AVLnet: Learning Audio-Visual Language Representations from Instructional Videos

Andrew Rouditchenko[1*] **Angie Boggust**[1*] **David Harwath**[1] **Dhiraj Joshi**[2,3]
Samuel Thomas[2,3] **Kartik Audhkhasi**[2,3†] **Rogerio Feris**[2,3] **Brian Kingsbury**[2,3]
Michael Picheny[4] **Antonio Torralba**[1] **James Glass**[1]
[1]MIT CSAIL, [2]IBM Research AI, [3]MIT-IBM Watson AI Lab, [4]NYU-Courant CS & CDS
{roudi,aboggust,dharwath,torralba,glass}@mit.edu,
{djoshi,sthomas,rsferis,bedk}@us.ibm.com, kartikaudhkhasi@gmail.com, map22@nyu.edu

Abstract

Current methods for learning visually grounded language from videos often rely on time-consuming and expensive data collection, such as human annotated textual summaries or machine generated automatic speech recognition transcripts. In this work, we introduce Audio-Video Language Network (AVLnet), a self-supervised network that learns a shared audio-visual embedding space directly from raw video inputs. We circumvent the need for annotation and instead learn audio-visual language representations directly from randomly segmented video clips and their raw audio waveforms. We train AVLnet on publicly available instructional videos and evaluate our model on video clip and language retrieval tasks on three video datasets. Our proposed model outperforms several state-of-the-art text-video baselines by up to 11.8% in a video clip retrieval task, despite operating on the raw audio instead of manually annotated text captions. Further, we show AVLnet is capable of integrating textual information, increasing its modularity and improving performance by up to 20.3% on the video clip retrieval task. Finally, we perform analysis of AVLnet's learned representations, showing our model has learned to relate visual objects with salient words and natural sounds.

1 Introduction

Humans learn to understand language, recognize objects, and identify the correspondences between the two by recognizing patterns in what they see and what they hear, often with very weak supervision. In this paper, we develop machine learning methods for this kind of audio-visual learning. Researchers have already developed models capable of learning language concepts from paired images and spoken audio captions describing the images [1, 2]. However, these approaches require a supervised data collection procedure where annotators are paid to describe images. Recent work on learning language concepts using text instead leverages instructional videos that are freely available on the internet [3, 4], but these videos require expensive and time-consuming annotation such as human-generated textual summaries or Automatic Speech Recognition (ASR) transcripts.

In this paper, we circumvent the need for text annotation by learning from naturally occurring audio-visual correspondences in instructional videos. We introduce the Audio-Video Language Network (AVLnet) architecture — a self-supervised model that learns a shared audio-visual embedding space directly from raw video. We leverage the HowTo100M dataset [3] to train models on publicly available

*These authors contributed equally to this work
†Now at Google

Preprint. Under review.

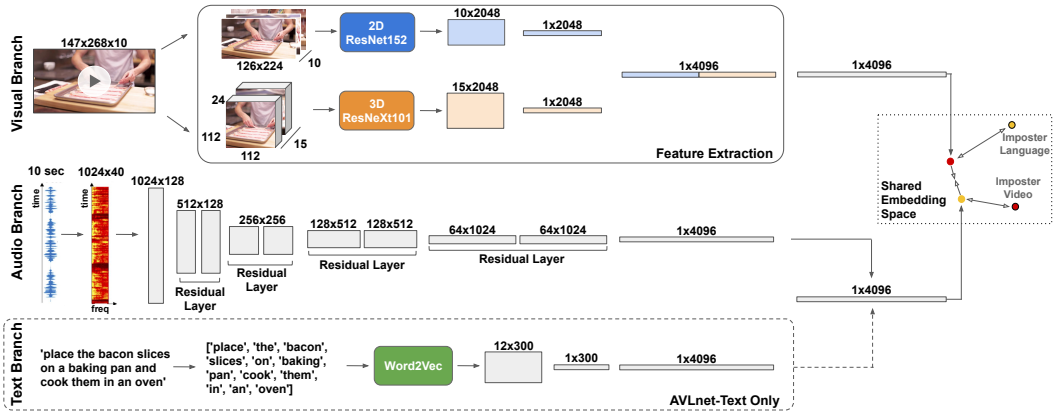


Figure 1: The Audio-Video Language Network (AVLnet) model.

instructional videos. In contrast to prior work using annotated data and supervised techniques to define video clips, AVLnet uses randomly sampled video clips to learn audio-visual representations from raw video. AVLnet can be further extended to integrate text as a third modality (AVLnet-Text), demonstrating that it can learn language representations from both raw audio and text.

Our AVLnet model achieves state-of-the-art performance on the YouCook2 [5] video clip and language retrieval tasks. Further, we demonstrate the transferability of our models to non-instructional video datasets: MSR-VTT [6] and LSDMC [7]. Integrating text captions in the AVLnet-Text model outperforms the prior state-of-the-art results on all three datasets. Finally, we show our models are able to semantically relate the audio and visual modalities to learn static concepts such as "flour", action words like "chop", and salient natural sounds such as sizzling.

2 Related Work

Most closely related to this paper is the work combining paired image and speech information in an unsupervised setting [1, 2, 8–17]. These models attempt to leverage the correlations between visual objects in images with spoken words as a grounding signal for learning visual semantics directly from speech. Our work builds upon recent results that demonstrate an ability to uncover concepts from images paired with spoken descriptions [1, 2] or video frames paired with raw audio [13] by learning a joint audio-visual latent space that reflects the underlying semantics of both modalities. While the aforementioned work relies on still image inputs, our proposed architecture learns from entire video clips. Further we eliminate the need for human-generated captions by applying our models to publicly available instructional videos.

Recently there has been an influx of instructional video datasets including How2 [18], Inria Instructional Videos [19], CrossTask [20], YouCook2 [5], and HowTo100M [3]. A variety of tasks, focused primarily on text-video modelling, have been applied to these datasets including: task segmentation [19–21], reference resolution [22, 23], action segmentation [5], video clip ordering [24–27], and action recognition [28–30]. More related to our task is text-video modelling focused on learning a joint multimodal embedding space [3, 4, 31–36]. We build upon this work and remove the need for human generated textual summaries or ASR transcripts by learning from videos and their raw audio.

Much of the prior work on learning from video and audio has been focused on correlating objects in videos with the sounds they produce (e.g., sight and sound of musical instruments) as a signal for self-supervised learning of both audio and visual features [37–42]. This idea has been further developed for visually-guided audio source separation [43–48], sound generation from silent videos [39, 49], sound localization in video frames [50, 51], face and voice association [52, 53], and video-based audio localization [54, 55]. While our work has a similar focus on audio and video, instead of learning the source of audio in the visual domain, we focus on learning the semantic correlations between objects and their spoken language descriptions.

3 Methods

3.1 Unsupervised Audio-Visual Video Clip Sampling

Current approaches to learning language representations from videos rely on text annotations and do not primarily leverage the existing audio in videos [3, 4, 29, 30]. Formally, these approaches start with a corpus of videos $C = \{(A^i, V^i)\}_{i=1}^n$, where A^i and V^i denote the audio samples and visual sequence in the i^{th} video. Since videos can be several minutes long, they are further segmented into shorter clips using supervision such as human annotation or the silence boundaries from ASR transcripts. Once the clip boundaries are decided, this results in a corpus of clips $C_{\texttt{T-V}} = \{\{(T_j^i, V_j^i)\}_{j=1}^{m_i}\}_{i=1}^n$, where T_j^i and V_j^i denote the text caption and visual sequence in the j^{th} clip of the i^{th} video. The text caption is written by a human annotator or generated from ASR transcripts and replaces the audio in each clip.

In our work, we generate training samples from the corpus C without supervision by randomly segmenting each video into m clips of length t (which may overlap) to obtain a corpus of clips $C_{\texttt{A-V}} = \{\{(A_j^i, V_j^i)\}_{j=1}^{m}\}_{i=1}^n$. Unlike previous methods, we do not replace the raw audio with text. This procedure allows us to sample clips without supervised annotation and 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 (Section 4.4) show our models perform comparably whether trained on randomly sampled clips or on clips determined by ASR boundaries.

While our main contribution is developing self-supervised models that learn from $C_{\texttt{A-V}}$, we also show how to extend our models to use the textual summaries that exist in many video datasets. We use the text annotations from the corpus C to generate a corpus of clips $C_{\texttt{A-T-V}} = \{\{(A_j^i, T_j^i, V_j^i)\}_{j=1}^{m_i}\}_{i=1}^n$, where A_j^i, T_j^i, and V_j^i denote the audio samples, text caption, and visual sequence in the j^{th} clip of the i^{th} video.

3.2 Audio-Video Language Network Architecture

In this work, we introduce AVLnet — a self-supervised model architecture that learns the correlation between semantically related visual objects and audio, including speech, from video clips in $C_{\texttt{A-V}}$. The AVLnet architecture consists of parallel audio and visual branches as shown in Figure 1. The audio branch consists of a convolutional model with residual layers as proposed in Harwath et al. [2]. It takes in spectrograms as input and first outputs a temporal feature map with dimensions (L, D), where L is the downsampled temporal dimension and D is the dimension of the joint audio-visual embedding space. The feature map is then mean-pooled over the time dimension to obtain a D-dimensional vector $\mathbf{a}$. The visual branch consists of a 2D and 3D CNN feature extraction pipeline as in Miech et al. [3] (further described in in Section 4.2) and outputs a D-dimensional vector $\mathbf{v}$.

After the audio and visual features are extracted, we apply nonlinear gating [56] to both modalities:

$$f(\mathbf{a}) = (W_1^a \mathbf{a} + b_1^a) \circ \sigma(W_2^a(W_1^a \mathbf{a} + b_1^a) + b_2^a) \tag{1}$$

$$g(\mathbf{v}) = (W_1^v \mathbf{v} + b_1^v) \circ \sigma(W_2^v(W_1^v \mathbf{v} + b_1^v) + b_2^v), \tag{2}$$

where $f(\mathbf{a})$ and $g(\mathbf{v})$ represent the output language and visual embedding vectors respectively, $W_1^a, W_2^a, W_1^v, W_2^v$ 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.

The AVLnet architecture is able to learn visually grounded language without text captions. However, our model is also capable of incorporating the text captions in $C_{\texttt{A-T-V}}$, enabling it to utilize the textual information that exists in many instructional datasets. To incorporate text in AVLnet, we add a third branch that processes the text caption into a D-dimensional vector $\mathbf{t}$. Due to the complementary language information in the raw audio and text, we fuse the outputs of the audio and text branches before non-linear gating. Specifically, we modify Equation 1 as follows:

$$f(\mathbf{a}, \mathbf{t}) = (W_1^a \mathbf{a} + W_1^t \mathbf{t} + b_1^{a+t}) \circ \sigma(W_1^{a+t}(W_1^a \mathbf{a} + W_1^t \mathbf{t} + b_1^{a+t}) + b_2^{a+t}) \tag{3}$$

where $f(\mathbf{a}, \mathbf{t})$ 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. The visual embedding vector

(Equation 2) remains unchanged. We refer to this variant of AVLnet as AVLnet-Text. In Section 4.4, we show that our proposed fusion approach outperforms an alternative architecture in which audio and text are processed in independent branches.

3.3 Self-Supervised Training Procedure

Due to the self-supervised nature of AVLnet and AVLnet-Text, we use a contrastive loss function that maximizes the similarity between audio and video from the same clip while minimizing the similarity of audio paired with imposter video from another clip or video paired with imposter audio from another clip. Here we define the similarity between audio and video as the dot product of their learned embedding vectors. In particular, we utilize the Masked Margin Softmax (MMS) loss function [17], and explore other loss functions in Section 4.4.

Unlike the triplet loss function used in prior unsupervised audio-image modeling [2] that samples imposter pairs randomly or using negative mining, the MMS loss enables comparisons of positives with a wider range of negative samples. During training, we use a batch size of N videos and sample M clips per video, resulting in $B = MN$ video clips per batch. The MMS loss trains the model to discriminate between the true audio-visual embedding pairs ($\mathbf{a}_i$, $\mathbf{v}_i$), and all imposter pairs where either the audio $\mathbf{a}_i$ is paired with a visual imposter $\mathbf{v}_j^{imp}$, or the visuals $\mathbf{v}_i$ are paired with an audio imposter $\mathbf{a}_k^{imp}$. The indices (i, j, k) indicate the index of the video clip in the batch. The loss is defined as:

$$L = -\frac{1}{B}\sum_{i=1}^{B}\left(\log\frac{e^{f(\mathbf{a}_i)\cdot g(\mathbf{v}_i)-\delta}}{e^{f(\mathbf{a}_i)\cdot g(\mathbf{v}_i)-\delta}+\sum\limits_{\substack{k=1\\k\neq i}}^{B} e^{f(\mathbf{a}_k^{imp})\cdot g(\mathbf{v}_i)}}+\log\frac{e^{f(\mathbf{a}_i)\cdot g(\mathbf{v}_i)-\delta}}{e^{f(\mathbf{a}_i)\cdot g(\mathbf{v}_i)-\delta}+\sum\limits_{\substack{j=1\\j\neq i}}^{B} e^{f(\mathbf{a}_i)\cdot g(\mathbf{v}_j^{imp})}}\right) \quad (4)$$

To train AVLnet-Text, $f(\mathbf{a}_i)$ is replaced with $f(\mathbf{a}_i, \mathbf{t}_i)$. In other words, the audio sample and text caption from each clip are treated as inseparable and are sampled together. In our experiments, we fixed the margin hyperparameter $\delta = 0.001$, $N = 64$ videos, and $M = 32$ video clips.

4 Experiments

4.1 Experimental Details

We train AVLnet and AVLnet-Text on the 1.2 million instructional YouTube videos from the HowTo100M [3] 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. Since AVLnet-Text requires textual input, we train it on the video, audio, and text captions corresponding to the given clips (denoted by $C_{\texttt{A-T-V}}$ in Section 3.1); however, to reduce the amount of supervision in our method, we train AVLnet on the video and audio from randomly segmented clips (denoted by $C_{\texttt{A-V}}$ in Section 3.1).

After training on HowTo100M, we evaluate and fine-tune our models on three established video and language datasets: YouCook2 [5], MSR-VTT [6], and LSMDC [7]. Each dataset provides human-annotated video clip boundaries and text summaries of the clips (details in the Appendix). We evaluate our models on the video clip and language retrieval tasks, in which a language query (audio or audio and text) is used to retrieve video and vice versa. In contrast to prior models applied to video and text annotations, our AVLnet model operates on the raw audio available in the clips. For both retrieval tasks, we use standard recall metrics R@1, R@5, R@10, and the median rank (Md. R).

4.2 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. During training we use 10 seconds of audio per video clip for HowTo100M, 50 seconds for YouCook2, and 30 seconds from MSR-VTT and LSMDC due to variation in clip length per dataset.

In the AVLnet visual branch, the 2D features are extracted at 1 feature per second using a ResNet-152 model [57] pretrained on ImageNet [58]. The 3D features are extracted at 1.5 features per second

Table 1: Video clip and language retrieval results. The best bi-modal and tri-modal results are bolded.

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. [3]	HT100M	T→V	6.1	17.3	24.8	46	V →T	5.3	16.5	25.2	42
Miech et al. [4]	HT100M	T→V	15.1	38.0	51.2	10	—	—	—	—	—
Miech et al. [3]	HT100M + YC2	T→V	8.2	24.5	35.3	24	V →T	7.2	22.8	34.3	24
Ours, AVLnet	HT100M	A→V	20.7	43.7	54.3	8	V →A	20.0	42.5	52.1	9
Ours, AVLnet	HT100M + YC2	A→V	**25.8**	**51.9**	**63.0**	**5**	V →A	**26.9**	**52.4**	**62.0**	**5**
Ours, AVLnet-Text	HT100M	T+A→V	25.6	52.7	64.4	5	V →T+A	29.3	55.3	65.5	4
Ours, AVLnet-Text	HT100M + YC2	T+A→V	**33.2**	**61.0**	**71.5**	**3**	V →T+A	**34.0**	**62.4**	**72.5**	**3**

(a) YouCook2

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. [3]	HT100M	T→V	7.5	21.2	29.6	38	V →T	8.4	21.3	28.9	42
Amrani et al. [36]	HT100M	T→V	8.0	21.3	29.3	33	—	—	—	—	—
Miech et al. [4]	HT100M	T→V	9.9	24.0	32.4	29.5	—	—	—	—	—
Miech et al. [3]	HT100M + MSR-VTT	T→V	14.9	40.2	52.8	9	V →T	16.8	41.7	**55.1**	**8**
Amrani et al. [36]	HT100M + MSR-VTT	T→V	17.4	**41.6**	**53.6**	**8**	—	—	—	—	—
Ours, AVLnet	HT100M	A→V	14.9	32.0	40.3	17	V →A	15.9	32.9	41.4	17
Ours, AVLnet	HT100M + MSR-VTT	A→V	**18.7**	40.9	51.0	9	V →A	**20.9**	**43.3**	51.8	8.5
JSFusion [34]	MSR-VTT	T→A+V	10.2	31.2	43.2	13	—	—	—	—	—
JPoSE [35]	MSR-VTT	T→A+V	14.3	38.1	53.0	9	V+A →T	16.4	41.3	54.4	8.7
CE [33]	MSR-VTT	T→A+V	20.9	48.8	62.4	6	V+A →T	20.6	50.3	64.0	5.3
Ours, AVLnet-Text	HT100M	T+A→V	19.6	40.8	50.7	9	V →T+A	19.7	43.0	54.9	8
Ours, AVLnet-Text	HT100M + MSR-VTT	T+A→V	**27.1**	**55.6**	**66.6**	**4**	V →T+A	**28.5**	**54.6**	**65.2**	**4**

(b) MSR-VTT

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. [3]	HT100M	T→V	4.0	9.8	14.0	137	V →T	2.4	8.1	11.8	154
Amrani et al. [36]	HT100M	T→V	4.2	11.6	17.1	119	—	—	—	—	—
Miech et al. [3]	HT100M + LSMDC	T→V	7.1	19.6	27.9	40	V →T	**6.6**	17.8	25.9	50
Amrani et al. [36]	HT100M + LSMDC	T→V	**6.4**	**19.8**	**28.4**	**39**	—	—	—	—	—
Ours, AVLnet	HT100M	A→V	2.4	7.6	10.1	167	V →A	1.7	6.7	11.7	207.5
Ours, AVLnet	HT100M + LSMDC	A→V	5.0	18.1	26.1	**39**	V →A	5.9	**19.7**	**27.4**	**40**
JSFusion [34]	LSMDC	T→A+V	9.1	21.2	34.1	36	—	—	—	—	—
CE [33]	LSDMC	T→A+V	11.2	26.9	34.8	25.3	—	—	—	—	—
Ours, AVLnet-Text	HT100M	T+A→V	4.4	10.6	15.3	105.5	V →T+A	3.8	11.3	15.9	109
Ours, AVLnet-Text	HT100M + LSMDC	T+A→V	**17.0**	**38.0**	**48.6**	**11**	V →T+A	**16.5**	**37.6**	**47.6**	**13**

(c) LSMDC

using a ResNeXt-101 model [59] pretrained on Kinetics [60]. For both architectures, we use the pretrained models from PyTorch [61] and feature extraction implementation provided by Miech et al. [3]. The output of each model is max-pooled over the time dimension and concatenated, resulting in a single 4096-dimensional visual embedding vector for each clip. When training AVLnet, we do not update the weights of the 2D and 3D feature extractors due to GPU memory limitations.

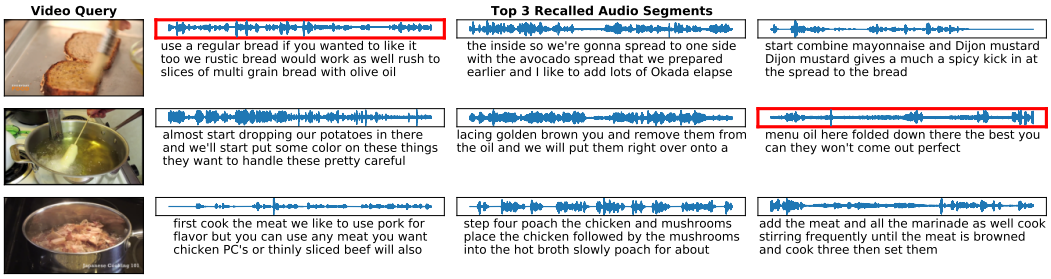
In the AVLnet-Text branch, we generate text features using a feature extraction pipeline [3] that generates word embeddings from a GoogleNews pretrained Word2vec model [62] and max-pools over the embeddings of the words in each clip's text caption. Although this text model is shallower than our audio model, a study of deeper text models for learning a text-video embedding found little improvement over this simple text model [4].

We minimize the MMS loss (Equation 4) with Adam [63] using a learning rate of $1e{-}3$ for AVLnet and $1e{-}4$ for AVLnet-Text. We trained each model on two V100 GPUs for 15 epochs, where each epoch took 10 hours.

4.3 Video Clip and Language Retrieval

As described in Section 4.1, we evaluate AVLnet and AVLnet-Text on the video clip retrieval and language retrieval tasks on three datasets: YouCook2 [5], MSR-VTT [6], and LSMDC [7]. For video

(a) Video clip retrieval examples. Each row displays the top recalled video clips (right) to the given audio (left).



(b) Audio retrieval examples. Each row displays the top recalled audio (right) to the given video clip query (left). For readability, each transcript is truncated to be at most 138 characters.

Figure 2: Retrieval results from AVLnet fine-tuned on YouCook2. We display video clips as their center frame. We display audio clips as their waveform and ASR transcript (transcribed using the IBM Speech-To-Text service [64]). The ASR transcripts contain mistakes, but are only used for visualization given AVLnet operates on raw audio. The correct match is highlighted.

clip retrieval, AVLnet retrieves video clips given input audio and AVLnet-Text retrieves video clips given input audio and text. We compare our models to state-of-the-art text-video models that retrieve video clips given text [3, 4, 36] and text-video models that additionally leverage audio [33–35]. In contrast to our work, these methods encode audio jointly with video frames instead of with text. Further, Liu et al. [33] use a pre-trained audio branch and additional models such as ASR, while our audio branch is not pre-trained. For language retrieval, given a video query, AVLnet retrieves audio and AVLnet-Text retrieves paired audio and text. We compare our models to state-of-the-art text-video models that retrieve text given input video and audio [35], or given video, audio, and ASR transcripts [33]. Since prior language retrieval results were only available on MSR-VTT, we also evaluated the text-video model provided by Miech et al. [3] on the language retrieval task.

Our models' video and language retrieval results are shown in Table 1. AVLnet outperforms all prior models on YouCook2 (zero-shot and fine-tune), with a 11.8% and 27.7% absolute increase in performance at R@10 over the previous state-of-the-art for the video clip and language retrieval tasks respectively. Compared with state-of-the-art text-video results [3, 4, 36], AVLnet achieves higher zero-shot and similar fine-tune performance on MSR-VTT, and similar performance to the baselines on LSMDC. Overall, our results indicate that AVLnet has learned powerful language representations despite never seeing text and only using raw audio.

The AVLnet-Text model further improves performance by leveraging text captions along with the raw audio and establishes state-of-the-art performance on all datasets. In the video clip retrieval task, AVLnet-Text improves the previous state-of-the-art R@10 score by 20.3% on YouCook2, 4.2% on MSR-VTT, and 13.8% on LSMDC. AVLnet-Text further improves the R@10 score on the language retrieval task by 38.2% on YouCook2, 1.2% on MSR-VTT, and 21.7% on LSMDC.

To better understand the performance gains our models achieve over state-of-the-art, we analyze retrieval examples from our AVLnet model fine-tuned on YouCook2. We show video and language retrieval examples from the YouCook2 validation set in Figure 2 (additional examples are shown in the Appendix). We find the retrieved results display high semantic similarity to salient content in the query. For example, in the top row in Figure 2b, the query video clip shows oil spread on bread

Table 2: Analysis of AVLnet and AVLnet-Text design choices. Performance of each compared model is reported as video clip retrieval R@10.

Experiment	Configuration	YouCook2		MSR-VTT		LSMDC	
		Zero-Shot	Fine-Tune	Zero-Shot	Fine-Tune	Zero-Shot	Fine-Tune
AVLnet Baseline	Random clips, no text, MMS loss	54.3	63.0	**40.3**	51.0	10.1	26.1
(a) Downstream Text	Text + audio for fine-tuning/evaluation	49.3	**66.3**	37.0	**59.7**	**10.4**	**44.4**
(b) Clip Sampling	HowTo100M ASR clips	**57.6**	62.8	38.5	49.4	7.7	26.1
(c) Loss Function	Max-Margin Ranking Loss	27.4	39.1	29.8	39.3	6.7	24.2
	NCE Loss	51.6	60.5	**40.3**	49.0	10.0	26.8
AVLnet-Text Baseline	Audio & Text Branch Fusion	**64.4**	**71.5**	50.7	**66.6**	15.3	**48.6**
(d) No Fusion	Independent Audio & Text Branches	57.0	65.5	**50.9**	64.9	**17.0**	48.0

and the retrieved audio contains the words 'bread' and 'spread'. This semantic relationship persists even when the correct clip is not the top result or is not in the top five results. For instance, in the bottom row of Figure 2a, the correct clip is not recalled in the top five results; however, the video and retrieved audio are related to chopping green onions. Further, we find our model has learned to relate natural sounds to salient video clips. The middle row of Figure 2a shows an example audio query containing only sizzling sounds; the ASR system fails as there was no speech, yet our model retrieved video clips of frying oil. These results suggest our model has learned the semantic relationships between speech, natural sounds, and visual clips.

4.4 Analysis of Proposed Methods

To verify the effectiveness of our proposed methods, we report the results of several additional experiments on AVLnet and AVLnet-Text in Table 2. For AVLnet, we investigate the effect of adding text to the audio-visual model during evaluation and fine-tuning, changing the HowTo100M clip sampling method, and varying the loss function. For AVLnet-Text, we investigate the effect of using independent audio and text branches as compared to our language fusion technique. We report each model's video clip retrieval R@10 on all three evaluative datasets in the zero-shot setting and after fine-tuning.

Adding text from the downstream datasets. We evaluate the performance of AVLnet trained on audio and video from HowTo100M and fine-tuned/evaluated on audio, video, and text from YouCook2, MSR-VTT, and LSMDC. This experiment represents the scenario where obtaining text annotations during training is expensive, but text exists or can be obtained for smaller evaluative datasets or real world applications. In Table 2(a), we observe that the fine-tuned performance is higher than audio-video AVLnet, but lower than AVLnet-Text, indicating that using ASR text captions during training on HowTo100M is beneficial. Further, this result suggests that AVLnet learns language representations from speech, not just natural sounds or voice characteristics.

HowTo100M clip selection. We compare our approach using randomly sampled HowTo100M video clips to train AVLnet to the approach of prior work [3, 4, 36] that used video clips segmented at the ASR speech boundaries. Table 2(b) shows AVLnet performs similarly on downstream tasks regardless of sampling method, suggesting our approach reduces the need for supervised clip sampling.

Loss functions. In Table 2(c), we compare the use of Masked Margin Softmax (MMS) loss [17], Max-Margin Ranking loss [3], and Noise Contrastive Estimation (NCE) loss [65, 66] to train AVLnet. We find the MMS loss and the NCE loss to outperform the Max-Margin Ranking loss, prompting us to use the MMS loss in our experiments.

Processing text with an independent branch. We study an alternative AVLnet-Text architecture that processies text in an independent branch instead of fusing the text and audio branches as in Equation 3. The MMS loss is applied 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. During evaluation, we use the sum of the audio and text embedding vectors to retrieve video clips. Table 2(d) shows that this approach performs worse than AVLnet-Text.

4.5 Audio-Visual Concept Discovery

To understand the audio-visual concepts learned by our models, we employ the unit visualization technique introduced by Zhou et al. [67]. In this procedure, we calculate the audio and visual purity

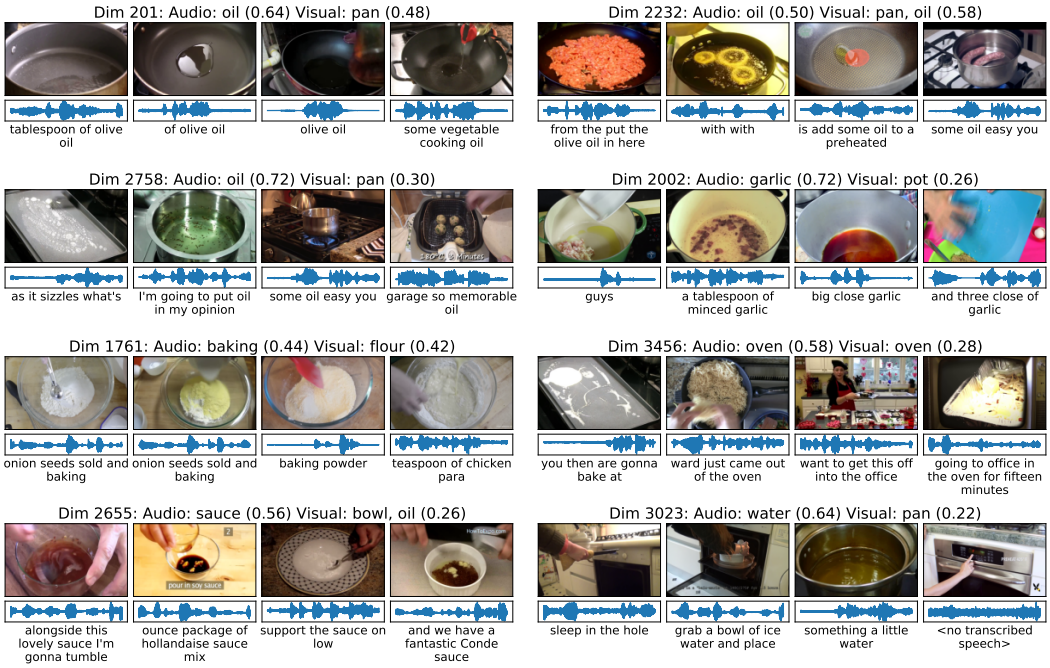


Figure 3: Top 8 dimensions sorted by geometric mean of their audio and visual purity displayed in row major order. Each dimension is represented as its top four visual features (shown as the clip's center frame) and top four frame-level audio features (shown as the frame's waveform and ASR transcript). The transcripts are shown for display purposes as AVLnet operates on video and raw audio.

of each dimension in AVLnet's learned audio-visual embedding space. We pass each YouCook2 validation clip through the AVLnet model trained on HowTo100M and fine-tuned on YouCook2 to extract its video and frame-level audio features. To extract frame-level audio features, we remove the temporal pooling layer from the audio branch. Each audio frame is mapped to the 2 seconds of audio surrounding it and the corresponding words during that time using the ASR transcripts. Each video clip is mapped to its set of food object labels given by the YouCook2 dataset [68]. Each dimension is given an audio label — defined as the word that occurs in the largest number of the dimension's top 50 maximally activating audio frames — and a visual label — defined as the food label that occurs in the largest number of the dimension's top 50 maximally activating video clips. We calculate audio purity and visual purity as the fraction of the dimension's top 50 maximally activating audio frames or video clips, respectively, that contain the dimension's label.

To identify dimensions that have learned audio-visual concepts, we sort all dimensions by the geometric mean of their audio and visual purity scores. The top 8 dimensions are shown in Figure 3 (additional dimensions are shown in the Appendix). Although the maximally activating video clips are chosen independently of the maximally activating audio, we find correspondences between the audio and visual content. For example, dimension 201's audio and visual labels are 'oil' and 'pan', and its maximally activating clips show pans of oil. Similarly, dimension 1761's labels are 'baking' and 'flour' and its top audio and visual frames contain language and visuals related to flour mixtures, and dimension 2655's audio label, visual label, and maximally activating clips are all related to bowls of sauce. These results suggest AVLnet has learned to align semantically related audio and visual features to particular dimensions of the embedding space.

5 Conclusion

In this paper, we present a novel self-supervised approach for learning audio-visual language representations from instructional videos. We circumvent the need for expensive and time consuming data annotation by introducing the AVLnet model that learns from audio naturally present in these

videos. This work establishes audio-video benchmarks on the YouCook2, MSR-VTT, and LSMDC video and language retrieval tasks and outperforms several text-video baselines. Further, we extend the AVLnet model to learn from audio, video, and text, leading to state-of-the-art performance on all downstream tasks. Finally, we show that training on natural audio from video enables our models to learn salient words and natural sounds, such as chopping and sizzling. Future work may include training the AVLnet visual branch on video frames instead of extracted visual features as well as further developing tri-branch architectures to learn from audio, video, and text.

Broader Impact

We have demonstrated a method 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. Our work could help scale the advancements in speech technologies developed for these languages, which would enable a greater number of people to interact more effectively with computers.

Grounded learning of audio and visual concepts is a fundamental problem in machine learning. We believe that developing grounded learning is promising for addressing problems such as bias, accountability, and robustness because it will allow systems to learn from a much broader variety of data modalities, in a way more analogous to the multi-faceted way in which people learn from their environment. As such we think this will mitigate many of the problems that exist in current AI systems that present inconsistent behavior patterns and give the appearance of malevolence, but actually reflect nothing more than inadequate learning mechanisms. Such systems will learn in more natural and explainable ways in contrast to current approaches which pick up on (often irrelevant) minute differences in data characteristics.

Our work here relies heavily on video datasets curated from YouTube (e.g., HowTo100M, YouCook2, MSR-VTT). 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. Further, there are ethical considerations using these datasets since YouTubers did not opt in to having their videos included in the datasets and videos used in this research may no longer exist publicly.

Acknowledgments and Disclosure of Funding

The authors are grateful for the support from the MIT-IBM Watson AI Lab.

References

[1] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2016, pp. 1858–1866.

[2] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 649–665.

[3] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 2630–2640.

[4] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[5] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, pp. 7590–7598.

[6] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5288–5296.

[7] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele, "Movie description," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 94–120, 2017.

[8] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 613–622.

[9] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 237–244.

[10] ——, "Learning word-like units from joint audio-visual analysis," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 506–517.

[11] K. Leidal, D. Harwath, and J. Glass, "Learning modality-invariant representations for speech and images," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2017, pp. 424–429.

[12] G. Synnaeve, M. Versteegh, and E. Dupoux, "Learning words from images and speech," *Neural Information Processing Systems (NeurIPS) Workshop on Learning Semantics*, 2014.

[13] A. Boggust, K. Audhkhasi, D. Joshi, D. Harwath, S. Thomas, R. Feris, D. Gutfreund, Y. Zhang, A. Torralba, M. Picheny, and J. Glass, "Grounding spoken words in unlabeled video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Sight and Sound Workshop*, 2019, pp. 29–32.

[14] D. Merkx, S. L. Frank, and M. Ernestus, "Language learning using speech to image retrieval," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.

[15] O. Scharenborg, L. Besacier, A. W. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stüker, P. Godard, M. Müller, L. Ondel, S. Palaskar, P. Arthur, F. Ciannella, M. Du, E. Larsen, D. Merkx, R. Riad, L. Wang, and E. Dupoux, "Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "Speaking Rosetta" JSALT 2017 workshop," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018.

[16] H. Kamper, G. Shakhnarovich, and K. Livescu, "Semantic speech retrieval with a visually grounded model of untranscribed speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, pp. 1–1, 09 2018.

[17] G. Ilharco, Y. Zhang, and J. Baldridge, "Large-scale representation learning from visually grounded untranscribed speech," in *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, 2019, pp. 55–65.

[18] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: a large-scale dataset for multimodal language understanding," *arXiv preprint arXiv:1811.00347*, 2018.

[19] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien, "Unsupervised learning from narrated instruction videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4575–4583.

[20] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic, "Cross-task weakly supervised learning from instructional videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3537–3545.

[21] F. Sener and A. Yao, "Unsupervised learning and segmentation of complex activities from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8368–8376.

[22] D.-A. Huang, J. J. Lim, L. Fei-Fei, and J. Carlos Niebles, "Unsupervised visual-linguistic reference resolution in instructional videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2183–2192.

[23] D.-A. Huang, S. Buch, L. Dery, A. Garg, L. Fei-Fei, and J. Carlos Niebles, "Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5948–5957.

[24] B. Fernando, H. Bilen, E. Gavves, and S. Gould, "Self-supervised video representation learning with odd-one-out networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3636–3645.

[25] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang, "Unsupervised representation learning by sorting sequences," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 667–676.

[26] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 527–544.

[27] D. Xu, J. Xiao, Z. Zhao, J. Shao, D. Xie, and Y. Zhuang, "Self-supervised spatiotemporal learning via video clip order prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 334–10 343.

[28] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 046–12 055.

[29] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 7464–7473.

[30] C. Sun, F. Baradel, K. Murphy, and C. Schmid, "Contrastive bidirectional transformer for temporal representation learning," *arXiv preprint arXiv:1906.05743*, 2019.

[31] A. Miech, I. Laptev, and J. Sivic, "Learning a text-video embedding from incomplete and heterogeneous data," *arXiv preprint arXiv:1804.02516*, 2018.

[32] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury, "Learning joint embedding with multimodal cues for cross-modal video-text retrieval," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018, pp. 19–27.

[33] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," *arXiv preprint arXiv:1907.13487*, 2019.

[34] Y. Yu, J. Kim, and G. Kim, "A joint sequence fusion model for video question answering and retrieval," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 471–487.

[35] M. Wray, D. Larlus, G. Csurka, and D. Damen, "Fine-grained action retrieval through multiple parts-of-speech embeddings," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 450–459.

[36] E. Amrani, R. Ben-Ari, D. Rotman, and A. Bronstein, "Noise estimation using density estimation for self-supervised multimodal learning," *arXiv preprint arXiv:2003.03186*, 2020.

[37] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 609–617.

[38] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2016, pp. 892–900.

[39] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman, "Visually indicated sounds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2405–2413.

[40] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba, "Ambient sound provides supervision for visual learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016, pp. 801–816.

[41] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2018, pp. 7763–7774.

[42] K. Yang, B. Russell, and J. Salamon, "Telling left from right: Learning spatial correspondence of sight and sound," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9932–9941.

[43] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 570–586.

[44] R. Gao, R. Feris, and K. Grauman, "Learning to separate object sounds by watching unlabeled video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 35–53.

[45] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[46] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba, "Self-supervised audio-visual co-segmentation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 2357–2361.

[47] H. Zhao, C. Gan, W.-C. Ma, and A. Torralba, "The sound of motions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1735–1744.

[48] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3879–3888.

[49] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, "Visual to sound: Generating natural sound for videos in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3550–3558.

[50] R. Arandjelovic and A. Zisserman, "Objects that sound," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 435–451.

[51] C. Gan, H. Zhao, P. Chen, D. Cox, and A. Torralba, "Self-supervised moving vehicle tracking with stereo sound," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7053–7062.

[52] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8427–8436.

[53] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2018, pp. 276–292.

[54] P. Morgado, N. Nvasconcelos, T. Langlois, and O. Wang, "Self-supervised generation of spatial audio for 360 video," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2018, pp. 362–372.

[55] R. Gao and K. Grauman, "2.5d visual sound," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 324–333.

[56] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on YouTube-8M Large-Scale Video Understanding*, 2017.

[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[58] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[59] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555.

[60] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.

[61] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8024–8035.

[62] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

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

[65] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010, pp. 297–304.

[66] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," *arXiv preprint arXiv:1602.02410*, 2016.

[67] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[68] L. Zhou, N. Louis, and J. J. Corso, "Weakly-supervised video object grounding from text by loss weighting and object interaction," in *British Machine Vision Conference*, 2018. [Online]. Available: http://bmvc2018.org/contents/papers/0070.pdf

A Appendix

A.1 Dataset Details

We train AVLnet and AVLnet-Text on instructional videos from the HowTo100M dataset [3] and evaluate our models on the YouCook2 instructional cooking video dataset [5], the MSR-VTT video dataset [6], and the LSMDC movie dataset [7][3].

HowTo100M The HowTo100M dataset [3] contains instructional YouTube videos from domains such as *home and garden*, *computers and electronics* and *food and entertaining*. At the time of download 1,166,089 videos were available on YouTube.

YouCook2 The YouCook2 dataset [5] 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 (e.g., *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. As in prior work [3], we evaluate on the validation clips because the test set does not contain text annotations. Following Miech et al. [3], we use 9,586 training clips and 3,350 validation clips.

MSR-VTT The MSR-VTT [6] 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 were available, resulting in 7,751 video clips. 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 [3, 34]. For consistency, we count the 32 test clips without audio as mistakes in our retrieval calculations.

LSMDC The LSMDC dataset [7] 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. [3], we use 101,079 training clips and 1000 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.

[3] We downloaded the HowTo100M and MSR-VTT datasets from YouTube between Dec. 2019 - Mar. 2020. The numbers we report reflect the videos that were available at the time of download.

A.2 Additional Video and Language Retrieval Examples

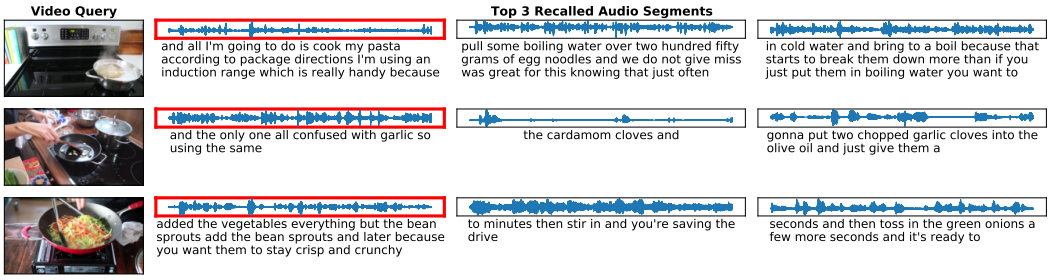
In Section 4.3, we analyze the video and language retrieval results of our model and show qualitative retrieval examples in Figure 2. We show additional video and language retrieval examples in Figures A1 and A2, respectively. These examples were generated using AVLnet trained on HowTo100M and fine-tuned on YouCook2. Consistent with our findings in Section 4.3, we find the recalled clips are often semantically related to the query clip.

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

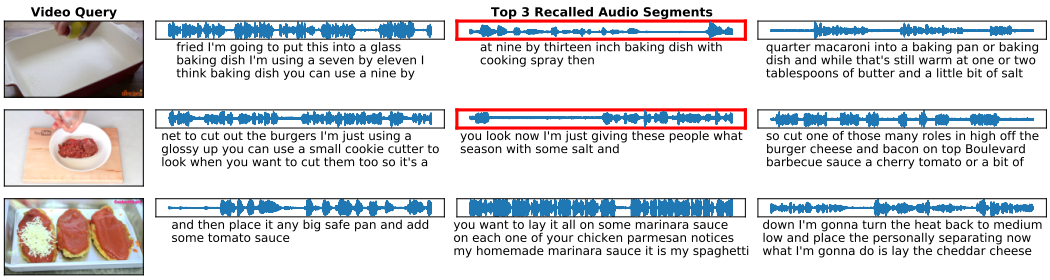
(b) Video clip retrieval examples for clips retrieved in the top 5 results (R@5).

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

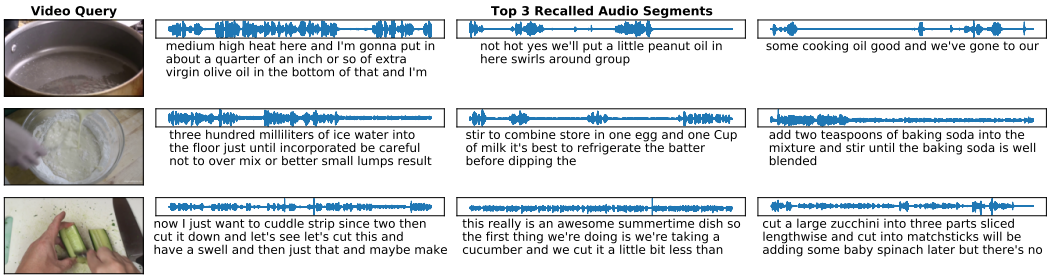
Figure A1: Additional video clip retrieval examples from the YouCook2 validation set. Each row displays the top recalled video clips (shown as each clip's center frame) to the given audio (shown as its waveform and ASR transcript). The ASR transcripts contain mistakes, but are only used for visualization 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).

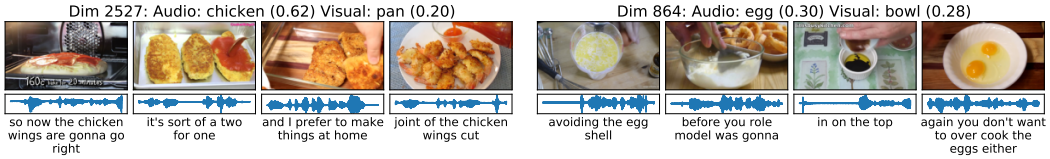


(c) Language retrieval examples for clips not retrieved in the top 5 results ($R > 5$).

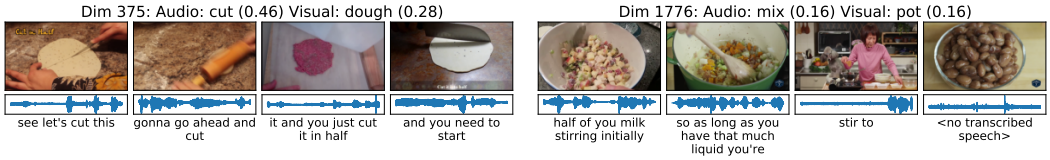
Figure A2: Additional language retrieval examples from the YouCook2 validation set. Each row displays the top recalled audio segments (shown as each segment's waveform and ASR transcript) to the given video (shown as its center frame). The ASR transcripts contain mistakes, but are only used for visualization given AVLnet operates on raw audio. The correct match is highlighted.

A.3 Additional Concept Discovery Examples

In Section 4.5, we show AVLnet learns to relate semantically related audio and visual features to dimensions of the shared embedding space. In Figure A3, we show six additional dimensions that exhibit salient relationships between their maximally activating audio and visual segments. In particular, Figure A3a shows dimensions that activate on words such as 'chicken' and 'egg' and Figure A3b shows dimensions that actions such as 'cut' and 'stir'. In Figure A3c we show dimensions that activate on natural sounds (e.g., sizzling and chopping) as opposed to speech.



(a) Word activated dimensions.



(b) Action activation dimensions.

(c) Natural sound activated dimensions.

Figure A3: Additional dimensions whose maximally activating audio and video features are semantically related. Each dimension is represented as its top four visual features (shown as the clip's center frame) and top four frame-level audio features (shown as the frame's waveform and ASR transcript). The transcripts are shown for display purposes as AVLnet operates on video and raw audio.