Spoken ObjectNet: A Bias-Controlled Spoken Caption Dataset

Ian Palmer, Andrew Rouditchenko[†], Andrei Barbu, Boris Katz, James Glass[†]

MIT CSAIL, USA

{iapalm,roudi,abarbu,boris,glass}@mit.edu

Abstract

Visually-grounded spoken language datasets can enable models to learn cross-modal correspondences with very weak supervision. However, modern audio-visual datasets contain biases that undermine the real-world performance of models trained on that data. We introduce Spoken ObjectNet, which is designed to remove some of these biases and provide a way to better evaluate how effectively models will perform in real-world scenarios. This dataset expands upon ObjectNet, which is a bias-controlled image dataset that features similar image classes to those present in ImageNet.

We detail our data collection pipeline, which features several methods to improve caption quality, including automated language model checks. Lastly, we show baseline results on image retrieval and audio retrieval tasks. These results show that models trained on other datasets and then evaluated on Spoken ObjectNet tend to perform poorly due to biases in other datasets that the models have learned. We also show evidence that the performance decrease is due to the dataset controls, and not the transfer setting.

Index Terms: spoken captions, dataset, speech, bias, retrieval

1. Introduction

Prior work in unsupervised spoken language learning has shown that neural models can learn meaningful audio-visual correspondences from visually grounded speech [1–3]. This mode of learning is inspired by humans in early childhood, who learn to use speech to describe the world before learning any written language. In practice, this could allow audio-visual models to learn from vast corpora of unlabeled images and videos.

However, many datasets in existence today, including audio-visual datasets, contain intrinsic biases that the models trained on those datasets then learn, which in turn degrades their performance on real-world data. For example, most images and videos uploaded to the Internet are nicely lit, well-framed, and contain objects in their usual settings. In turn, image captioning models are biased towards describing people on beaches as happy and image classification models don't recognize wolves outside of a snowy backdrop [4].

ObjectNet, a large-scale bias-controlled object classification dataset, addressed these problems by collecting a corpus of entirely new images instead of relying on those already uploaded to the Internet in some form [5]. Workers were asked to position a variety of household objects in a certain way against a specified background. The viewpoint of the camera was also controlled. In this way, ObjectNet has systematic controls in place for some of the biases that most other datasets exhibit.

In this work, we introduce Spoken ObjectNet (SON), a large-scale corpus of spoken image descriptions based on the ObjectNet dataset. Our dataset addresses some of the

biases present in existing audio-visual datasets. We introduce our data collection pipeline, which includes a novel language modeling step that increases the general quality and acceptance rate of worker submissions. Lastly, we conduct retrieval experiments to demonstrate that audio-visual models struggle to transfer to this bias-controlled domain, and the decreased performance is due to the controls and not just the transfer setting. We will release the dataset publicly at https://groups.csail.mit.edu/sls/downloads/placesaudio, and our code is available at https://github.com/iapalm/Spoken-ObjectNet.

2. Related Work

2.1. Spoken Caption Datasets

Spoken captions of images were originally collected to build models that learn to recognize words from semantic-level supervision without any form of automatic speech recognition [6]. Many spoken caption datasets have since been collected, and we show a comparison between them in Table 1. We focus on human-recorded captions in English, although spoken captions have been collected in Hindi [7] and Japanese [8, 9]. The datasets mainly vary in modalities (images or videos) and whether the speech is spontaneous or read from text captions.

The Places Audio Captions dataset [3] was the first large-scale dataset of spontaneous captions, and it contains over 400k spoken captions based on the Places 205 image dataset [10]. The captions were collected via Amazon Mechanical Turk, with an average caption length of 20 words and audio duration of 10 seconds. Spoken ObjectNet shares a similar data collection framework and is approximately equal in average sequence length and duration. However, Spoken ObjectNet is smaller in scale than Places Audio Captions and also features the controls for bias within the ObjectNet dataset. Spoken ObjectNet, then, can function as a test set for models trained on Places Audio Captions. Figure 1 shows samples from both datasets, demonstrating the stark differences in images and captions.

Localized Narratives [11] is also a recently collected large-scale spoken caption dataset with spontaneous speech. Several other datasets exist, but they are based on previous text captions and therefore contain fewer words on average per caption. The Flickr Audio Caption dataset was one of the first datasets and contains 40k read spoken captions based on the Flickr8k dataset and captions [6, 12]. Spoken captions have been collected for the Microsoft Common Objects in Context (MS-COCO) dataset as well, which features over one million text captions for a variety of images [13]. The SpokenCOCO [14] dataset contains human-recorded captions, while other datasets contain synthetic captions generated using text-to-speech systems [15, 16] which are less natural. Synthetic spoken captions have also been collected in other contexts [17, 18].

Researchers have recently collected spoken captions for videos. QuerYD [19] features audio descriptions for video seg-

---

[†] Corresponding author.

Spoken ObjectNet

a jar of honey being held on its side of the yellow lid in front of old flowers in a vase on a bathroom sink countertop

a straw like sun hat upside down on a brown sectional with a gray piece of clothing in the corner

someone holding a bottle of extra virgin olive oil upside down over a white washer with a plug-in behind it

Places Audio Captions

picture is a bridge runs along a body of water to the right is a tree with no leaves on a hill

a flower garden with a stone structure wooden gates and trees in the background

at the bottom of the ocean floor we see several yellow small fish they are swimming in between the green coral

Figure 1: *Samples of images, spoken captions, and ASR transcripts from Spoken ObjectNet and Places Audio [3].*

Table 1: *Comparison of human-recorded spoken caption datasets in English. Mod.=Modality; W./C.=Words per Caption; S.=Spontaneous; C.=Control for Bias.*

Dataset	Mod.	Samples	W./C.	S.	C.
QuerYD [19]	Video	31,441	19.9	✓	✗
Spoken Moments [20]	Video	515,912	18	✓	✗
Flickr Audio [6]	Image	40,000	10.9	✗	✗
SpokenCOCO [14]	Image	605,495	10	✗	✗
Loc. Narratives [11]	Image	873,107	36.5	✓	✗
Places Audio [3]	Image	402,385	21.9	✓	✗
Spoken ObjectNet	Image	50,273	20.5	✓	✓

ments annotated via YouDescribe, a volunteer-based community that adds narration to existing YouTube videos. Spoken Moments [20] contains spoken captions of 3s action video clips from Moments [21].

Lastly, several video datasets exist without spoken captions but can still be used to learn audio-visual representations. The How2 [22] and HowTo100M [23] instructional video datasets naturally contain spoken descriptions of visual events. AudioSet [24] and VGG-Sound [25] are both video datasets useful for audio-visual recognition tasks. Videos in VGG-Sound were selected and filtered using CNNs to curate a dataset with several hundred object classes. Similarly, we use automated language modeling checks in our data collection pipeline.

2.2. ObjectNet Dataset

The ObjectNet dataset is an object detection test set collected in a way which explicitly controls for object viewpoints, rotations, and backgrounds. Removing these priors from the images results in significant performance drops (approximately 45% for most models) versus performance on the ImageNet test set. The purpose of ObjectNet is to enable models that are more robust to real-world scenarios where objects may be in unusual contexts; similarly, the purpose of Spoken ObjectNet is to provide a test set for audio-visual models to measure how robustly they generalize to real-world situations.

2.3. Audio-Visual Models

Models for learning audio-visual correspondences typically learn an embedding space where true visual inputs and spo-

ken caption pairs are close, while non-matching pairs are far apart. In this work, we consider the ResDAVEnet [26] architecture which combines CNN-based audio and image models. The ResDAVEnet-VQ [27] architecture adds configurable vector quantization layers to the audio model. Several other models that learn audio-visual correspondences from both images and videos have been presented in recent work [28–32].

3. Spoken ObjectNet Dataset Collection

To collect samples for this dataset, we extended the approach used to collect the Places Audio Caption dataset [3]. We released an Amazon Mechanical Turk (AMT) Human Intelligence Task (HIT), which allowed workers to submit captions for four images in the ObjectNet dataset at a time. Workers were compensated $0.20 for four recordings that passed our validation steps. Workers were prohibited from submitting more than 3,000 HITs to prevent speaker bias from impacting the dataset.

During data collection, workers were given an image and asked to record themselves as if they were describing the image to someone who could not see it. Workers were told they could describe shapes, objects, locations, colors, and anything else of interest as they saw fit. After each recording was completed, we ran several validation steps on the recorded audio to ensure that it met our requirements. If a worker failed a validation step, they were immediately asked to redo the recording. We found that providing this feedback in real time, as opposed to rejecting the HIT outright hours or days later, increased the rate at which we could collect high-quality samples and improved the experience for workers. After four recordings were completed, workers could submit the assignment and proceed to the next HIT.

3.1. Validation

Each recording had to pass three checks in order for the worker to proceed to the next image. First, the recording had to be at least 1s in duration. This prevented workers from simply clicking through the screens as fast as possible in order to complete the task. The recording was also run through the Google Speech Recognition API to generate an ASR transcript of the recording. We required that each recording have at least four words in the transcript to be accepted. This prevented workers from recording silence or other non-speech sounds.

Lastly, we introduced a new step in which the ASR tran-

Table 2: *Vocabulary comparison of Spoken ObjectNet (SON-50k) and Places Audio (including our 50k split Places-50k).*

Dataset	Audio	Words	Nouns	Verbs	Adj.	Adv.
SON-50k	155h	18,780	11,666	3,252	2,324	478
Places-50k	115h	20,140	11,212	4,332	2,963	620
Places-400k	944h	51,764	27,074	11,293	8,271	1,660

script was fed into a BERT model with a language modeling head. We used this model to produce a numerical score to approximate how well-formed the ASR text was. The model was a `BertForMaskedLM` model from the Python Huggingface library [33], and our score is based on the cross entropy loss between the model's predictions based on the masked input tokens and the ground truth tokens. Any transcript that scores above a certain threshold (where higher scores are predicted to be less grammatical) failed the validation step. Given the unusual contexts of the objects and the potential for ASR errors, a low cutoff score could frustrate workers who were attempting to complete the task properly, so we used existing collected samples to measure a cutoff score that would prevent blatantly non-grammatical captions from passing. Overall, this approach increased average caption quality, increased our HIT acceptance rate, and reduced the amount of manual validation that was required.

3.2. Speaker Information

In practice, audio-visual models may learn information about the speaker's microphone instead of the content of the speech signal. If a worker annotated examples of primarily one class (e.g. all images from the `measuring_cup` class), the model could exploit that correlation during training to make predictions without ever learning from the spoken words. To combat this, speakers were presented with images in a random order, so models cannot exploit speaker information to predict class identity. The train, validation, and test splits were also constructed such that there is no speaker overlap between any of the sets.

3.3. Finalizing Splits

In total, we collected over 70,000 samples. One sample per image in ObjectNet was selected to form the Spoken ObjectNet-50k dataset, with a total of 50,273 samples. 48,273 are included in the training set, and 1,000 are included in both the validation and test sets. We compare the vocabulary of the dataset with Places Audio [3] in Table 2, where we find that our captions contain a similar part-of-speech distribution to a split of Places with 50k captions.

4. Retrieval Experiments

4.1. Experimental Setup

Because Spoken ObjectNet is best understood as a test set relative to a dataset like Places Audio [3], our experiments are based on these two datasets. We also created a split of Places Audio that is the same size as Spoken ObjectNet, with 48,273 training samples randomly selected from the original 400k and the same 1,000 sample validation set as the original Places Audio. This split allows us to control for data quantity in our experiments, and we refer to the split as Places-50k and to the full training set as Places-400k. All experiments used the datasets' training sets to train models and the validation sets to evaluate models (as Places Audio does not have a predefined test set).
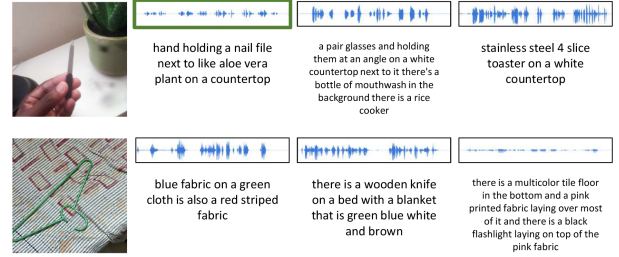

Figure 2: *Top 3 retrieved audio captions for two sample images. The true caption for the image is boxed in green.*

We conducted both audio to image and image to audio retrieval experiments, where a model is tasked with retrieving the most similar images to a given audio caption, and vice versa. We report two results, recall at 1 and recall at 10 (R@1 and R@10, respectively). For R@N, the model is successful if any of the top N recalled images are the correct match to the given audio caption (and vice versa for image to audio retrieval).

We tested frozen and trainable image branches, where freezing the image branch means prohibiting the parameters from being modified via backpropagation so that only the audio model and embedding layers are trained. We define embedding layers as the light-weight layers (ie. linear or non-linear projections) after the backbone layers (ie. convolutional or attention layers). The ObjectNet license prohibits training model parameters on images in the dataset, and the frozen setting complies with this restriction. We also compared image branches pretrained on ImageNet [34] versus randomly initialized.

4.2. Implementation Details

All models were trained for 150 epochs with a batch size of 64 using the Adam optimizer [35]. Every epoch we evaluated the model on the validation set, and the best performance out of the 150 epochs is reported. The learning rate depended on whether the image branch was frozen or not. For trainable image branch experiments, we use a learning rate of $2 \cdot 10^{-4}$ following [27]. For frozen image branch experiments, a larger learning rate of 10^{-3} produced the best results. Also, the learning rate exponentially decayed by a factor of 0.95 every 3 epochs. For all other hyperparameters and data processing, we followed [27].

4.3. Transfer from Places Audio to Spoken ObjectNet

To understand how the bias controls in Spoken ObjectNet impact transfer performance from other spoken caption datasets, we ran transfer learning experiments with a model trained on Places Audio [27]. We used the ResDAVEnet model trained on Places-400k which achieved a mean R@10 of 0.735 on the validation set. There are two ways in which Spoken ObjectNet can be used as a test set: the first is for evaluating zero-shot performance (where the model undergoes no fine-tuning on Spoken ObjectNet), and the second is for evaluating performance after fine-tuning with a frozen image branch (where only the audio and embedding layers are fine-tuned). We also report the results of an experiment in which the entire image branch was made trainable and thus fine-tuned, strictly for comparison, as this setting will be prohibited due to ObjectNet's license.

The results are shown in Table 4. In the zero-shot setting, the model's mean R@10 performance decreases from 0.735 on Places to 0.118 on Spoken ObjectNet. This shows that the

Table 3: *Comparison of training on Spoken ObjectNet (SON) versus Places-50k.*

| | | Frozen image branch | | | | | | Trainable image branch | | | | | |
| | | I $\longrightarrow$ A | | A $\longrightarrow$ I | | Mean | | I $\longrightarrow$ A | | A $\longrightarrow$ I | | Mean | |
Dataset	Pretraining	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
SON	ImageNet	0.064	0.291	0.060	0.268	0.062	0.279	0.066	0.315	0.089	0.332	0.077	0.324
Places-50k	ImageNet	0.093	0.364	0.079	0.360	0.086	0.362	0.067	0.306	0.081	0.335	0.074	0.321
SON	None	-	-	-	-	-	-	0.017	0.123	0.016	0.132	0.017	0.128
Places-50k	None	-	-	-	-	-	-	0.027	0.139	0.026	0.145	0.026	0.142

Table 4: *Transfer learning experiments from Places-400k to Spoken ObjectNet in the following settings: (1) No Fine-tuning (Zero-shot); (2) Fine-tuning (Frozen image branch); (3) Fine-tuning (Trainable image branch).*

| | I $\longrightarrow$ A | | A $\longrightarrow$ I | | Mean | |
Setting	R@1	R@10	R@1	R@10	R@1	R@10
(1)	0.019	0.096	0.033	0.140	0.026	0.118
(2)	0.040	0.216	0.048	0.213	0.044	0.214
(3)	0.102	0.391	0.115	0.416	0.108	0.403

model trained on Places can be directly applied to Spoken ObjectNet, but the performance is much lower. Fine-tuning the model with a frozen image branch recovers some of the performance, up to a 0.214 mean R@10. When the image branch is made trainable, the performance increases to a mean R@10 of 0.403. These experiments show that the controls for viewpoint, rotation, and background make it difficult for the image model (trained on Places-400k) to meaningfully featurize the images in Spoken ObjectNet, as fine-tuning the embedding layers and audio model without fine-tuning the entire image model was not enough to recover the performance of the fully-trainable model.

4.4. Comparison of Spoken ObjectNet and Places Audio

Table 3 compares the relative difficulties of Places-50k and Spoken ObjectNet (SON), where the datasets are matched in size. By running these experiments, we provide additional evidence that the difficulty of Spoken ObjectNet (and the performance drop shown in the transfer setting) is due to the controls for bias. In the frozen image branch setting, the model trained on Spoken ObjectNet performs significantly worse than the model trained on Places-50k based on mean R@10. These results indicate that the ImageNet-pretrained image model is more effective for Places-50k than Spoken ObjectNet when it is kept frozen.

In the second half of Table 3, we show the results of two pairs of experiments in which the image branch was trainable. While this setting will be prohibited due to ObjectNet's license, we show the results to give insight on the difficulty of Spoken ObjectNet versus Places. In the first experiment, the image branch was pretrained on ImageNet. In this experiment, the performance of the model trained on Spoken ObjectNet increases by approximately 20% relative to its frozen counterpart. However, the model trained on Places-50k with a trainable image branch actually decreases in performance compared to the frozen image branch model. This decrease in performance is surprising, and as a result the mean R@10 scores of both models are roughly equivalent. This is likely due to the class overlap between ImageNet and ObjectNet. With a relatively small number of training samples, the model is able to learn enough about the viewpoint, rotation, and background controls applied to objects it already knows about to increase its performance. On

the other hand, when the parameters of the Places-50k image model are adjusted on a relatively small set of Places images it results in a featurizer that performs worse than the original frozen ImageNet-pretrained model.

In the second experiment, the image branch was still fully trainable, but not pretrained on ImageNet. The model trained on Places-50k slightly outperforms the model trained on Spoken ObjectNet, but by a small margin. This shows that without any prior training on any other datasets, and thus without leveraging biases learned from other datasets, Spoken ObjectNet and Places-50k are comparable in difficulty to learn from.

Finally, we show qualitative retrieval results in Figure 2 for the model trained and evaluated on Spoken ObjectNet. In these examples, the model retrieves several relevant captions for each image, including the true caption for the first image. However, the relevance of the other captions could be improved.

5. Conclusion

We introduce Spoken ObjectNet as a bias-controlled spoken language dataset designed to function as a "test set" for audio-visual models. To use the dataset, we suggest training an audio-visual model on some other dataset first. To evaluate the performance of the model in a bias-controlled setting, evaluate the model on the provided 1,000 sample evaluation set. To account for the different classes in ObjectNet and to therefore improve performance slightly, the model's embedding layers and audio model may be fine-tuned on the Spoken ObjectNet training set. As with the original ObjectNet dataset, training model parameters on the images is prohibited.

Spoken ObjectNet exposes the performance gains that models gain from the priors baked into many modern datasets. Our hope is that Spoken ObjectNet can provide inspiration for researchers to explore model architectures that are more robust to priors in data and therefore more likely to generalize to real-world scenarios.

6. Acknowledgements

7. References

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

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

[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] Z. Zhu, L. Xie, and A. L. Yuille, "Object recognition with and without objects," in *IJCAI*, 2017.

[5] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. Katz, "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," in *NeurIPS*, 2019.

[6] D. Harwath and J. Glass, "Deep multimodal semantic embeddings for speech and images," in *ASRU*, 2015.

[7] D. Harwath, G. Chuang, and J. Glass, "Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech," in *ICASSP*, 2018.

[8] W. N. Havard, J.-P. Chevrot, and L. Besacier, "Models of visually grounded speech signal pay attention to nouns: A bilingual experiment on english and japanese," in *ICASSP*, 2019.

[9] Y. Ohishi, A. Kimura, T. Kawanishi, K. Kashino, D. Harwath, and J. Glass, "Trilingual semantic embeddings of visually grounded speech with self-attention mechanisms," in *ICASSP*, 2020.

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

[11] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, "Connecting vision and language with localized narratives," in *ECCV*, 2020.

[12] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's Mechanical Turk," in *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[13] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *CoRR*, 2015.

[14] W.-N. Hsu, D. Harwath, C. Song, and J. Glass, "Text-free image-to-speech synthesis using learned segmental units," *NeurIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*, 2020.

[15] W. Havard, L. Besacier, and O. Rosec, "Speech-coco: 600k visually grounded spoken captions aligned to mscoco data set," *arXiv preprint arXiv:1707.08435*, 2017.

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

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

[18] D. Suris, A. Recasens, D. Bau, D. Harwath, J. Glass, and A. Torralba, "Learning words by drawing images," in *CVPR*, 2019.

[19] A.-M. Oncescu, J. F. Henriques, Y. Liu, A. Zisserman, and S. Albanie, "Queryd: A video dataset with high-quality text and audio narrations," in *ICASSP*, 2021.

[20] M. Monfort, S. Jin, A. Liu, D. Harwath, R. Feris, J. Glass, and A. Oliva, "Spoken moments: Learning joint audio-visual representations from video descriptions," in *CVPR*, 2021.

[21] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfreund, C. Vondrick *et al.*, "Moments in time dataset: one million videos for event understanding," *TPAMI*, 2019.

[22] 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.

[23] 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.

[24] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.

[25] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP*, 2020.

[26] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," *IJCV*, 2020.

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

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

[29] 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.

[30] A. Rouditchenko, A. Boggust, D. Harwath, D. Joshi, S. Thomas, K. Audhkhasi, R. Feris, B. Kingsbury, M. Picheny, A. Torralba, and J. Glass, "Avlnet: Learning audio-visual language representations from instructional videos," *arXiv preprint arXiv:2006.09199*, 2020.

[31] 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.

[32] L. Wang, X. Wang, M. Hasegawa-Johnson, O. Scharenborg, and N. Dehak, "Align or attend? toward more efficient and accurate spoken word discovery using speech-to-image retrieval," in *ICASSP*, 2021.

[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," in *EMNLP: System Demonstrations*, 2020.

[34] 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.

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