

See, Hear, and Read: Deep Aligned Representations

Yusuf Aytar, Carl Vondrick, Antonio Torralba
Massachusetts Institute of Technology
`{yusuf, vondrick, torralba}@csail.mit.edu`

Abstract

We capitalize on large amounts of readily-available, synchronous data to learn a deep discriminative representations shared across three major natural modalities: vision, sound and language. By leveraging over a year of sound from video and millions of sentences paired with images, we jointly train a deep convolutional network for aligned representation learning. Our experiments suggest that this representation is useful for several tasks, such as cross-modal retrieval or transferring classifiers between modalities. Moreover, although our network is only trained with image+text and image+sound pairs, it can transfer between text and sound as well, a transfer the network never observed during training. Visualizations of our representation reveal many hidden units which automatically emerge to detect concepts, independent of the modality.

1. Introduction

Invariant representations are core for vision, audio, and language models because they abstract our data. For example, we desire viewpoint and scale invariance in vision, reverberation and background noise invariance in audio, and synonym and grammar invariance in language. Discriminative, invariant representations learned from large datasets have enabled machines to understand unconstrained situations to huge success [20, 26, 13, 1].

The goal of this paper is to create representations that are robust in another way: we learn representations that are aligned across modality. Consider the sentence “she jumped into the pool.” This same concept could also appear visually or aurally, such as the image of a pool or the sound of splashing. Representations are robust to modality if there is alignment in the representation across modalities. The pool image, the splashing sound, and the above sentence should have similar representations.

We believe aligned cross-modal representations will have a large impact in computer vision because they are fundamental components for machine perception to understand relationships between modalities. Cross-modal perception

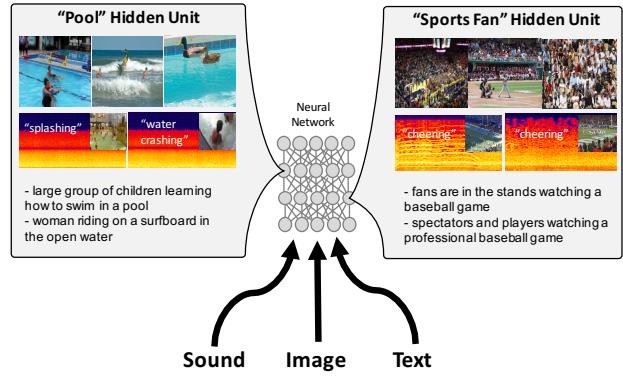


Figure 1: **Aligned Representations:** We present a deep cross-modal convolutional network that learns a representation that is aligned across three senses: seeing, hearing, and reading. Above, we show inputs that activate a hidden unit the most. Notice that units fire on concepts independent of the modality. See Figure 5 for more.

plays key roles in the human perceptual system to recognize concepts in different modalities [2, 9]. Cross-modal representations also have many practical applications in recognition and graphics, such as transferring learned knowledge between modalities.

In this paper, we learn rich deep representations that are aligned across the three major natural modalities: vision, sound, and language. We present a deep convolutional network that accepts as input either a sound, a sentence, or an image, and produces a representation shared across modalities. We capitalize on large amounts of in-the-wild data to learn this aligned representation across modalities. We develop two approaches that learn high-level representations that can be linked across modalities. Firstly, we use an unsupervised method that leverages the natural synchronization between modalities to learn an alignment. Secondly, we design an approach to transfer discriminative visual models into other modalities.

Our experiments and visualizations show that a representation automatically emerges that detects high-level concepts independent of the modality. Figure 1 visualizes this

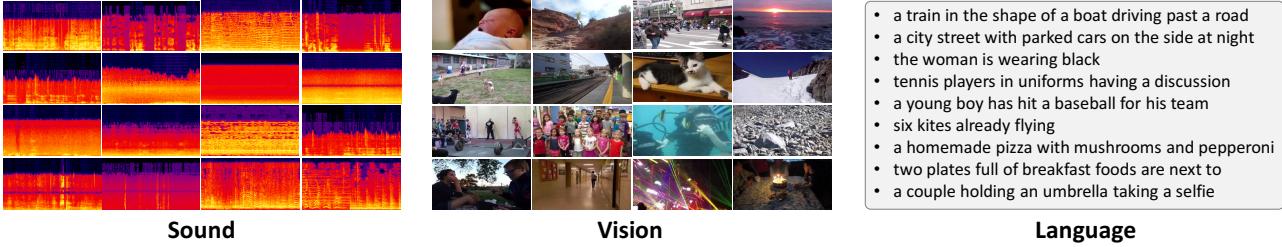


Figure 2: **Dataset:** We learn deep, aligned representations by capitalizing on large amounts of raw, unconstrained data.

learned representation: notice how units in the upper layers have learned automatically to detect some objects agnostic of the modality. We experiment with this representation for several multi-modal tasks, such as cross-modal retrieval and classification. Moreover, although our network is only trained with image+text and image+sound pairs, our representation can transfer between text and sound as well, a transfer the network never saw during training.

Our primary contribution is showing how to leverage massive amounts of synchronized data to learn a deep, aligned cross-modal representation. While the methods in the paper are standard, their application on a large-scale to the three major natural modalities is novel to our knowledge. In the remainder of this paper, we describe the approach and experiments in detail. In section 2, we discuss our datasets and modalities. In section 3, we present a model for learning deep aligned cross-modal representations. In section 4, we present several experiments to analyze our representations.

1.1. Related Work

Vision and Sound: Understanding the relationship between vision and sound has been recently explored in the computer vision community. One of the early works, [24], explored the cross-modal relations between “talking head” images and speech through CCA and cross-modal factor analysis. [43] applied CCA between visual and auditory features, and used common subspace features for aiding clustering in image-audio datasets. [37] explored interaction between visual and audio modalities through human behavior analysis using Kernel-CCA and Multi-view Hidden CRF. [27] investigates RBM auto-encoders between vision and sound. [22] investigated the relations between materials and their sound in a weakly-paired settings. Recent work [29] has capitalized on material properties to learn to regress sound features from video, learn visual representations [30], and [5] analyzes small physical vibrations to recover sounds in video. We learn cross-modal relations from large quantities of unconstrained data.

Sound and Language: Even though the relation between sound and language is mostly studied in the line of speech recognition [31], in this paper we are interested in

matching sentences with auditory signals. This problem is mainly studied in the audio retrieval setting. Early work [35] performs semantic audio retrieval by aligning sound clusters with hierarchical text clusters through probabilistic models. [4] applies a passive-aggressive model for content-based audio retrieval from text queries. [40] uses probabilistic models for annotating novel audio tracks with words and retrieve relevant tracks given a text-based query. However, we seek to learn the relationship between sound and language using vision as an intermediary, i.e. we do not use audio+text pairs.

Language and Vision: Learning to relate text and images has been extensively explored in the computer vision community. Pioneering work [7, 32, 28, 21] explore image-captioning as a retrieval task. More recently, [42, 15, 6] developed deep large-scale models to generate captions from images. In this paper, rather than generating sentences, we instead seek to learn a representation that is aligned with images, audio, and text. [7] explores aligned representations, but does not learn the representation with a deep architecture. Moreover, rather than using recurrent networks [42], we use convolutional networks for text. [45] learns to align books and movies. [10, 11] learn joint image-tag embeddings through several CCA variations. We instead seek to align three natural modalities using readily-available large-scale data. While [10] harnesses clusters of tags as a third view of the data, we instead obtain clusters from images through state-of-the-art visual categorization models. This is crucial since only the image modality is shared in both image+sound and image+text pairs.

2. Datasets and Modalities

We chose to learn aligned representations for sound, vision, and language because they are frequently used in everyday situations. Figure 2 shows a few examples of the data we use.

Sound: We are interested in natural environmental sounds. We download videos from videos on Flickr [39] and extract their sounds. We downloaded over 750,000 videos from Flickr, which provides over a year (377 days) of continuous audio, as well as their corresponding video

frames. The only pre-processing we do on the sound is to extract the spectrogram from the video files and subtract the mean. We extract spectrograms for approximately five seconds of audio, and keep track of the video frames for both training and evaluation. We use 85% of the sound files for training, and the rest for evaluation.

Language: We combine two of the largest image description datasets available: COCO [25], which contains 400,000 sentences and 80,000 images, and Visual Genome [19], which contains 4,200,000 descriptions and 100,000 images. The concatenation of these datasets results in a very large set of images and their natural language descriptions, which cover various real-world concepts. We pre-process the sentences by removing English stop words, and embedding each word with word2vec [26].

Images: We use the frames from our sound dataset [39] and the images from our language datasets [25, 19]. In total, we have nearly a million images which are synchronized with either sound or text (but not both). The only pre-processing we do on the images is subtracting the channel-wise mean RGB value. We use the same train/test splits as their paired sounds/descriptions.

Synchronization: We use the synchronous nature of these modalities to learn the relationships between them. We have pairs of images and sound (from videos) and pairs of images and text (from caption datasets). Note we lack pairs of sound and text during training. Instead, we hope our network will learn to map between sound and text by using images as a bridge (which our experiments suggest happens). To evaluate this, we also collected 1,000 text descriptions of videos (image/sound) from workers on Amazon Mechanical Turk [38], which we only use for testing the ability to transfer between sound and text.

3. Cross-Modal Networks

We design a model that can accept as input either an image, a sound, or a sentence, and produces a common representation shared across modalities. Let x_i be a sample from modality x , and y_i be the corresponding sample from modality y . For example, x_i may be an image, and y_i may be a sound of that image. For clarity, here we describe how to align two modalities x and y , but the method easily generalizes to any number of modalities (we do three).

Our goal is to learn representations for x_i and y_i that are aligned. We write $f_x(x_i)$ to be the representation in modality x , and $f_y(y_i)$ to be the representation in modality y . Representations $f_x(x_i)$ and $f_y(y_i)$ are aligned if they are close to each other under some distance metric, e.g. cosine similarity. However, similarity alone is not enough because there is a trivial solution to ignore the input and produce a constant. Instead, we desire the representation to be both aligned and discriminative. We explore two approaches.

3.1. Alignment by Model Transfer

We take advantage of discriminative visual models to teach a student model to have an aligned representation. Let $g(x_i)$ be a teacher model that estimates class probabilities for a particular modality. For example, $g(x_i)$ could be any image classification model, such as AlexNet [34]. Since the modalities are synchronized, we can train $f_y(y_i)$ to predict the class probabilities from the teacher model $g(x_i)$ in another modality. We use the KL-divergence as a loss:

$$\sum_i^N D_{\text{KL}}(g(x_i) \parallel f_y(y_i)) \quad (1)$$

where $D_{\text{KL}}(P \parallel Q) = \sum_j P_j \log \frac{P_j}{Q_j}$. This objective by itself will enable alignment to emerge at the level of categories predicted by g . However, the internal representations of f would not be aligned since each student model is disjoint.

To enable an alignment to emerge in the internal representation, we therefore constrain the upper layers of the network to have shared parameters across modalities, visualized in Figure 3. While the early layers of f are specific to modality, the upper layers will now be shared. This encourages an internal representation to emerge that is shared across modalities. Interestingly, as we show in the experiments, visualizations suggest that hidden units emerge internally to detect some objects independent of modality.

Student-teacher models have been explored in transfer learning before [1, 12]. In this work, we are instead transferring into an aligned representation, which is possible by constraining the learned parameters to be shared across the upper levels of representation.

3.2. Alignment by Ranking

We additionally employ a ranking loss function to obtain both aligned and discriminative representations:

$$\sum_i^N \sum_{j \neq i} \max\{0, \Delta - \psi(x_i, y_i) + \psi(x_i, y_j)\} \quad (2)$$

where Δ is a margin hyper-parameter, ψ is a similarity function, and j iterates over negative examples. Note that, for clarity and in slight abuse of notation, f may be a different layer in the network from above.

This loss seeks to push paired examples close together in representation space, and mismatched pairs further apart, up to some margin Δ . We use cosine similarity in representation space:

$$\psi(x, y) = \cos(f_x(x), f_y(y)) \quad (3)$$

where \cos is the cosine of the angle between the two representation vectors.

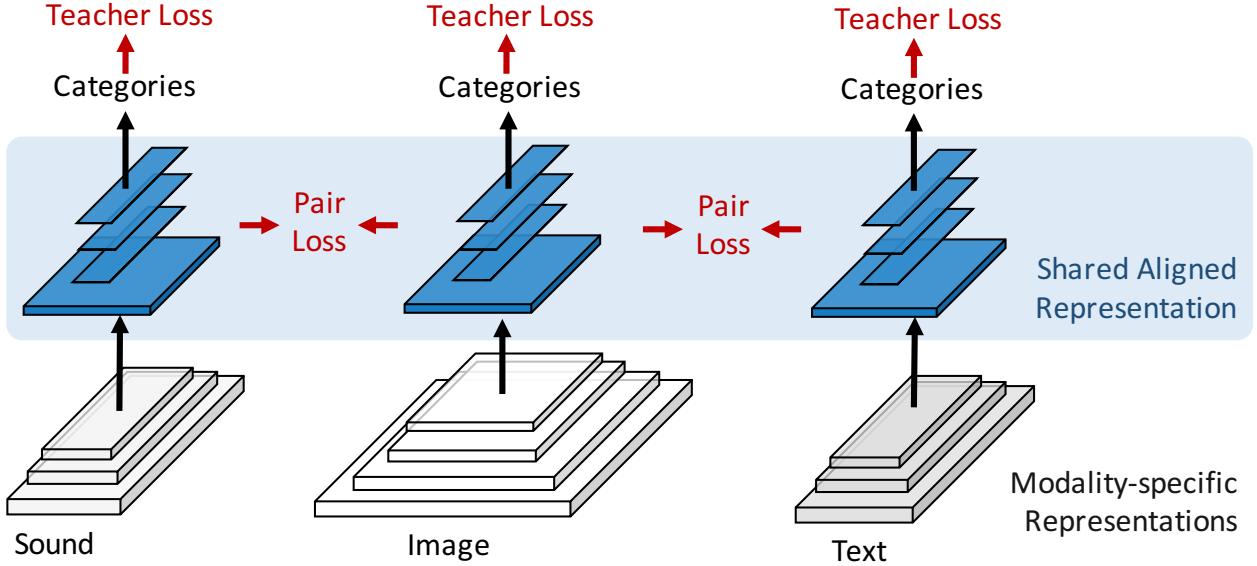


Figure 3: **Learning Aligned Representations:** We design a network that accepts as input either an image, a sound, or a text. The model produces a common shared representation that is aligned across modality (blue) from modality-specific representations (grays). We train this model using both a model transfer loss, and a ranking pair loss. The modality-specific layers are convolutional, and the shared layers are fully connected.

Ranking loss functions are commonly used in vision to learn cross-modal embeddings in images and text [18, 36, 15, 41, 8]. Here, we are leveraging them to learn aligned, discriminative representations across three major natural modalities using in-the-wild data.

3.3. Learning

To train the network, we use the model transfer loss in Equation 1 and the ranking loss in Equation 2 on different layers in the network. For example, we can put the model transfer loss on the output layer of the network, and the ranking loss on all shared layers in the network. The final objective becomes a sum of these losses.

Model transfer: We train student models for sound, vision, and text to predict class probabilities from a teacher ImageNet model. We constrain the upper weights to be shared in the student models. Since vision is a rich modality with strong recognition models, it is an attractive resource for transfer.

Ranking: We apply the ranking loss for alignment between vision → text, text → vision, vision → sound, and sound → vision on the last three hidden activations of the network. Since we do not have large amounts of sound/text pairs, we do not supervise those pairs. Instead, we expect the model to learn a strong enough alignment using vision as a bridge to enable transfer between sound/text (which our experiments suggest).

3.4. Network Architecture

Our network has three different inputs, depending on the modality of the data. We design each input to have its own disjoint pathway in the beginning in the network. In the end, however, the pathways converge to common layers that are shared across all modalities. Our intention is that the disjoint pathways can adapt to modal-specific features (such as shapes, audible notes, or text phrases), while the shared layers can adapt to modal-robust features (such as objects and scenes).

Sound Network: The input to our sound pathway are spectrograms. Since sound is a one-dimensional signal, we use a four-layer one-dimensional convolutional network to transform the spectrogram into a higher-level representation. The output of the sound network is then fed into the modal-agnostic layers.

Text Network: The input to our text pathway are sentences where each word is embedded into a word representation using word2vec [26]. By concatenating each word together, we can use a deep one-dimensional convolutional network on the sentence, similar to [16]. We again use a four-layer network. While the earlier layers in the network have a small receptive field and can only detect simple n -grams, by stacking convolutions, the later layers have a larger receptive field. In contrast to [16] which uses convolutions of varying kernel size to handle long-range dependencies, we instead use convolutions with fixed kernel sizes, but go deeper to capture long-range dependencies be-

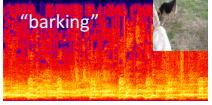
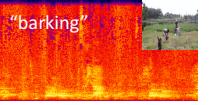
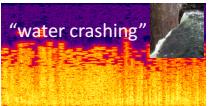
| Input Query | Sound Retrievals | | Text Retrievals | | Image Retrievals | |
|---|---|---|---|---|---|--|
|  |  |  | - A dog lying down on the beach - The dog belongs to the homeowner |  |  | |
|  |  |  | - Steel tracks under the train. - The train platform |  |  | |
| The choppy water the man is riding |  |  | - A person stands on water skis in the water - A couple of kayakers paddling through water |  |  | |

Figure 4: **Example Cross-Modal Retrievals:** We show top retrievals for cross-modal retrieval between sounds, images, and text using our deep representation.

| Method | IMG | SND | IMG | TXT |
|----------------------|-------------|-------------|------------|------------|
| | ↓ SND | ↓ IMG | ↓ TXT | ↓ IMG |
| Random | 500.0 | 500.0 | 500.0 | 500.00 |
| Linear Reg. | 345.8 | 319.8 | 14.2 | 18.0 |
| CCA [33] | 313.6 | 316.1 | 17.0 | 16.2 |
| Normalized CCA [10] | 295.6 | 296.0 | 14.2 | 12.8 |
| Ours: Model Transfer | 144.6 | 143.8 | 8.5 | 10.8 |
| Ours: Ranking | 49.0 | 47.8 | 8.6 | 8.2 |
| Ours: Both | 47.5 | 49.5 | 5.8 | 6.0 |

Table 1: **Cross Modal Retrieval:** We evaluate average median rank for cross-modal retrieval on our held-out validation set. Lower is better. See Section 4.2 for details.

tween words. The output of this network is finally fed into the modal-agnostic layers.

Vision Network: Our visual network follows the standard Krizhevsky architecture [20]. We use the same architecture up until pool5, which is flattened and directly fed into the modal-agnostic layers.

Shared Network: The outputs from the sound, text, and vision networks are fixed length vectors with the same dimensionality. In order to create a representation that is independent of the modality, we then feed this fixed length vector into a network that is shared across all modalities, similar to [3]. We visualize this sharing in Figure 3. While the weights in the earlier layers are specific to their modality, the weights in the upper layers are shared across all modalities. We use two fully connected layers of dimensionality 4096 with rectified linear activations as this shared network. The output is 1000 dimensional with a softmax activation function.

3.5. Implementation Details

Optimization: We optimize the network using mini-batch stochastic gradient descent and back propagation

| Method | TXT → SND | SND → TXT |
|----------------------|--------------|--------------|
| Random | 500.0 | 500.0 |
| Linear Reg. | 315.0 | 309.0 |
| Ours: Model Transfer | 140.5 | 142.0 |
| Ours: Ranking | 190.0 | 189.5 |
| Ours: Both | 135.0 | 140.5 |

Table 2: **Cross Modal Retrieval for Sound and Text:** We evaluate average median rank for retrievals between sound and text. See Section 4.3

[23]. We use the Adam solver [17] with a learning rate of 0.00001. We initialize all parameters with Gaussian white noise. We train with a batch size of 200 for a fixed number of iterations (50,000). We train the network in Caffe [14] and implement a new layer to perform the cosine similarity. Training typically takes a day on a GPU.

Sound Details: The input spectrogram is a 500×257 signal, which can be interpreted as 257 channels over 500 time steps. We use three one-dimensional convolutions with kernel sizes 11, 5, and 3 and 128, 256, 256 filters respectively. Between each convolutional layer, we use rectified linear units, and downsample with one-dimensional max-pooling by a factor of 5. The output of these convolutions is a 4×256 feature map. Since these convolutions are over time and the other modalities do not have time (e.g., images are spatial), we finally project this feature map to a 9216 dimensional vector with a fully connected layer, which is fed into the modality-agnostic layers.

Text Details: The pretrained model for word2vec embeds each word into a 300 dimensional vector. We concatenate words in a sentence into a fixed length matrix of size 16×300 for 16 words. We pad shorter sentences with zeros, and crop longer sentences, which we found to be effective in practice. We then have three one-dimensional convolutions with 300 filters and kernel size of 3 with rectified linear activation functions. We have max-pooling after the second

| Train Modality: | IMG | | | SND | | | TXT | | |
|----------------------|-------------|------------|-------------|-------------|------------|-------------|-------------|------------|-------------|
| Test Modality: | IMG | SND | TXT | IMG | SND | TXT | IMG | SND | TXT |
| Chance | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 | 2.3 |
| Linear Reg. | 26.5 | 3.3 | 23.1 | 3.0 | 6.6 | 2.9 | 18.3 | 3.4 | 34.3 |
| CCA TXT↔IMG | 23.8 | - | 22.2 | - | - | - | 18.5 | - | 35.6 |
| CCA SND↔IMG | 21.1 | 3.0 | - | 2.7 | 6.8 | - | - | - | - |
| Ours: Ranking | 23.5 | 5.7 | 21.3 | 6.6 | 5.7 | 6.3 | 11.3 | 5.2 | 32.9 |
| Ours: Model Transfer | 30.9 | 5.6 | 32.0 | 8.7 | 9.0 | 12.3 | 26.5 | 5.1 | 39.0 |
| Ours: Both | 32.6 | 5.8 | 33.8 | 12.8 | 9.0 | 15.2 | 22.6 | 6.2 | 40.3 |

Table 3: **Classifier Transfer:** We experiment with training scene classifiers in one modality, but testing on a different modality. Since our representation is aligned, we can transfer the classifiers without any labeled examples in the target modality. The table reports classification accuracy and the dash indicates a comparison is not possible because CCA only works with two views. The results suggest that our representation obtains a better alignment than baseline methods. Moreover, this shows that the representation is both aligned and discriminative.

and third convolutions to down-sample the input by a factor of two. We finally have a fully connected layer to produce a 9216 dimensional vector that is fed into the shared layers.

4. Experiments

We present three main experiments to analyze our aligned representation. Firstly, we experiment with cross-modal retrieval that, given a query in one modality, find similar examples in all modalities. Secondly, we show discriminative classifiers trained on our representation transfer to novel modalities. Finally, we visualize the internal representation, and show that some object detectors independent of modality are automatically emerging.

4.1. Experimental Setup

We split our data into disjoint training, validation, and testing sets. We learn all models on training, fit hyper-parameters on the validation set, and report performance on the testing set. For both validation and testing, we use 5,000 video/sound pairs and similarly 5,000 image/text pairs. The rest is used for training. For the image descriptions, we use the standard training/validation split from COCO [25]. Since Visual Genome [19] did not release a standard training/validation split, we randomly split the dataset. However, because Visual Genome has some overlap with the images in COCO, if an image belongs in both COCO and Visual Genome, then we assigned it to the same training/validation/testing split as COCO in order to keep the splits disjoint. We train all networks from scratch (random initialization).

4.2. Cross Modal Retrieval

We quantify the learned alignment by evaluating our representations at a cross-modal retrieval task. Given a query input in one modality, how well can our representation retrieve its corresponding pair from a different modality? For

our method, we input the example from the query modality into our network, and extract the features from the last hidden layer. We then normalize the query features to be zero mean and unit variance. Finally, we find examples in the target modality with the nearest cosine similarity.

We compare against two baselines for this task.

CCA: Firstly, we compare against CCA [33], which is a state-of-the-art method for cross-modal retrieval. Using our training set, we compute CCA between images and text, and CCA between images and sound. To do this, we need to operate over a feature space. For images, we use fc7 features from [20]. For sentences, we use a concatenation of words embedded with word2vec [26]. For sound, we reduce the dimensionality of the spectrograms to 512 dimensions using PCA, which we found improved performance. We do retrieval using the joint latent space learned by CCA.

Linear Regression: Secondly, we compare against a linear regression trained from the query modality to visual features, and use vision as the common feature space. We use the same features for linear regression as we did in the CCA baseline. Note we add a small isotropic prior to the transformation matrix which acts as a regularizer. We then perform retrieval using the regressed target features using cosine similarity.

Results: Using our test set, we report the average median rank over five splits of 1,000 each, following [41]. Table 1 shows that our representation learns a significantly better alignment across vision, sound, and text than baselines. However, for retrieval between text and images, our method marginally outperforms CCA. Since our network is capable of learning deep features, our method can learn to align spectrograms using features higher level than what is possible with CCA. On the other hand, since text is already high-level, our method only provides a slight advantage over CCA for text/images. In general, the task of retrieving between images and sound appears to be a more challenging task than between images and text, perhaps be-

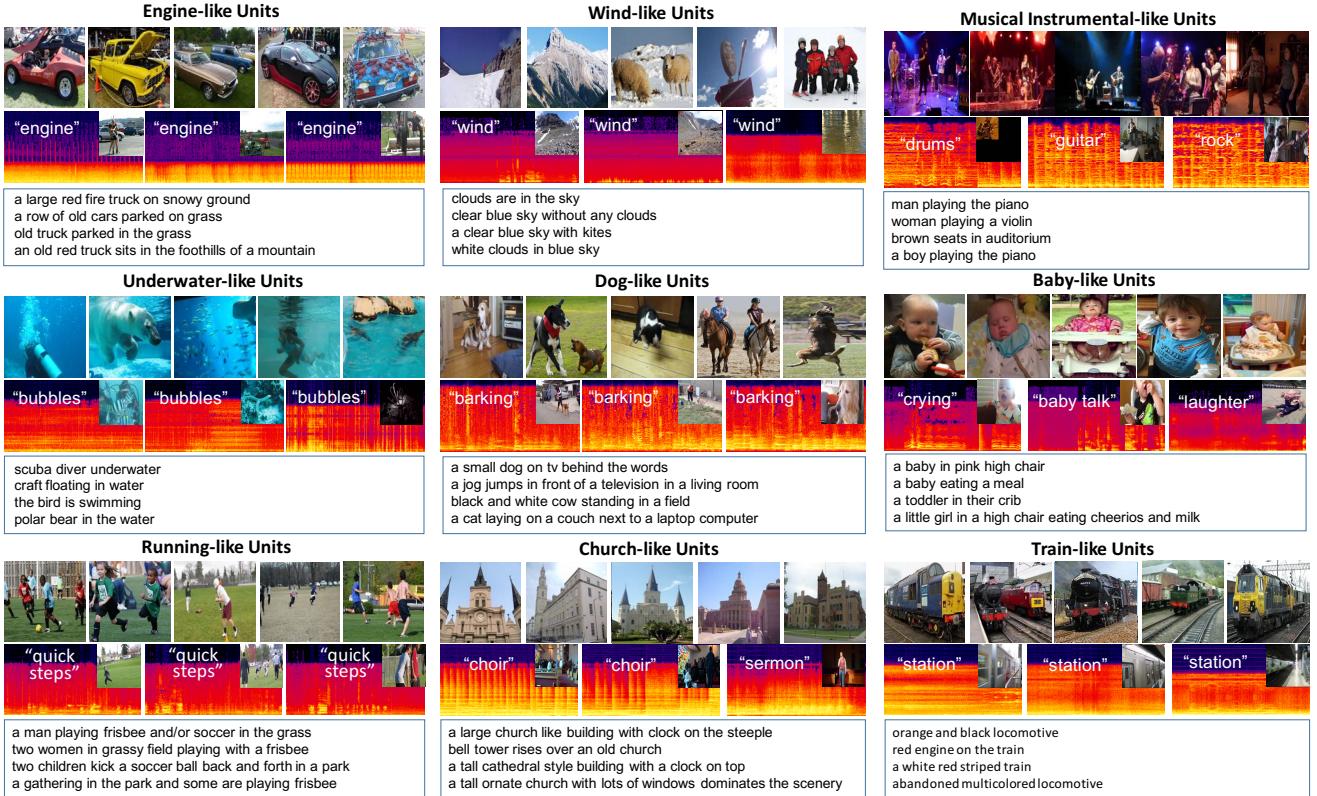


Figure 5: **Hidden Unit Visualization:** We visualize a few units from the last hidden layer of our cross modal network. Note that, on top of the spectrograms (yellow/red heatmaps), we also show the original video and a blurb to describe the sound, which is only included for visualization purposes. See Section 4.5.

cause less information is available in sound and the original features are not as high level. We show some qualitative top retrievals in Figure 4.

4.3. Sound and Text Transfer

Although our network was trained using only image/sound and image/text pairs, we also experiment with transfer between sound/text. The network never saw sound/text pairs during training. This task is particularly challenging because the network would need to develop a strong enough alignment between modalities such that it can exploit images as a bridge between sound and text.

Baseline: Although we cannot train a linear regression between sound and text (because there are no pairs), we can train linear regressions from spectrograms to image features, and text features and image features. We can use the regressed image features as the common space to do retrieval.

For our method, we simply perform retrieval using cosine similarity using our learned representations. Given a sound query, we compute its representation, and retrieve text that is near it, and similarly for the reverse direction.

Results: Table 2 reports the average median rank for

sound/text retrievals. Our experiments suggest that deep cross-modal representations outperform both cluster CCA and a linear regression by considerable margins (over 100 points). We believe this is the case because our network is capable of learning high-level features, which are easier to align across modalities. Interestingly, our network can transfer between sound/text only slightly worse than sound/images, suggesting that our network is capable of learning alignment between modalities even in the absence of synchronized data.

4.4. Zero Shot Classifier Transfer

We explore using the aligned representation as a means to transfer classifiers across modalities. If the representation obtains a strong enough alignment, then an object recognition classifier trained in a source modality should still be able to recognize objects in a different target modality, even though the classifier never saw labeled examples in the target modality.

Dataset: To quantify performance on this task, we collected a new medium size dataset for transferring classifiers across vision, sound, and text modalities. We annotate held-out videos into 42 categories consisting of objects and

scenes using Amazon Mechanical Turk. The training set is 2,799 videos and the testing set is 1,050 videos, which is balanced. We additionally annotated each video with a short text description, similar to sentences from COCO. This results in a dataset where we have paired data across all modalities.

Classifier: We experiment with training a linear one-vs-all SVM to recognize the categories where we train and test on different modalities using our aligned representation as the feature space. Note to pick hyper-parameters, we use two-fold cross validation on the training set.

Results: Table 3 reports classification accuracy for the classifier across modalities. We compare the representation from our approach versus a representation obtained by CCA and Linear regression, similar to before. Our experiments suggest that our representation learns a stronger discriminative alignment than CCA and Linear regression, obtaining up to 10% gain over baselines.

Particularly the cross-modal columns in table 3, where train and test modalities are different, shows that even without seeing any example from the target modality our methods can achieve significant classification accuracies. The most challenging source modality for training is sound, which makes sense as vision and text are very rich modalities. However, our approach still learns to align sound with vision and text. By combining both a paired ranking objective and a model transfer objective, our representation is both discriminative and aligned.

4.5. Visualization

To better understand what our model has learned, we visualize the hidden units in the shared layers of our network, similar to [44]. Using our validation set, we find which inputs activate a unit in the last hidden layer the most, for each modality. We visualize the highest scoring inputs for several hidden units in Figure 5. We observe two properties. Firstly, although we do not supervise semantics on the hidden layers, many units automatically emerge that detect high-level concepts. Secondly, many of these units seem to detect objects independently of the modality, suggesting the representation is learning an alignment at the object level.

5. Conclusion

Invariant representations enable computer vision systems to operate in unconstrained, real-world environments. We believe aligned, modality-robust representations are crucial for the next generation of machine perception as the field begins to leverage cross-modal data, such as sound, vision, and language. In this work, we present a deep convolutional network for learning cross-modal representations from over a year of video and millions of sentences. Our experiments show an alignment emerges that improves both

retrieval and classification performance for challenging in-the-wild situations. Although the network never saw pairs of sounds and text during training, our experiments empirically suggest it has learned an alignment between them, possibly by using images as a bridge internally. Our visualizations reveal that units for high-level concepts emerge in our representation, independent of the modality.

References

- [1] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *NIPS*, 2016. 1,3
- [2] G. A. Calvert, R. Campbell, and M. J. Brammer. Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current biology*, 2000. 1
- [3] L. Castrejon, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba. Learning aligned cross-modal representations from weakly aligned data. In *CVPR*, 2016. 5
- [4] G. Chechik, E. Ie, M. Rehn, S. Bengio, and D. Lyon. Large-scale content-based audio retrieval from text queries. In *CMIR*, 2008. 2
- [5] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand, and W. T. Freeman. The visual microphone: passive recovery of sound from video. 2014. 2
- [6] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *CVPR*, 2015. 2
- [7] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*. 2010. 2
- [8] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 4
- [9] M. H. Giard and F. Peronnet. Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of cognitive neuroscience*, 1999. 1
- [10] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 2014. 2,5
- [11] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*, 2014. 2
- [12] S. Gupta, J. Hoffman, and J. Malik. Cross modal distillation for supervision transfer. *arXiv*, 2015. 3
- [13] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv*, 2014. 1
- [14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Conference on Multimedia*, 2014. 5

- [15] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 2, 4
- [16] Y. Kim. Convolutional neural networks for sentence classification. *arXiv*, 2014. 4
- [17] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv*, 2014. 5
- [18] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv*, 2014. 4
- [19] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016. 3, 6
- [20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 5, 6
- [21] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. Berg. Babytalk: Understanding and generating simple image descriptions. *PAMI*, 2013. 2
- [22] C. H. Lampert and O. Krömer. Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In *ECCV 2010*. 2010. 2
- [23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998. 5
- [24] D. Li, N. Dimitrova, M. Li, and I. K. Sethi. Multimedia content processing through cross-modal association. In *Conference on Multimedia*, 2003. 2
- [25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*. 2014. 3, 6
- [26] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 1, 3, 4, 6
- [27] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011. 2
- [28] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 2
- [29] A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. *arXiv*, 2015. 2
- [30] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016. 2
- [31] L. Rabiner and B.-H. Juang. Fundamentals of speech recognition. 1993. 2
- [32] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *NAACL HTL*, 2010. 2
- [33] N. Rasiwasia, J. Costa Pereira, E. Covillo, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ICM*, 2010. 5, 6
- [34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014. 3
- [35] M. Slaney. Semantic-audio retrieval. In *ICASSP*, 2002. 2
- [36] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *ACL*, 2014. 4
- [37] Y. Song, L.-P. Morency, and R. Davis. Multimodal human behavior analysis: learning correlation and interaction across modalities. In *International conference on Multimodal interaction*, 2012. 2
- [38] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. 2008. 3
- [39] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv*, 2015. 2, 3
- [40] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *TASL*, 2008. 2
- [41] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun. Order-embeddings of images and language. *arXiv*, 2015. 4, 6
- [42] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 2
- [43] H. Zhang, Y. Zhuang, and F. Wu. Cross-modal correlation learning for clustering on image-audio dataset. In *Conference on Multimedia*, 2007. 2
- [44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *arXiv*, 2014. 8
- [45] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015. 2