Spoken ObjectNet: A Bias-Controlled Spoken Caption Dataset - Extended Abstract

Ian Palmer, Andrew Rouditchenko[†], Andrei Barbu, Boris Katz, James Glass[†]
MIT Computer Science and Artificial Intelligence Laboratory
{iapalm,roudi,abarbu,boris,glass}@mit.edu [†]Contact Authors

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 with 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. We encourage readers to check our full paper, which has been accepted to Interspeech 2021 and will be available publicly soon, for the full details and more experiments.

1. Introduction

Prior work has shown that neural models can learn meaningful audio-visual correspondences from visually grounded speech [4, 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 [6].

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 [1]. 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.

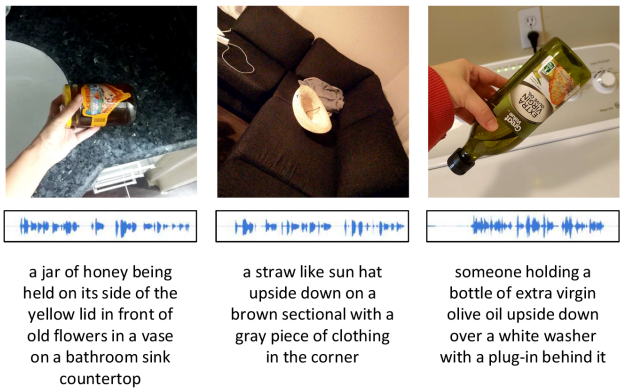

Figure 1. Samples of images, spoken captions, and ASR transcripts from Spoken ObjectNet.

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 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.

2. 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.

2.1. 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 ensured the recording setup was functioning properly and that workers were actually speaking.

Lastly, we introduced a new step in which the ASR transcript 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 [5], 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.

2.2. 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. Samples are shown in Figure 1.

Table 1. Transfer learning experiments from a model trained on Places Audio. (1) No Fine-tuning (Zero-shot); (2) Fine-tuning (Frozen image branch); (3) Fine-tuning (Trainable image branch). I $\longrightarrow$ A = image to audio; A $\longrightarrow$ I = audio to image.

	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

3. Retrieval Experiments

3.1. Transfer from Places to Spoken ObjectNet

Because Spoken ObjectNet is best understood as a test set relative to a dataset like Places Audio, we ran transfer learning experiments with a model trained on Places audio [2]. The original model is the best ResDAVEnet-VQ model (without any VQ layers enabled) that was trained on Places Audio Captions. This model achieved a mean R@10 of 0.735 [2] on the Places Audio 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 1. 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 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 demonstrate that the controls for viewpoint, rotation, and background make it difficult for the image model (trained on Places Audio) 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. 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.

Acknowledgements This research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. Andrei Barbu and Boris Katz were also supported by the Center for Brains, Minds and Machines, NSF STC award 1231216, and the Office of Naval Research under Award Number N00014-20-1-25. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. We thank Rami Manna for helpful discussions.

References

[1] 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. 1

[2] D. Harwath, W.-N. Hsu, and J. Glass. Learning hierarchical discrete linguistic units from visually-grounded speech. In *ICLR*, 2020. 2

[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. 1

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

[5] 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. 2

[6] Z. Zhu, L. Xie, and A. L. Yuille. Object recognition with and without objects. In *IJCAI*, 2017. 1