MIT Open Access Articles

# *Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input*

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

| | |
|---|---|
| **As Published** | https://doi.org/10.1007/s11263-019-01205-0 |
| **Publisher** | Springer US |
| **Version** | Author's final manuscript |
| **Citable link** | https://hdl.handle.net/1721.1/131515 |
| **Terms of Use** | Creative Commons Attribution-Noncommercial-Share Alike |
| **Detailed Terms** | http://creativecommons.org/licenses/by-nc-sa/4.0/ |

Massachusetts Institute of Technology

DSpace@MIT

AUTHOR ACCEPTED MANUSCRIPT

# Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input

**Cite this article as:** David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba and James Glass, Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input, International Journal of Computer Vision https://doi.org/10.1007/s11263-019-01205-0

This Author Accepted Manuscript is a PDF file of an unedited peer-reviewed manuscript that has been accepted for publication but has not been copyedited or corrected. The official version of record that is published in the journal is kept up to date and so may therefore differ from this version.

Terms of use and reuse: academic research for non-commercial purposes, see here for full terms. https://www.springer.com/aam-terms-v1

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT

**Noname manuscript No.**
(will be inserted by the editor)

# Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input

**David Harwath · Adrià Recasens · Dídac Surís · Galen Chuang**
**Antonio Torralba · James Glass**

Received: date / Accepted: date

**Abstract** In this paper, we explore neural network models that learn to associate segments of spoken audio captions with the semantically relevant portions of natural images that they refer to. We demonstrate that these audio-visual associative localizations emerge from network-internal representations learned as a by-product of training to perform an image-audio retrieval task. Our models operate directly on the image pixels and speech waveform, and do not rely on any conventional supervision in the form of labels, segmentations, or alignments between the modalities during training. We perform analysis using the Places 205 and ADE20k datasets demonstrating that our models implicitly learn semantically coupled object and word detectors.

**Keywords** Vision and language, sound, speech, multimodal learning, language acquisition, visual object discovery, unsupervised learning, self-supervised learning

The authors would like to thank Toyota Research Institute, Inc. for supporting this work.

D. Harwath
E-mail: dharwath@csail.mit.edu

A. Recasens
E-mail: recasens@csail.mit.edu

D. Surís
E-mail: didacsuris@gmail.com

G. Chuang
E-mail: galen.chuang@gmail.com

A. Torralba
E-mail: torralba@csail.mit.edu

J. Glass
E-mail: glass@mit.edu

Fig. 1: The input to our models: images paired with waveforms of speech audio.

## 1 Introduction

Babies face an impressive learning challenge: they must learn to visually perceive the world around them, and to use language to communicate. They must discover the objects in the world and the words that refer to them. They must solve this problem when both inputs come in raw form: unsegmented, unaligned, and with enormous appearance variability both in the visual domain (due to pose, occlusion, illumination, etc.) and in the acoustic domain (due to the unique voice of every person, speaking rate, emotional state, background noise, accent, pronunciation, etc.). Babies learn to understand speech and recognize objects in an extremely weakly supervised fashion, aided not by ground-truth annotations, but by observation, repetition, multimodal context, and environmental interaction [12,55]. In this paper, we do not attempt to model the cognitive development of humans, but instead ask whether a machine can jointly learn spoken language and visual perception when faced with similar constraints; that is, with inputs in the form of unaligned, unannotated raw speech audio and images (Figure 1). To that end, we present models capable of jointly discovering words in

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



raw speech audio, objects in raw images, and associating them with one another.

There has recently been a surge of interest in bridging the vision and natural language processing (NLP) communities, in large part thanks to the ability of deep neural networks to effectively model complex relationships within multimodal data. These visual-linguistic models have immense potential to address challenging problems within both communities. Language offers a far more flexible and naturalistic way of annotating visual data that goes beyond rigidly defined class labels. It also opens the door for completely new problems, such as caption generation and visual question answering (VQA). Because human language is grounded in the real world, the linguistic representations that can be learned with the benefit of visual context have the potential to be far more semantically rich than text-only models.

Current work bringing together vision and language [2, 14, 16, 29, 33, 39, 40, 46, 47, 58, 59, 62] relies on written text. In this situation, the linguistic information is presented in a pre-processed form in which words have been segmented and clustered. The text word *car* has no variability between sentences (other than synonyms, capitalization, etc.), and it is already segmented apart from other words. This is dramatically different from how children learn language. The speech signal is continuous, noisy, unsegmented, and exhibits a wide number of non-lexical variabilities. The problem of segmenting and clustering the raw speech signal into discrete words is analogous to the problem of visual object discovery in images - the goal of this paper is to address both problems jointly.

Recent work has focused on cross modal learning between vision and sounds [3, 4, 42, 43]. This work has focused on using ambient sounds and video to discover sound generating objects in the world. In our work we will also use both vision and audio modalities except that the audio corresponds to speech. In this case, the problem is more challenging as the portions of the speech signal that refer to objects are shorter, creating a more challenging temporal segmentation problem, and the number of categories is much larger. Using vision and speech was first studied in [23], but it was only used to relate full speech signals and images using a global embedding. Therefore the results focused on image and speech retrieval. Here we introduce a model able to segment both words in speech and objects in images without supervision.

The premise of this paper is as follows: given an image and a raw speech audio recording describing that image, we propose a neural model which can highlight the relevant regions of the image as they are being described in the speech. What makes our approach unique is the fact that we do not use any form of conventional speech recognition or transcription, nor do we use any conventional object detection or recognition models. In fact, both the speech and images are completely unsegmented, unaligned, and unannotated during training, aside from the assumption that we know which images and spoken captions belong together as illustrated in Figure 1. We train our models to perform semantic retrieval at the whole-image and whole-caption level, and demonstrate that detection and localization of both visual objects and spoken words emerges as a by-product of this training.

## 2 Prior Work

### 2.1 Visual Object Recognition and Discovery

Classification of visual objects (or other patterns) is a long-standing problem within the computer vision community, with the MNIST [35] handwritten digit task being a classic and widely known example. Recent progress in the field has been driven in part by recurring challenge competitions such as ISLVRC [51]. Since 2012, the task has been dominated by deep convolutional neural networks (CNNs), popularized by [34]. Since that time, improved variants of the basic CNN architecture have continued to push the state of the art [24, 54]. While classification asks the question of "what", object detection and localization (also part of the ISLVRC suite of tasks) address the problem of "where". State of the art systems are trained using bounding box annotations for the training data [19, 45], however other works investigate weakly-supervised or unsupervised object localization [5, 7, 9, 65]. A large body of research has also focused on unsupervised visual object discovery, in which case there is no labeled training dataset available. One of the first works within this realm is [60], which utilized an iterative clustering and classification algorithm to discover object categories. Further works borrowed ideas from textual topic models [52], assuming that certain sets of objects generally appear together in the same image scene. More recently, CNNs have been adapted to this task [10, 20], for example by learning to associate image patches which commonly appear adjacent to one another.

### 2.2 Unsupervised Speech Processing

Automatic speech recognition (ASR) systems have recently made great strides thanks to the revival of deep neural networks. Training a state-of-the-art ASR system requires thousands of hours of transcribed speech audio, along with expert-crafted pronunciation lexicons and text corpora covering millions, if not billions of words for language model training. The reliance on expensive, highly supervised training paradigms has restricted the application of ASR to the major languages of the world, accounting for a small fraction of the more than 7,000 human languages spoken worldwide [37]. Within the

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



speech community, there is a continuing effort to develop algorithms less reliant on transcription and other forms of supervision. Generally, these take the form of segmentation and clustering algorithms whose goal is to divide a collection of spoken utterances at the boundaries of phones or words, and then group together segments which capture the same underlying unit. Popular approaches are based on dynamic time warping [26,28,44], or Bayesian generative models of the speech signal [31,36,41]. Neural networks have thus far been mostly utilized in this realm for learning frame-level acoustic features [30,48,56,64].

### 2.3 Fusion of Vision with Language and Sound

Joint modeling of images and natural language text has gained rapidly in popularity, encompassing tasks such as image captioning [14,33,29,58,62], visual question answering (VQA) [2,16,39,40,47], multimodal dialog [59], and text-to-image generation [46]. While most work has focused on representing natural language with text, there are a growing number of papers attempting to learn directly from the speech signal. A major early effort in this vein was the work of Roy [50,49], who learned correspondences between images of objects and the outputs of a supervised phoneme recognizer. Recently, it was demonstrated by Harwath et al [23] that semantic correspondences could be learned between images and speech waveforms at the signal level, with subsequent works providing evidence that linguistic units approximating phonemes and words are implicitly learned by these models [1,8,11,21,32]. This paper follows in the same line of research, introducing the idea of "matchmap" networks which are capable of directly inferring semantic alignments between acoustic frames and image pixels.

A number of recent models have focused on integrating other acoustic signals to perform unsupervised discovery of objects and ambient sounds [3,4,42,43]. Our work concentrates on speech and word discovery. But combining both types of signals (speech and ambient sounds) opens a number of opportunities for future research beyond the scope of this paper.

## 3 Spoken Captions Dataset

For training our models, we use the Places Audio Caption dataset [23,21]. This dataset contains approximately 200,000 recordings collected via Amazon Mechanical Turk of people verbally describing the content of images from the Places 205 [67] image dataset. We augment this dataset by collecting an additional 200,000 captions, resulting in a grand total of 402,385 image/caption pairs for training and a held-out set of 1,000 additional pairs for validation.

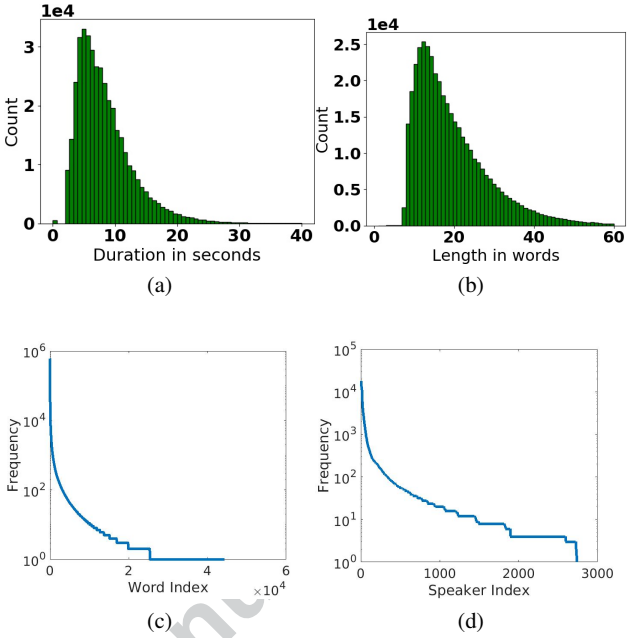


Fig. 2: Statistics of the 400k spoken captions. From left to right, the plots represent (a) the histogram over caption durations in seconds, (b) the histogram over caption lengths in words, (c) the estimated word frequencies across the captions, and (d) the number of captions per speaker. Note that the rapid dropoff in the tail of (d) is assocated with the speakers who only provided a single caption.

In order to perform a fine-grained analysis of our models ability to localize objects and words, we collected an additional set of captions for 9,895 images from the ADE20k dataset [68] whose underlying scene category was found in the Places 205 label set. The ADE20k data contains pixel-level object labels, and when combined with acoustic frame-level ASR hypotheses, we are able to determine which underlying words match which underlying objects. In all cases, we follow the original Places audio caption dataset and collect 1 caption per image. Aggregate statistics over the data are shown in Figure 2.

While we do not have exact ground truth transcriptions for the spoken captions, we use the Google ASR engine to derive hypotheses which we use for experimental analysis (but not training, except in the case of the text-based models). A vocabulary of 44,342 unique words were recognized within all 400k captions, which were spoken by 2,683 unique speakers. The distributions over both words and speakers follow a power law with a long tail (Figure 2). We also note that the free-form nature of the spoken captions generally results in longer, more descriptive captions than exist in text captioning datasets. While MSCOCO [38] contains an average of just over 10 words per caption, the places audio captions are on average 20 words long,

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



with an average duration of 10 seconds. The extended Places 205 audio caption corpus, the ADE20k caption data, and a PyTorch implementation of the model training code are available at `http://groups.csail.mit.edu/sls/downloads/placesaudio/`.

## 4 Models

Our model (Figure 3) is similar to that of Harwath et al [23], in which a pair of convolutional neural networks (CNN) [35] are used to independently encode a visual image and a spoken audio caption into a shared embedding space. What differentiates our models from prior work is the fact that instead of mapping entire images and spoken utterances to fixed points in an embedding space, we learn representations that are *distributed* both spatially and temporally, enabling our models to directly co-localize within both modalities.

In this section, we begin by describing the model architectures used for the vision and audio branches of our model (Sections 4.1 and 4.2). Next, we describe the various ways we can compute a similarity score between an image and an audio caption from the outputs of both branches (Section 4.3). Finally, we describe the loss functions and optimization methods used to train the models (Section 4.4).

### 4.1 Image Modeling

For the purpose of modeling images, we make use of two different CNN architectures: the VGG16 network [54] as well as the ResNet50 [24] network. In the majority of prior work on two-branched neural models of visually grounded speech, the image branch utilized the VGG16 network [54] [23,21,17,8,1,32]. In all of these cases, the weights of the image network were pre-trained on ImageNet, and thus had a significant amount of visual discriminative ability built-in from the start. In this work, we demonstrate how both branches could be trained end-to-end in a completely unsupervised fashion, without the need for ImageNet pre-training. Additionally in these prior works, the entire network below the classification layer was utilized to derive a single, global image embedding. One problem with this approach is that coupling the output of the final convolutional layer to a fully connected involves a flattening operation, which makes it difficult to recover associations between any neuron above the final convolution and the spatially localized stimulus which was responsible for its output. We address this issue here by retaining only the convolutional banks of the networks. For VGG16, we keep all layers up through `conv5`, discarding `pool5` and everything above it. For ResNet50, we keep all layers up through the final residual block, discarding the global average pooling and fully connected layer.

For a 224 by 224 pixel input image, the output of the network would be a 14 by 14 feature map across 512 channels (for VGG16), or a 7 by 7 feature map across 2048 channels (for ResNet50). In either case, each location within the map possesses a receptive field that can be related directly back to the input. In order to map an image into an embedding space of the same dimension as the output of the audio branch, we apply a final 1024-channel linear convolution with no nonlinearity. In the case of ResNet50, we use a 1x1 convolution, while for VGG16 we use a 3x3 convolution due to the its output feature map is of higher resolution than ResNet50.

For both network architectures, image pre-processing for training and retrieval evaluation consists of resizing the smallest dimension to 256 pixels, taking a random 224 by 224 crop (the center crop is taken for validation), and normalizing the pixels according to a global pixel mean and variance. When producing the matchmap visualizations, such as those depicted in Figures 14 and 15, we resize the smallest image dimension to 256, but do not perform any cropping.

### 4.2 Audio Modeling

To model the spoken audio captions, we use two model architectures: the DAVEnet (Deep Audio-Visual Embedding network) 5-layer model (detailed in [22]), and a residual version, ResDAVEnet, which is inspired by the ResNet [24] architecture. The 5 layer DAVEnet is similar to that of [21], but modified to output a feature map across the audio during training, rather than a single embedding vector. The audio waveforms are represented as log Mel filter bank spectrograms. Computing these involves first removing the DC component of each recording via mean subtraction, followed by pre-emphasis filtering. The short-time Fourier transform is then computed using a 25 ms Hamming window with a 10 ms shift. We take the squared magnitude spectrum of each frame and compute the log energies within each of 40 Mel filter bands. We treat these final spectrograms as 1-channel images, and model them with the CNN displayed in Figure 3. [23] utilized truncation and zero-padding of each spectrogram to a fixed length of 2048 frames, or approximately 20 seconds. We then truncate the output feature map of each caption on an individual basis to remove the frames corresponding to zero-padding - although surprisingly, we found that doing this padding compensation made very little difference in terms of the retrieval recall scores compared to a model which did not truncate the output at the beginning of the padding. Rather than manually normalizing the spectrograms, we employ a BatchNorm [25] layer at the front of the network.

The ResDAVEnet model features a cascade of four ResNet-style residual blocks, but which in our case are designed to model 1-dimensional inputs (i.e. a temporal sequence of features). Because each of the four ResDAVEnet residual

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



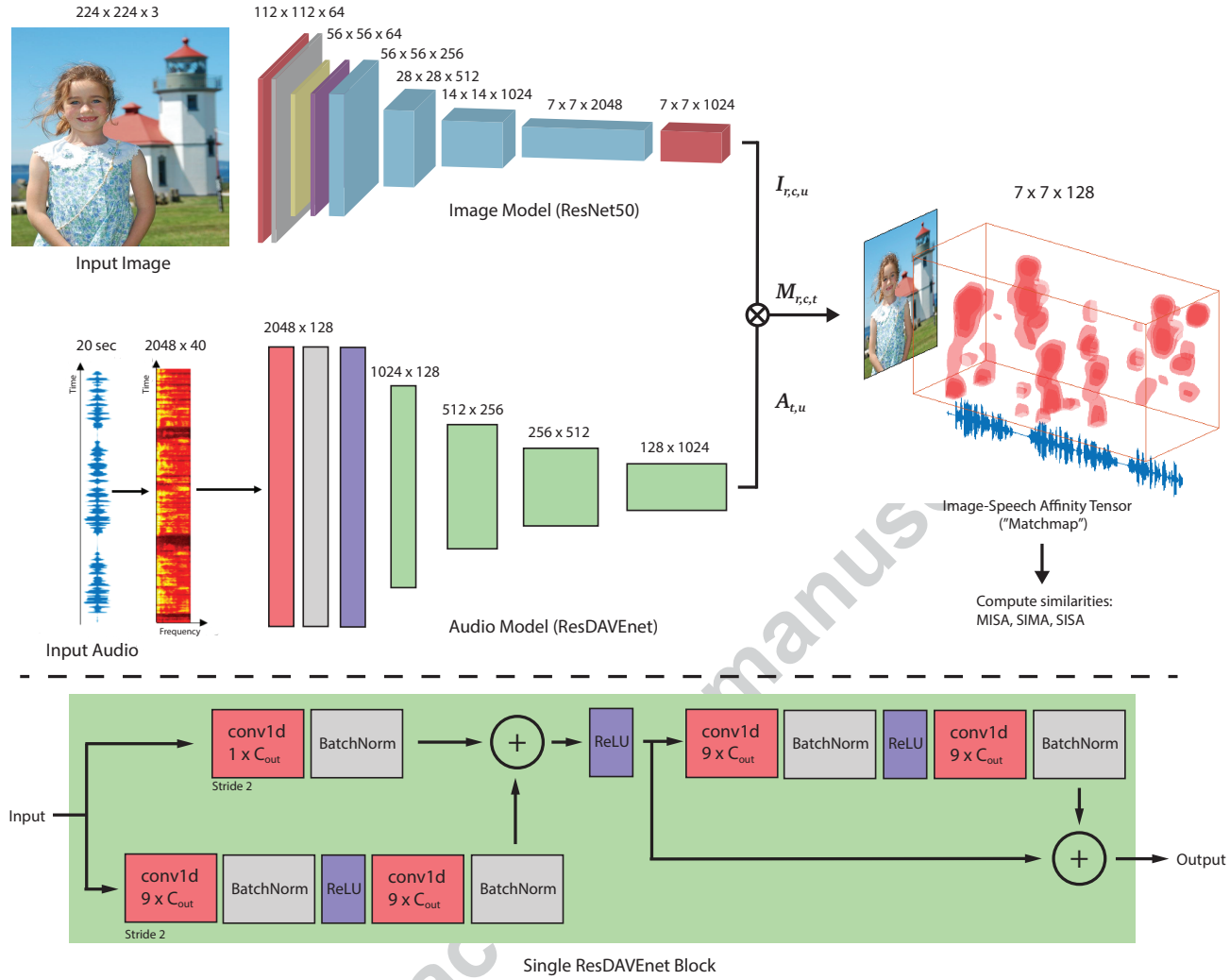


Fig. 3: The ResNet-ResDavenet variant of our model architecture (upper left), along with an example matchmap output (upper right), displaying a 3-D density of spatio-temporal similarity. The image branch is based on the ResNet network architecture, while the audio branch depicted is the ResDAVENet model. Red blocks represent convolutional layers, gray blocks indicate BatchNorm layers, yellow block MaxPooling layers, and purple blocks ReLU activations. The four blue blocks in the image branch represent the four bottleneck residual blocks in the ResNet50 model, while the four green blocks in the speech branch represent ResDAVEnet blocks. A schematic diagram of a single ResDAVEnet block is shown in the bottom half of the figure.

blocks involves an overall downsampling factor of two, the final temporal resolution of the ResDAVEnet outputs is half that of the DAVEnet-5 model.

Next, we discuss methods for relating the visual and auditory feature maps to one another.

### 4.3 Computing Image-Speech Similarity

Many cross-modal grounding models operate by independently encoding each of their inputs into an embedding vector representation [13, 14, 33]. These vectors are constrained to live within the same space, enabling arithmetic operations to be applied between the representations, despite the fact that the inputs may have originated in very different modalities (such as visual images and written text - or in our case, speech audio). Semantic similarity between cross-modal inputs is typically assumed to correlate with vector space similarities, such as cosine similarity, dot product similarity, inverse Euclidean distance, etc. Under this formulation, semantic nearest neighbors can be efficiently computed across modalities, enabling applications such as semantic image search based on natural language queries. In our case, we are only tangentially interested in semantic cross-modal retrieval; our ultimate goal is to co-segment visual and audio

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



inputs into object-like and word-like patterns. In this section, we describe how we can adapt retrieval-inspired cross modal fusion techniques for this purpose. We observe that there is a an interesting similarity between inferring latent semantic alignments in our case and other vision-and-language tasks such as captioning and VQA (which is often accomplished through an attention mechanism [53,61])

Zhou et al [66] demonstrate that global average pooling applied to the `conv5` layer of several popular CNN architectures not only provides good accuracy for image classification tasks, but also enables the recovery of spatial activation maps for a given target class at the `conv5` layer, which can then be used for object localization. The idea that a pooled representation over an entire input used for training can then be unpooled for localized analysis is powerful because it does not require localized annotation of the training data, or even any explicit mechanism for localization in the objective function or network itself, beyond what already exists in the form of convolutional receptive fields. Although our models perform a ranking task and not classification, we can apply similar ideas to both the image and speech feature maps in order to compute their pairwise similarity, in the hopes to recover localizations of objects and words.

Let $I$ represent the output feature map output of the image network branch, $A$ be the output feature map of the audio network branch, and $\bar{I}$ and $\bar{A}$ be their globally average-pooled counterparts:

$$\bar{I} = \frac{1}{N_r N_c} \sum_{r=1}^{N_r} \sum_{c=1}^{N_c} I_{r,c,:} \tag{1}$$

$$\bar{A} = \frac{1}{N_t} \sum_{t=1}^{N_t} A_{t,:} \tag{2}$$

Here we use the colon (:) to indicate selection of all elements across an indexing plane; in other words, $I_{r,c,:}$ is a 1024-dimensional vector representing the $(r, c)$ coordinate of the image feature map, and $A_{t,:}$ is a 1024-dimensional vector representing the $t^{th}$ frame of the audio feature map. One straightforward choice of similarity function between and image and audio caption is the dot product between the global average pooled embeddings,

$$S(I, A) = \bar{I}^T \bar{A} \tag{3}$$

Substituting Equations 1 and 2 into Equation 3, we have that

$$S(I, A) = (\frac{1}{N_r N_c} \sum_{r=1}^{N_r} \sum_{c=1}^{N_c} I_{r,c,:})^T (\frac{1}{N_t} \sum_{t=1}^{N_t} A_{t,:}) \tag{4}$$

By distributing between the summations and collecting the coefficients, we can write the similarity as

$$S(I, A) = \frac{1}{N_r N_c N_t} \sum_{r=1}^{N_r} \sum_{c=1}^{N_c} \sum_{t=1}^{N_t} I_{r,c,:}^T A_{t,:} \tag{5}$$

We can see from Equation 5 that the combination of global average pooling and the dot product results in the similarity score taking on large values when *all* local regions of the image feature map exhibit a large dot product with *all* local regions of the audio feature map. We also notice that implicit in this computation is a 3rd order tensor $M$, where $M_{r,c,t} = I_{r,c,:}^T A_{t,:}$. Because $M$ reflects the localized similarity between a small image region (possibly containing an object, or part of an object) and a segment of speech audio (possibly containing a word or short phrase), we dub $M$ the "matchmap" tensor between and image and an audio caption. Explicitly computing $M$ ideally enables us to learn a latent semantic alignment between matching objects and words. Under this view, the similarity between the global average pooled image and audio representations can be found by averaging the similarity between *all* audio frames and *all* image regions. We call this similarity scoring function SISA (sum image, sum audio):

$$\text{SISA}(M) = \frac{1}{N_r N_c N_t} \sum_{r=1}^{N_r} \sum_{c=1}^{N_c} \sum_{t=1}^{N_t} M_{r,c,t} \tag{6}$$

For the sake of computational efficiency, at training time we compute the SISA scoring function by using global average pooling and a dot product. In our experiments exploring object and word discovery (detailed in Section 5.1), we explicitly utilize the matchmap $M$. If we are willing to incur the extra computational cost of computing $M$ at train time, there are a multitude of ways in which we can reduce a matchmap to a single scalar-valued score, two of which we describe here.

Because it is not completely realistic to expect all words within a caption to simultaneously match all objects within an image, we consider computing the similarity between an image and an audio caption using several alternative functions of the matchmap density. By replacing the averaging summation over image patches with a simple maximum, MISA (max image, sum audio) effectively matches each frame of the caption with the most similar image patch, and then averages over the caption frames:

$$\text{MISA}(M) = \frac{1}{N_t} \sum_{t=1}^{N_t} \max_{r,c}(M_{r,c,t}) \tag{7}$$

By preserving the sum over image regions but taking the maximum across the audio caption, SIMA (sum image, max audio) matches each image region with only the audio frame with the highest similarity to that region:

$$\text{SIMA}(M) = \frac{1}{N_r N_c} \sum_{r=1}^{N_r} \sum_{c=1}^{N_c} \max_{t}(M_{r,c,t}) \tag{8}$$

Next, we describe the how these similarities are integrated into the loss functions used to train our models.

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



### 4.4 Training

Our models are trained to optimize a ranking-based criterion [6], such that images and captions that belong together are more similar in the embedding space than mismatched image/caption pairs. Specifically, across a batch of $B$ image/caption pairs $(I_j, A_j)$ (where $I_j$ represents the output of the image branch of the network for the $j^{th}$ image, and $A_j$ the output of the audio branch for the $j^{th}$ caption) we first randomly select impostor samples according to

$$\hat{A}_j \sim \text{UniformCategorical}(\{A_1, \ldots, A_B\} \setminus A_j) \tag{9}$$

$$\hat{I}_j \sim \text{UniformCategorical}(\{I_1, \ldots, I_B\} \setminus I_j) \tag{10}$$

We then compute the sampling-based triplet loss as:

$$\mathcal{L}_s = \sum_{j=1}^{B} \Big( \max(0, S(I_j, \hat{A}_j) - S(I_j, A_j) + \eta) + \max(0, S(\hat{I}_j, A_j) - S(I_j, A_j) + \eta) \Big), \tag{11}$$

where $S(I, A)$ represents the similarity score between an image $I$ and audio caption $A$ and $\eta$ is a margin hyperparameter.

Hard negative mining has been shown to offer substantial improvements over the standard triplet loss formulation in the context of cross-modal retrieval [13]. Rather than randomly sampling impostors (or summing over all possible impostors within a batch), only the impostor sample with the largest similarity with respect to the anchor is considered when computing the loss. *Semi*-hard negative mining [27] is a variant of hard negative mining in which the impostors are constrained to be *less* similar to the anchor than its paired sample (Figure 4). Semi-hard negative mining can help to mitigate the detrimental effect of label noise on regular hard negative mining. We chose to use semi-hard negative mining because in our experience, we found standard negative mining to be highly unstable during training.

Mathematically, we first select the candidate image negatives $\bar{\mathbf{I}}_\mathbf{j}$ and candidate audio negatives $\bar{\mathbf{A}}_\mathbf{j}$ to be the set of all images (or audio captions) less similar to the anchor image (or caption) than the anchor's paired caption (or image):

$$\bar{\mathbf{A}}_\mathbf{j} = \{A \in \{A_1, \ldots, A_N\} | S(I_j, A) < S(I_j, A_j)\}, \tag{12}$$

$$\bar{\mathbf{I}}_\mathbf{j} = \{I \in \{I_1, \ldots, I_N\} | S(I, A_j) < S(I_j, A_j)\}. \tag{13}$$

Then, we construct the semi-hard negative triplet loss by maximizing over all candidate negatives:

$$\mathcal{L}_h = \sum_{j=1}^{B} \Big( \max(0, \max_{A \in \bar{\mathbf{A}}_\mathbf{j}}(S(I_j, A)) - S(I_j, A_j) + \eta) + \max(0, \max_{I \in \bar{\mathbf{I}}_\mathbf{j}}(S(I, A_j)) - S(I_j, A_j) + \eta) \Big), \tag{14}$$

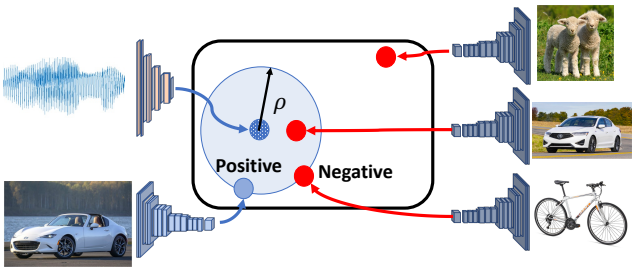


Fig. 4: We utilize a training scheme inspired by [27], where the negative sample is selected as the hardest sample in the batch which is, at most, as similar to the positive sample than the ground truth. This strategy avoids training instabilities due to noise in the training data.

In the case that there are no potential semi-hard negatives that satisfy Equations 12 or 13, we default to randomly sampling the negatives. Empirically, we found that semi-hard negative training on its own was unstable to train, and worked much better when combined with the sampling-based triplet loss $\mathcal{L}_s$. For our models which utilize semi-hard negative training, the loss function becomes:

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_h \tag{15}$$

Although in theory the two losses could be assigned different weights, in our experiments we weight them equally with good results.

For both the randomly sampled and semi-hard negative mined loss functions, the imposter images and captions for each image/caption pair are selected from the same mini-batch. We also fix $\eta$ to 1 in all of our experiments. The choice of similarity function $S(I, A)$ is flexible, which we explore in Section 4.3. This criterion directly enables semantic retrieval of images from captions and vice versa, but in this paper much of our focus is to explore how object and word *co-localization* naturally emerges as a by-product of this training scheme.

An important issue to consider with hard negative mining in the context of our models is computational complexity. Several of our matchmap similarity functions (MISA and SIMA) require explicit computation of the full matchmap between an image-caption pair, which requires $O(T * H * W * D)$ multiply-adds, where $T$ is the caption duration, $H$ and $W$ are the image height and width, and $D$ is the embedding dimension. Semi-hard negative mining using full matchmap-based similarity scores increases this complexity by a factor of $B^2$; in practice, we found that even with parallel training across multiple GPUs, this was computationally impractical. The exception to this is the SISA loss computed via global average pooling, for which the within-batch similarity matrix can be computed in $O(D * B^2)$ time. For this

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



reason, all of our models which rely on semi-hard negative mining utilize the SISA matchmap similarity function.

### 4.5 Pre-Training Methods

A core issue which we investigate is the manner in which various forms of pre-training influence our model's ability to learn. Many previously published works on visually grounded speech utilized an audio network which was trained from a random initialization, but used a vision model which underwent supervised pre-training e.g. on ImageNet [23,21,17,8,1,32]. This leads to the question of whether the model able to learn new concepts by grounding speech to images, or if the audio network is simply learning to predict the image features that were originally derived from a supervised classification task. To that end, we consider three methods for initializing our models:

1. **Fully random initialization.** Under this condition, the weights of both the image and audio branches of the model are randomly initialized at the start of training.
2. **Unsupervised pre-training on Flickr Natural Sounds.** Under this condition, the models are pre-trainied without labels using a database of videos containing natural sounds [57]. Similar to [4], we use videos from Flickr selected by querying popular words and tags. We take the audio track and sample image frames from these videos and then use the semi-hard negative triplet loss (Section 4.4) to train our model to recognize pairs of audio-image from the same video (positive exampled) and audio-image pairs from different videos (negative examples). We use 2,146,055 Flickr videos for pre-training, and achieve an average Recall@10 score of 0.441 on this task, using 500 validation samples.
3. **Fully supervised pre-training on ImageNet and AudioSet.** In this case, both the image and audio branches of the network are pre-trained in a supervised fashion. We use ImageNet classification to pre-train the image branch, and AudioSet sound classification [18] to pre-train the audio branch. For the AudioSet classification, we subsample a class-balanced subset of the total training set. We take the global maxed pooled outputs of the audio branch and add one final fully connected layer with a softmax activation on top of it. We use a Cross-Entropy loss for training randomly sampling the training class at every iteration; average per class AUC was found to be 0.891 on the validation set.

### 4.6 Training Details

All models were trained using stochastic gradient descent with a batch size of 80, a fixed momentum of 0.9. We use learning of 0.001 for the randomly initialized ResNet50 + ResDAVEnet models and all VGG16 + DAVENet models, 0.01 for the ResNet50 + ResDAVEnet models with AudioSet + ImageNet initialization and 0.03 for the ResNet50 + ResDAVEnet models initialized with Natural Sounds. Anecdotally, we found that the higher learning rates could lead to instability for randomly initialized models, but not for the models which had already undergone pre-training. We decayed the learning rate by a factor of 10 every 30 epochs and initially trained for a minimum of 90 epochs; however, we found that some of the models (especially the randomly initialized models) began to overfit the training data in later epochs. For this reason, all of the results presented in this paper were computed with models that were subject to early stopping at 40 epochs. In the models trained using a blend of the sampled and semi-hard negative triplet losses, we simply weighted the loss terms equally.

## 5 Experiments

### 5.1 Image and Caption Retrieval

We first present experiments detailing the performance of our models for an image/caption retrieval task. We use a held-out set of 1,000 image/caption pairs from the Places audio caption dataset to validate the models on the image/caption retrieval task, similar to the one described in [23,21,8,1]. This task serves to provide a single, high-level metric which captures how well the model has learned to semantically bridge the audio and visual modalities. While providing a good indication of a model's overall ability, it does not directly examine which specific aspects of language and visual perception are being captured, which we later investigate in Sections 5.3, 5.4, 5.5, and 5.6.

The core retrieval results for our unsupervised models are summarized in Table 1. A comparison against previously published baselines, as well as our supervised pre-training results, are shown in Table 2. Finally, we show retreival results for text-based models which operate on the text transcripts of the spoken captions (estimated using the Google public speech recognition API) rather than the speech audio in Table 3.

In Table 1, we anchor our analysis to our best-performing unsupervised model (last row, bold) and ablate the model in a variety of ways. The main takeaways from these results are detailed below:

1. Pre-training on natural sounds dramatically helps retrieval performance. In the second-to-last row of Table 1, we compare a randomly initialized version of the ResNet50 + ResDAVEnet model trained using the SISA objective with semi-hard negative mining to a version of the same model pre-trained on the Flickr natural sound videos.

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



Table 1: **Ablation study for unsupervised models**: Recall scores on the held out set of 1,000 images/captions for our various ablations of the speech-image grounding model. SHN stands for semi-hard negative training, while the RN prefix indicates the use of a ResNet50 image branch and a ResDAVEnet audio branch. VGG refers to the model using the VGG16 architecture in the image branch, and the audio branch on the DAVEnet-5 architecture.

| | | | Speech to Image | | | Image to Speech | | |
|---|---|---|---|---|---|---|---|---|
| Model | Loss | Pretrained | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| VGG | SISA-SHN | *Natural Sounds* | .145 | .382 | .503 | .115 | .352 | .471 |
| RN | SIMA | *Natural Sounds* | .118 | .331 | .463 | .126 | .347 | .461 |
| RN | SISA | *Natural Sounds* | .132 | .376 | .490 | .112 | .318 | .445 |
| RN | MISA | *Natural Sounds* | .143 | .364 | .514 | .096 | .311 | .458 |
| RN | SISA-SHN | No | .147 | .375 | .512 | .099 | .328 | .452 |
| **RN** | **SISA-SHN** | **Natural Sounds** | .268 | .545 | .684 | .211 | .528 | .660 |

Table 2: **Supervised baseline comparison**: Recall scores on the held out set of 1,000 images/captions comparing the unsupervised pre-training approaches (top two rows) against supervised models (bottom three rows). The top three rows in this table use the ResNet50/ResDAVEnet architecture.

| | Speech to Image | | | Image to Speech | | |
|---|---|---|---|---|---|---|
| Method | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| Random | .147 | .375 | .512 | .099 | .328 | .452 |
| Natural Sounds | .268 | .545 | .684 | .211 | .528 | .660 |
| ImageNet/AudioSet | .276 | .584 | .716 | .218 | .551 | .690 |
| [23] | .148 | .403 | .548 | .121 | .335 | .463 |
| [21] | .161 | .404 | .564 | .130 | .378 | .542 |

Table 3: **Text-based models:** Recall scores on the held out set of 1,000 images/captions for our various text-image grounding models. SHN stands for semi-hard negative training. All models use a ResNet50 image branch and a 2-layer text branch with 1-D convolutional layers that operate on input sequences of word embeddings.

| | | Text to Image | | | Image to Text | | |
|---|---|---|---|---|---|---|---|
| Loss | Pretrained | R@1 | R@5 | R@10 | R@1 | R@5 | R@10 |
| SIMA | No | .018 | .135 | .294 | .071 | .217 | .325 |
| SISA | No | .105 | .309 | .419 | .064 | .220 | .332 |
| MISA | No | .100 | .283 | .395 | .048 | .185 | .308 |
| SISA-SHN | No | .206 | .481 | .632 | .138 | .398 | .558 |
| SISA-SHN | ImageNet | .322 | .659 | .782 | .235 | .551 | .719 |

We see that the average Recall@10 score increases from .482 to .672 when pre-training with natural sounds, representing a 39.4% relative improvement over the exact same model with a random initialization.

2. Semi-hard negative mining is also immensely beneficial for the model. Even when retaining the residual architecture and natural sound pre-training, a model trained without semi-hard negative mining (third row) achieves an average Recall@10 of .468.
3. The residual architecture (ResNet50 + ResDAVEnet) significantly outperforms VGG16 + DAVEnet. We trained a VGG16 + DAVEnet model with the SISA semi-hard negative loss and natural sound pre-training (first row), which resulted in an average Recall@10 of .487.
4. For models trained without semi-hard negative mining, MISA outperforms SISA which outperforms SIMA - but the differences between these models are small compared to the impact of natural sound pre-training, semi-hard negative mining, and the residual model architecture.

In Table 2, we examine the ResNet50 + ResDAVEnet model trained with the SISA-SHN loss under random initialization, natural sound pre-training, and supervised classification pre-training. We see a clear ranking between the methods, with natural sound pre-training outperforming random initialization but supervised pre-training coming out on top. What is interesting to note, however, is the fact that the gap between the average Recall@10 score for the natural sound pre-trained model and the supervised pre-trained model is much smaller than between the random model and the natural sound model. While natural sound pre-training offers a nearly 40% relative improvement, supervised pre-training offers only an additional 4.6% relative improvement. This suggests that the performance gap between our pre-trained and non-pre-trained models is not solely due to supervised labeling information leaking into the network weights, but instead is more likely a function of the total amount of training data seen by the model. This is an extremely encouraging result not only because it implies that we have not yet exhausted the learning capacity of our models, but also because it indicates that synergies between different domains within the same modality (natural sounds vs. speech audio) can be exploited to our benefit.

We also compare our models against reimplementations of two previously published speech-to-image models (both of which utilized ImageNet pre-trained image branches) in Table 2. Both previously published baselines we compare to used the full VGG16 network, deriving an embedding for the entire image from the `fc2` outputs. By contrast, all of our models output spatial and temporal feature maps. The

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



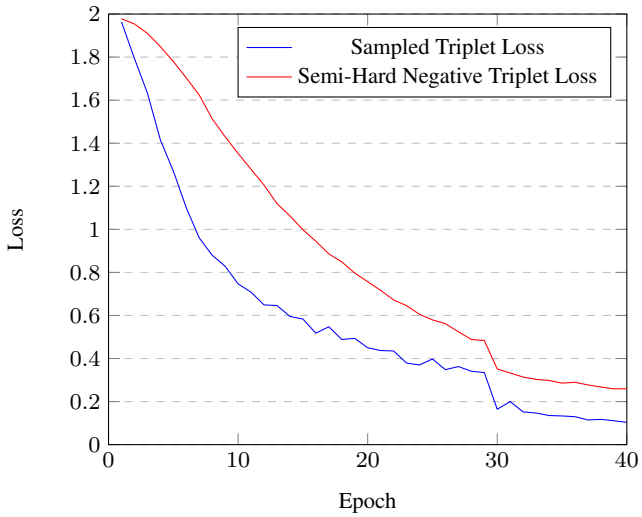


Fig. 5: Value of sampled and semi-hard triplet losses as a function of training epoch. The sampled loss term decays more quickly than the semi-hard negative loss term, but does not disappear completely.

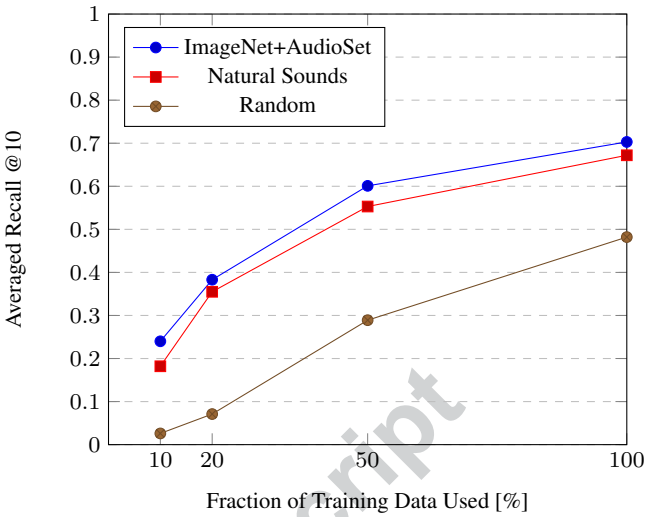


Fig. 6: Performance as a function of training data amount for 3 different pre-training scenarios. The same ResNet50-ResDaveNet model architecture is used throughout. We evaluate three different methods of model initialization: random, an image branch pretrained on ImageNet and the audio branch in AudioSet, and both the image and audio branches pretrained on videos with natural sounds.

fact that all of our models either outperform or perform comparably to these baselines suggests that there is not much to be lost when doing away with the fully connected layers that hamper localization.

In Table 3, we compare against baselines that operate on automatic speech recognition (ASR) derived text transcriptions of the spoken captions. The text-based model we used is based on the two-branch topology of the speech and image model, but replaces the speech audio branch with a CNN that operates on word sequences. The ASR text network uses a 200-dimensional word embedding layer, followed by a 512 channel, 1-dimensional convolution across windows of 3 words with a ReLU nonlinearity. A final convolution with a window size of 3 and no nonlinearity maps these activations into the 1024 multimodal embedding space. Because the use of text as an input effectively solves half the problem faced by our models (recognizing words in raw speech signals), the retrieval scores are unsurprisingly higher relative to the speech-based models, representing an approximate upper bound on the performance we can expect from the speech audio-based models.

In Figure 5, we plot the values of the randomly sampled and semi-hard negative triplet losses as a function of training epoch. It is reasonable to hypothesize that at some point during training, the model would become powerful enough that the sampled loss would vanish (or plateau at a very small value) and the gradient would become dominated by the semi-hard negative loss; however, we did not observe this during the first 40 epochs of training (where we perform early stopping).

### 5.2 Varying the Amount of Training Data

Here, we examine varying the amount of training data influences the performance of our model under the various pre-training regimes. In Figure 6, we display the learning curves of 3 different models in terms of the average of the caption to image and image to caption Recall@10 on the Places audio validation set. The models were trained on subsets comprised of 10%, 20%, and 50% of the full 400k training set. We note that the trends observed in Table 2 are reflected in Figure 6 for all training set sizes. Namely, both supervised and unsupervised pre-training consistently improves the performance of the model regardless of how much training data is available. Without any pre-training, the model struggles to reach 0.1 R@10 with 20% of the training data (corresponding to 80,000 examples), even with semi-hard negative training. The fact that none of the curves have levelled off suggests that even larger training datasets would be helpful for achieving further performance improvements.

### 5.3 Speech-Prompted Object Detection and Localization

To evaluate our models' ability to detect and segment visual objects given a spoken prompt, we use the spoken captions for the ADE20k [68] dataset. The ADE20k images contain pixel-level object masks and labels - in conjunction with a time-aligned transcription produced via ASR (we use the

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT




Fig. 7: Comparison of speech-prompted object localization heatmaps for 8 different word/object pairs and the three pre-training conditions (Random, Natural Sounds, ImageNet+AudioSet), using the ResNet50 + ResDAVEnet model and the SISA-SHN loss function.

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



public Google Speech Recognition API for this purpose), we can associate each matchmap cell with a specific visual object label as well as a word label. These labels enable us to analyze which words are being associated with which objects. We do this by performing speech-prompted object detection and localization, which we evaluate separately.

Because there are a very large number of different words appearing in the speech, and no one-to-one mapping between words and ADE20k objects exists, we manually define a set of 100 word-object pairings. We choose commonly occurring (at least 9 occurrences) pairs that are unambiguous, such as the word "building" and object "building," the word "man" and the "person" object, etc. For each word type, we isolate all occurrences of that word in the ADE20k spoken captions and compute an embedding vector for each one by feeding the isolated words into the audio branch of our model and averaging the output across the time dimension. We then compute a single embedding representing the word category by averaging the individual embeddings for all instances of the word.

To perform word-prompted object detection for a given word-object pair, we compute a score for every ADE20k image by taking the dot product of the aggregate word embedding with each spatial position of the image branch's output feature map. We then apply a global max pooling operation to this score map to derive a single score for each image. Using these scores, we compute the average precision for each word-object pairing, and take the mean average precision (mAP) across the 100 word-object pairs.

To evaluate object localization separately from object detection, we select only the subset of the ADE20k images which contain the target object for a given word-object pairing. Next, we compute a heatmap over each image by taking the dot product of the word embedding with each spatial output of the image branch. We normalize this heatmap to sit within the interval [0, 1], upsample it to the same size as the ADE20k pixel-level segmentation, apply a threshold (0.5 in all of our experiments), and then compute intersection over union (IoU), intersection over detection (IoD), and intersection over target (IoT) with respect to the target object label.

The results for both object detection and localization are summarized in Table 4. We evaluate all of the unsupervised models from Table 1, as well as the highest performing overall model which underwent supervised pre-training on ImageNet and AudioSet from Table 2. We also compare to a full-frame baseline, which assumes that the target object is always present in every image, and hypothesizes the entire image frame for the segmentation. We found that generally speaking, all of our models perform much better at detecting the presence of objects than segmenting them, as indicated by the fact that the mAP scores are several times higher than the full-frame baseline, but the mIoU scores are only 45% higher than the full-frame baseline in the best case. We note that the relative performance differences between the models in terms of object detection mAP closely mirror the retrieval results shown in Tables 1 and 2.

Table 4: Speech-prompted object detection and localization scores on ADE20K for the 100 handcrafted word-object pairs and various models. For the model type, VGG indicates a model based on the VGG16 + DAVEnet architecture, while RN indicates a model based on the ResNet50 + ResDAVEnet architecture. For pre-training, NS indicates unsupervised pre-training on natural sounds, while IN+AS indicates supervised pre-training on ImageNet and AudioSet. To evaluate object detection, we report mean average precision (mAP) for predicting whether or not a particular object exists anywhere in an image. We evaluate segmentation performance using mean intersection over union (mIoU), mean intersection over detection (mIoD), and mean intersection over target (mIoT). Segmentation scores are computed for each word-object pair only on the subset of the ADE20k images that contain the target object. In all cases, the threshold was set at 0.5, which we found produced near-optimal results for IoU for all models.

| Model | Loss | Pre-trained | mAP | mIoU | mIoD | mIoT |
|---|---|---|---|---|---|---|
| Full Frame | N/A | N/A | .129 | .11 | .11 | 1.0 |
| VGG | SISA-SHN | NS | .329 | .12 | .16 | .69 |
| RN | MISA | NS | .224 | .11 | .12 | .68 |
| RN | SIMA | NS | .158 | .11 | .12 | .88 |
| RN | SISA | NS | .283 | .12 | .15 | .61 |
| RN | SISA-SHN | None | .297 | .13 | .15 | .64 |
| RN | SISA-SHN | NS | .368 | .15 | .21 | .62 |
| **RN** | **SISA-SHN** | **IN+AS** | .440 | .16 | .23 | .63 |

While the same rankings between the models hold in terms of object segmentation, e.g. with supervised pre-training outperforming natural sound pre-training which outperforms random initialization, the differences in the mIoU scores here are much smaller. We provide a visual comparison of the segmentation performance between these models in Figure 7. Generally speaking, the three models appear to focus on the same regions of each image, although all of them suffer from similar problems. In the case of smaller objects, like chandeliers and laptops, all of the models tend to undersegment, capturing a significant amount of background pixels around the target object. In the case of larger objects, like fields, mountains, and bridges, the models tend to oversegment, focusing on a few small regions of the target object. Although the pre-trained models subjectively appear to do a better job of capturing a fuller extent of these large objects, it is interesting to note that the highest scoring regions of each image tend to be consistent across the models. In Figure 8, we present many more segmentation examples for our best unsupervised model.

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



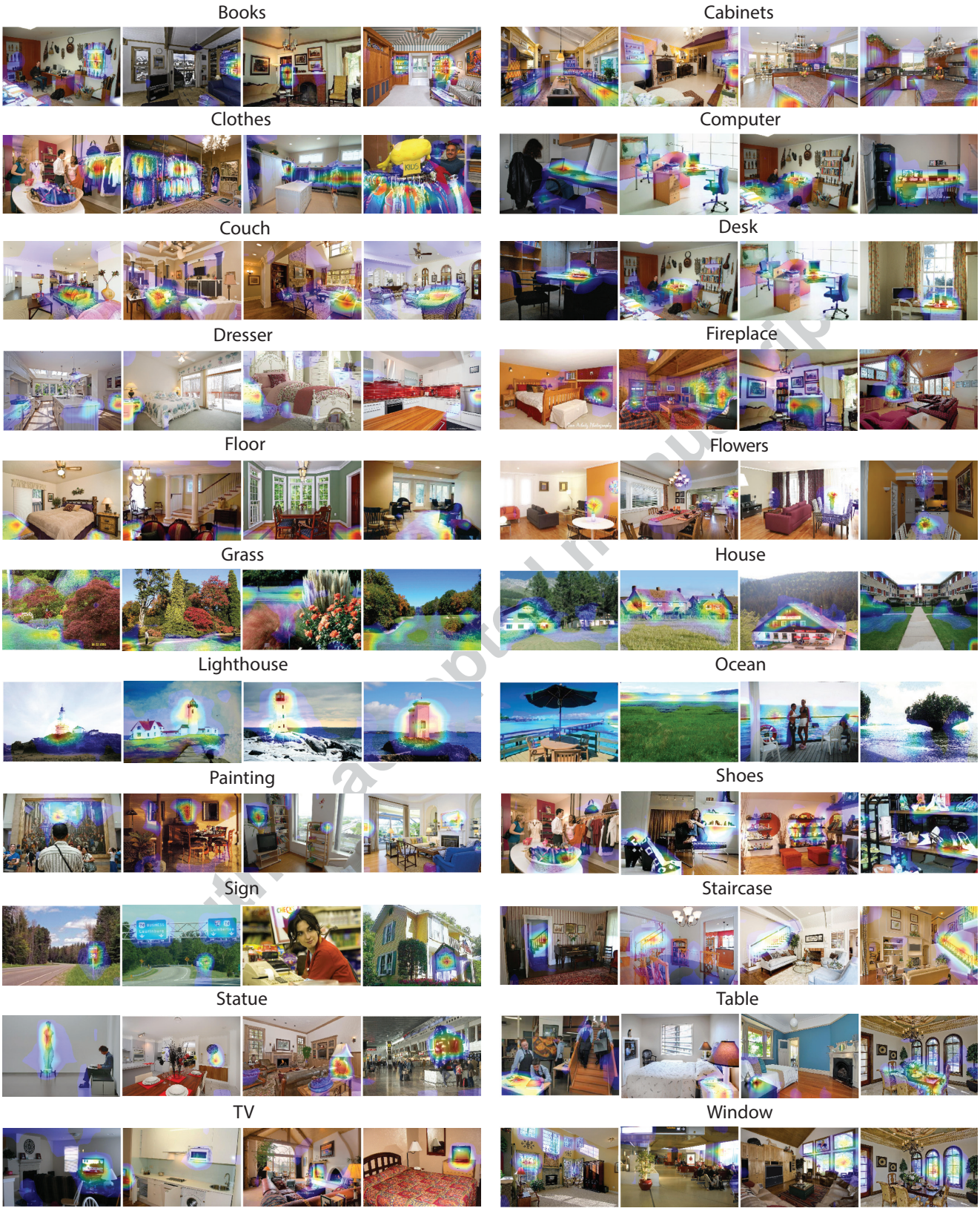


Fig. 8: Example speech-prompted object localization heatmaps for several word/object pairs using the natural sounds pre-trained ResNet50 + ResDAVEnet model, using the SISA-SHN loss function.

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



### 5.4 Clustering of Audio-Visual Patterns

The next experiment we consider is automatic discovery of audio-visual clusters from the ADE20k matchmaps using our best unsupervised model (ResNet50 + ResDAVEnet, SISA-SHN, natural sounds pre-training). Once a matchmap has been computed for an image and caption pair, we binarize it according to an absolute score threshold. While we use a threshold of 400 here, we achieved good results in the range of 200 to 450. Next, we extract volumetric connected components and their associated masks over the image and audio. We average pool the image and audio feature maps within these masks, producing a pair of vectors for each component. Because we found the image and speech representations to exhibit different dynamic ranges, we first rescale them by the average L2 norms across all derived image vectors and speech vectors, respectively. We concatenate the image and speech vectors for each component, and finally perform hierarchical clustering using the Birch algorithm [63] which resulted in 423 final clusters. To derive labels for each cluster, we take the most frequent word label as overlapped by the components belonging to a cluster. To generate the object labels, we compute the number of pixels belonging to each ADE20k class assigned to a particular cluster, and take the most common label. We display the labels and their purities for the top 100 most pure clusters in Figure 9.

### 5.5 Concept discovery: building an image-word dictionary

The clustering results displayed in Figure 9 indicate that the audio and image networks are able to agree to a common representation of knowledge, clustering similar concepts together. An interesting property of our models is the fact that because the dot product between embeddings is used to compute similarity scores, both the image and speech networks must learn to agree on the meaning of the different dimensions of the embedding space. To further explore this phenomenon, we decided to visualize the concepts associated with each of these dimensions for both image and audio networks separately and then find a quantitative strategy to evaluate the agreement.

To visualize the visual concepts associated with each of the dimensions in the image output, we use the unit visualization technique introduced in [65]. A set of images is run through the image network and the ones that most activate a particular dimension are selected. We then visualize the spatial activations in these top images. The same procedure can be done for the audio network, where we search for the set of audio captions that maximally activate the same neuron. Finally, we extract the segment of the audio caption that maximally activated the neuron in question. For both modalities, we perform segmentation by first normalizing the activations for each dimension to have zero mean and unit variance across the entire dataset. Then, we threshold the activations within each image at 1.2 and activations within the caption at 1.3.

We then treat the set of neurons in the embedding layer as a "picture dictionary," in which each dimension has the potential to capture a single concept. A dimension in this embedding space which has properly learned a concept should satisfy three requirements. First, it should strongly and reliably activate on image regions containing a specific object type. Second, it should strongly and reliably activate on spoken caption regions containing a specific word or phrase. Third, there should exist a semantic agreement between the word and object which activate this dimension. We cannot expect every dimension in the embedding space to perfectly capture a concept, but we would like to be able to find those that do. To that end, we devise an automatic selection method for finding the neurons which have captured a concept.

To quantify the quality each dimension in the picture dictionary, we rely on the object segmentation labels as well as the ASR-derived text transcripts for the spoken captions from the ADE20k dataset [68]. Using these, we can rank the most strongly detected objects for each neuron. We pass through the image branch of the network approximately 10,000 images from the ADE20k dataset and check for each neuron which classes are most activated for that particular dimension. As a result, we have a set of object labels associated with the image neuron (coming from the segmentation classes). We do the same with the time-aligned text transcripts of the spoken captions to derive a set of words associated with each neuron in the audio branch's output layer. To estimate the semantic agreement between words from the caption transcript and ADE20k object labels, we use the shortest path distance along the WordNet [15] hyponym-hypernym tree. We then define the following concept score metric:

$$c_j = \sum_{i=1}^{|O^{\text{im}}|} w_i Sim_{\text{wup}}(o_i^{\text{im}}, o_j^{\text{au}}), \tag{16}$$

with $o_i^{\text{im}} \in O^{\text{im}}$, where $O^{\text{im}}$ is the set of classes present in the TOP5 segmented images, $Sim_{\text{wup}}(.,.)$ is the Wu and Palmer WordNet-based similarity, with range [0,1] (higher is more similar), and $o_j^{\text{au}}$ is a word from the top audio activations. We weight the similarity with $w_i$, which is proportional to intersection over union of the pixels for that class into the masked region of the image. Using this metric, we can then assign one value per pair of word and image activation. To assign one single value to the whole dimension, we take the maximum among all the concept values $c_j$ for the different audio words. In our experiments, we take at most 2 words from the audio, only considering words that

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



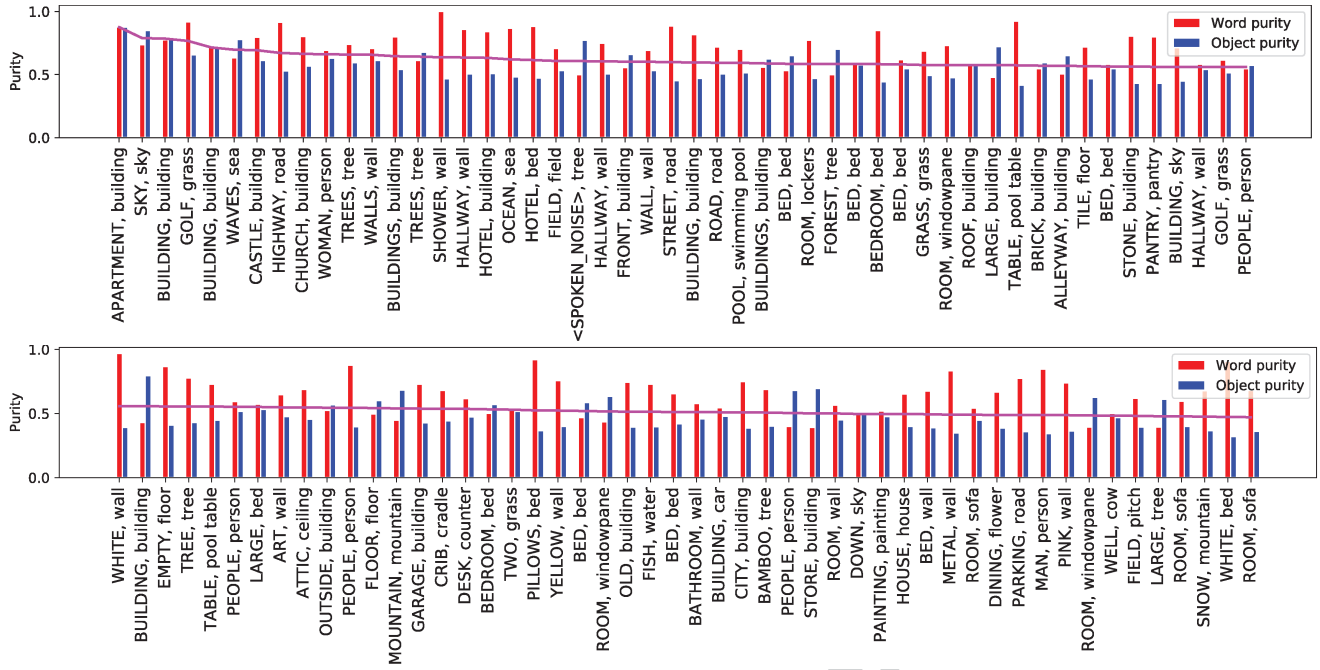


Fig. 9: Some clusters (speech and visual) found by our approach. Each cluster is jointly labeled with the most common word (capital letters) and object (lowercase letters). For each cluster we show the precision for both the word (blue) and object (red) labels, as well as their harmonic mean (magenta). The average cluster size across the top 100 clusters was 81.

at least repeat in the 5 audio pieces we consider. The final concept value $c = \max_j c_j$ measures how well both the audio network and the image network agree on that particular concept. Interestingly, the concepts are represented by two words (if two words are more than one time in the most activated region) or by one single words. Examples for many concepts are shown in Figure 10. Anecdotally, we found $c > 0.7$ to be a good indicator that a concept has been learned, and it is the threshold we use to count the number of concepts learned by the models, shown in Figure 12 as a function of the training epoch. We display some of the concepts learned at various stages during training in Figure 11.

The pairs image-word allow us to explore multiple questions. First, can we build an image-word dictionary by only listening to descriptions of images? As we show in Figure 10, we do. It is important to remember that these pairs are learned in a completely unsupervised fashion, without any concept previously learned by the network. Furthermore, in the scenario of a language without written representation, we could just have an image-audio dictionary using exactly the same technique.

Another important question is whether a better audio-visual dictionary is indicative of a better model architecture. We would expect that a better model should learn more total concepts. In this section we propose a metric to quantify this dictionary quality. This metric will help us to compute the quality of each individual neuron and of each particular model.

Table 5: The number of concepts learned by the different networks with different losses. We find it is consistently highest when using semi-hard negative mining and various forms of pre-training.

| Model | Loss | Pre-trained | Concepts |
|---|---|---|---|
| VGG | SISA-SHN | Natural Sounds | 91 |
| RN | MISA | Natural Sounds | 99 |
| RN | SIMA | Natural Sounds | 96 |
| RN | SISA | Natural Sounds | 74 |
| RN | SISA-SHN | No | 58 |
| **RN** | **SISA-SH** | **Natural Sounds** | 109 |
| **RN** | **SISA-SH** | **ImageNet/Audioset** | 126 |

Finally, we analyze the relation between the concepts learned and the architecture used in Table 5. Interestingly, the four maintain the same order in the three different cases, indicating that the architecture does influence the number of concepts learned.

### 5.6 Matchmap Visualizations and Videos

We can visualize the matchmaps produced by our models in several ways. The 3-dimensional density shown in Figure 3

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



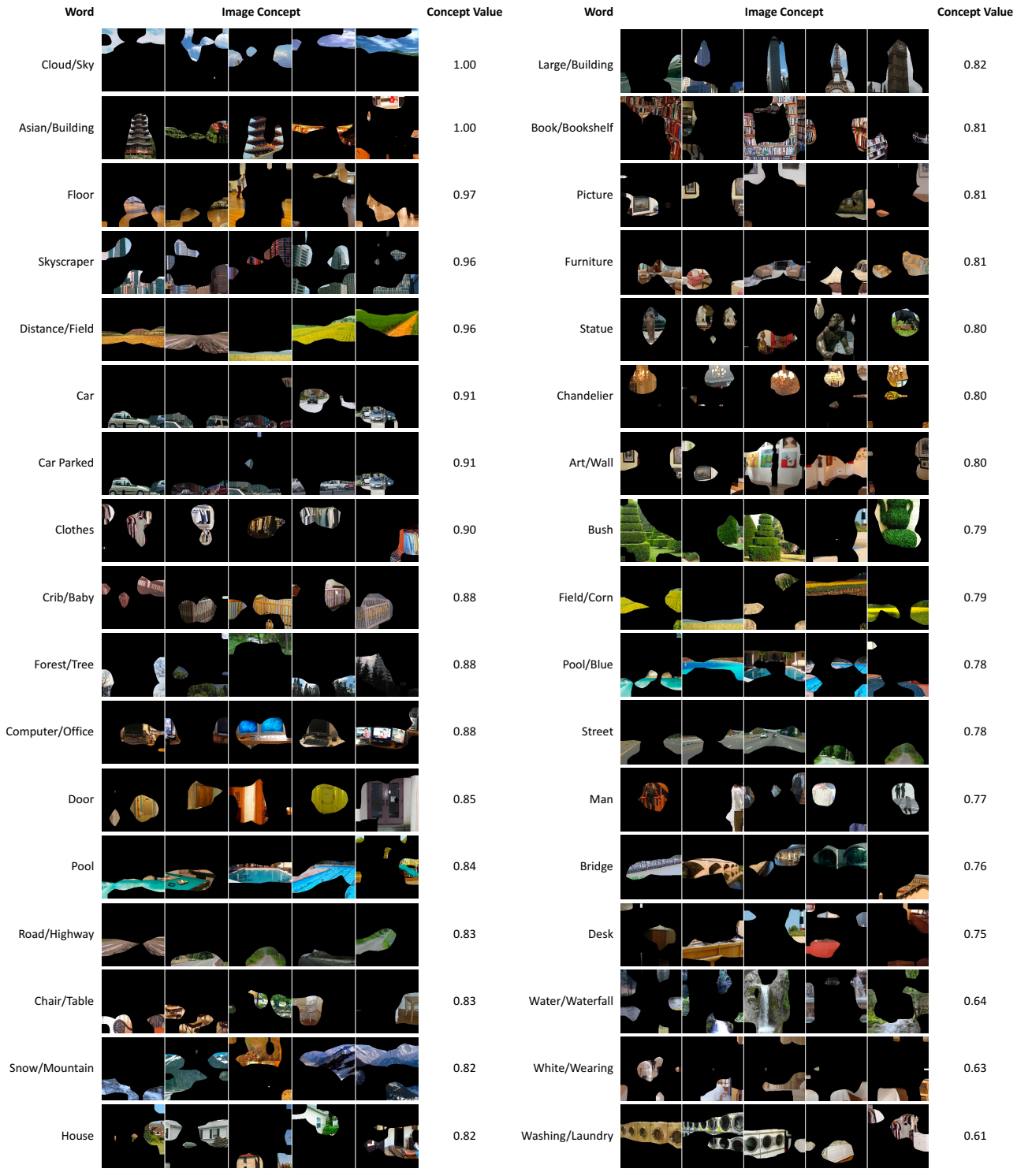


Fig. 10: Matching the most activated images in the image network and the activated words in the audio network we can establish pairs of image-word, as shown in the figure. We also define a concept value, which captures the agreement between both networks and ranges from 0 (no agreement) to 1 (full agreement).

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



| Epoch 1 | Epoch 5 | Epoch 10 | Epoch 20 | Epoch 30 | Epoch 40 |
|---|---|---|---|---|---|
| 1.0 Building/Sky | 1.0 Tree | 1.0 Cloudy/Sky | 1.0 Building | 1.0 Cloud/Sky | 1.0 Asian/Building |
| .97 Building/City | 1.0 Building/Skyline | 1.0 Bed/Bedroom | 1.0 Corn/Field | 1.0 Building | 1.0 Cloud/Sky |
| .96 Tree | .98 Building/City | 1.0 Tree | 1.0 Bed/Hotel | 1.0 Road | 1.0 Truck |
| .95 Building | .98 Building/Door | .97 Brick/Building | 1.0 Asian/Building | 1.0 Corn/Field | .99 Building/City |
| .94 Building/Window | .97 Building/Red | .97 Room/Window | .96 City/Skyscraper | .96 Skyscraper | .97 Floor |
| .94 Street/Tree | .97 Building | .97 Window | .95 Floor | .96 Bed/Bedding | .96 Skyscraper |
| .89 Wall | .97 Building/Story | .97 Floor/Wall | .92 Blue/Structure | .96 Building/Glass | .96 Distance/Field |
| .88 Large/Structure | .95 Road | .96 Road | .92 Building/Structure | .96 Distance/Field | .93 Mountain/Snow |
| .88 House/Picture | .94 Cabinet/Kitchen | .96 Restaurant/Window | .92 House | .95 City/Skyscraper | .93 Building |
| .87 House/Lighthouse | .94 Corn/Field | .95 Bed | .91 Floor/Wall | .94 Window | .92 Glass/Skyscraper |
| .87 Mountain | .94 Bedroom/Wall | .94 Highway/Road | .91 Building/City | .94 Bed/Bedroom | .92 Building/Has |
| .86 Area/Tree | .93 Brick/Building | .93 Bed/Size | .89 Table/Wooden | .94 Bed | .92 Bed/Size |
| .83 House | .93 Has/Highway | .93 Grass/Green | .89 Skyscraper | .94 Highway/Road | .91 Car |
| .82 Water | .93 Bed/Bedroom | .92 Wall/White | .89 Bed | .93 Mountain/Snowy | .91 Window |
| .82 Wall/White | .92 Field | .92 House | .89 Three/Window | .93 Floor | .91 Car/Parked |
| .81 Building/House | .92 Cloudy/Sky | .91 Home/House | .88 Car | .93 Clothes/Clothing | .90 Bed |
| .81 Blue/Sky | .91 Floor/Wall | .91 Building/City | .88 Table | .92 Bed/Hotel | .90 Table |
| .80 Flower/Green | .89 Water | .90 Building/Old | .88 Pool | .92 Building/Has | .90 Clothes |
| .80 Sky | .89 Skyscraper | .90 City/Skyscraper | .87 Water | .91 Pool | .90 Highway/Road |
| .78 Green/Tree | .89 River/Road | .89 Several/Track | .87 Bedroom/Couch | .91 Door/Window | .90 Cabinet/Wooden |

Fig. 11: We show the top 20 concepts (by concept score) at various epochs during a single training run of the ResNet50 + ResDAVEnet SISA-SHN model with natural sound pre-training. The concepts containing words separated by a slash represent multi-word concepts. We subjectively observe that the concepts learned at earlier epochs tend to be simpler and larger objects (e.g. building, sky, water).

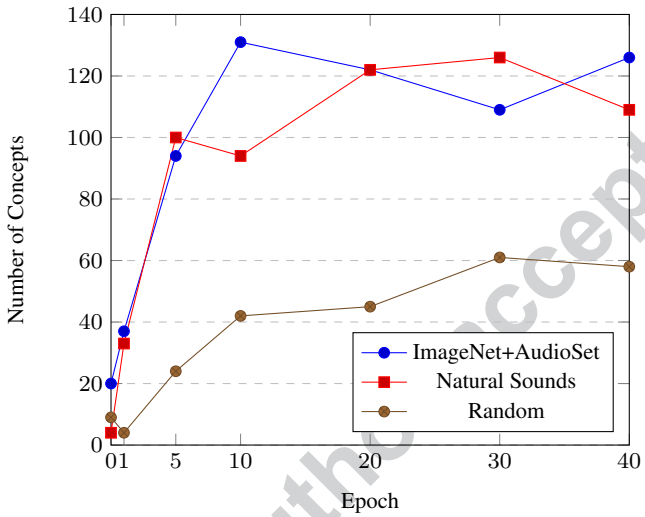


Fig. 12: The number of neurons whose concept value exceeds 0.7 as a function of training epoch for the ResNet50 + ResDAVEnet + SISA-SHN model using three different initializations.

is perhaps the simplest, although it can be difficult to read as a still image. Instead, we can treat it as a stack of masks overlayed on top of the image and played back as a video. We use the matchmap score to modulate the alpha channel of the image synchronously with the speech audio. The resulting video is able to highlight the salient regions of the images as the speaker is describing them.

We can also apply a threshold to the matchmaps and then extract volumetric connected components from the density. We then project them down onto the image and spectrogram axes, shown in Figure 13. More visualizations of this are shown in Figures 14 and 15. In practice, we found that an absolute score threshold between 100 and 400 generally produced attractive results, although the threshold required some hand-tuning between models. Future work should investigate better ways to normalize and segment the matchmaps. In Figure 14, we compare the segmented matchmaps computed with ResNet50 + ResDAVEnet SISA-SHN models under the three pre-training regimes. We find that they all do a good job co-segmenting the speech and image, although arguably the pre-trained models tend to be more precise than the random model. In Figure 15, we show many more example visualizations produced by the natural sound pre-trained model.

## 6 Conclusions

In this paper, we introduced audio-visual "matchmap" neural networks which are capable of directly learning the semantic correspondences between speech frames and image pixels without the need for annotated training data in either modality. We applied these networks for semantic image/spoken caption search, speech-prompted object localization, audio-visual clustering and concept discovery, and real-time, speech-driven, semantic highlighting. We examined the various ways in which factors such as the specific model architecture, training algorithm, and model pre-training influence the ability of our matchmap networks to learn spoken words, visual objects, and the semantics that link them. We also introduced an extended version of the Places audio

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



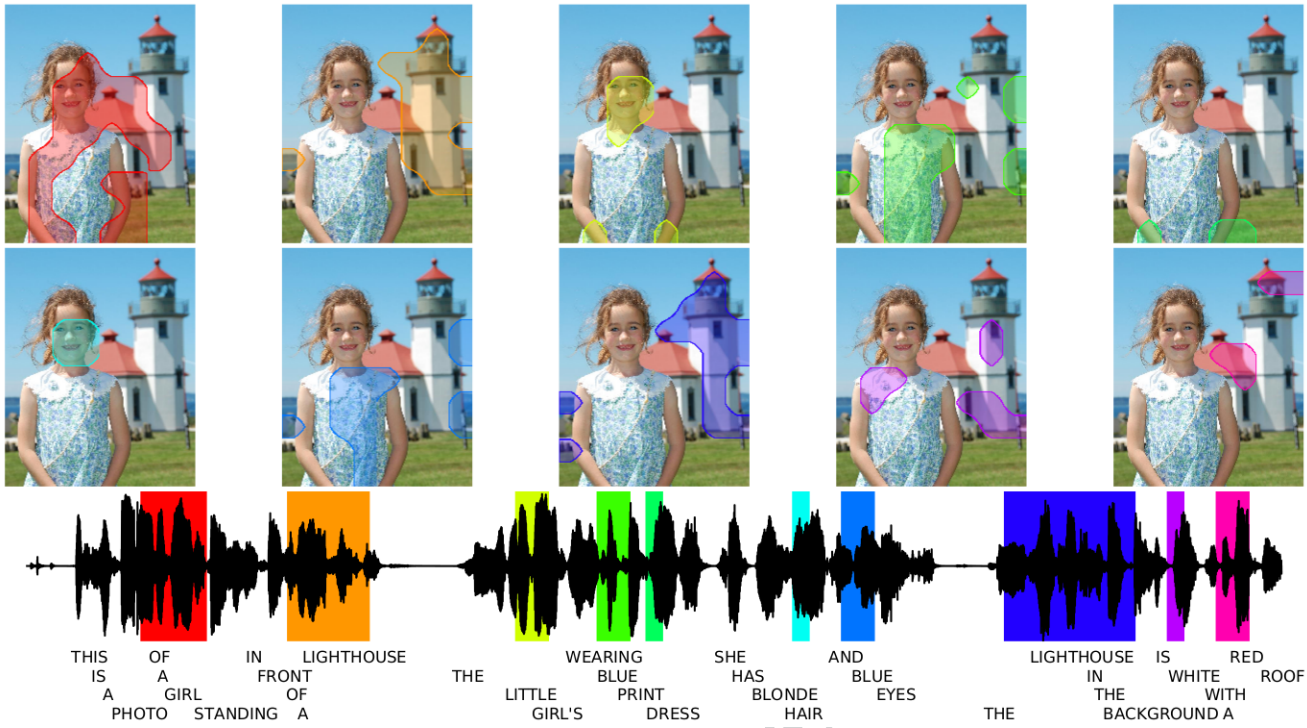


Fig. 13: Co-segmentation of the example image-caption pair shown in Figure 3.

caption dataset [23], doubling the total number of captions. Additionally, we introduced nearly 10,000 captions for the ADE20k dataset.

There are numerous avenues for future work, including expansion of the models to handle videos, additional languages, richer modeling of environmental sounds, etc. It may possible to directly generate images given a spoken description, or generate artificial speech describing a visual scene. More focused datasets that go beyond simple spoken descriptions and explicitly address relations between objects within the scene could be leveraged to learn richer linguistic representations. We are also excited by the potential that this line of work offers for embodied learning agents. One of the central difficulties faced by embodied agents in the real world is learning where their attention should be directed in the first place. Speech and language offer a way for agents to share social cues with one another to direct this attention. Finally, and related to this, a crucial element of human language learning is the dialog feedback loop, and future work should investigate the addition of that mechanism to the models.

## References


1. Alishahi, A., Barking, M., Chrupala, G.: Encoding of phonology in a recurrent neural model of grounded speech. In: Proc. ACL Conference on Natural Language Learning (CoNLL) (2017)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence, Z., Parikh, D.: VQA: Visual question answering. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2015)
3. Arandjelovic, R., Zisserman, A.: Look, listen, and learn. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2017)
4. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Proc. Neural Information Processing Systems (NeurIPS) (2016)
5. Bergamo, A., Bazzani, L., Anguelov, D., Torresani, L.: Self-taught object localization with deep networks. CoRR **abs/1409.3964** (2014). URL `http://arxiv.org/abs/1409.3964`
6. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. In: J.D. Cowan, G. Tesauro, J. Alspector (eds.) Advances in Neural Information Processing Systems 6, pp. 737–744. Morgan-Kaufmann (1994)
7. Cho, M., Kwak, S., Schmid, C., Ponce, J.: Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
8. Chrupala, G., Gelderloos, L., Alishahi, A.: Representations of language in a model of visually grounded speech signal. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL) (2017)
9. Cinbis, R., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **39**(1), 189–203 (2016)
10. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. CoRR **abs/1505.05192** (2015). URL `http://arxiv.org/abs/1505.05192`
11. Drexler, J., Glass, J.: Analysis of audio-visual features for unsupervised speech recognition. In: Proc. Grounded Language Understanding Workshop (2017)


© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



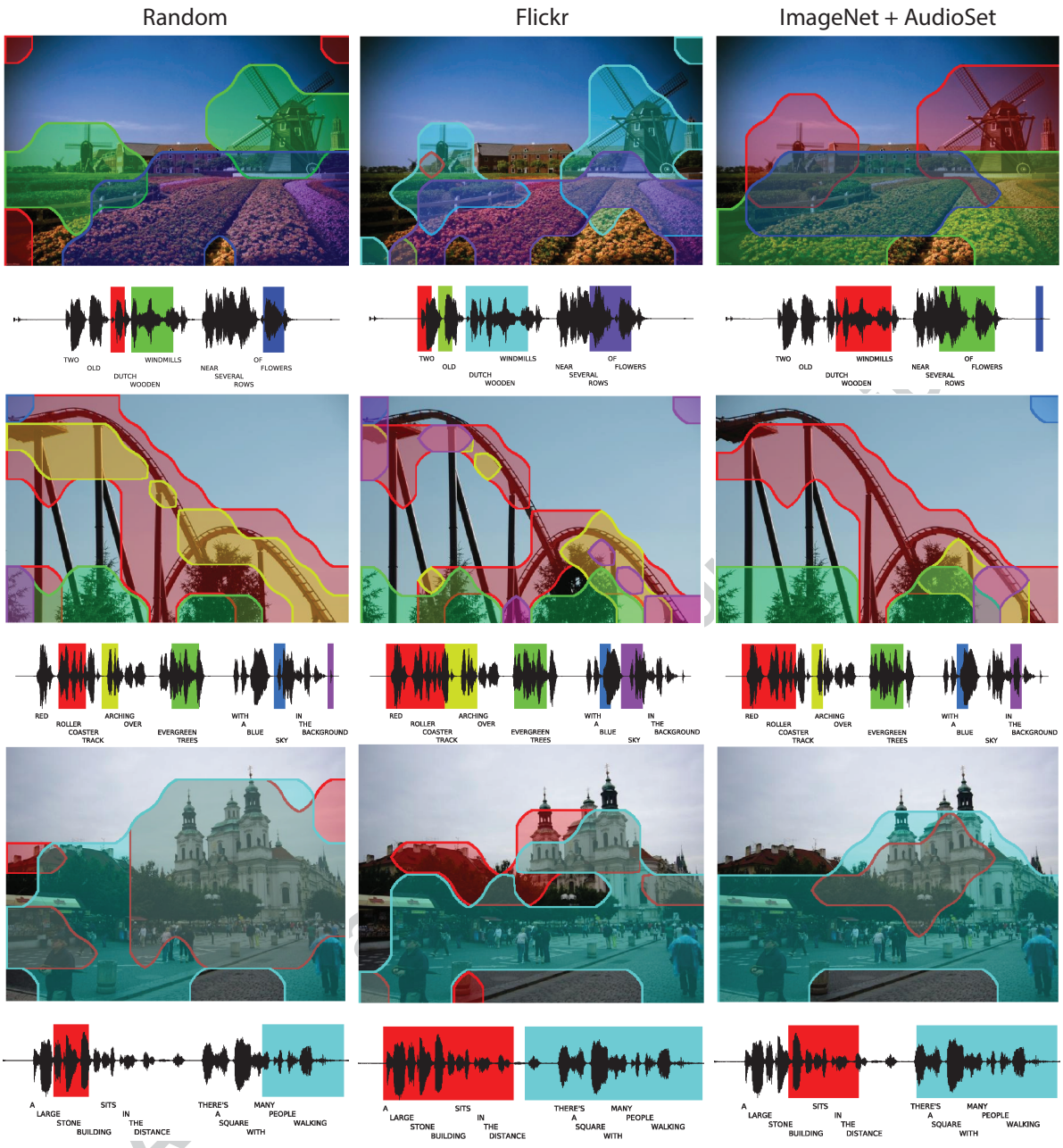


Fig. 14: Co-segmentation of images and their spoken captions using thresholded matchmaps produced by the ResNet50 + ResDAVEnet model, using the SISA-SHN loss. We compare three versions of this model with various pre-training conditions.

12. Dupoux, E.: Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. In: Cognition (2018)
13. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. In: Proc. British Machine Vision Conference (BMVC) (2018)
14. Fang, H., Gupta, S., Iandola, F., Rupesh, S., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., C., P.J., Zitnick, C.L., Zweig, G.: From captions to visual concepts and back. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
15. Fellbaum, C.: WordNet: An Electronic Lexical Database. Bradford Books (1998)
16. Gao, H., Mao, J., Zhou, J., Huang, Z., Yuille, A.: Are you talking to a machine? dataset and methods for multilingual image question answering. In: Proc. Neural Information Processing Systems (NeurIPS) (2015)
17. Gelderloos, L., Chrupaa, G.: From phonemes to images: levels of representation in a recurrent neural model of visually-grounded language learning. In: arXiv:1610.03342 (2016)
18. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017)

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



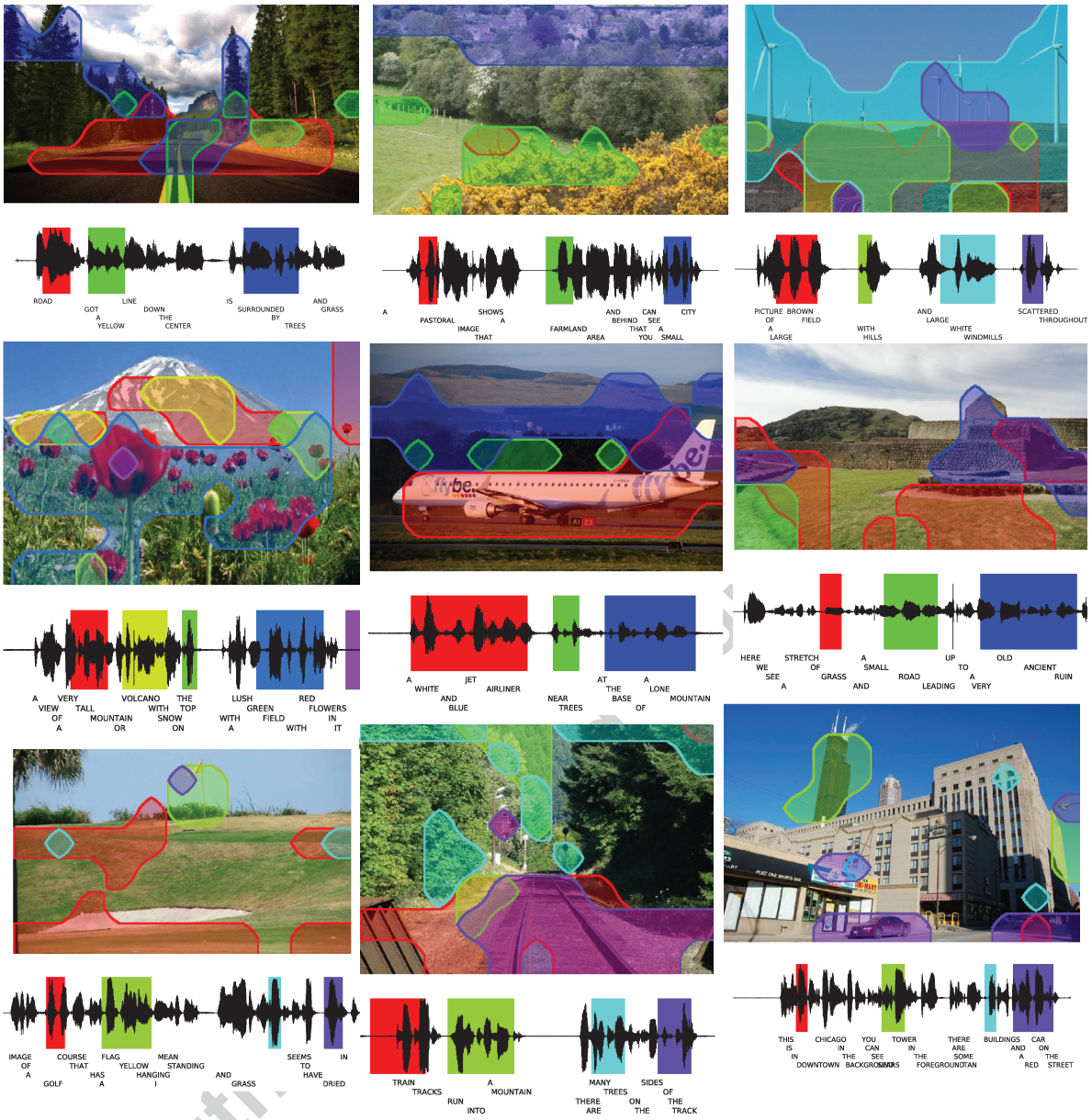


Fig. 15: Additional co-segmentation examples using the SISA-SHN ResNet50 + ResDAVEnet model, pre-trained on the natural sound Flickr videos.


19. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
20. Guérin, J., Gibaru, O., Thiery, S., Nyiri, E.: CNN features are also great at unsupervised classification. CoRR **abs/1707.01700** (2017). URL `http://arxiv.org/abs/1707.01700`
21. Harwath, D., Glass, J.: Learning word-like units from joint audio-visual analysis. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL) (2017)
22. Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., Glass, J.: Jointly discovering visual objects and spoken words from raw sensory input. In: Proc. IEEE European Conference on Computer Vision (ECCV) (2018)
23. Harwath, D., Torralba, A., Glass, J.R.: Unsupervised learning of spoken language with visual context. In: Proc. Neural Information Processing Systems (NeurIPS) (2016)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015). URL `http://arxiv.org/abs/1512.03385`
25. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Journal of Machine Learning Research (JMLR) (2015)
26. Jansen, A., Church, K., Hermansky, H.: Toward spoken term discovery at scale with zero resources. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2010)
27. Jansen, A., Plakal, M., Pandya, R., Ellis, D.P., Hershey, S., Liu, J., Moore, R.C., Saurous, R.A.: Unsupervised learning of seman-


© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



tic audio representations. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018)
28. Jansen, A., Van Durme, B.: Efficient spoken term discovery using randomized algorithms. In: Proc. IEEE Workshop on Automfatic Speech Recognition and Understanding (ASRU) (2011)
29. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
30. Kamper, H., Elsner, M., Jansen, A., Goldwater, S.: Unsupervised neural network based feature extraction using weak top-down constraints. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2015)
31. Kamper, H., Jansen, A., Goldwater, S.: Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. IEEE Transactions on Audio, Speech and Language Processing **24**(4), 669–679 (2016)
32. Kamper, H., Settle, S., Shakhnarovich, G., Livescu, K.: Visually grounded learning of keyword prediction from untranscribed speech. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2017)
33. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
34. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proc. Neural Information Processing Systems (NeurIPS) (2012)
35. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
36. Lee, C., Glass, J.: A nonparametric Bayesian approach to acoustic model discovery. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL) (2012)
37. Lewis, M.P., Simon, G.F., Fennig, C.D.: Ethnologue: Languages of the World, Nineteenth edition. SIL International. Online version: http://www.ethnologue.com (2016)
38. Lin, T., Marie, M., Belongie, S., Bourdev, L., Girshick, R., Perona, P., Ramanan, D., Zitnick, C.L., Dollar, P.: Microsoft COCO: Common objects in context. In: arXiv:1405.0312 (2015)
39. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: Proc. Neural Information Processing Systems (NeurIPS) (2014)
40. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: Proc. IEEE International Conference on Computer Vision (ICCV) (2015)
41. Ondel, L., Burget, L., Cernocky, J.: Variational inference for acoustic unit discovery. In: 5th Workshop on Spoken Language Technology for Under-resourced Language (2016)
42. Owens, A., Isola, P., McDermott, J.H., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
43. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: Proc. IEEE European Conference on Computer Vision (ECCV) (2016)
44. Park, A., Glass, J.: Unsupervised pattern discovery in speech. IEEE Transactions on Audio, Speech and Language Processing **16**(1), 186–197 (2008)
45. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
46. Reed, S.E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. CoRR **abs/1605.05396** (2016). URL `http://arxiv.org/abs/1605.05396`
47. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: Proc. Neural Information Processing Systems (NeurIPS) (2015)
48. Renshaw, D., Kamper, H., Jansen, A., Goldwater, S.: A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2015)
49. Roy, D.: Grounded spoken language acquisition: Experiments in word learning. IEEE Transactions on Multimedia **5**(2), 197–209 (2003)
50. Roy, D., Pentland, A.: Learning words from sights and sounds: a computational model. Cognitive Science **26**, 113–146 (2002)
51. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115**(3), 211–252 (2015). DOI 10.1007/s11263-015-0816-y
52. Russell, B., Efros, A., Sivic, J., Freeman, W., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2006)
53. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
54. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
55. Spelke, E.S.: Principles of object perception. Cognitive Science **14**(1), 29 – 56 (1990). DOI https://doi.org/10.1016/0364-0213(90)90025-R. URL `http://www.sciencedirect.com/science/article/pii/036402139090025R`
56. Thiolliere, R., Dunbar, E., Synnaeve, G., Versteegh, M., Dupoux, E.: A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH) (2015)
57. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.: The new data and new challenges in multimedia research. CoRR **abs/1503.01817** (2015). URL `http://arxiv.org/abs/1503.01817`
58. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
59. de Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.C.: Guesswhat?! visual object discovery through multi-modal dialogue. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
60. Weber, M., Welling, M., Perona, P.: Towards automatic discovery of object categories. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
61. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proc. International Conference on Machine Learning (ICML) (2015)
62. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: Proc. International Conference on Machine Learning (ICML) (2015)
63. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: an efficient data clustering method for very large databases. In: ACM SIGMOD international conference on Management of data, pp. 103–114 (1996)
64. Zhang, Y., Salakhutdinov, R., Chang, H.A., Glass, J.: Resource configurable spoken query detection using deep boltzmann machines. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2012)

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.

AUTHOR ACCEPTED MANUSCRIPT



65. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene CNNs. In: Proc. International Conference on Learning Representations (ICLR) (2015)
66. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
67. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Proc. Neural Information Processing Systems (NeurIPS) (2014)
68. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

Author accepted manuscript

© 2019 Springer Science+Business Media, LLC, part of Springer Nature.