This ECCV 2018 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ECCV 2018
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/eccv

Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input

David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass

Massachusetts Institute of Technology
[dharwath,recasens,didac,torralba]@csail.mit.edu, glass@mit.edu

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, convolutional networks, multi-modal learning, unsupervised learning

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, multi-modal context, and environmental interaction [12, 47]. 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 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 multi-modal data. Current work bringing together vision and language [2, 13, 14, 23, 28, 33, 34, 40, 41, 49, 50, 52]

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, 36, 37]. 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 [19], 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.

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

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

Visual Object Recognition and Discovery. State of the art systems are trained using bounding box annotations for the training data [16, 39], however other works investigate weakly-supervised or unsupervised object localization [5, 7, 9, 56]. 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 [51], which utilized an iterative clustering and classification algorithm to discover object categories. Further works borrowed ideas from textual topic models [45], assuming that certain sets of objects generally appear together in the same image scene. More recently, CNNs have been adapted to this task [10, 17], for example by learning to associate image patches which commonly appear adjacent to one another.

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 [31]. Within the 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 [21, 22, 38], or Bayesian generative models of the speech signal [25, 30, 35]. Neural networks have thus far been mostly utilized in this realm for learning frame-level acoustic features [24, 42, 48, 54].

Fusion of Vision and Language. Joint modeling of images and natural language text has gained rapidly in popularity, encompassing tasks such as image captioning [13, 28, 23, 49, 52], visual question answering (VQA) [2, 14, 33, 34, 41], multimodal dialog [50], and text-to-image generation [40]. 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 [44, 43], who learned correspondences between images of objects and the outputs of a supervised phoneme recognizer. Recently, it was demonstrated by Harwath et al [19] 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, 18, 26]. 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.

Fusion of Vision and Sounds. A number of recent models have focused on integrating other acoustic signals to perform unsupervised discovery of objects and ambient sounds [3, 4, 36, 37]. 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 [19, 18]. This dataset contains approximately 200,000 recordings collected via Amazon Mechanical Turk of people verbally describing the content of images from the Places 205 [58] 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. 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 [59] 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 [32] contains an average of just over 10 words per caption, the places audio captions are on average 20 words long, 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 is similar to that of Harwath et al [19], in which a pair of convolutional neural networks (CNN) [29] 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. Our models are trained to optimize a ranking-based criterion [6, 27, 19], 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 compute the loss:

$$\begin{aligned} \mathrm{L} = \sum_{j=1}^{B} \Big(& \max(0, S(I_j, A_j^{imp}) - S(I_j, A_j) + \eta) \\ & + \max(0, S(I_j^{imp}, A_j) - S(I_j, A_j) + \eta) \Big), \end{aligned} \tag{1}$$

where $S(I, A)$ represents the similarity score between an image I and audio caption A, I_j^{imp} represents the j^{th} randomly chosen imposter image, A_j^{imp} the j^{th} imposter caption, and η is a margin hyperparameter. We sample the imposter image and caption for each pair from the same minibatch, and fix η to 1 in our experiments. The choice of

similarity function 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 our focus is to explore how object and word *localization* naturally emerges as a by-product of this training scheme. An illustration of our two-branch matchmap networks is shown in Figure 3. Next, we describe the modeling for each input mode.

4.1 Image Modeling

We follow [19, 18, 15, 8, 1, 26] by utilizing the architecture of the VGG16 network [46] to form the basis of the image branch. In all of these prior works, however, the weights of the VGG network were pre-trained on ImageNet, and thus had a significant amount of visual discriminative ability built-in to their models. We show that our models do not require this pre-training, and can be trained end-to-end in a completely unsupervised fashion. Additionally in these prior works, the entire VGG 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 `conv5` to `fc1` involves a flattening operation, which makes it difficult to recover associations between any neuron above `conv5` and the spatially localized stimulus which was responsible for its output. We address this issue here by retaining only the convolutional banks up through `conv5` from the VGG network, and discarding `pool5` and everything above it. For a 224 by 224 pixel input image, the output of this portion of the network would be a 14 by 14 feature map across 512 channels, with each location within the map possessing a receptive field that can be related directly back to the input. In order to map an image into the shared embedding space, we apply a 3 by 3, 1024 channel, linear convolution (no nonlinearity) to the `conv5` feature map. Image pre-processing 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 mean and variance.

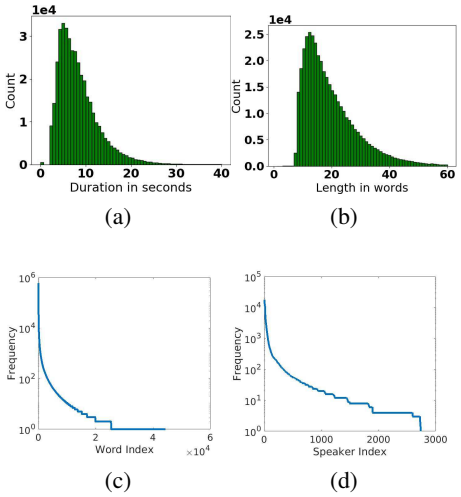

Fig. 2: Statistics of the 400k spoken captions. From left to right, the plots represent (a) the histogram over caption durations in seconds, (b) the histogram over caption lengths in words, (c) the estimated word frequencies across the captions, and (d) the number of captions per speaker.

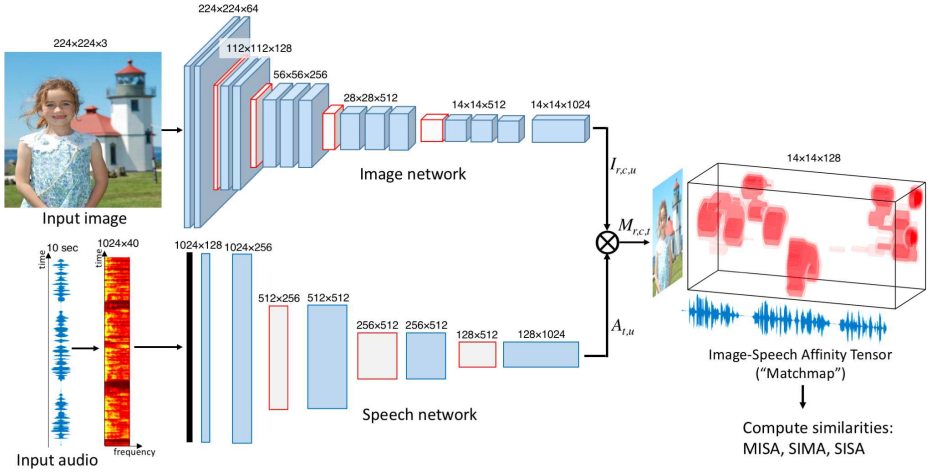

Fig. 3: The audio-visual matchmap model architecture (left), along with an example matchmap output (right), displaying a 3-D density of spatio-temporal similarity. Conv layers shown in blue, pooling layers shown in red, and BatchNorm layer shown in black. Each conv layer is followed by a ReLU. The first conv layer of the audio network uses filters that are 1 frame wide and span the entire frequency axis; subsequent layers of the audio network are hence 1-D convolutions with respective widths of 11, 17, 17, and 17. All maxpool operations in the audio network are 1-D along the time axis with a width of 3. An example spectrogram input of approx. 10 seconds (1024 frames) is shown to illustrate the pooling ratios.

4.2 Audio Caption Modeling

To model the spoken audio captions, we use a model similar to that of [18], 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. [19] utilized truncation and zero-padding of each spectrogram to a fixed length. While this enables batched inputs to the model, it introduces a degree of undesirable bias into the learned representations. Instead, we pad to a length long enough to fully capture the longest caption within a batch, and truncate the output feature map of each caption on an individual basis to remove the frames corresponding to zero-padding. Rather than manually normalizing the spectrograms, we employ a BatchNorm [20] layer at the front of the network. Next, we discuss methods for relating the visual and auditory feature maps to one another.

4.3 Joining the Image and Audio Branches

Zhou et al [57] 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 I^p and A^p be their globally average-pooled counterparts. One straightforward choice of similarity function is the dot product between the pooled embeddings, $S(I, A) = I^{pT} A^p$. Notice that this is in fact equivalent to first computing a 3rd order tensor M such that $M_{r,c,t} = I^T_{r,c,:} A_{t,:}$, and then computing the average of all elements of M. 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. In this regard, the similarity between the global average pooled image and audio representations is simply the average similarity between *all* audio frames and *all* image regions. We call this similarity scoring function SISA (sum image, sum audio):

$$\text{SISA}(M) = \frac{1}{N_r N_c N_t} \sum_{r=1}^{N_r} \sum_{c=1}^{N_c} \sum_{t=1}^{N_t} M_{r,c,t} \tag{2}$$

Because M reflects the localized similarity between a small image region (possibly containing an object) and a small segment of audio (possibly containing a word), we dub M the "matchmap" tensor between and image and an audio caption. As 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:

$$\text{MISA}(M) = \frac{1}{N_t} \sum_{t=1}^{N_t} \max_{r,c}(M_{r,c,t}) \tag{3}$$

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:

$$\text{SIMA}(M) = \frac{1}{N_r N_c} \sum_{r=1}^{N_r} \sum_{c=1}^{N_c} \max_{t}(M_{r,c,t}) \tag{4}$$

In the next section, we explore the use of these similarities for learning semantic correspondences between objects within images and spoken words within their captions.

5 Experiments

5.1 Image and Caption Retrieval

All models were trained using the sampled margin ranking objective outlined in Equation 1, using stochastic gradient descent with a batch size of 128. We used a fixed momentum of 0.9 and an initial learning rate of 0.001 that decayed by a factor of 10 every 70 epochs; generally our models converged in less than 150 epochs. 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 [19, 18, 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. Table 1 displays the image/caption recall scores achieved when training a matchmap model using the SISA, MISA, and SIMA similarity functions, both with a fully randomly initialized network as well as with an image branch pre-trained on ImageNet. In all cases, the MISA similarity measure is the best performing, although all three measures achieve respectable scores. Unsurprisingly, using a pre-trained image network significantly increases the recall scores. In Table 1, we compare our models against reimplementations of two previously published speech-to-image models (both of which utilized pre-trained VGG16 networks). We also 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 architecture 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. Both previously published baselines we compare to used the full VGG network, deriving an embedding for the entire image from the `fc2` outputs. In the pre-trained case, our best recall scores for the MISA model outperform [19] overall as well as [18] on image recall; the caption recall score is slightly lower than that of [18]. This demonstrates that there is not much to be lost when doing away with the fully connected layers of VGG, and much to be gained in the form of the localization matchmaps.

5.2 Speech-Prompted Object Localization.

To evaluate our models' ability to associate spoken words with visual objects in a more fine-grained sense, we use the spoken captions for the ADE20k [59] dataset. The ADE20k images contain pixel-level object masks and labels - in conjunction with a time-aligned transcription produced via ASR (we use the public Google SpeechRecognition API for this purpose), we can associate each matchmap cell with a specific visual

Table 1: Recall scores on the held out set of 1,000 images/captions for the three matchmap similarity functions. We also show results for the baseline models which use automatic speech recognition-derived text captions. The (P) indicates the use of an image branch pre-trained on ImageNet

	Speech						ASR Text					
	Caption to Image			Image to Caption			Caption to Image			Image to Caption		
Model	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SISA	.063	.191	.274	.048	.166	.249	.136	.365	.503	.106	.309	.430
MISA	.079	.225	.314	.057	.191	.291	.162	.417	.547	.113	.309	.447
SIMA	.073	.213	.284	.065	.168	.255	.134	.389	.513	.145	.336	.459
SISA(P)	.165	.431	.559	.120	.363	.506	.230	.525	.665	.174	.462	.611
MISA(P)	.200	.469	.604	.127	.375	.528	.271	.567	.701	.183	.489	.622
SIMA(P)	.147	.375	.506	.139	.367	.483	.215	.518	.639	.220	.494	.599
[19](P)	.148	.403	.548	.121	.335	.463	-	-	-	-	-	-
[18](P)	.161	.404	.564	.130	.378	.542	-	-	-	-	-	-

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 localization. Given a word in the speech beginning at time t_1 and ending at time t_2, we derive a heatmap across the image by summing the matchmap between t_1 and t_2. We then normalize the heatmap to sit within the interval [0,1], threshold the heatmap, and evaluate the intersection over union (IoU) of the detection mask with the ADE20k label mask for whatever object was referenced by the word.

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-object pair, we compute an average IoU score across all instances of the word-object pair appearing together in an ADE20k image and its associated caption. We then average these scores across all 100 word-object pairs and report results for each model type in Table 2. We also report the IoU scores for the ASR text-based baseline models described in Section 5.1. Figure 4 displays a sampling of localization heatmaps for several query words using the non-pretrained speech MISA network.

5.3 Clustering of Audio-Visual Patterns

The next experiment we consider is automatic discovery of audio-visual clusters from the ADE20k matchmaps using the fully random speech MISA network. Once a matchmap has been computed for an image and caption pair, we smooth it with an average or max pooling window of size 7 across the temporal dimension before binarizing it according to a threshold. In practice, we set this threshold on a matchmap-specific basis to be 1.5 standard deviations above the mean value of the smoothed matchmap. Next, we extract

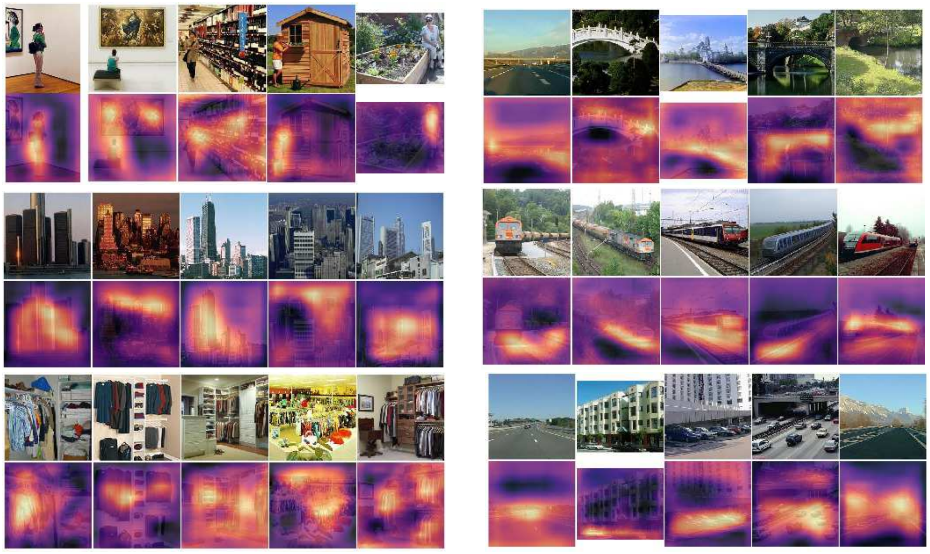

Fig. 4: Speech-prompted localization maps for several word/object pairs. From top to bottom and from left to right, the queries are instances of the spoken words "WOMAN," "BRIDGE,", "SKYLINE", "TRAIN", "CLOTHES" and "VEHICLES" extracted from each image's accompanying speech caption.

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 Birch clustering [53] with 1000 target clusters for the first step, and an agglomerative final step that resulted in 135 clusters. To derive word 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 50 most pure clusters in Figure 5.

5.4 Concept discovery: building an image-word dictionary

Figure 5 shows the clusters learned by our model. Interestingly, the audio and image networks are able to agree to a common representation of knowledge, clustering similar concepts together. Since both representations are directly multiplied by a dot product, both networks have to agree on the meaning of these different dimensions. 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.

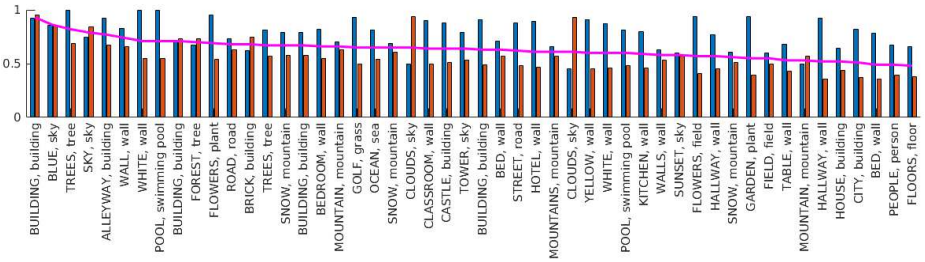

Fig. 5: Some clusters (speech and visual) found by our approach. Each cluster is jointly labeled with the most common word (capital letters) and object (lowercase letters). For each cluster we show the precision for both the word (blue) and object (red) labels, as well as their harmonic mean (magenta). The average cluster size across the top 50 clusters was 44.

To visualize the concepts associated with each of the dimensions in the image path, we use the unit visualization technique introduced in [55]. A set of images is run through the image network and the ones that activate the most that particular dimension get selected. Then, we can visualize the spatial activations in the top activated images. The same procedure can be done for the audio network, where we get a set of descriptions that maximally activate that neuron. Finally, with the temporal map, we can find which part of the description has produced that activation. Some most activated words and images can be found in Figure 6. We show four dimensions with their associated most activated word in the audio neuron, and the most activated images in the image neuron. Interestingly, these pairs of concepts have been found completely independently, as we did not use the final activation (after the dot product) to pick the images.

Table 2: Speech-prompted and ASR-prompted object localization IoU scores on ADE20K, averaged across the 100 word-object pairs. 'Rand.' indicates a randomly initialized model, while 'Pre.' indicates an image branch pre-trained on ImageNet. The full-frame baseline IoU was 0.16

	Speech		ASR Text	
Sim. Func.	Rand.	Pre.	Rand.	Pre.
SIMA	.1607	.1857	.1743	.1995
SISA	.1637	.1970	.1750	.2161
MISA	.1795	.2324	.2060	.2413

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 6, 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.

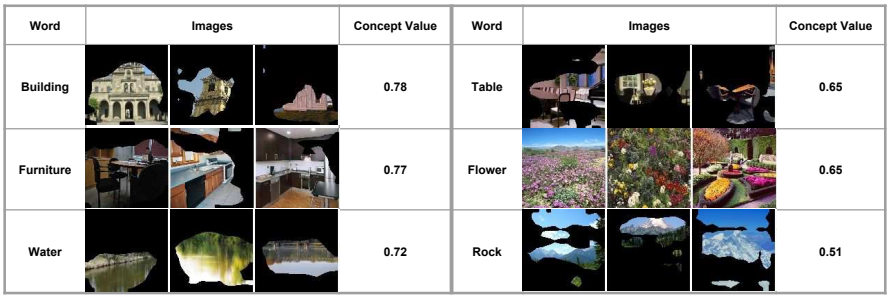

Fig. 6: Matching the most activated images in the image network and the activated words in the audio network we can establish pairs of image-word, as shown in the figure. We also define a concept value, which captures the agreement between both networks and ranges from 0 (no agreement) to 1 (full agreement).

This metric will help us to compute the quality of each individual neuron and of each particular model.

To quantify the quality of the dictionary, we need to find a common space between the written descriptions and the image activations. Again, this common space comes from a segmentation dataset. Using [59], we can rank the most detected objects by each of the neurons. We pass through the network approx. 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), and a word associated with the audio neuron. Using the WordNet tree, we can compute the word distance between these concepts and define the following metric:

$$c = \sum_{i=1}^{|O^{\text{im}}|} w_i Sim_{\text{wup}}(o_i^{\text{im}}, o^{\text{au}}), \tag{5}$$

with $o_i^{\text{im}} \in O^{\text{im}}$, where O^{im} is the set of classes present in the TOP5 segmented images and $Sim_{\text{wup}}(.,.)$ is the Wu and Palmer WordNet-based similarity, with range [0,1] (higher is more similar). 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 dimension, which measures how well both the audio network and the image network agree on that particular concept. The numerical values for six concept pairs are shown in Figure 6. We see how neurons with higher value are cleaner and more related with its counterpart. The bottom right neuron shows an example of low concept value, where the audio word is "rock" but the neuron images show mountains in general. Anecdotally, we found $c > 0.6$ to be a good indicator that a concept has been learned.

Finally, we analyze the relation between the concepts learned and the architecture used in Table 3. Interestingly, the four maintain the same order in the three different cases, indicating that the architecture does influence the number of concepts learned.

5.5 Matchmap Visualizations and Videos

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

Table 3: The number of concepts learned by the different networks with different losses. We find it is consistently highest for MISA.

	Speech		ASR Text	
Sim. Func.	Rand.	Pre.	Rand.	Pre.
SIMA	166	124	96	96
SISA	210	192	103	102
MISA	242	277	140	150

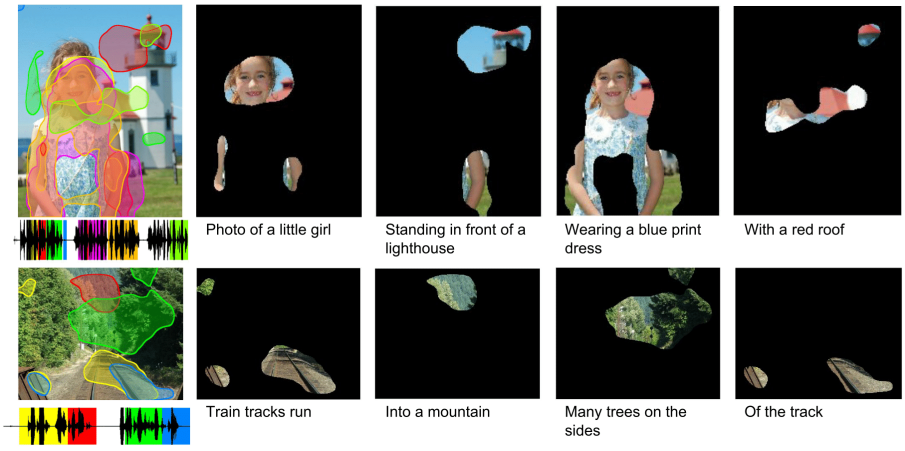

Fig. 7: On the left are shown two images and their speech signals. Each color corresponds to one connected component derived from two matchmaps from a fully random MISA network. The masks on the right display the segments that correspond to each speech segment. We show the caption words obtained from the ASR transcriptions below the masks. Note that those words were never used for learning, only for analysis.

We can also extract volumetric connected components from the density and project them down onto the image and spectrogram axes; visualizations of this are shown in Figures 7 and 8. We apply a small amount of thresholding and smoothing to prevent the matchmaps from being too fragmented. We use a temporal max pooling window with a size of 7 frames, and normalize the scores to fall within the interval [0, 1] and sum

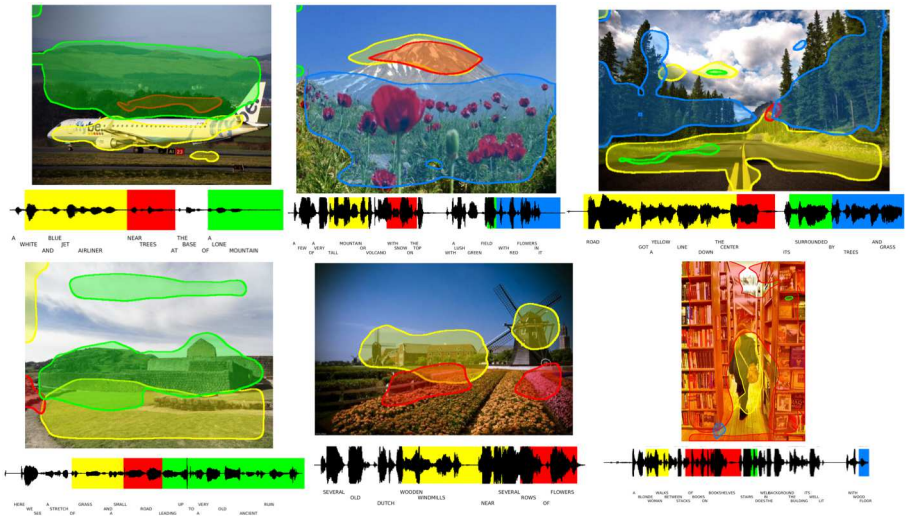

Fig. 8: Additional examples of discovered image segments and speech fragments using the fully random MISA speech network.

to 1. We zero out all the cells outside the top p percentage of the total mass within the matchmap. In practice, p values between 0.15 and 0.3 produced attractive results.

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 also introduced an extended version of the Places audio caption dataset [19], 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, environmental sounds, additional languages, 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. Finally, 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.

Acknowledgments

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

References

1. Alishahi, A., Barking, M., Chrupala, G.: Encoding of phonology in a recurrent neural model of grounded speech. In: 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: ICCV (2017)
4. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems 29, pp. 892–900 (2016)
5. Bergamo, A., Bazzani, L., Anguelov, D., Torresani, L.: Self-taught object localization with deep networks. CoRR **abs/1409.3964** (2014), 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: Cowan, J.D., Tesauro, G., Alspector, J. (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: 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), http://arxiv.org/abs/1505.05192
11. Drexler, J., Glass, J.: Analysis of audio-visual features for unsupervised speech recognition. In: Grounded Language Understanding Workshop (2017)
12. Dupoux, E.: Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. In: Cognition (2018)
13. 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)
14. 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: NIPS (2015)
15. 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)
16. 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)
17. Guérin, J., Gibaru, O., Thiery, S., Nyiri, E.: CNN features are also great at unsupervised classification. CoRR **abs/1707.01700** (2017), http://arxiv.org/abs/1707.01700
18. Harwath, D., Glass, J.: Learning wor d-like units from joint audio-visual analysis. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL) (2017)
19. Harwath, D., Torralba, A., Glass, J.R.: Unsupervised learning of spoken language with visual context. In: Proc. Neural Information Processing Systems (NIPS) (2016)
20. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Journal of Machine Learning Research (JMLR) (2015)
21. 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)

22. Jansen, A., Van Durme, B.: Efficient spoken term discovery using randomized algorithms. In: Proc. IEEE Workshop on Automfatic Speech Recognition and Understanding (ASRU) (2011)
23. 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)
24. 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)
25. 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 (Apr 2016)
26. Kamper, H., Settle, S., Shakhnarovich, G., Livescu, K.: Visually grounded learning of keyword prediction from untranscribed speech. In: INTERSPEECH (2017)
27. Karpathy, A., Joulin, A., Fei-Fei, L.: Deep fragment embeddings for bidirectional image sentence mapping. In: Proc. Neural Information Processing Systems (NIPS) (2014)
28. Karpathy, A., Li, F.F.: Deep visual-semantic alignments for generating image descriptions. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
29. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
30. Lee, C., Glass, J.: A nonparametric Bayesian approach to acoustic model discovery. In: Proc. Annual Meeting of the Association for Computational Linguistics (ACL) (2012)
31. 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)
32. 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)
33. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: NIPS (2014)
34. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: ICCV (2015)
35. Ondel, L., Burget, L., Cernocky, J.: Variational inference for acoustic unit discovery. In: 5th Workshop on Spoken Language Technology for Under-resourced Language (2016)
36. Owens, A., Isola, P., McDermott, J.H., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. pp. 2405–2413 (2016)
37. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient Sound Provides Supervision for Visual Learning, pp. 801–816 (2016)
38. Park, A., Glass, J.: Unsupervised pattern discovery in speech. IEEE Transactions on Audio, Speech and Language Processing **16**(1), 186–197 (2008)
39. 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)
40. Reed, S.E., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. CoRR **abs/1605.05396** (2016), http://arxiv.org/abs/1605.05396
41. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: NIPS (2015)
42. 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)

43. Roy, D.: Grounded spoken language acquisition: Experiments in word learning. IEEE Transactions on Multimedia **5**(2), 197–209 (2003)
44. Roy, D., Pentland, A.: Learning words from sights and sounds: a computational model. Cognitive Science **26**, 113–146 (2002)
45. 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)
46. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
47. Spelke, E.S.: Principles of object perception. Cognitive Science **14**(1), 29 – 56 (1990). https://doi.org/https://doi.org/10.1016/0364-0213(90)90025-R, http://www.sciencedirect.com/science/article/pii/036402139090025R
48. 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)
49. 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)
50. 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)
51. Weber, M., Welling, M., Perona, P.: Towards automatic discovery of object categories. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
52. 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: ICML (2015)
53. 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)
54. 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)
55. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene CNNs. arXiv preprint arXiv:1412.6856 (2014)
56. 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)
57. 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)
58. 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 (NIPS) (2014)
59. 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)