INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

AVLnet: Learning Audio-Visual Language Representations from Instructional Videos

Andrew Rouditchenko[1*], *Angie Boggust*[1*], *David Harwath*[2], *Brian Chen*[3], *Dhiraj Joshi*[4], *Samuel Thomas*[4], *Kartik Audhkhasi*[5], *Hilde Kuehne*[4], *Rameswar Panda*[4], *Rogerio Feris*[4], *Brian Kingsbury*[4], *Michael Picheny*[6], *Antonio Torralba*[1], *James Glass*[1]

[1]MIT CSAIL, USA
[2]UT Austin, USA
[3]Columbia University, USA
[4]IBM Research AI, USA
[5]Google, USA
[6]NYU, USA

roudi@mit.edu

Abstract

Current methods for learning visually grounded language from videos often rely on text annotation, such as human generated captions or machine generated automatic speech recognition (ASR) transcripts. In this work, we introduce the Audio-Video Language Network (AVLnet), a self-supervised network that learns a shared audio-visual embedding space directly from raw video inputs. To circumvent the need for text annotation, we learn audio-visual representations from randomly segmented video clips and their raw audio waveforms. We train AVLnet on HowTo100M, a large corpus of publicly available instructional videos, and evaluate on image retrieval and video retrieval tasks, achieving state-of-the-art performance. Finally, we perform analysis of AVLnet's learned representations, showing our model utilizes speech and natural sounds to learn audio-visual concepts.
Index Terms: audio-visual, multimodal learning, self-supervised learning, video retrieval, spoken captions

1. Introduction

Humans learn to understand language, recognize objects, and identify correspondences between the two by recognizing patterns in what they see and what they hear. Researchers have developed machine learning models similarly capable of relating spoken words to semantically relevant images [1–8]. 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. However, these methods require the collection of recorded spoken captions, limiting their scalability to other languages and visual contexts.

Videos 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 [9–15] have been applied to cross-modal tasks focused on identifying non-speech sounds and localizing the objects that produced them. We instead focus on relating spoken words to visual entities in videos such as objects and actions, which is a challenging task since human speech is semantically complex and the objects of interest do not produce the sound. Towards this goal, we use instructional videos which provide opportunities to learn semantic relationships between raw speech and visual entities given the narration naturally present in them.

[*] Equal contribution.

A common approach for learning from instructional videos is to develop text-video models that learn a multi-modal embedding space. These models typically do not incorporate the audio signal, but even models that do [16–22] still require text captions. To collect captions, some methods rely on humans to generate visual descriptions [23]. Unlike raw audio which can be noisy and nondescript, human-generated text provides a clean, visually salient signal; however, collecting text descriptions is time-consuming and 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 [24–28]. However, ASR transcripts process the continuous speech signal into discrete words, which limits words to a certain vocabulary and misses the opportunity to learn from visually relevant non-speech sounds. Further, models trained on ASR transcripts are inapplicable to the 98% of languages for which ASR is unavailable [29]. For these reasons, our goal is to learn from the raw audio and visual channels in videos without any additional annotation or ASR transcripts.

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. We train AVLnet on HowTo100M [25], 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 [3]. We propose video retrieval tasks on three video datasets, YouCook2 [23], CrossTask [30], and MSR-VTT [31]. 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. Our code, data, and trained models will be released at http://avlnet.csail.mit.edu

2. Technical Approach

2.1. Audio-Video Models

The AVLnet architecture (Figure 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

http://dx.doi.org/10.21437/Interspeech.2021-1312

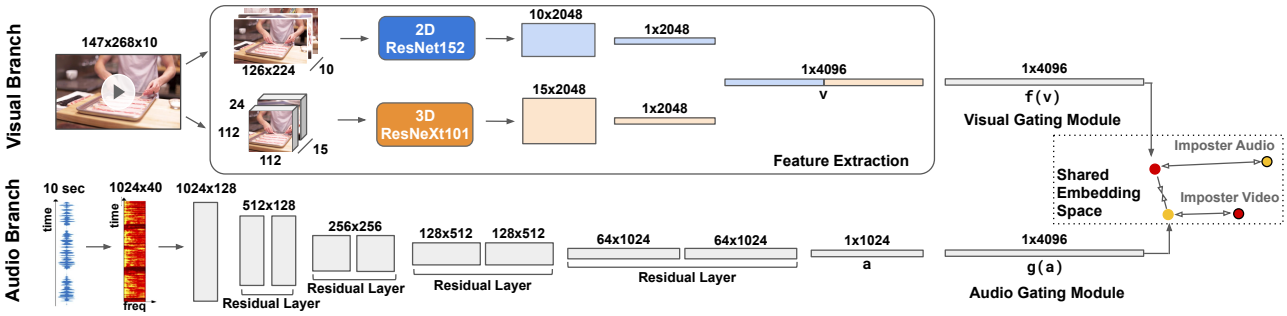

Figure 1: *The Audio-Video Language Network (AVLnet) model consists of video and audio branches, non-linear feature gating, and an audio-video embedding space. The model is trained through self-supervision and applied to image and video retrieval tasks.*

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 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 [33] and 3D video features to obtain 1.5 features per second using a ResNeXt-101 model [34] pretrained on Kinetics [35]. 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 [3] 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 [24, 25], our audio model is not pretrained, so it can be applied to videos in any language, including those for which ASR is not available.

2.2. 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 [36] 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. In Section 4, we analyze the embedded dimensions and show the model indeed activates for similar concepts along the same dimension in both audio and video modalities. Non-linear gating is defined as:

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

$$g(\mathbf{a}) = (W_1^a \mathbf{a} + b_1^a) \circ \sigma(W_2^a (W_1^a \mathbf{a} + b_1^a) + b_2^a) \quad (2)$$

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.

2.3. Contrastive Loss for Audio-Video Retrieval

Due to the self-supervised nature of AVLnet, we use the Masked Margin Softmax (MMS) loss [5], 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 $(\mathbf{a}_j^{\text{imp}}, \mathbf{v}_i)$. Unlike the triplet loss used in prior unsupervised audio-image modeling [3] that samples imposter pairs randomly or via negative mining, the MMS loss enables comparisons of positives with a wider range of negatives. The loss is defined as $L(f(\mathbf{v}), g(\mathbf{a})) + L(g(\mathbf{a}), f(\mathbf{v}))$. We modify the loss to exclude the masking component, because it is inapplicable to our procedure where each clip contains only one ground truth audio-video pair. During training, we use a batch of N videos and sample M clips per video, resulting in $B = MN$ video clips per batch. Since L (Eq. 3) is applied post non-linear gating, we pass the gated embeddings $f(\mathbf{v})$ and $g(\mathbf{a})$ to the function.

$$L(\mathbf{x}, \mathbf{y}) = -\frac{1}{B} \sum_{i=1}^{B} \left(\log \frac{e^{\mathbf{x}_i \cdot \mathbf{y}_i - \delta}}{e^{\mathbf{x}_i \cdot \mathbf{y}_i - \delta} + \sum_{\substack{j=1 \\ j \neq i}}^{B} e^{\mathbf{x}_i \cdot \mathbf{y}_j^{\text{imp}}}} \right) \quad (3)$$

We note that the MMS loss function can be seen as two applications of InfoNCE [37] (with a margin), however, the negatives are sampled from both within the same video and from others.

2.4. Video Clip Sampling

Given a corpus of unlabeled instructional videos, we generate training samples by randomly segmenting each video into M clips of length t (which may overlap) to obtain a corpus of clips. 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 (Section 3.4) show our model performs comparably whether trained on randomly sampled clips or on clips determined by ASR boundaries.

3. Experiments

3.1. Implementation Details

We train AVLnet on the 1.2 million instructional YouTube videos in the HowTo100M [25] dataset using our random clip sampling technique. The audio input is represented as a log Mel spectrogram (16 kHz sampling rate, 25 ms Hamming window, 10 ms window stride, 40 Mel filters). We extract 2D and 3D visual features following Miech et al. [25]. During training, we do not update the feature extractor weights due to GPU memory limitations. We use a batch of $N = 128$ videos and sample $M = 32$

Table 1: *Video clip and language retrieval results on YouCook2, CrossTask, and MSR-VTT. Models trained on: (1) target dataset only (no pretraining); (2) HowTo100M only (zero-shot); (3) HowTo100M and target dataset (pretrain and fine-tune). All models use pretrained visual features. Baseline models are from Boggust et al. [38] and Arandjelović et al. [11].*

Method	YouCook2						CrossTask						MSR-VTT					
	Video Clip (A→V)			Language (V→A)			Video Clip (A→V)			Language (V→A)			Video Clip (A→V)			Language (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) [38]	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) [11]	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) **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) [38]	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) [11]	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) **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) [38]	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) [11]	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) **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**

clips per video, each $t = 10$ seconds long. We minimize the MMS loss with Adam [39] using a learning rate of $1e{-}3$ and margin hyperparameter of $\delta = 0.001$. We train each model on 2 V100 GPUs for 30 epochs, which takes $\sim$2 days.

3.2. Experimental Setup

Audio-Image Retrieval. Since instructional videos and spoken captions of images both contain descriptive speech of visual scenes, learning from instructional videos could 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 in the Places Audio Caption [3]. The dataset contains 400k images from the Places205 dataset [40] paired with 1,000 hours of unscripted spoken captions. We evaluate the performance on audio to image and image to audio retrieval tasks. Following the prior work, results are reported on the validation set. We use the standard recall metrics R@1, R@5, and R@10.

Audio-Video Retrieval. We fine-tune and evaluate our model on two instructional video datasets: YouCook2 [23] and CrossTask [30]. While YouCook2 contains cooking videos, CrossTask contains a wider range of instructional videos. We also fine-tune and evaluate on MSR-VTT [31] 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. 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 [38]. This procedure tests our model's capability for video search using audio and spoken queries, without needing to transcribe speech. We report results in the no-pretraining, zero-shot, and fine-tuned settings.

3.3. Comparison to State-of-the-art

Audio-Image Retrieval. In this experiment, we train AVLnet on HowTo100M using the 2D CNN features so that it 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 2, we compare prior models trained only on Places-400k [2–4, 41, 42] 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. We also show the results of concurrent work [8] achieving similar results with different audio features and pretraining datasets.

Table 2: *Retrieval on Places using 400k training set. ‡Results found in [3]. †Obtained using official code. *Concurrent work.*

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. [2]‡	14.8	40.3	54.8	12.1	33.5	46.3
Harwath et al. [41]‡	16.1	40.4	56.4	13.0	37.8	54.2
DAVEnet [3]	20.0	46.9	60.4	12.7	37.5	52.8
ResDAVEnet [4]	27.6	58.4	71.6	21.8	55.1	69.0
ResDAVEnet-VQ [42]†	34.9	70.2	79.4	32.7	65.6	77.0
MILAN [8]*	**53.4**	**79.1**	86.3	**53.0**	**78.2**	**85.6**
Ours, AVLnet	44.8	76.9	**86.4**	42.8	76.2	84.8

Audio-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. [38] only uses the center image frame from each video clip during training and inference. The model from Arandjelović et al. [11] 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 1 shows the retrieval results on YouCook2, CrossTask, and MSR-VTT in the no-pretraining, zero-shot, and fine-tuned settings. The performances on video clip retrieval (A→V) and language retrieval (V→A) are similar for the same target dataset. When trained only on the target dataset, the models all perform comparably. Training on HowTo100M significantly improves the performance in the zero-shot and fine-tuned settings, suggesting that large-scale pretraining is essential. This is true across all datasets, including on YouCook2 and CrossTask which contain instructional videos similar in content to HowTo100M videos, and on MSR-VTT which contains general videos. AVLnet outperforms the baseline models, especially in the zero-shot and fine-tuned settings, and achieves significant performance on all datasets regardless of the domain.

3.4. Ablation Studies

We evaluate our design choices via ablation studies comparing each model's video clip retrieval on YouCook2 and

Table 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	2D features only	51.6	57.9	32.6	37.9
	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 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

CrossTask (Table 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 [43].

Next, we evaluate loss functions. MMS [5] outperforms MIL-NCE [24], Binary Cross Entropy [10], Max-Margin Ranking [25], and InfoNCE [37]. 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 [3,4] use spoken captions that are typically 20s, and text-video models [24] 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.

3.5. 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 [44]. 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-241 (Table 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 (4.1%), demonstrating AVLnet also detects relevant cues in natural sounds.

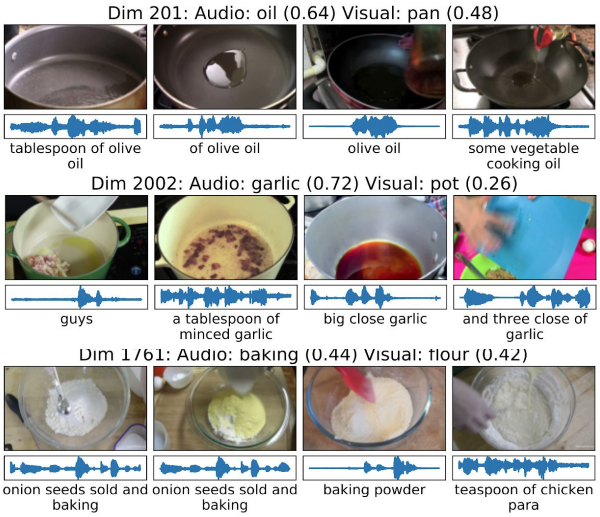

Figure 2: *AVLnet aligns audio-visual concepts to latent dimensions. The top 3 dimensions are shown as their maximally activating visuals (center frame) and audio (waveform and transcript).*

4. Audio-Visual Concept Discovery

To discover audio-visual concepts learned by AVLnet, we apply unit visualization [45] to the multi-modal embedding space and identify dimensions that activate for semantically similar audio and visual inputs. We first extract audio and visual embeddings for each YouCook2 validation clip, and identify the top 50 inputs that maximally activate each dimension. We remove the temporal pooling layer from AVLnet's audio branch to get word-level audio embeddings. Each audio embedding is mapped to the ASR-detected words in the surrounding 2 seconds, and each visual embedding is mapped to a set of food labels provided by YouCook2 [46]. To systematically identify dimensions that activate for similar concepts, we label each dimension with the most frequent food label and word in its maximally activating inputs. We then compute each dimension's audio and visual purity as the fraction of its maximally activating inputs that contain the correct label. We rank dimensions by geometric mean of their audio and visual purity (Figure 2). Although the maximally activating visuals and audio are chosen independently, we find strong semantic correlations, suggesting AVLnet has learned audio-visual concepts from raw instructional video.

5. Conclusion

We present a self-supervised method for learning audio-video representations from instructional videos with the goal of relating 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. We plan to investigate the model's ability to learn representations in other languages as future work.

6. Acknowledgements

This research was supported by the MIT-IBM Watson AI Lab.

7. References

[1] G. Synnaeve, M. Versteegh, and E. Dupoux, "Learning words from images and speech," *NeurIPS Workshop on Learning Semantics*, 2014.

[2] D. Harwath, A. Torralba, and J. Glass, "Unsupervised learning of spoken language with visual context," in *NeurIPS*, 2016.

[3] 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 *ECCV*, 2018.

[4] ——, "Jointly discovering visual objects and spoken words from raw sensory input," *IJCV*, 2020.

[5] G. Ilharco, Y. Zhang, and J. Baldridge, "Large-scale representation learning from visually grounded untranscribed speech," in *CoNLL*, 2019.

[6] G. Chrupała, L. Gelderloos, and A. Alishahi, "Representations of language in a model of visually grounded speech signal," in *ACL*, 2017.

[7] D. Merkx, S. L. Frank, and M. Ernestus, "Language learning using speech to image retrieval," in *INTERSPEECH*, 2019.

[8] R. Sanabria, A. Waters, and J. Baldridge, "Talk, don't write: A study of direct speech-based image retrieval," *arXiv preprint arXiv:2104.01894*, 2021.

[9] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *NeurIPS*, 2016.

[10] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *ICCV*, 2017.

[11] ——, "Objects that sound," in *ECCV*, 2018.

[12] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *ECCV*, 2018.

[13] B. Korbar, D. Tran, and L. Torresani, "Cooperative learning of audio and video models from self-supervised synchronization," in *NeurIPS*, 2018.

[14] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *ECCV*, 2018.

[15] A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, and A. Torralba, "Self-supervised audio-visual co-segmentation," in *ICASSP*, 2019.

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

[17] 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 *ICMR*, 2018.

[18] M. Wray, D. Larlus, G. Csurka, and D. Damen, "Fine-grained action retrieval through multiple parts-of-speech embeddings," in *ICCV*, 2019.

[19] Y. Yu, J. Kim, and G. Kim, "A joint sequence fusion model for video question answering and retrieval," in *ECCV*, 2018.

[20] 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.

[21] N. Holzenberger, S. Palaskar, P. Madhyastha, F. Metze, and R. Arora, "Learning from multiview correlations in open-domain videos," in *ICASSP*, 2019.

[22] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman, "Self-supervised multimodal versatile networks," *NeurIPS*, vol. 33, 2020.

[23] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *AAAI*, 2018.

[24] 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 *CVPR*, 2020.

[25] 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 *ICCV*, 2019.

[26] C. Sun, F. Baradel, K. Murphy, and C. Schmid, "Learning video representations using contrastive bidirectional transformer," *arXiv preprint arXiv:1906.05743*, 2019.

[27] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *ICCV*, 2019.

[28] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: a large-scale dataset for multimodal language understanding," in *Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS, 2018.

[29] M. Prasad, D. van Esch, S. Ritchie, and J. F. Mortensen, "Building large-vocabulary asr systems for languages without any audio training data." in *INTERSPEECH*, 2019.

[30] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic, "Cross-task weakly supervised learning from instructional videos," in *CVPR*, 2019.

[31] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *CVPR*, 2016.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[34] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *CVPR*, 2018.

[35] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.

[36] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," *CVPR Workshop on YouTube-8M Large-Scale Video Understanding*, 2017.

[37] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[38] 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 *CVPR Sight and Sound Workshop*, 2019.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[40] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NeurIPS*, 2014.

[41] D. Harwath and J. Glass, "Learning word-like units from joint audio-visual analysis," in *ACL*, 2017.

[42] D. Harwath, W.-N. Hsu, and J. Glass, "Learning hierarchical discrete linguistic units from visually-grounded speech," in *ICLR*, 2020.

[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

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

[45] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," in *ICLR*, 2015.

[46] L. Zhou, N. Louis, and J. J. Corso, "Weakly-supervised video object grounding from text by loss weighting and object interaction," in *BMVC*, 2018.