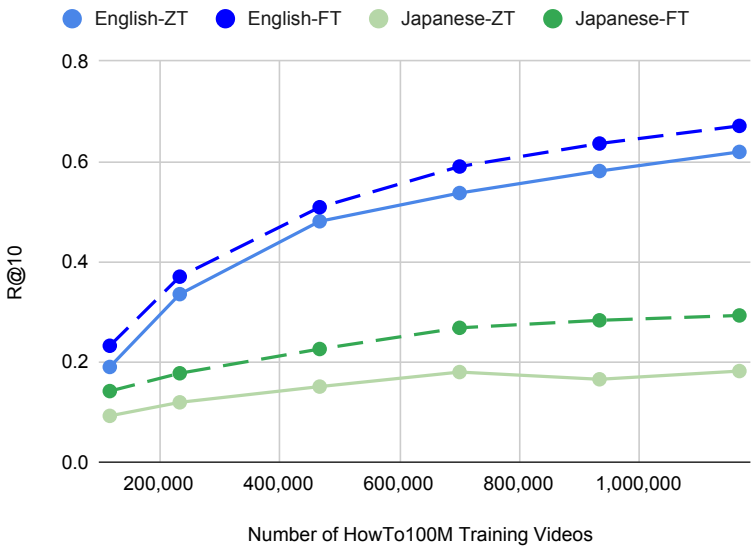
5.5 Chapter Summary

In this chapter, we propose a cascaded approach to learn multilingual audio-visual representations. Given the AVLnet model trained on English HowTo100M videos, we fine-tuned and evaluated it on YouCook-Japanese videos and the images and spoken captions in the Places Audio Caption dataset in Japanese and Hindi. The representations learned from HowTo100M serve as a strong initialization for fine-tuning on Japanese videos through our cascaded approach, which improves performance by nearly 10x compared to training on the Japanese videos solely. Our approach could hypothetically work for instructional videos in any language. One direction that we

plan to explore in the future is cross-lingual retrieval in videos, for example, to retrieve Japanese audio from English audio, potentially using video clips as an intermediate modality.



(a) {2,4,6,8,10}% of HowTo100M.



(b) {10,20,40,60,80,100}% of HowTo100M.

Figure 5-3: Video retrieval performance when varying the % of HowTo100M videos. ZT=Zero-Shot, FT=Fine-tune.

Chapter 6

Conclusion

6.1 Summary of Contributions

The goal of this thesis is to develop methods that learn to ground speech to visual content in instructional videos, which contain spoken descriptions of actions and objects and are freely available on the internet in large quantities. Since these methods learn directly from videos without requiring annotation, they can be applied to videos in any language, including the many low-resource languages which do not have speech recognition capabilities. In Chapter 3, we proposed a self-supervised model, the audio-video language network (AVLnet), that learns from the raw audio and visual channels in unlabeled instructional videos. The model is trained through a contrastive loss to discriminate between temporally aligned audio-video pairs and temporally mismatched pairs. This results in an audio-video embedding space which colocates semantically similar audio and visual inputs. We trained AVLnet on the largest available dataset of instructional videos containing 1.2 million videos and evaluated on image retrieval and video retrieval tasks, achieving state-of-the-art performance. We demonstrated that the model learned semantic correspondences between speech, natural sounds, and visual content by successfully applying it to video retrieval using spoken queries and audio, without needing to transcribe speech in the query to text. In Chapter 4, we proposed a tri-modal model, AVLnet-Text, that additionally learns from the text narration which already exists in many instructional video datasets. The training

method results in a multi-modal embedding space useful for text to video retrieval. In Chapter 5, we explore the multilingual capabilities of the audio-video AVLnet model and described our collection of the YouCook-Japanese dataset of Japanese cooking videos. We proposed a cascaded approach that applies our model trained on English videos to the videos in Japanese, improving retrieval performance by nearly 10x. This thesis establishes benchmarks on audio to video retrieval on several datasets for future work on self-supervised learning from videos.

6.2 Future Directions

The research in this thesis can be expanded in many directions, and we discuss a few of them here.

6.2.1 Handling Misalignment in Instructional Videos

A key challenge of learning from instructional videos is handling misalignment between what a speaker describes and what is on-screen, since they may describe an object before or after showing it. Given that speech in instructional videos is not always visually aligned, we trained AVLnet to aggregate the context over long clips from HowTo100M (10s), which seems to resolve the misalignment problem to some extent. However, representing 10s of audio and video as single embedding vectors is potentially discarding the fine-grained nuances of the video clip. To learn more complex relationships between spoken words and visual content, it would be beneficial to model the relationships between audio and video samples more densely, perhaps with attention. However, while this approach works well in supervised multi-modal domains [78], the contrastive loss used in this work for self-supervised learning computes a similarity matrix of space $O(N^2)$, where N is the batch size. Therefore, future work is required to improve the modeling in this aspect, while managing the memory limits of current computing machines.

6.2.2 Object Grounding and Spatial Reasoning

Previously, Harwath et al. [28,29] demonstrated that image and spoken caption models trained for retrieval could learn to highlight the objects related to particular spoken words. Boggust et al. [10] applied the image model to videos and presented initial results on object grounding using speech in videos, however, the model often missed semantically relevant pixels or highlighted pixels unrelated to the audio. The current AVLnet model can retrieve relevant video clips given an input audio query, but it cannot highlight or ground specific objects related to input words or sounds. This is primarily due to the visual CNN encoders being frozen and the loss of spatial activations from the spatio-temporal max-pooling. Therefore, further work is required to enable more complex object grounding abilities.

6.2.3 Improving Video Dataset Reproducibility

Our work here relies heavily on video datasets curated from YouTube (e.g., HowTo100M, YouCook2, MSR-VTT, YouCook-Japanese). To comply with YouTube's terms of service, these video datasets are typically distributed via URL, and each research group must scrape the videos independently. Over time, as YouTube and YouTubers remove videos from the platform, the original datasets shrink, making it challenging to reproduce, expand upon, and compare to our results. Solving this challenge will require the entire research community to come together and propose new solutions.

6.3 Parting Discussion

In this thesis, we introduced methods to learn correspondences between video and speech using video content naturally generated by humans instead of using manually annotated data. This enables the possibility of learning correspondences in any language in the world with such video content. As less than 2% of the world's languages have Automatic Speech Recognition (ASR) capability, this presents a significant opportunity. Given the rapid adoption of video platforms by users globally,

we expect that our methods could help scale the advancements in speech technologies developed for these languages. This would enable a greater number of people to interact more effectively with computers.

Bibliography

[1] `https://www.ibm.com/watson/services/speech-to-text/`.

[2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *CVPR*, 2016.

[3] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In *NeurIPS*, 2020.

[4] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020.

[5] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. *arXiv preprint arXiv:2003.03186*, 2020.

[6] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017.

[7] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*, 2018.

[8] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NeurIPS*, 2016.

[9] Emmanuel Azuh, David Harwath, and James R Glass. Towards bilingual lexicon discovery from visually grounded speech audio. In *INTERSPEECH*, 2019.

[10] Angie Boggust, Kartik Audhkhasi, Dhiraj Joshi, David Harwath, Samuel Thomas, Rogerio Feris, Dan Gutfreund, Yang Zhang, Antonio Torralba, Michael Picheny, and James Glass. Grounding spoken words in unlabeled video. In *CVPR Sight and Sound Workshop*, 2019.

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[13] Jaejin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiesner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiat, Shinji Watanabe, and Takaaki Hori. Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling. In *SLT*, 2018.

[14] Grzegorz Chrupała. Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. *arXiv preprint arXiv:2104.13225*, 2021.

[15] Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *ACL*, 2017.

[16] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[18] David Doukhan, Jean Carrive, Félicien Vallet, Anthony Larcher, and Sylvain Meignier. An open-source speaker gender detection framework for monitoring gender equality. In *ICASSP*, 2018.

[19] Ryan Eloff, Herman A Engelbrecht, and Herman Kamper. Multimodal one-shot learning of speech and images. In *ICASSP*, 2019.

[20] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018.

[21] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *CVPR*, 2019.

[22] Arnab Ghoshal, Pawel Swietojanski, and Steve Renals. Multilingual training of deep neural networks. In *ICASSP*, 2013.

[23] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018.

[24] David Harwath, Galen Chuang, and James Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *ICASSP*, 2018.

[25] David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In *ASRU*, 2015.

[26] David Harwath and James Glass. Learning word-like units from joint audio-visual analysis. In *ACL*, 2017.

[27] David Harwath, Wei-Ning Hsu, and James Glass. Learning hierarchical discrete linguistic units from visually-grounded speech. In *ICLR*, 2020.

[28] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *ECCV*, 2018.

[29] David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *IJCV*, 2020.

[30] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *NeurIPS*, 2016.

[31] William N Havard, Jean-Pierre Chevrot, and Laurent Besacier. Models of visually grounded speech signal pay attention to nouns: A bilingual experiment on english and japanese. In *ICASSP*, 2019.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[33] Nils Holzenberger, Shruti Palaskar, Pranava Madhyastha, Florian Metze, and Raman Arora. Learning from multiview correlations in open-domain videos. In *ICASSP*, 2019.

[34] Wei-Ning Hsu and James Glass. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264*, 2018.

[35] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *CVPR*, 2019.

[36] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *ICASSP*, 2013.

[37] Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. Large-scale representation learning from visually grounded untranscribed speech. In *CoNLL*, 2019.

[38] Herman Kamper, Aristotelis Anastassiou, and Karen Livescu. Semantic query-by-example speech search using visual grounding. In *ICASSP*, 2019.

[39] Herman Kamper and Michael Roth. Visually grounded cross-lingual keyword spotting in speech. *arXiv preprint arXiv:1806.05030*, 2018.

[40] Martin Karáfidt, Murali Karthick Baskar, Karel Veselỳ, František Grézl, Lukáš Burget, and Jan Černockỳ. Analysis of multilingual blstm acoustic model on low and high resource languages. In *ICASSP*, 2018.

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[42] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018.

[43] Hilde Kuehne, Ahsan Iqbal, Alexander Richard, and Juergen Gall. Mining youtube-a dataset for learning fine-grained action concepts from webly supervised video data. *arXiv preprint arXiv:1906.01012*, 2019.

[44] Kenneth Leidal, David Harwath, and James Glass. Learning modality-invariant representations for speech and images. In *ASRU*, 2017.

[45] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019.

[46] Danny Merkx, Stefan L. Frank, and Mirjam Ernestus. Language learning using speech to image retrieval. In *INTERSPEECH*, 2019.

[47] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020.

[48] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. In *CVPR Workshop on YouTube-8M Large-Scale Video Understanding*, 2017.

[49] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

[50] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.

[51] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[52] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ICMR*, 2018.

[53] Mathew Monfort, S Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *CVPR*, 2021.

[54] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *NeurIPS*, 2018.

[55] Masood S Mortazavi. Speech-image semantic alignment does not depend on any prior classification tasks. In *INTERSPEECH*, 2020.

[56] Yasunori Ohishi, Akisato Kimura, Takahito Kawanishi, Kunio Kashino, David Harwath, and James Glass. Pair expansion for learning multilingual semantic embeddings using disjoint visually-grounded speech audio datasets. In *INTERSPEECH*, 2020.

[57] Yasunori Ohishi, Akisato Kimura, Takahito Kawanishi, Kunio Kashino, David Harwath, and James Glass. Trilingual semantic embeddings of visually grounded speech with self-attention mechanisms. In *ICASSP*, 2020.

[58] Andreea-Maria Oncescu, Jõao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality textual and audio narrations. *arXiv preprint arXiv:2011.11071*, 2020.

[59] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[60] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018.

[61] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016.

[62] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[63] Manasa Prasad, Daan van Esch, Sandy Ritchie, and Jonas Fromseier Mortensen. Building large-vocabulary asr systems for languages without any audio training data. In *INTERSPEECH*, 2019.

[64] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. In *IJCV*, 2017.

[65] Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020.

[66] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *ICASSP*, 2019.

[67] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. In *Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS, 2018.

[68] Ramon Sanabria, Austin Waters, and Jason Baldridge. Talk, don't write: A study of direct speech-based image retrieval. *arXiv preprint arXiv:2104.01894*, 2021.

[69] Gunnar A Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, João Carreira, Phil Blunsom, and Andrew Zisserman. Visual grounding in video for unsupervised word translation. In *CVPR*, 2020.

[70] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019.

[71] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019.

[72] Didac Surís, Amanda Duarte, Amaia Salvador, Jordi Torres, and Xavier Giró-i Nieto. Cross-modal embeddings for video and audio retrieval. In *ECCV*, 2018.

[73] Didac Suris, Adria Recasens, David Bau, David Harwath, James Glass, and Antonio Torralba. Learning words by drawing images. In *CVPR*, 2019.

[74] Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. Learning words from images and speech. In *NeurIPS Workshop on Learning Semantics*, 2014.

[75] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *CVPR*, 2019.

[76] Samuel Thomas, Sriram Ganapathy, and Hynek Hermansky. Multilingual mlp features for low-resource lvcsr systems. In *ICASSP*, 2012.

[77] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018.

[78] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, 2019.

[79] Liming Wang and Mark Hasegawa-Johnson. A dnn-hmm-dnn hybrid model for discovering word-like units from spoken captions and image regions. In *INTERSPEECH*, 2020.

[80] Liming Wang and Mark A Hasegawa-Johnson. Multimodal word discovery and retrieval with phone sequence and image concepts. In *INTERSPEECH*, 2019.

[81] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019.

[82] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *ICCV*, 2019.

[83] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[84] Karren Yang, Bryan Russell, and Justin Salamon. Telling left from right: Learning spatial correspondence of sight and sound. In *CVPR*, 2020.

[85] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018.

[86] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018.

[87] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *NeurIPS*, 2014.

[88] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018.

[89] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019.