

Spoken ObjectNet: Creating a Bias-Controlled Spoken Caption Dataset

by

Ian A. Palmer

B.S. Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2020

Submitted to the Department of Electrical Engineering and Computer
Science

in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

© Massachusetts Institute of Technology 2021. All rights reserved.

Author
Department of Electrical Engineering and Computer Science
May 14, 2021

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

Accepted by
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Spoken ObjectNet: Creating a Bias-Controlled Spoken Caption Dataset

by

Ian A. Palmer

Submitted to the Department of Electrical Engineering and Computer Science
on May 14, 2021, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

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 large-scale image dataset that features controls for biases encoded into many other common image datasets.

We detail our data collection pipeline, which features several methods to improve caption quality, including automated language model checks. We also present an analysis of the vocabulary of our collected captions. Lastly, we show baseline results on several audio-visual machine learning tasks, including retrieval and machine captioning. 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 intend to make our dataset openly available to the general public to encourage new lines of work in training models that are better equipped to operate in the real world.

Thesis Supervisor: James R. Glass
Title: Senior Research Scientist

Acknowledgments

I am extremely grateful to everyone who made my time in the Spoken Language Systems Group so rewarding. I would like to thank my advisor, James Glass, whose advice and guidance made this project possible. I am lucky to have been a part of the group that he works tirelessly to develop, and I will always appreciate the positive culture within SLS. Thanks as well to the students of SLS for their constant support and collaboration. I would also like to thank the AIA meeting group, including the members of the CSAIL Infolab group, Air Force Research Lab, and Lincoln Lab for their insights and advice. Thanks as well to the Air Force Institute of Technology for sponsoring my research.

Lastly, I would like to thank my family, whose constant encouragement and support has made my journey to where I am today possible.

Contents

1	Introduction	13
1.1	Motivation	13
1.2	Problem Description	14
1.3	Contributions	15
1.4	Outline	15
2	Related Works	17
2.1	Audio-Visual Caption Datasets	17
2.2	Visual Datasets	19
2.3	Captioning Datasets	20
2.4	Audio-Visual Models	21
2.5	Transformers	22
2.6	Captioning Models	22
3	Dataset Collection	25
3.1	Crowdworking Task	25
3.2	Validation	27
3.3	Evaluation Server	31
3.4	Transcript Correction	33
3.4.1	Comparison of AMT and Rev	34
3.5	Additional Samples	36
4	Dataset Analysis	37

4.1	Spoken ObjectNet-50k	37
4.2	Spoken ObjectNet-20k	40
4.3	Places-50k	42
4.4	Dataset Splits	42
5	Retrieval Experiments	43
5.1	Experimental Setup	43
5.2	Implementation Details	44
5.3	Transfer Experiments	45
5.4	Comparing Spoken ObjectNet & Places Audio	46
5.5	Analysis	49
6	Image Captioning Experiments	51
6.1	Experimental Setup	51
6.2	Implementation Details	52
6.3	Captioning From Scratch	52
6.4	Transfer Experiments	54
6.5	Additional Models	55
6.6	Generated Captions	57
6.7	Analysis	57
7	Conclusion	61
7.1	Dataset Release	62
7.2	Future Work	62
A	Instructions	65
A.1	Image Captioning Task	65
A.2	Transcript Correction Task	67

List of Figures

3-1	The Amazon Mechanical Turk interface that workers used to record captions for ObjectNet images. Workers could reference the example on the right as they captioned the image on the left.	26
3-2	A histogram of the language model scores of all of the samples collected for Spoken ObjectNet-50k (excluding those that were rejected).	29
3-3	Examples of several captions, their corresponding language model score, and the approximate percent of all captions that have lower language model scores than the one shown.	30
3-4	An example screenshot of our validation server page. The image is displayed along with its audio caption and ASR transcript. The sample shown here has the fourth-lowest language model score.	32
3-5	Example mistakes made in ASR captioning, including substitutions and deletions.	33
3-6	Examples of original transcripts and the corrected transcripts received from Amazon Mechanical Turk workers.	34
3-7	The interface workers used to correct transcripts. The text box automatically populated with the original transcript, and workers could see an example on the right hand side of the screen.	35
4-1	Samples of images and ASR captions from Spoken ObjectNet-50k.	39
4-2	Two examples from Spoken ObjectNet-20k, with five captions per image.	41
5-1	Top 5 retrieved audio captions for two sample images. The true caption for the image (if applicable) is boxed in green.	48

6-1	Examples of captions for COCO images produced by our model trained on COCO Captions.	58
6-2	Examples of captions for ObjectNet images produced by our model trained on COCO Captions. This is the zero-shot setting described in Table 6.2.	59

List of Tables

4.1	An analysis of the vocabulary of Spoken ObjectNet-50k, Spoken ObjectNet-20k, and the combined datasets. Each category shows the number of the unique speakers, words, etc. in each dataset.	37
4.2	A comparison of the vocabularies of Spoken ObjectNet, Places Audio, and several other popular audio-visual event localization datasets. . .	38
4.3	A comparison of the distribution of common parts of speech in Spoken ObjectNet and other audio-visual datasets.	38
5.1	Results of retrieval experiments based on transfer learning from a model trained on Places-400k.	45
5.2	Comparison of training on Spoken ObjectNet-50k versus Places-50k with frozen image branches.	46
5.3	Comparison of training on Spoken ObjectNet-50k versus Places-50k with trainable image branches.	47
6.1	Results of training captioning models from scratch on both COCO Captions and SON-20k, with and without self-critical sequence training (SCST).	53
6.2	The results of several transfer experiments in which a model that was originally trained on COCO Captions was evaluated on SON-20k under several different fine-tuning settings.	54
6.3	Results of our additional captioning experiments on both COCO Captions and SON-20k.	56

Chapter 1

Introduction

1.1 Motivation

Countless datasets have been developed and used at large scales to train machine learning models. Very few of these datasets, however, make an effort to control for implicit biases in the data. In turn, models learn these biases, and their performance often suffers in the real world as a result. The fundamental issue is that many popular datasets are constructed via large-scale web scraping, which selects for the sort of data that would be uploaded to the internet in the first place. As a result, the majority of image classification datasets are comprised of objects that are well-lit, in their usual contexts, and photographed from a familiar viewpoint. When presented with real-world examples that may not have any of those characteristics, models that achieve state-of-the-art performance on existing datasets are often reduced to pre-deep learning performance levels.

Simultaneously, neural networks have been shown to learn complex audio and visual representations in weakly supervised settings. As models learn more directly from data with less supervision, the removal of biases in the data becomes all the more important. One such setting is the pairing of images with spoken audio captions that describe the contents of the image. Just from that pairing, neural audio-visual models can learn to correspond objects and sounds in the waveform. A visual reference or grounding is in many ways essential to developing a true understanding of language:

after all, a model could read thousands of books and never understand what the color red is. It is important, then, that there exists a bias-controlled audio-visual dataset from which models can learn these correspondences with as few intrinsic biases as possible.

1.2 Problem Description

There has been prior work focused on creating bias-controlled image classification datasets. Separately, there are a number of popular spoken caption datasets. However, very little work has focused on the intersection of these two endeavors. Because there is no available bias-controlled large-scale dataset in this space, it is extremely difficult to evaluate how effectively a model will perform in the real world.

Our primary contribution is a novel spoken captions dataset based on ObjectNet, an existing bias-controlled image dataset. Like several other datasets in existence, it features both images and a spoken language audio track in which humans describe what they see in the image. Unlike any other audio-visual datasets in existence, our dataset features controls for biases in the image domain. This has the potential to serve as a test set that is indicative of performance in the real world, which is not always true of held-out data.

Collecting data at this scale poses a significant challenge, so we developed several tools to improve the automated evaluation of worker submissions and increase the rate at which we could manually inspect parts of the dataset. We also developed another large-scale crowdworking task that improved the quality of the text captions associated with each spoken language recording.

To ensure that neural models perform as expected on the dataset, we conducted a series of retrieval and machine captioning experiments and compared the results against equivalent experiments on other datasets. This comparison provides a proof-of-concept that the dataset is difficult for the right reasons: it has controls in place for the biases present in other datasets, not because the data is low-quality. We show results in a variety of settings to explore the effect that different training regimes and

model restrictions have on performance.

1.3 Contributions

Our contributions towards this challenge include the following:

- We present Spoken ObjectNet, a novel large-scale bias-controlled spoken captions dataset collected via a crowdworking task.
- We describe novel contributions to our data collection and evaluation pipeline that improved the quality of captions and reduced the amount of manual inspection that was required to create the dataset.
- We present an analysis of the audio samples and their corresponding text transcripts, giving an overview of the dataset and comparing it against other existing datasets.
- We describe an additional crowdworking task used in conjunction with caption collection in which workers were asked to correct the automatic speech recognition transcripts of spoken language samples in our dataset.
- To evaluate the dataset, we show a series of retrieval and machine captioning experiments and compare the results against equivalent experiments on other datasets.

1.4 Outline

In Chapter 2, we discuss works related to several common audio-visual machine learning tasks. This includes audio-visual models, datasets, and evaluation techniques. In Chapter 3, we discuss the techniques used to collect the dataset via a crowdworking task. This chapter includes an analysis of a novel language modeling step used for data validation. Chapter 4 presents an analysis of our collected dataset and a comparison of it and other popular audio-visual datasets. In Chapter 5, we continue

our analysis of the dataset via a series of image and audio retrieval experiments. We compare results on our dataset with results on other audio-visual datasets to show the unique debiasing that our dataset possesses. Chapter 6 presents additional machine captioning experiments comparing our dataset and other datasets. Lastly, Chapter 7 presents future directions for this work.

Chapter 2

Related Works

2.1 Audio-Visual Caption Datasets

The Places Audio Captions dataset (Harwath *et al.*, 2018) is the most similar existing dataset to Spoken ObjectNet, and features spontaneous speech about images in the Places 205 dataset (Zhou *et al.*, 2014). While many other datasets ask workers to describe the contents of images via text descriptions (and then have other workers read those text fragments out loud), the creators of Places Audio Captions find that asking crowdworkers to describe an image verbally with a prompt to simply “describe what you see” produces captions that have significantly more detail. As a result, the average number of words in this dataset is significantly higher than in other datasets with non-spontaneous speech.

Beyond the dataset itself, we used some of the infrastructure developed by the creators of Places Audio Captions to develop our Amazon Mechanical Turk (AMT) task. Their toolkit uses a Flask web server to host a task on AMT, then run a series of validation scripts and save the completed recordings to disk. Our extensions to this software toolkit are described in a later section.

Microsoft Common Objects in Context (MS-COCO) (Lin *et al.*, 2014) is one of the most popular image classification and image captioning datasets. Its captioning dataset features 5 captions per image for each of the 330,000 images in the dataset. While MS-COCO is one of the largest-scale image captioning datasets in existence,

its captions are text-based, making it of limited use to recent work in directly learning audio-visual correspondences from paired speech and images. There have been several efforts to develop spoken captions for MS-COCO, however. The first, called SPEECH-COCO (Havard *et al.*, 2017), uses text-to-speech software to create a corresponding audio recording for each of the text captions. The authors go to some lengths to make the automatically generated captions more natural, adding disfluencies and changing the rate of speech for some recordings. These captions are still less natural than human recordings, however. Over 600,000 human recordings were collected in SpokenCOCO (Hsu *et al.*, 2020), and while these recordings improve upon the text-to-speech results, they are fundamentally limited by their non-spontaneity. As a result, SpokenCOCO is the most recent and most natural spoken caption dataset based on MS-COCO, but it has fewer words per caption than Places Audio Captions.

Flickr8k Audio Captions (Harwath & Glass, 2015) is similar to SpokenCOCO in that the captions are non-spontaneous, and the dataset is significantly smaller in scale than SpokenCOCO at 5 captions each for 8,000 images. It is based on the Flickr8k caption dataset (Rashtchian *et al.*, 2010). The authors collected those captions via Amazon Mechanical Turk, asking workers to read a text caption (collected in a prior work by different authors) aloud. While the non-spontaneity and limited size of the dataset make it of limited use to our particular project, the dataset has use in pretraining and testing the transfer performance of audio-visual models.

The Localized Narratives (Pont-Tuset *et al.*, 2020) dataset augments spoken captions with a visual grounding in the form of mouse pointer tracking. Workers are asked to simultaneously describe an image out loud and move their mouse pointer to the region of the image that they are referencing in their speech. This provides a strong grounding for audio-visual models and requires less human annotation than bounding boxes or other manual geometric labels. Overall, the creators of the Localized Narratives dataset find that this technique produces detailed captions and high-quality visual grounding for each of the 873,107 samples in the dataset.

QuerYD (Oncescu *et al.*, 2021) is a large-scale dataset containing videos with two audio tracks: the original audio, and a spoken description task collected via volunteers

on the YouDescribe service. The service is designed to assist visually-impaired people use YouTube by providing a spoken closed-captioning track to existing YouTube videos. Because the captions are created by volunteers, QuerYD contains high-quality descriptions. Unlike Places Audio Captions, the visual domain has videos. This makes the dataset most useful for video retrieval and localization tasks.

There are also datasets, such as AudioSet (Gemmeke *et al.*, 2017) and VGG-Sound (Chen *et al.*, 2020), that are useful primarily for audio-visual prediction tasks and audio recognition tasks. AudioSet is curated from YouTube videos, which are split into 10 second clips that contain a certain sound. VGG-Sound is also collected from open-source video libraries, but the authors use a series of convolutional neural network-based filters to select for high-quality samples. Although our implementation is very different, Spoken ObjectNet also uses neural networks as filters for incoming data, as described in Chapter 3.2. In general, though, Spoken ObjectNet is intended to be used to train models for different tasks than AudioSet or VGG-Sound.

2.2 Visual Datasets

ImageNet (Deng *et al.*, 2009) is a large-scale image classification dataset that is widely used as a benchmark for image classification models as well as a pretraining dataset for convolutional neural networks. It has several million images across 1,000 object classes. The test set contains 50 samples for each of the 1,000 classes. In our experiments, we use ImageNet primarily as a pretraining dataset for audio-visual models that are later trained on the Places 205 dataset or directly evaluated on Spoken ObjectNet.

The ObjectNet (Barbu *et al.*, 2019) dataset is an object detection dataset designed to be of a similar size to the ImageNet test set, with 50,273 images. Because it was collected entirely via crowdworking, the authors could explicitly control for object viewpoints, rotations, and backgrounds in the collected image. Workers would be instructed to find a household object out of several hundred classes, then place it in a specified area of the house (like washroom, bedroom, or kitchen). Explicit instruc-

tions on the rotation of the object and the viewpoint of the camera were also provided. The dataset therefore avoids the potential biases introduced into the dataset by collecting existing images from the Internet. To fully realize the dataset’s intended use as a test set, the authors take the additional step of prohibiting models from learning from the images in ObjectNet. Users can instead measure the transfer performance of models trained on other datasets and then evaluated on ObjectNet. In this setting, the authors find that most state-of-the-art image classification models suffer a performance drop of about 40% versus their performance on the ImageNet test set. Overall, ObjectNet provides a more realistic assessment of how well an image classifier will work in the real world.

2.3 Captioning Datasets

As mentioned previously, MS-COCO Captions (Chen *et al.*, 2015) is one of the most popular datasets for image captioning. It has 5 captions per image and 120,000 images, with an average length of 10 words per caption. The captions were produced by paid annotators on a crowdworking platform. The scale of this dataset makes it one of the most popular sources of image captioning training data, although it has a relatively small number of object categories and has been overtaken in recent years by larger-scale datasets.

The Google Conceptual Captions Dataset (Sharma *et al.*, 2018) is one of the largest-scale captioning datasets in existence. It was gathered by scraping existing images and alt-texts from the Internet, so the images are weakly labeled (using a computer vision system) and the alt-texts vary in quality depending on the website. However, they use a novel pipeline that extracts and filters the raw image/caption pairs before creating the final dataset. This approach improves the image/caption correspondences in the final dataset. They report an average of 10.3 tokens per caption in the training set.

Lastly, the VizWiz dataset (Gurari *et al.*, 2020) uses images taken by vision-impaired users. Because many image captioning services are deployed to assist vision-

impaired people navigate the world, it is important that there exists a dataset with images that are representative of the real world. In that sense, the images in the VizWiz dataset look similar to those in ObjectNet: they are sometimes blurry, poorly-lit, or feature objects in unusual orientations. VizWiz has 5 captions per image for each of the 39,181 images in the dataset, so it is approximately the same size as ObjectNet. The authors of VizWiz split the data into training, validation, and test sets. They report an average of 13 words per caption in the dataset, greater than those of MS-COCO Captions or the Google Conceptual Captions Dataset.

2.4 Audio-Visual Models

There are several neural models that can directly learn audio-visual correspondences from paired visuals (either images or videos) and spoken language. ResDAVENet (Harwath *et al.*, 2018) uses two convolutional neural networks to jointly embed images and audio captions into a shared embedding space. Their unique contribution is that the images and audio are embedded spatially as well as temporally, enabling models to co-localize image and audio features. While this allows the model to simply retrieve corresponding images and audio captions, the authors of Harwath *et al.* (2018) show that the model learns object and word localization as a result of its training objective. The image branch of ResDAVENet is adapted from the ResNet50 (He *et al.*, 2016) architecture, but in ResDAVENet the final softmax and fully-connected layers are removed and replaced with a 1x1 convolutional layer. This layer projects the model’s features into the desired embedding dimension. The audio branch is a 17-layer fully convolutional model with residual connections. Audio samples are converted into log Mel-frequency spectrograms.

The model used in this paper is an extension of the ResDAVENet architecture that adds multiple vector quantization (VQ) layers (Harwath *et al.*, 2020). VQ layers act as a bottleneck, constraining the amount of information that is passed on to the next layers. ResDAVENet-VQ has a total of five VQ layers, and each can be independently activated or deactivated. ResDAVENet-VQ is trained with a triplet

loss, which combines random sampling of negative examples and semi-hard negative mining (Jansen *et al.*, 2018). For simplicity, we do not explore the use of any of the VQ layers in our experiments.

AVLnet can learn correspondences between spoken words and visual content in videos, all in a self-supervised setting. In that way, it extends models like ResDAVEnet-VQ by adding a temporal dimension to the image branch. It can learn audio-visual correspondences from videos, specifically instructional videos from YouTube and other open source video websites, via datasets like HowTo100m (Miech *et al.*, 2019) and YouCook2 (Zhou *et al.*, 2018).

2.5 Transformers

In recent years, transformer-based models have exceeded the performance of traditional recurrent or convolutional neural network-based models in almost every major NLP task. BERT (Bidirectional Encoder Representations from Transformers) (Devlin *et al.*, 2019) is one of the most popular transformer models. It has been shown to be surprisingly versatile, performing at state-of-the-art (at the time) in question answering and natural language inference challenges. BERT introduces a masked language modeling training objective, which allows the transformer to be trained bidirectionally. Word embeddings generated by BERT are useful in a variety of downstream tasks, and BERT can be extended to many other tasks by simply replacing the final layer with a new layer and fine-tuning.

2.6 Captioning Models

Based on BLEU (Papineni *et al.*, 2002) score, the current state-of-the-art model as tested on MS-COCO Captions is Oscar (Li *et al.*, 2020), a multi-layer transformer-based vision and language system. The models are pretrained on a large image-text corpus, where each sample is a triple of the sequence of words, a set of object tags, and image features. The authors find that the object tags, which are generated using a

simple convolutional neural network, provide an additional visual grounding that the model can take advantage of during training. This association results in better learned representations and, as a result, good performance on a number of downstream tasks, including retrieval, image captioning, and visual question answering.

However, there are several challenges associated with this model in particular. Because it is transformer-based, it contains many more parameters than traditional CNN/RNN-based captioning models. This makes it more difficult and time-consuming to train. Additionally, while the authors have released source code for the model and CNN-generated object tags for MS-COCO, they have not released preprocessing source code. This makes it very difficult to replicate their methodology on another dataset. Because we were more interested in measuring performance differences across datasets rather than achieving state-of-the-art performance on one dataset, we chose not to use Oscar in our experiments.

Instead, we implemented a CNN/LSTM-based model with self-critical sequence training, as described in Rennie *et al.* (2017) and Luo *et al.* (2018). This model achieves a slightly lower but comparable BLEU score to Oscar on MS-COCO captions. It is also smaller, easier to train, and requires less preprocessing than Oscar, making it ideal for our experiments. The authors introduce a technique for optimizing image captioning systems using reinforcement learning. In self-critical sequence training, a REINFORCE (Williams, 1992)-like algorithm (with some modifications) is used to optimize for CIDEr (Vedantam *et al.*, 2015) score during decoding. At the time, this technique was the state-of-the-art on the MS-COCO captions test set, and remains one of the best non-transformer models. Because of its relatively high performance, availability of source, and ease of training, we decided to use this model in our image captioning experiments.

Chapter 3

Dataset Collection

3.1 Crowdfunding Task

Our intent for the first round of data collection for Spoken ObjectNet was to collect one spoken caption per image in the dataset, for a total of 50,273 captions. Because of the scale of the dataset and the importance of having a variety of speakers for models to learn from, we chose to use Amazon Mechanical Turk to collect the data. Amazon Mechanical Turk (AMT) is an online marketplace that allows requesters to submit jobs to the global public. Each job has a certain pre-determined pay rate and an estimated work load. Requesters can set certain requirements for workers to be able to accept the jobs, such as geographic requirements, previous assignment acceptance rates, and specialized training that workers must complete before attempting the task. Each task is referred to as a Human Intelligence Task, or HIT. Workers often complete many HITs at the same time, and requesters commonly submit thousands of HITs to AMT at one time. For our purposes, AMT provided the scale and platform from which we could collect samples for Spoken ObjectNet.

The toolkit we used to host and review the tasks was based on the tool used by the authors of Harwath *et al.* (2018) to collect audio samples for Places Spoken Captions. The tool was built on a WSGI and nginx server to enable multithreading and maintain high performance even as many users are simultaneously submitting HITs, which was important in our use case. The server itself was written in Python

and used Flask and the AWS boto3 library to serve questions to workers. The worker interface on AMT is shown in Figure 3-1, and the instructions (shown above the interface) are included in Appendix A.1.

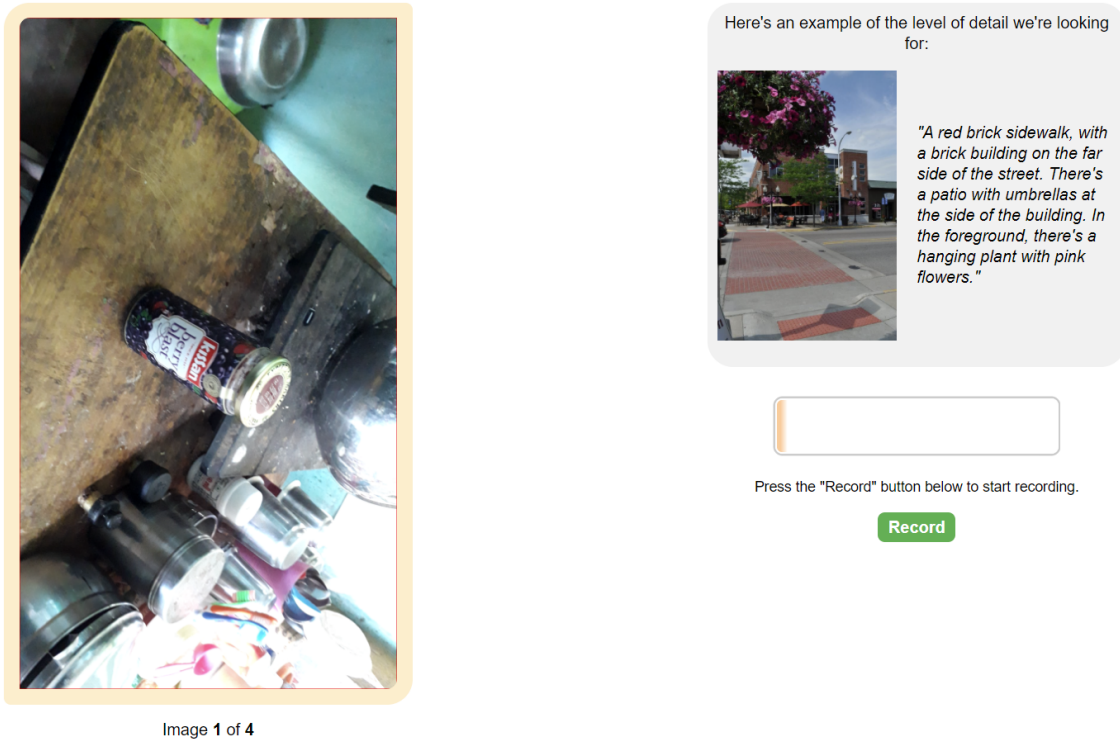


Figure 3-1: The Amazon Mechanical Turk interface that workers used to record captions for ObjectNet images. Workers could reference the example on the right as they captioned the image on the left.

First, each task was assigned a unique URL. Each task was comprised of four total images, and those assignments were deterministically generated when the Flask server initialized. The webpage itself consisted of a set of instructions, followed by an image to caption and a recording prompt. One example was provided so workers would have a general idea of the amount of detail we were looking for in recordings. The unique URL determined the images that would be loaded, so no two tasks loaded the same images.

The tasks were then submitted to AMT using the boto3 library. Using the task to URL mapping generated when the webserver initialized, each task's URL was embedded into a boto3 ExternalQuestion and launched as a HIT on AMT. As workers

completed the HITs, their recordings were validated in real time (as described in Section 3.2). If a worker did not pass the validation steps, they were immediately asked to retry the recording. If they did pass the validation steps, they were either prompted to complete the next recording (out of four total) or prompted to submit their completed HIT. When a worker submitted a HIT, the audio, automatic speech recognition transcript, and metadata were serialized into a JSON package and returned to the Flask server.

Automatic speech recognition (ASR) was applied to every recording before the validation step took place. We used the Google SpeechRecognition library to generate approximate transcripts of the speech samples. While imperfect, these ASR transcripts provided us with a useful piece of information during the validation step. By bundling the transcript with the rest of the data in the final JSON package, we also saved time on having to run ASR later in the process.

Once submitted, the HITs were held for a manually-initiated review step. Every collected sample could either be inspected manually, or passed/failed based on the validation steps computed at the time of the recording (see Section 3.2). In practice, the number of samples being collected at a time meant manual inspection was impractical, and instead we focused on tuning the automatic validation steps to pass only samples that met our acceptance criteria. Once accepted, the audio files and ASR transcripts were written to disk, and sorted by image. Metadata, including the worker ID and HIT ID, was added to a database. If a sample was rejected during review for any reason, the recording was moved to another separate folder on disk and the metadata was still retained.

3.2 Validation

After workers submitted a recording for an image during the HIT, the recording was passed to our validation engine. Rather than evaluate responses after the entire HIT was submitted, our data collection tool runs a validation step in real time. This improves the experience for workers, who are given more direct feedback about their

work and can adjust their work if they are not meeting our criteria for acceptance. It also ensures they are not disappointed if, for example, their work is rejected several days later without an explanation. Real-time feedback also benefits data collectors, who can accept a much higher percentage of submitted work.

We added a novel language modeling feature to the data collection tool’s original validation procedure, which further improved the quality of data that we collected. The motivation for this language modeling step came in the early stages of data collection. Some of the HIT submissions we received were from workers who found a way to submit HITs that would be automatically accepted but that did not actually solve the task. Usually, these submissions contained garbled English words in a seemingly random order. These words would be detected in the ASR step, and any submission that had at least four words in its ASR transcript and was at least one second long would pass the original validation steps. We therefore needed a way to distinguish between well-formed English text and the malformed submissions that we would sometimes receive.

To solve this challenge, we added a language model to the validation pipeline. The language model is a BERT (Devlin *et al.*, 2019) transformer with a language modeling head from the huggingface (Wolf *et al.*, 2020) library. At a high level, the language modeling step masks one token at a time in the input and measures the probability that the model predicts the correct word in the masked location. We then compute the cross entropy loss between the ground truth and predictions for every token, and lastly we sum the log probabilities together to compute a final scalar loss. Given a sequence of tokens s and a model \mathbf{M} , we compute the following:

$$loss = - \sum_{i=0}^{|s|} \log \left(\frac{e^{M(s)_i}}{\sum_j e^{M(s)_j}} \right) \quad (3.1)$$

To obtain a value that lies within the range of $[0, \infty)$, we use an exponential. Our final score is computed as follows:

$$score = e^{loss} = - \sum_{i=0}^{|s|} \left(\frac{e^{M(s)_i}}{\sum_j e^{M(s)_j}} \right) \quad (3.2)$$

Because of the exponential, higher language model scores indicate that a sample is less grammatical. To determine a cutoff score for this step in the validation process, we scored all of the captions that we had collected so far and sorted them by increasing language model score. We were then able to qualitatively see how the quality of the caption varied with the language model score. The lowest score was about 1, the mean score was approximately 20, and the max score was over 1,000. A histogram of all scores is shown in Figure 3-2. We decided on a cutoff score of 80, which was a higher threshold than 90% of the existing data. The remaining 10% was mostly the low-quality data that we had set out to remove in the first place, which confirmed that this was a useful addition to the validation procedure.

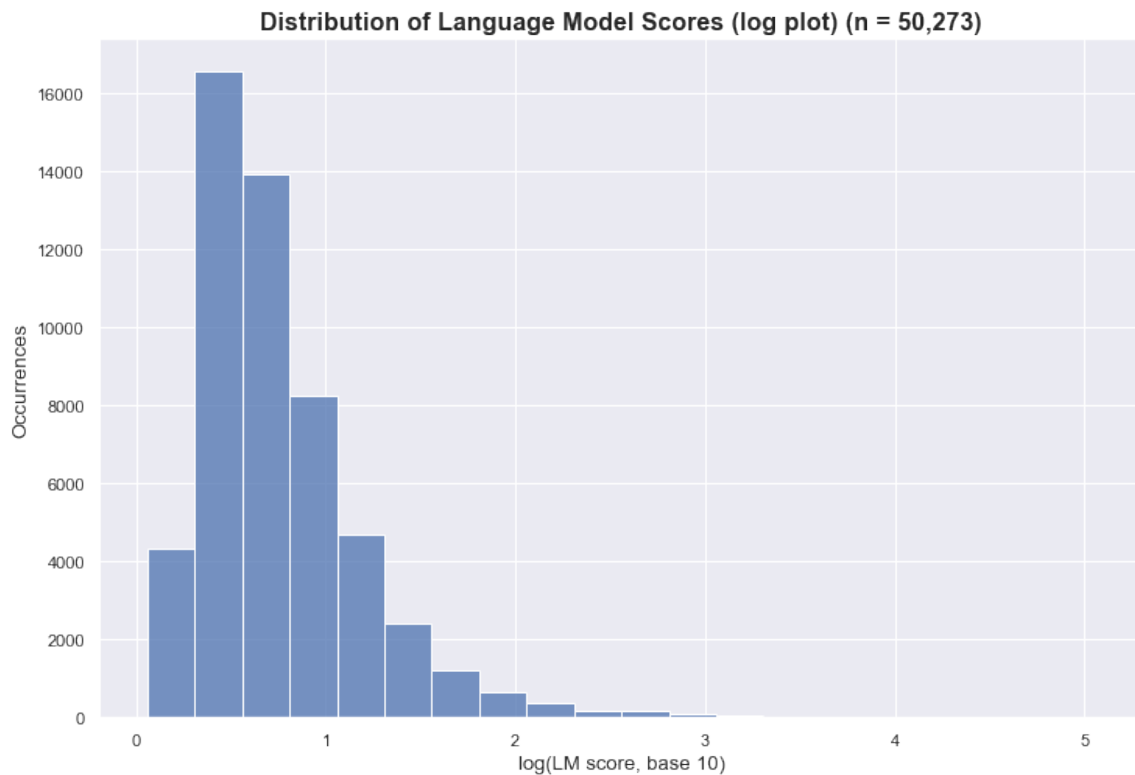


Figure 3-2: A histogram of the language model scores of all of the samples collected for Spoken ObjectNet-50k (excluding those that were rejected).

Figure 3-3 shows examples of captions and their corresponding language model scores. Higher scoring captions are generally longer, contain more detailed descriptions, and may have fewer ASR errors than lower scoring captions.

	LM Score	Caption	Percentile
Better captions - longer - more grammatical - detailed descriptions	1.14	there's a human hand which is holding a jewelry piece and it's holding it over at the floor the person is wearing a ring the tile itself is the color of off-white the jewelry is gold the person's holding it between their two fingers and they're wearing a ring	1%
	2.08	is a white and black marble counter there are several objects including a red lipstick to a blue bottle to glass bottles in the back one containing a green plant in a jar wrapped in brown twine with white seashell cascading from the top	10%
	2.72	a pair of silver colored tweezers is sitting on top of a brown surface perhaps a table or desk	25%
Worse captions - often short - less grammatical - ASR errors	4.21	a black laptop with the that is open with the screen on sitting on a white tile floor	50%
	6.26	bathroom sink sitting on top of that is a mug of coffee mug red and white designed up behind that there's a door with a knob on it and to the right there's some towels hanging	75%
	17.21	chairs in front of dishwasher inside the kitchen	90%
	74.62	a man holds up a knit rug	98%

Figure 3-3: Examples of several captions, their corresponding language model score, and the approximate percent of all captions that have lower language model scores than the one shown.

In the current form, the validation procedure consists of all three checks, performed one after another. Samples must be at least one second long, contain at least four words in the ASR transcript, and score less than 80 on the language modeling score. Each check prevents a different type of low-quality submission from being accepted, and in practice we have found that having all three present results in high-quality collected captions with minimal manual analysis required.

3.3 Evaluation Server

In practice, it was very helpful to inspect some of the incoming data to ensure that the instructions were clear and the collected data was high-quality. While this was made easier by the automated validation steps described in Section 3.2, manually inspecting the data still provided us feedback on our task and the approach to data collection that we had taken.

Manually copying data from the file system it was saved to at any significant scale was infeasible, so we designed a web server built using the same Flask server that the Amazon Mechanical Turk task was based off of. It is password protected, so any potentially personal information like worker IDs and assignment IDs is protected. It allows any number of trusted users to inspect any of the over 50,000 samples that have been collected.

To improve the organization of the data and support our goal of being able to focus our manual inspection efforts on the lowest-quality data, the samples on the server are sorted by language model score. As input, the server takes in a text file with paths to all of the audio samples that should be displayed on the server along with the corresponding language model score of the ASR transcript of each audio sample. Upon startup, the server creates a mapping to every image, audio sample, and audio transcript. When requesting a page, the server fills in a Flask template page with the corresponding image, text, and language model score. An example of the Flask page is shown in Figure 3-4.

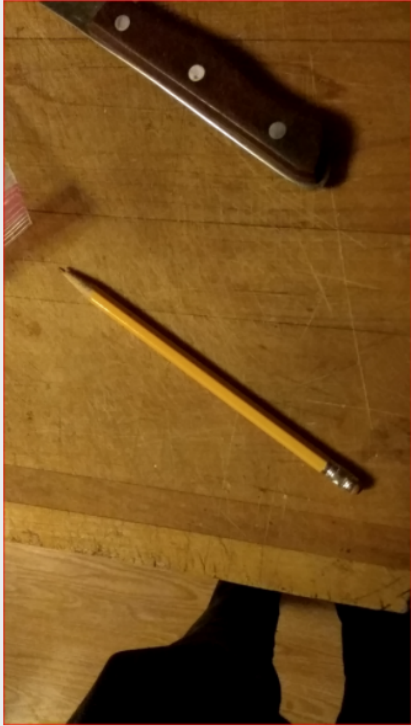
Starting from the sample with the highest language model score and working down, we were able to review about 2,500 samples. Of these, some were low-quality and removed from the final dataset. New HITs were posted to re-collect data for these images. Most, however, were of a reasonable quality. We observed that transcription errors could cause the language model score to increase significantly, and that was often the reason why an otherwise acceptable caption would be scored highly.

Overall, designing this evaluation platform greatly increased the speed at which we could manually inspect the data. It also made sharing the data with other col-

ObjectNet Spoken Captions: Samples

First Previous 4 Next Last

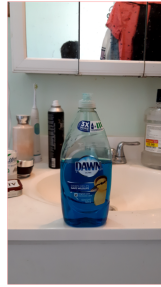
Worker ID: [REDACTED]
Assignment ID: [REDACTED]
Language model score: 1.233419418334961



there's a pencil on top of wood and there's the end of a knife in the top of the image and there's a person's legs in the bottom

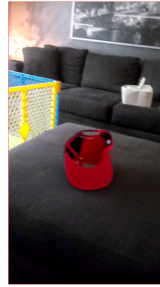
Play

Figure 3-4: An example screenshot of our validation server page. The image is displayed along with its audio caption and ASR transcript. The sample shown here has the fourth-lowest language model score.



ASR Transcript
"a bottle of dishwashing soap blue liquid brand of dawn sitting on top of a white sink surrounding it is a Listerine bottle to the right can of hairspray to the left electric toothbrush bottle of soap and a bottle **ferments** on top hanging on the wall which is blue is a white medicine cabinet with a mirror"

Ground Truth
"a bottle of dishwashing soap blue liquid brand of dawn sitting on top of a white sink surrounding it is a Listerine bottle to the right can of hairspray to the left electric toothbrush bottle of soap and a bottle **for mints** on top hanging on the wall which is blue is a white medicine cabinet with a mirror"



ASR Transcript
"a red baseball cap on a hammock that is colored gray **and** gray Sofas in the background there appears to be a basket of **some time** sitting on the sofa there is a large picture hanging up on the sofa **wall**. On the left hand side that is colored yellow blue and white"

Ground Truth
"a red baseball cap on a hammock that is colored gray **a** gray Sofa is in the background there appears to be a basket of **some type** sitting on the sofa there is a large picture hanging up on the sofa wall. **There appears to be a baby playpen** on the left hand side that is colored yellow blue and white"

Figure 3-5: Example mistakes made in ASR captioning, including substitutions and deletions.

laborators easier, as they could quickly look at as many samples as they wanted to.

3.4 Transcript Correction

The transcripts generated via automatic speech recognition were generally sufficient for our validation steps, but they often contained errors. Most frequently, words would be replaced with other, phonetically similar words and phrases. Figure 3-5 shows several examples of mistakes that the ASR engine makes. In one instance, the ASR engine transcribes the phrase "for mints" as "ferments", which is phonetically similar. In the other image, the ASR engine omits the first part of a sentence, resulting in the deletion of seven words from the transcript. While this doesn't affect the performance of any audio-visual models that are trained directly on the audio waveform, it can impact the performance of image captioning models that are trained on the contents of text transcript. The omission of key phrases has a negative impact on the quality of the representations that those models learn from the text and image pairs.

To address this shortcoming, we decided to create another crowdsourcing task in which workers would play the audio file and correct the transcripts. To simplify things, we also built this on top of the existing Flask server. It works in a similar way to the image captioning task, with a few key differences. First, the input to the server is a list of all of the existing transcript files, not a list of all image files. We divide the list of files into tasks in a similar way as before, with four transcripts per task.

The original (uncorrected) transcripts are displayed one at a time and placed in an editable text field. The worker can click a button to play the original audio file, then edit the transcripts as the audio file plays. Once they have made all of their desired edits to the captions, the workers can submit the transcript and proceed to the next transcript, if applicable. Example submissions from workers are shown in Figure 3-6, and the interface is shown in Figure 3-7. The instructions, which were again shown above the correction interface, are included in Appendix A.2

Original transcript	Corrected transcript
a lamp that has a black neck and looks like a Candlestick holding a glass lamp shade that is green colored upside down around carpet	A lamp that has a black neck and looks like a candlestick holding a glass lamp shade that is cream colored upside down on a brown carpet.
how to plaid shirt hanging on a Ledge in a kitchen nursing chairs in the wall	it's plaid shirt hanging on a ledge in a kitchen there's some chairs and a the wall
held up by a woman	A tube of toothpaste, held up by a woman.

Figure 3-6: Examples of original transcripts and the corrected transcripts received from Amazon Mechanical Turk workers.

Just as in the image captioning task, we run inline validation steps to reduce the acceptance rate of low-quality submissions. If the worker attempts to submit a corrected transcript without listening to the audio clip, they are asked to retry and an error popup appears. Similarly, workers are asked to try again if their submitted transcript has fewer than four words. Upon submission, we mark the HIT as potentially being fraudulent if the worker completed the task in less than 15 seconds. This flags the submission for manual review later. The rest of the infrastructure is similar to that used in the image captioning task, which makes review simple.

3.4.1 Comparison of AMT and Rev

After collecting corrected captions for 500 samples on Amazon Mechanical Turk, we tested Rev, a video transcription service. They offered transcription services starting at \$1.25 per minute (at the time of our experiments). The primary metrics we were

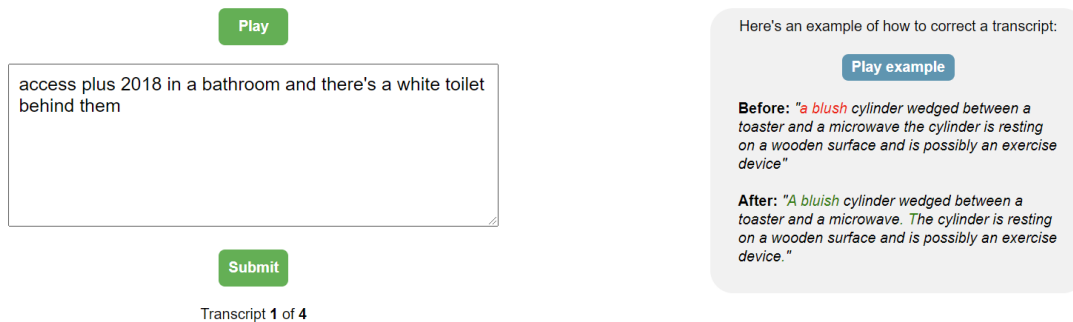


Figure 3-7: The interface workers used to correct transcripts. The text box automatically populated with the original transcript, and workers could see an example on the right hand side of the screen.

evaluating the two services on were cost and accuracy. Rev.com has a fixed rate and hires professional transcription workers, while AMT workers are not professionals but the price is adjustable. However, there is a minimum payment for AMT HITs for both ethical (Hara *et al.*, 2019) and economic reasons, as low-paying HITs will rarely be accepted.

For comparison, we used the same 500 samples that were corrected on AMT to create a 2.5 hour long video. The audio track of this video was a concatenation of all of the original audio captions, with one second of silence added in between. The video track was comprised of the still images corresponding to the active audio track, with one second of an empty screen added in between. Although Rev accepts audio-only files in its transcription service, we anticipated that the transcription workers would be aided by visuals. Sometimes listening to the audio alone wasn't enough to tell what a worker was saying, and in those cases a visual reference helped.

Once we received our captions back from Rev, we qualitatively compared the AMT and Rev corrected captions for each sample. Rev was faster and had a slightly lower word error rate, but AMT was significantly less expensive and produced transcripts of an acceptable quality. Based on these results, we decided to continue correcting transcripts on AMT instead of switching to Rev. In the future, however, if quality, convenience, or turnaround time are more important than cost, Rev would be a good option.

3.5 Additional Samples

Most popular image captioning datasets have 4 or 5 captions per image, as described in Section 2.3. In order to create an image captioning challenge based on ObjectNet, we needed to collect additional samples to reach that quantity of data. However, collecting 4 or 5 samples for the entire set of 50,273 ObjectNet images was outside the scope of this project. As a result, we decided to create a subset of 20,000 images that would receive the full set of captions. We call this subset Spoken ObjectNet-20k, or SON-20k.

SON-20k was constructed after the first set of captions was completed, so we used the collected captions to inform its creation. Because one of the distinguishing features of Spoken ObjectNet is that its captions are spontaneous speech and therefore feature significantly more words per caption than other captioning datasets, we started by looking at the existing captions for each image. A longer caption might suggest that the image has more details, while a shorter caption might suggest that the image is difficult to describe in some way (due to an unusual object orientation making object identification difficult, for example). We decided to sample most of the images in the 20k subset from the images that produced longer captions, as these images are more likely to be interesting from a captioning perspective. We also sampled from the more difficult images, though, in order to ensure that the split was reflective of the difficulty of the larger dataset. The 20k subset is balanced across classes in ObjectNet, too.

We used the same Amazon Mechanical Turk task to collect additional samples for the 20k subset. We first collected one additional caption for each image, then another, etc., until we reached our desired quantity of data. In its final form, Spoken ObjectNet-20k will contain 5 samples for each of the 20,159 images in the subset, for a total of 100,795 captions.

Chapter 4

Dataset Analysis

4.1 Spoken ObjectNet-50k

An analysis of the 50k split of Spoken ObjectNet, with a total of 50,273 spoken language captions, is presented here. Many captioning datasets, including QuerYD (Onicescu *et al.*, 2021), report the total size of the vocabulary as well as the number of unique nouns, verbs, adjectives, and adverbs used in the captions. These metrics provide insight into the distribution of captions, where a larger vocabulary generally indicates that there is a greater diversity of audio and visual content for models to learn from.

We analyzed the ASR transcripts of each of the 50,273 audio captions in Spoken ObjectNet-50k to find the size of the vocabulary. To compute part-of-speech tags for each caption, we used the Python spaCy library. Results are shown in Table 4.1.

Dataset Vocabulary							
Dataset	Speakers	Words	Nouns	Verbs	Adjectives	Adverbs	Avg. Length
SON-50k	1,030	18,780	11,666	3,252	2,324	478	21.2
SON-20k	1,710	25,768	15,910	4,506	3,236	630	23.6
Combined	1,792	27,554	16,959	4,812	3,453	667	22.6

Table 4.1: An analysis of the vocabulary of Spoken ObjectNet-50k, Spoken ObjectNet-20k, and the combined datasets. Each category shows the number of the unique speakers, words, etc. in each dataset.

Vocabulary Comparison						
Dataset	Total Audio	Words	Nouns	Verbs	Adjectives	Adverbs
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 Audio [11]	944h	51,764	27,074	11,293	8,271	1,660
QuerYD [25]	74h	28,515	8,825	3,551	3,128	907
DiDeMo [15]	67h	7,865	3,475	1,316	841	339
ACT [32]	31h	12,413	5,218	2,162	1,590	534

Table 4.2: A comparison of the vocabularies of Spoken ObjectNet, Places Audio, and several other popular audio-visual event localization datasets.

POS Distribution					
Dataset	% Nouns	% Verbs	% Adjectives	% Adverbs	% Other
SON-50k	62.1	17.3	12.4	2.5	5.6
Places-50k	55.7	21.5	14.7	3.1	5.0
Places Audio [11]	52.3	21.8	16.0	3.2	6.7
QuerYD [25]	30.9	12.5	11.0	3.2	4.2
DiDeMo [15]	44.2	16.7	10.7	4.3	24.1
ACT [32]	42.0	17.4	12.8	4.3	23.4

Table 4.3: A comparison of the distribution of common parts of speech in Spoken ObjectNet and other audio-visual datasets.

Spoken ObjectNet-50k has a comparable vocabulary to many other audio-visual datasets, as shown in Table 4.2. Given that it is smaller than most of the other datasets, the relatively large vocabulary speaks to the high average length of each caption and the diversity of objects and poses in ObjectNet. When the images have a lot of detail and the objects are in unusual settings, workers spontaneously captioning the images with speech are inclined to use more words than they would otherwise. Based on the analysis in Table 4.3, Spoken ObjectNet has a similar POS distribution to Places Audio, which is likely due to the similarity in data collection strategies. These datasets have significantly more nouns than the other audio-visual datasets, which feature similar frequencies of verbs, adjectives, and adverbs, but significantly fewer nouns.

Qualitatively, looking at examples of collected captions can generally indicate how well workers understood the instructions and how relevant the captions are to

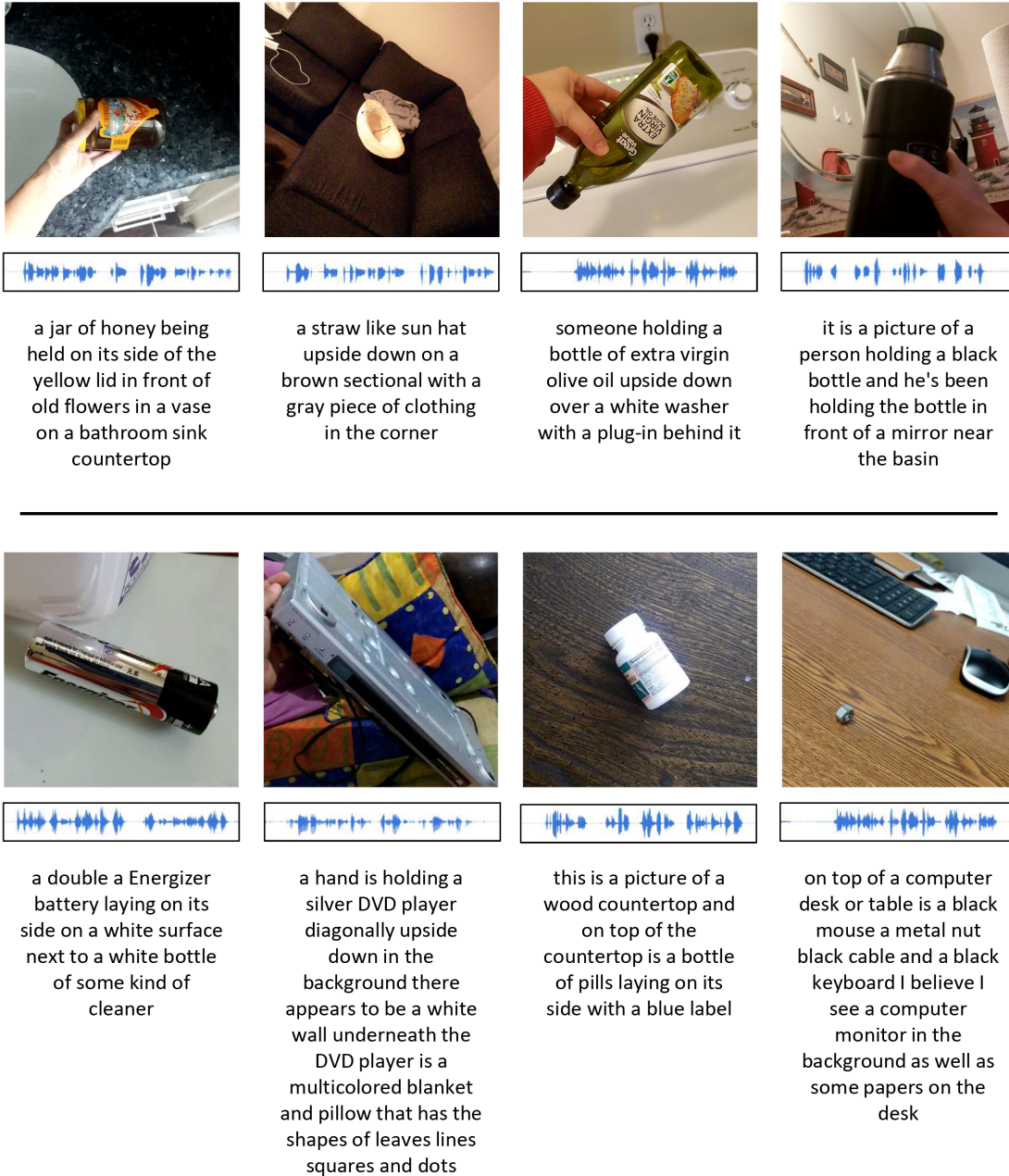


Figure 4-1: Samples of images and ASR captions from Spoken ObjectNet-50k.

the images. Figure 4-1 shows a number of samples from Spoken ObjectNet-50k. In these examples, the caption describes the image in high detail. In most examples, the caption describes both the main object of interest (the dataset class to which the image belongs) and the background. Descriptions that include both the foreground and the background (and therefore contain more words) allow models to learn richer representations. All in all, Spoken ObjectNet-50k contains rich descriptions useful to models that learn audio-visual correspondences.

4.2 Spoken ObjectNet-20k

Spoken ObjectNet-20k is intended to be used primarily for image captioning experiments, so while caption richness is just as important to this subset of the data as it is to Spoken ObjectNet-50k, the correspondences between the multiple captions for each image are also important. Qualitatively, selecting images from this subset and checking all of the captions submitted for that image to make sure they say roughly the same thing is a simple but effective way to gauge how useful this dataset is for image captioning. Several examples from Spoken ObjectNet-20k are shown in Figure 4-2. In these examples, the captions share many words and phrases, and in most cases the class to which the image belongs (`hairtie` and `razor`, respectively) is explicitly named within the caption. The shared vocabulary between captions is important in evaluating BLEU scores as well, which is particularly relevant in the image captioning experiments described in Chapter 6.

Vocabulary statistics for this subset are shown in Table 4.1. Because we specifically selected for captions that have a high average number of words in their ASR transcripts, Spoken ObjectNet-20k has on average more words per caption than Spoken ObjectNet-50k. It is also larger-scale than SON-50k, so it has a larger vocabulary.



Image	Captions
	<p>there is a hand holding what looks like a black ponytail holder in front of it in the upper right corner is a pocketbook with the egg in the top left corner there's a stand for camera</p> <p>is holding out a black rubber band for the for the hair his hand is laying on the yellow counter</p> <p>white hand holding a hair tie on top of a table with Hello Kitty purse on table</p> <p>somebody holding a rubber band in their hand on a wood surface and I see a purse that has like people on it</p> <p>I see a black hair tie being held near a purse on a wooden counter</p>
	<p>a person standing in a bathroom holding a blue and white disposable razor there is a tan rug on the floor with a pink floral pattern around the border</p> <p>a blue and white handled razor being held out above a brown rug that has flowers at the rib</p> <p>this picture of person is holding a razor in their hand looking down you can see there and they're holding razor bathroom as one with flowers on it down around the toilet</p> <p>somebody holding a razor over a rug that has flowers around the edge</p> <p>somebody holding up a blue and white razor above a beige rug that has like red and pink flowers around the rim of it</p>

Figure 4-2: Two examples from Spoken ObjectNet-20k, with five captions per image.

4.3 Places-50k

In order to compare the relative difficulties of training on Places Audio Captions versus Spoken ObjectNet, we created a split of the original Places Audio Captions dataset containing 48,273 training images (the same number of training samples as in Spoken ObjectNet). We called this split Places-50k, and called the original Places-400k. The validation set remains the same so we can directly compare the performance of models trained on Places-50k with that of models trained on the full dataset.

This subset contains a comparable vocabulary to Spoken ObjectNet-50k. It has a slightly higher number of unique words, but fewer unique nouns. The average length of ASR captions in Places-50k is 19.1, versus 21.2 in Spoken ObjectNet-50k.

4.4 Dataset Splits

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 three sets.

Chapter 5

Retrieval Experiments

With Spoken ObjectNet-50k now complete, we wanted to conduct a series of experiments using the data in order to compare the relative difficulty of training on Spoken ObjectNet-50k and Places-50k. In addition, we wanted to model how the dataset can be used in practice and measure the performance of models in a transfer learning setting.

5.1 Experimental Setup

The retrieval tasks that we conducted were from image to audio and audio to image. In the image to audio setting, the model was presented with an image and tasked with retrieving the most relevant audio captions from the dataset for that image. The audio to image setting was the reverse. We report two results, recall at 1 and recall at 10 (R@1 and R@10, respectively). For image to audio R@N, the model is successful if any of the top N recalled audio waveforms are the correct match to the given image (and vice versa for audio to image retrieval). Additionally, all experiments were conducted with the **ResDAVEnet-VQ** Harwath *et al.* (2018) model with all quantization layers turned off.

The license of ObjectNet ¹ prohibits model parameters from being adjusted based on the information in the dataset’s images. This is in line with the dataset’s intended

¹See <https://objectnet.dev/>

use as a test set for audio-visual models. As a result, we tested a setting in which the image branch of the model (used to embed images into a joint embedding space) was frozen. In this setting, only the parameters of the audio model and the final embedding layer could be trained. We also conducted experiments to measure the effect of image model pretraining in the **ResDAVEnet-VQ** models. In some experiments, the image branch ResNet was initialized with weights from a model trained on ImageNet (Deng *et al.*, 2009), while in other experiments the weights were randomly initialized.

5.2 Implementation Details

All models were trained for 150 epochs with a batch size of 64 using the Adam (Kingma & Ba, 2015) optimizer. We chose different learning rates depending on the setting, however. During experiments with fully trainable image and audio branches, we used the original learning rate of $2 \cdot 10^{-4}$. When the image branch was frozen, we increased the learning rate to $1 \cdot 10^{-3}$. We chose this learning rate after conducting a sweep over several different learning rates, ranging from $1 \cdot 10^{-5}$ to $1 \cdot 10^{-2}$. Models trained with this learning rate had the highest validation recall scores, all other parameters held constant.

For other parameters, we largely based our decisions on the hyperparameters specified in the original **ResDAVEnet-VQ** paper (Harwath *et al.*, 2018). Our learning rate exponentially decayed by a factor of 0.95 every 3 epochs. During training, images were resized such that their smallest dimension was 256 pixels, then a random 224 by 224 crop was taken from the image. During validation, the center 224 by 224 crop was always taken. Images were also randomly flipped with a probability of 0.5.

We computed validation recall scores after every epoch of training. We report the maximum validation recall score out of the 150 epochs as the score of the model. Because Places Audio Captions does not have a pre-defined test set, this procedure allowed us to make the best comparison between Spoken ObjectNet (which does have 1,000 samples reserved as a test set) and Places Audio Captions.

5.3 Transfer Experiments

To better 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. The original model was the best ResDAVEnet-VQ model (without any vector quantization layers enabled) that was trained on Places-400k. This model achieved a mean R@10 of 0.735 on the Places-400k validation set (Harwath *et al.*, 2020). 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 is prohibited due to ObjectNet’s license.

Transfer from Places-400k to Spoken ObjectNet						
Setting	I \rightarrow A		A \rightarrow I		Mean	
	R@1	R@10	R@1	R@10	R@1	R@10
No Fine-tuning (Zero-shot)	0.019	0.096	0.033	0.140	0.026	0.118
Fine-tuning (Frozen image branch)	0.040	0.216	0.048	0.213	0.044	0.214
Fine-tuning (Trainable image branch)	0.102	0.391	0.115	0.416	0.108	0.403

Table 5.1: Results of retrieval experiments based on transfer learning from a model trained on Places-400k.

The results are shown in Table 5.1. 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. For comparison, the chance R@10 (achievable by guessing uniformly at random from the entire dataset) is 0.010. 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 con-

trols 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.

5.4 Comparing Spoken ObjectNet & Places Audio

Table 5.2 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. However, this is a relatively small amount of data to train these audio-visual models on. In general, ResDAVENet-VQ is trained on over 400,000 samples, not the 48,000 training samples in these datasets. Training from scratch results in models that are still likely under-trained, so performance is heavily based on how well the pretrained image branch transfers to the respective datasets.

Frozen image branch							
		I \rightarrow A		A \rightarrow I		Mean	
Dataset	Pretraining	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
Places-50k	ImageNet	0.093	0.364	0.079	0.360	0.086	0.362

Table 5.2: Comparison of training on Spoken ObjectNet-50k versus Places-50k with frozen image branches.

In Table 5.3, 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.

		Trainable image branch					
		I \rightarrow A		A \rightarrow I		Mean	
Dataset	Pretraining	R@1	R@10	R@1	R@10	R@1	R@10
SON	ImageNet	0.066	0.315	0.089	0.332	0.077	0.324
Places-50k	ImageNet	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 5.3: Comparison of training on Spoken ObjectNet-50k versus Places-50k with trainable image branches.

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.

The top 5 retrieved audio captions for two Spoken ObjectNet samples are shown in Figure 5-1. The correct caption (boxed in green) is retrieved for the first example, but none of the top 5 retrieved captions are the true caption for the second example.

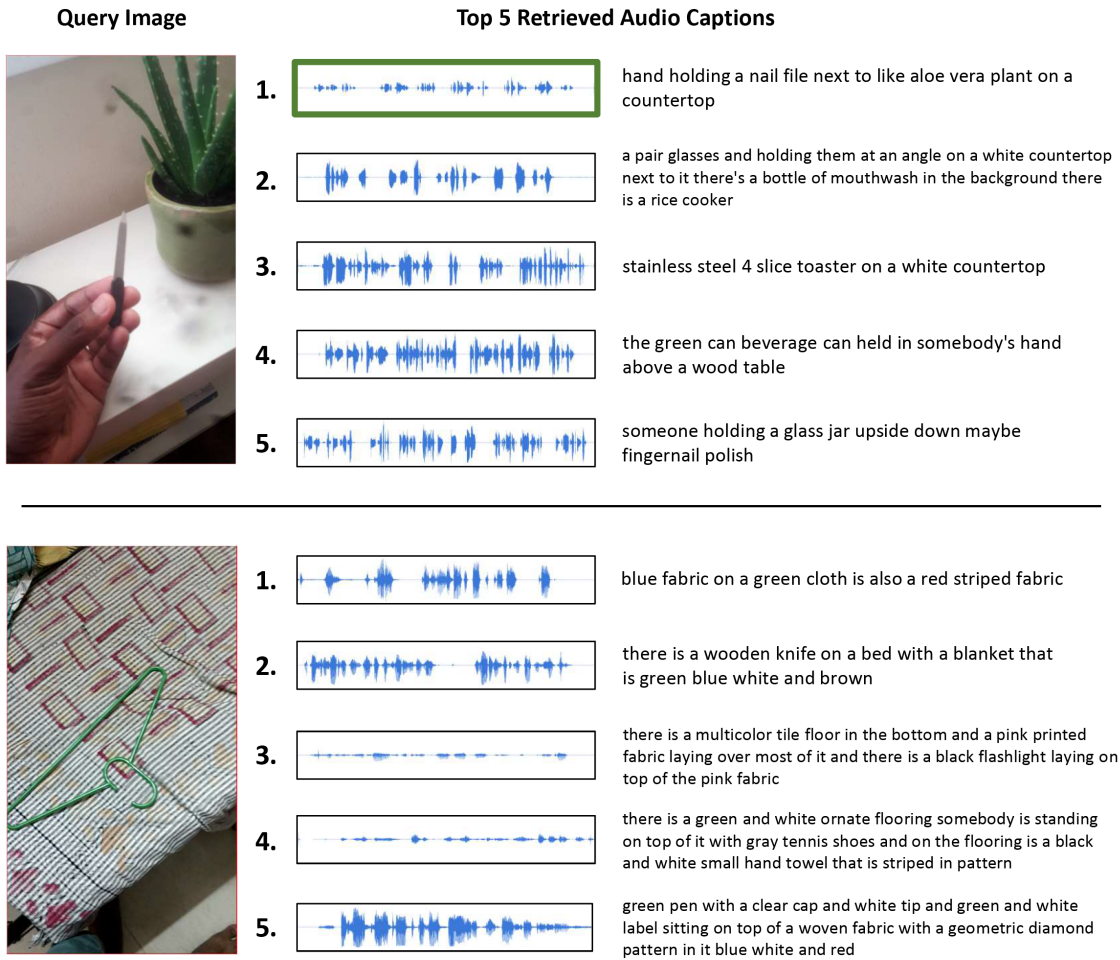


Figure 5-1: Top 5 retrieved audio captions for two sample images. The true caption for the image (if applicable) is boxed in green.

Qualitatively, looking at the retrieved captions (both correct and incorrect) in this example can provide insight into the representations that the model is learning. In the first example, the term "white countertop" appears multiple times in the retrieved captions. The word "countertop" appears in the true caption as well. This suggests that one of the most salient features in the image is the countertop. Additionally, several captions use the word "holding". This suggests that the model has learned some representation of when objects are held in a person's hand, which occurs frequently within ObjectNet.

Similarly, the retrieved captions for the second image provide insight into the model's learned representations, even though none of them are the true caption for

the image. All of the top 5 captions mention color in some way - the green color of the coat hanger, the red accents on the blanket, or the white color of the bedsheet. One of the captions references a green and white tile floor, which might bear some resemblance to the green coat hangar on top of a white blanket with a tile pattern. Even though the model was ultimately incorrect, the relevant features found in the impostor captions suggests that it is learning good cross-modal representations.

5.5 Analysis

In conclusion, the results of the retrieval experiments suggest that multimodal models have a lower performance on Spoken ObjectNet than a comparable subset of the Places dataset. This performance gap is due to the priors in the dataset, as when those priors are removed (as it is in the zero-shot transfer setting) performance drops dramatically, but when both models are trained from scratch on the same amount of data (and the model can learn better representations that are not dependent on dataset bias), the performance is almost equivalent. Next, we turn to another audio-visual task, image captioning, to further explore the effects of bias controls in Spoken ObjectNet.

Chapter 6

Image Captioning Experiments

While Spoken ObjectNet-50k, with one caption per image, can support retrieval experiments, Spoken ObjectNet-20k can support image captioning experiments because it contains multiple captions per image. We wanted to conduct captioning experiments to both compare Spoken ObjectNet-20k against other image captioning datasets and compare the change in performance of retrieval tasks with the change in performance of captioning tasks.

6.1 Experimental Setup

In these experiments, models are provided with an image and asked to produce a natural language caption describing the contents of that image. In our experiments, we use a CNN-LSTM-based model that employs self-critical sequence training (SCST), as described in Section 2.6. This model builds up feature vectors using a ResNet-101 convolutional neural network, then decodes those representations into natural language using a LSTM and an algorithm inspired by the field of reinforcement learning.

ObjectNet has a license that prohibits model parameters from being tuned on images in the dataset, but because SCST uses a pretrained ResNet-101 model to generate image features, no backpropagation is required on the image model. This cooperates nicely with the license of ObjectNet. As a result, SCST can be applied to Spoken ObjectNet as-is, although higher performance could perhaps be achieved by

training the image model on the target dataset.

6.2 Implementation Details

We used the SCST FC-2k features, in which each image is encoded with a ResNet-101 model pretrained on ImageNet without any cropping or resizing. The final convolutional layer is taken from this model and average pooling is applied to obtain a 2048-dimensional vector. During training, we used a batch size of 64 and a learning rate of $5 \cdot 10^{-4}$. The rest of the model parameters were the same as in the original SCST paper (Rennie *et al.*, 2017), with 512-dimensional LSTM embeddings, a learning rate decay factor of 0.8, and the Adam optimizer.

We trained this model in two stages. In the first warmup stage, the model is trained without self-critical sequence training. After 30 epochs, we decrease the learning rate to $5 \cdot 10^{-5}$ and decrease the batch size to 10. We train for 600,000 more iterations, after which the language evaluation metrics rise considerably.

We conducted validation steps during training in regular intervals based on the number of optimizer steps taken. In each of these validation steps, we computed a number of language evaluation metrics, including BLEU (Papineni *et al.*, 2002), METEOR (Lavie & Agarwal, 2007), ROUGE (Lin, 2004), CIDEr (Vedantam *et al.*, 2015), and SPICE (Anderson *et al.*, 2016). We report the evaluation results of the epoch in which the highest CIDEr score was achieved.

6.3 Captioning From Scratch

Table 6.1 shows language evaluation metrics for several different machine captioning experiments, trained on both COCO Captions and Spoken ObjectNet-20k. The experiments without SCST show the evaluation metrics of the model after the 30 epochs of warmup training, and the SCST results show the maximum results achieved during the 600k additional training iterations.

On COCO, we achieve near the maximum score reported in the original SCST

Language Evaluation Scores					
Setting: train on COCO, evaluate on COCO					
Setting	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
COCO (no SCST)	28.8	24.3	52.4	90.8	0.17
COCO (with SCST)	30.6	24.8	53.5	101.8	0.18

Setting: train on SON-20k, evaluate on SON-20k					
Setting	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
SON-20k (no SCST)	5.7	9.1	27.3	4.4	0.03
SON-20k (with SCST)	11.9	11.3	32.8	6.6	0.05

Table 6.1: Results of training captioning models from scratch on both COCO Captions and SON-20k, with and without self-critical sequence training (SCST).

paper (Rennie *et al.*, 2017), with a BLEU score of 30.6 and a CIDEr score of 101.8. The language evaluation metrics improve considerably during SCST, as the BLEU score rises 2 points and the CIDEr score improves by 11 points. Based primarily on the SCST experiments, which are the main experiments of interest in this chapter, this captioning model performs very well on the COCO Captions dataset.

However, there is a significant performance drop when the same model is trained from scratch on Spoken ObjecNet-20k. Without SCST, the model achieves a BLEU score of just 5.7 and a CIDEr score of 4.4 - a dramatic drop from the 28.8/90.8 scores when trained on COCO Captions, even without SCST. When trained with SCST, the language evaluation scores undergo a significant relative improvement, but the model still performs significantly worse than the model trained on COCO Captions. The BLEU score doubles to 11.9, and the CIDEr score increases to 6.6.

From these experiments, it is clear that the model is less effective at captioning ObjectNet images than COCO images under the same training regime, all else equal. As in the retrieval experiments, part of the decrease may be due to the use of a CNN pretrained on ImageNet. COCO images are often visually similar to ImageNet, and neither contains any debiasing controls. ObjectNet, on the other hand, has controls for biases and as such the CNN will not be as effective in producing useful features

from those images. Overall, these results mirror those in the retrieval experiments, specifically those in which the models used frozen image branches pretrained on ImageNet.

6.4 Transfer Experiments

Table 6.2 details our results of our transfer experiments. We present results for a zero-shot setting and a fine-tuning setting. In the zero-shot setting, the model trained on COCO Captions is then evaluated on Spoken ObjectNet-20k without fine-tuning on the dataset. In the fine-tuning setting, the model is allowed to train for 600k iterations with SCST before being evaluated. In these experiments, because the features are pre-generated, the model is updating the parameters in its decoder using the REINFORCE-like algorithm described in Section 2.6.

Transfer Experiments					
Setting	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Zero-shot	7.2	10.6	28.7	4.5	0.04
Fine-tuning	8.1	10.8	28.5	5.2	0.04

Table 6.2: The results of several transfer experiments in which a model that was originally trained on COCO Captions was evaluated on SON-20k under several different fine-tuning settings.

In the zero-shot setting, the model is not able to produce accurate captions for the images. The language evaluation metrics fall significantly from the evaluation results on MS-COCO, as shown in Table 6.1. In some ways, it is surprising that the model does so poorly on ObjectNet images. To humans, the images are not all that different. To the model, however, the novel object classes, unusual orientations, and backgrounds of the objects in ObjectNet are enough to substantially reduce performance.

In the fine-tuning setting, performance is slightly better than in the zero-shot setting. Just as in the retrieval experiments, training the model on the audio or textual components of Spoken ObjectNet, even without training the image branch,

recovers some of the performance lost in the zero-shot setting. We do not show results for a captioning experiment in which the image model is trained because the SCST model uses pre-processed features, but it would be expected to further increase the performance of the model at the expense of violating ObjectNet’s license.

Surprisingly, the fine-tuned model is not able to match the performance of the model trained from scratch. This is most likely due to the decayed learning rate, which may be too low to effectively learn new representations after learning representations tuned for COCO for over 600k steps. We leave improvement of the fine-tuning setting to future work, as our primary goal is to show that performance is poor in the zero-shot setting, but some performance can be recovered by fine-tuning.

6.5 Additional Models

To further explore the decrease in performance of models trained from scratch on Spoken ObjectNet, we conducted additional experiments using other image captioning models. The first of these models is an attention-based captioning model also introduced in Rennie *et al.* (2017), referred to in the paper as att2in. The key difference between this model and the FC model is that attention layers re-weight the CNN features produced by the ResNet101 during preprocessing. This allows the model to improve the features used as input for captioning.

We also test a simple baseline model that only attempts to train a LSTM decoder using constant input. While not useful in practice, this model gives a baseline idea of how much performance can be recovered from simply predicting tokens that commonly appear in the captions. Intuitively, n-grams like "a man", "a woman", and "is sitting" are likely to occur frequently in the dataset, so an above-chance performance is possible even without using image features.

Results from both of these models are shown in Table 6.3. In each experiment, the model was trained for 30 epochs, and we select the scores from the highest-performing epoch based on CIDEr scores. Unlike the above experiments, we did not use self-critical sequence training in these experiments.

Other Models: Language Evaluation Scores					
Setting: train on COCO, evaluate on COCO					
Model	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Att2in	29.2	24.6	52.7	93.1	0.18
Baseline LSTM	4.4	11.2	32.2	7.5	0.02

Setting: train on SON-20k, evaluate on SON-20k					
Setting	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
Att2in	8.2	10.8	28.7	6.2	0.05
Baseline LSTM	2.7	7.2	25.3	2.2	0.01

Table 6.3: Results of our additional captioning experiments on both COCO Captions and SON-20k.

In general, the Att2in model performs better than the FC model without self-critical sequence training, but worse than the FC model with SCST. The performance dropoff between COCO and Spoken ObjectNet-20k is approximately the same, further confirming the difficulty of Spoken ObjectNet-20k as compared to COCO Captions.

The baseline model performs poorly on both datasets, which is expected. Its performance is slightly lower on Spoken ObjectNet-20k than COCO Captions, but not by the same margin as in other experiments. Interestingly, the baseline is lower, but not all that different from the results of the FC and Att2in models on Spoken ObjectNet-20k. This suggests that a model that doesn't use image features can achieve about half of the performance of a model that does use image features on SON-20k. We hypothesize that the image features, then, are not that useful to the model during captioning. This aligns with our expectation that an ImageNet-pretrained feature extractor (as these models use) would be much less effective at extracting features from ObjectNet images than COCO images. The baseline experiments provide evidence to support that conclusion.

Overall, these additional experiments with several other models serve to provide further evidence that SON-20k is a more challenging dataset to train certain captioning models on than COCO Captions.

6.6 Generated Captions

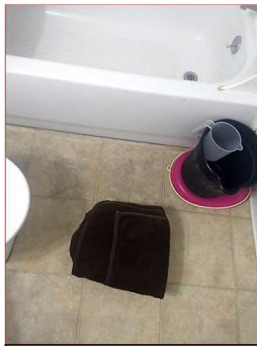
Lastly, we qualitatively evaluated the performance of the captioning models by having several of our trained models generate captions based on 100-image subsets of the validation sets of both COCO and ObjectNet. Figure 6-1 shows captions generated for the COCO dataset by the model trained on COCO Captions with SCST. Overall, these captions are very grammatical and accurately reflect the contents of the image. Figure 6-2 shows the captions generated by the same model on ObjectNet images. Unlike the first setting, the model struggles to produce accurate captions. In many cases, the generated caption doesn't seem to reference the contents of the image at all.

6.7 Analysis

These captioning experiments suggest that models that are trained on standard image captioning datasets then evaluated on Spoken ObjectNet suffer significant performance decreases, mirroring the result of our retrieval experiments in Chapter 5. In particular, the generated captions in Figures 6-1 and 6-2 are a reminder of how brittle many of the state-of-the-art machine learning models in existence today are. Overall, conducting these image captioning experiments provided us with additional evidence for the trends discovered in our retrieval experiments while simultaneously experimenting with the approximately 80,000 samples included in Spoken ObjectNet-20k that are not in Spoken ObjectNet-50k.



Figure 6-1: Examples of captions for COCO images produced by our model trained on COCO Captions.



a bathroom with a toilet and a sink



a statue of a snow is sitting in the snow



a person is sitting on a table with a plate of food



a pair of shoes sitting on the side of a pair of shoes



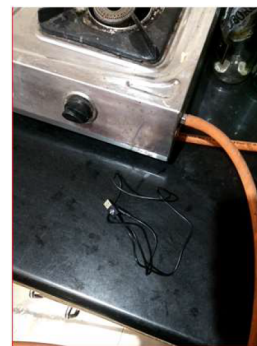
a black and white photo of a man standing on a waist



a dog is sitting on top of a bed



a person is sitting on a table with an umbrella



a pair of scissors sitting on top of a table



a view of a view of a bed in a room



a piece of cake on a bed with a table



a man laying on the back of a bed with a tie



a pair of scissors sitting on top of a table

Figure 6-2: Examples of captions for ObjectNet images produced by our model trained on COCO Captions. This is the zero-shot setting described in Table 6.2.

Chapter 7

Conclusion

In conclusion, we present Spoken ObjectNet, a novel large-scale bias-controlled spoken caption dataset. Spoken ObjectNet is best used as a test set for audio-visual models trained on other datasets. It can provide a better indication of how well a model will generalize to real-world data than held-out data from the original training dataset.

To collect samples for Spoken ObjectNet, we developed some novel components in the data collection pipeline, including the language modeling checks detailed in Chapter 3. This improved the overall quality of our captions and decreased the amount of manual validation required. In addition, we developed an evaluation server that allowed us to quickly view the entire corpus of collected captions and determine how effective our language modeling score was. In the last part of our data collection pipeline, we created a transcript correction task and evaluated several different transcription services on the basis of cost and accuracy.

An analysis of our dataset shows that it contains a comparable vocabulary to other popular large-scale audio-visual datasets. We present two splits of Spoken ObjectNet, called Spoken ObjectNet-50k and Spoken ObjectNet-20k. Spoken ObjectNet-50k contains one caption for each image in the ObjectNet dataset and is intended to test audio-visual models. Spoken ObjectNet-20k, on the other hand, is primarily intended for image captioning. It features 5 images per caption, but for a smaller subset of approximately 20,000 ObjectNet images.

In both retrieval experiments and image captioning experiments, models suffered

significant performance drops when trained on other datasets and then evaluated on Spoken ObjectNet. In both cases, the performance of the models was limited by their ability to produce meaningful representations of the ObjectNet images, which did not feature the same biases encoded into their previous training datasets. These experiments provided insight into the properties of this dataset and confirmed that controls for image biases can impact the performance of audio-visual models.

7.1 Dataset Release

We plan on releasing Spoken ObjectNet to the public under the Creative Commons Attribution 4.0 license. This is the same license that the original ObjectNet dataset uses, but while ObjectNet prohibits models from being trained on its images we do not make the same restriction on the audio files in Spoken ObjectNet. Models may be tuned on the audio waveforms in Spoken ObjectNet, but the restriction on training on images still applies. Models must test either zero-shot transfer performance or train with a frozen image branch on Spoken ObjectNet.

7.2 Future Work

Beyond simply providing researchers with a new dataset to use in their evaluation pipelines, we wish to inspire a broader conversation about the current limitations of modern machine learning through this work. While the sudden explosion of deep learning has improved the performance of machine learning models, the rate at which this change is occurring makes it incredibly difficult to carefully consider the benefits and limitations of every emerging technique. In general, too, promising new results produce more excitement than careful analyses of existing results. Given the increasing integration of deep learning-enabled systems into our daily lives, then, it may be time to take a step back and examine the current state of the field.

The limitations of some modern models can be seen in Figure 6-2. Here, a model that performs very well on a traditional image captioning task produces nonsensical

captions for images that are visually very similar to humans. More broadly, training models on large-scale datasets like MS-COCO produces models that produce great results for images in the training set, good results for held-out images that are still in the original distribution of data, and poor results for any images that are not in the same distribution of data. In image captioning, poor performance on that data might not have a significant effect; perhaps an incorrect alt-text is generated for an image on a social network. As machine learning is integrated into more sensitive fields like medicine, however, the risk is greater. Society's expectation that all people should receive equal treatment, all else equal, has the potential to be subverted by machine learning models that learn from fundamentally biased information and have no concept of ethical decision-making.

There is a solution, though. We believe that through careful engineering of the entire machine learning pipeline - from data collection to embedding in production systems - it is possible to construct a system that is capable of making predictions with a minimal adverse effect from learned bias. After all, systems that make unbiased predictions are more trustworthy, more explainable, and no less capable than other systems. Our work here is an attempt to de-bias the dataset development component of the pipeline. In the same vein, we hope that this work inspires future research in the other components of the pipeline.

Appendix A

Instructions

The following sections contain the instructions we used in our Amazon Mechanical Turk tasks to instruct workers on how to complete the image captioning and transcript correction HITs.

A.1 Image Captioning Task

Disclaimer: This HIT is part of a MIT scientific research project. Your decision to complete this HIT is voluntary, and your responses are anonymous. The results of the research may be presented at scientific meetings, published in scientific journals, or made publicly available to other researchers. Clicking on the 'Submit' button indicates that you are at least 18 years of age, you are a native English speaker, and you agree to complete this HIT voluntarily. Because this scientific research study requires a balanced amount of speech collected from many different individuals, we can only accept up to 3,000 HITs from any single worker.

Notice: If you encounter any issues or find any bugs, please email us **with the copy-pasted text of any error messages you receive**, and we will do our best to fix them. If you consistently encounter this error, please **do not continue to attempt to complete more HITs**, and instead email us with some information about your system configuration including your operating system and web browser version.

Requirements: To complete this task, you must be in a relatively quiet environment on a computer equipped with a microphone, using one of the following web browsers: Edge, Chrome, Firefox, Safari, or Opera. **You must have cookies enabled** or you will be unable to submit the HIT.

Instructions: You will be submitting audio recordings using the interface below.

1. When prompted, grant permission to the site to use your microphone for the duration of the HIT.
2. Use the volume meter in the bottom-right of the window to help ensure that your microphone is working properly, and that you are a proper distance away from it. The meter should move as you speak. **If the volume meter does not move, or if the recording button is disabled, please check to make sure that you have given permission to your web browser to access your microphone.**
3. Press the green "Record" button to start recording. After you press it, the button will turn into a red "Stop" button.
4. Complete the task as described below.
5. Press the red "Stop" button to stop recording. After you press it, your audio recording will be processed automatically.
6. If your recording is acceptable, you will be prompted with the next photo. Otherwise, you will be asked to try recording again.
7. Once you have submitted all the necessary recordings, press the green "Submit" button to submit the HIT.

Task: Throughout the task, on the left of the screen you will be presented with 4 different images, one at a time. **Please record yourself describing each image as if you were explaining it to someone who could not see it.** We're looking for a couple of sentences per image. You can talk about specific objects, locations, shapes, colors, etc. in the image. For help, refer to the example on the right.

A.2 Transcript Correction Task

Disclaimer: This HIT is part of a MIT scientific research project. Your decision to complete this HIT is voluntary, and your responses are anonymous. The results of the research may be presented at scientific meetings, published in scientific journals, or made publicly available to other researchers. Clicking on the 'Submit' button indicates that you are at least 18 years of age, you are a native English speaker, and you agree to complete this HIT voluntarily. Because this scientific research study requires a balanced amount of speech collected from many different individuals, we can only accept up to 3,000 HITs from any single worker.

Notice: If you encounter any issues or find any bugs, please email us **with the copy-pasted text of any error messages you receive**, and we will do our best to fix them. If you consistently encounter this error, please **do not continue to attempt to complete more HITs**, and instead email us with some information about your system configuration including your operating system and web browser version.

Requirements: To complete this task, you must be able to listen to short audio clips and use a keyboard. You must use one of the following web browsers: Edge, Chrome, Firefox, Safari, or Opera. You must have cookies enabled or you will be unable to submit the HIT.

Instructions: You will be listening to audio clips and correcting their automatically generated transcripts using the interface below.

1. You may use the "Play example" button to ensure that your audio playback system is working properly. If you click the button and do not hear a sound, please double check your system settings to ensure you can hear audio playback.
2. Press the green "Play" button to begin playback of the sound clip.
3. Use the text box on the left to transcribe the contents of the audio clip. To help you do so, the box is pre-filled with automatically generated text that may contain errors.

4. Press the "Submit" button when you have finished correcting the transcript.
5. Once you have submitted all the necessary transcripts, press the green "Submit" button to submit the HIT.

Task: Use the green "Play" button to listen to an audio clip. Using the text box on the left, correct the audio transcription. You should fix incorrectly transcribed words and add punctuation, but don't add words that are not in the recording. When you are done, click the green "Submit" button. For help, refer to the example on the right.

Bibliography

- [1] Anderson, Peter, Fernando, Basura, Johnson, Mark, & Gould, Stephen. 2016. SPICE: Semantic Propositional Image Caption Evaluation. *In: ECCV*.
- [2] Barbu, Andrei, Mayo, David, Alverio, Julian, Luo, William, Wang, Christopher, Gutfreund, Dan, Tenenbaum, Josh, & Katz, Boris. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Pages 9448–9458 of: Advances in Neural Information Processing Systems*, vol. 32.
- [3] Chen, Honglie, Xie, Weidi, Vedaldi, Andrea, & Zisserman, Andrew. 2020. VG-SSound: A Large-scale Audio-Visual Dataset. *In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.
- [4] Chen, Xinlei, Fang, Hao, Lin, Tsung-Yi, Vedantam, Ramakrishna, Gupta, Saurabh, Dollár, Piotr, & Zitnick, C. Lawrence. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *CoRR*.
- [5] Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, & Fei-Fei, Li. 2009. Imagenet: A large-scale hierarchical image database. *In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [6] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, & Toutanova, Kristina. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics.
- [7] Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. *Pages 776–780 of: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.
- [8] Gurari, Danna, Zhao, Yinan, Zhang, Meng, & Bhattacharya, Nilavra. 2020. Captioning images taken by people who are blind. *Pages 417–434 of: European Conference on Computer Vision*. Springer.
- [9] Hara, Kotaro, Adams, Abigail, Milland, Kristy, Savage, Saiph, Hanrahan, Benjamin V., Bigam, Jeffrey P., & Callison-Burch, Chris. 2019. Worker Demographics and Earnings on Amazon Mechanical Turk: An Exploratory Analysis. *Page 1–6 of:*

Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery.

- [10] Harwath, David, & Glass, James. 2015. Deep multimodal semantic embeddings for speech and images. *In: Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*.
- [11] Harwath, David, Recasens, Adria, Surís, Dídac, Chuang, Galen, Torralba, Antonio, & Glass, James. 2018. Jointly discovering visual objects and spoken words from raw sensory input. *In: Proceedings of the European Conference on Computer Vision (ECCV)*.
- [12] Harwath, David, Hsu, Wei-Ning, & Glass, James. 2020. Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech. *In: Proceedings of the International Conference on Learning Representations (ICLR)*.
- [13] Havard, William, Besacier, Laurent, & Rosec, Olivier. 2017. SPEECH-COCO: 600k Visually Grounded Spoken Captions Aligned to MSCOCO Data Set. *In: Proc. GLU 2017 International Workshop on Grounding Language Understanding*.
- [14] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian. 2016. Deep residual learning for image recognition. *Pages 770–778 of: Proceedings of the IEEE conference on computer vision and pattern recognition*.
- [15] Hendricks, Lisa Anne, Wang, Oliver, Shechtman, Eli, Sivic, Josef, Darrell, Trevor, & Russell, Bryan. 2017 (Aug.). Localizing Moments in Video with Natural Language. *In: ICCV*.
- [16] Hsu, Wei-Ning, Harwath, David, Song, Christopher, & Glass, James. 2020. Text-Free Image-to-Speech Synthesis Using Learned Segmental Units. *Conference on Neural Information Processing Systems (NeurIPS) Workshop on Self-Supervised Learning for Speech and Audio Processing*.
- [17] Jansen, Aren, Plakal, Manoj, Pandya, Ratheet, Ellis, Daniel PW, Hershey, Shawn, Liu, Jiayang, Moore, R Channing, & Saurous, Rif A. 2018. Unsupervised learning of semantic audio representations. *Pages 126–130 of: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE.
- [18] Kingma, Diederik P, & Ba, Jimmy. 2015. Adam: A method for stochastic optimization. *In: Proceedings of the International Conference on Learning Representations (ICLR)*.
- [19] Lavie, Alon, & Agarwal, Abhaya. 2007. Meteor: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. *Page 228–231 of: Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

- [20] Li, Xiujun, Yin, Xi, Li, Chunyuan, Hu, Xiaowei, Zhang, Pengchuan, Zhang, Lei, Wang, Lijuan, Hu, Houdong, Dong, Li, Wei, Furu, Choi, Yejin, & Gao, Jianfeng. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. *ECCV 2020*.
- [21] Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Pages 74–81 of: Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics.
- [22] Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollár, Piotr, & Zitnick, C Lawrence. 2014. Microsoft coco: Common objects in context. *Pages 740–755 of: European Conference on Computer Vision (ECCV)*. Springer.
- [23] Luo, Ruotian, Price, Brian, Cohen, Scott, & Shakhnarovich, Gregory. 2018. Discriminability objective for training descriptive captions. *arXiv preprint arXiv:1803.04376*.
- [24] Miech, Antoine, Zhukov, Dimitri, Alayrac, Jean-Baptiste, Tapaswi, Makarand, Laptev, Ivan, & Sivic, Josef. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. *In: ICCV*.
- [25] Oncescu, Andreea-Maria, Henriques, João F., Liu, Yang, Zisserman, Andrew, & Albanie, Samuel. 2021. QuerYD: A video dataset with high-quality text and audio narrations. *In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.
- [26] Papineni, Kishore, Roukos, Salim, Ward, Todd, & Zhu, Wei-Jing. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Page 311–318 of: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.
- [27] Pont-Tuset, Jordi, Uijlings, Jasper, Changpinyo, Soravit, Soricut, Radu, & Ferrari, Vittorio. 2020. Connecting Vision and Language with Localized Narratives. *In: ECCV*.
- [28] Rashtchian, Cyrus, Young, Peter, Hodosh, Micah, & Hockenmaier, Julia. 2010. Collecting Image Annotations Using Amazon’s Mechanical Turk. *Pages 139–147 of: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*.
- [29] Rennie, Steven J, Marcheret, Etienne, Mroueh, Youssef, Ross, Jerret, & Goel, Vaibhava. 2017. Self-critical sequence training for image captioning. *Pages 7008–7024 of: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- [30] Sharma, Piyush, Ding, Nan, Goodman, Sebastian, & Soricut, Radu. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. *In: Proceedings of ACL*.

- [31] Vedantam, Ramakrishna, Zitnick, C. Lawrence, & Parikh, Devi. 2015. CIDEr: Consensus-based image description evaluation. *Pages 4566–4575 of: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [32] Wang, Xiaolong, Farhadi, Ali, & Gupta, Abhinav. 2016. Actions ~ Transformations. *In: CVPR*.
- [33] Williams, Ronald J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 229–256.
- [34] Wolf, Thomas, Debut, Lysandre, Sanh, Victor, Chaumond, Julien, Delangue, Clement, Moi, Anthony, Cistac, Pierric, Rault, Tim, Louf, Rémi, Funtowicz, Morgan, Davison, Joe, Shleifer, Sam, von Platen, Patrick, Ma, Clara, Jernite, Yacine, Plu, Julien, Xu, Canwen, Scao, Teven Le, Gugger, Sylvain, Drame, Mariama, Lhoest, Quentin, & Rush, Alexander M. 2020. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*.
- [35] Zhou, Bolei, Lapedriza, Agata, Xiao, Jianxiong, Torralba, Antonio, & Oliva, Aude. 2014. Learning Deep Features for Scene Recognition using Places Database. *In: Advances in Neural Information Processing Systems*, vol. 27.
- [36] Zhou, Luowei, Xu, Chenliang, & Corso, Jason J. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. *Pages 7590–7598 of: AAAI Conference on Artificial Intelligence*.