See, Hear, and Read: Deep Aligned Representations

Hyun-Min Park

A.I Lab. Dept. of Computer Science and Engineering
Dongguk University. Seoul. Korea

About Paper

- See, Hear, and Read: Deep Aligned Representations
- Yusuf Aytar, Carl Vondrick, Antonio Torralba
  - Massachusetts Institute of Technology
- Submitted on 3rd, Jun, 2017

Abstract

- A deep discriminative representations shared across three major natural modalities:
  - Vision
  - Sound
  - Language
- Useful for several tasks, such as:
  - Cross-modal retrieval
  - Transferring classifiers between modalities
- Only trained with
  - Image + Text pairs
  - Image + Sound pairs

Introduction

- Invariant representations are core for vision, audio, and language models because they abstract our data.
  - Viewpoint and scale invariance in vision
  - Reverberation and background noise invariance in audio
  - Synonym and grammar invariance in language

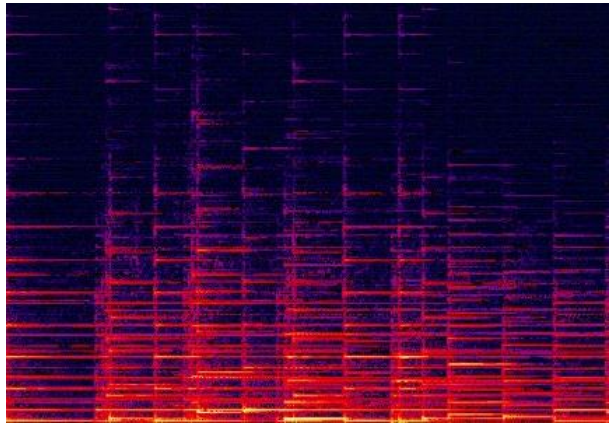
- The goal of this paper is to create representations that are robust in another way:
  - We learn representations that are aligned across modality.

Introduction

Consider the sentence *"she jumped into the pool."*

Same Concept could also appear visually or aurally.

Image of pool

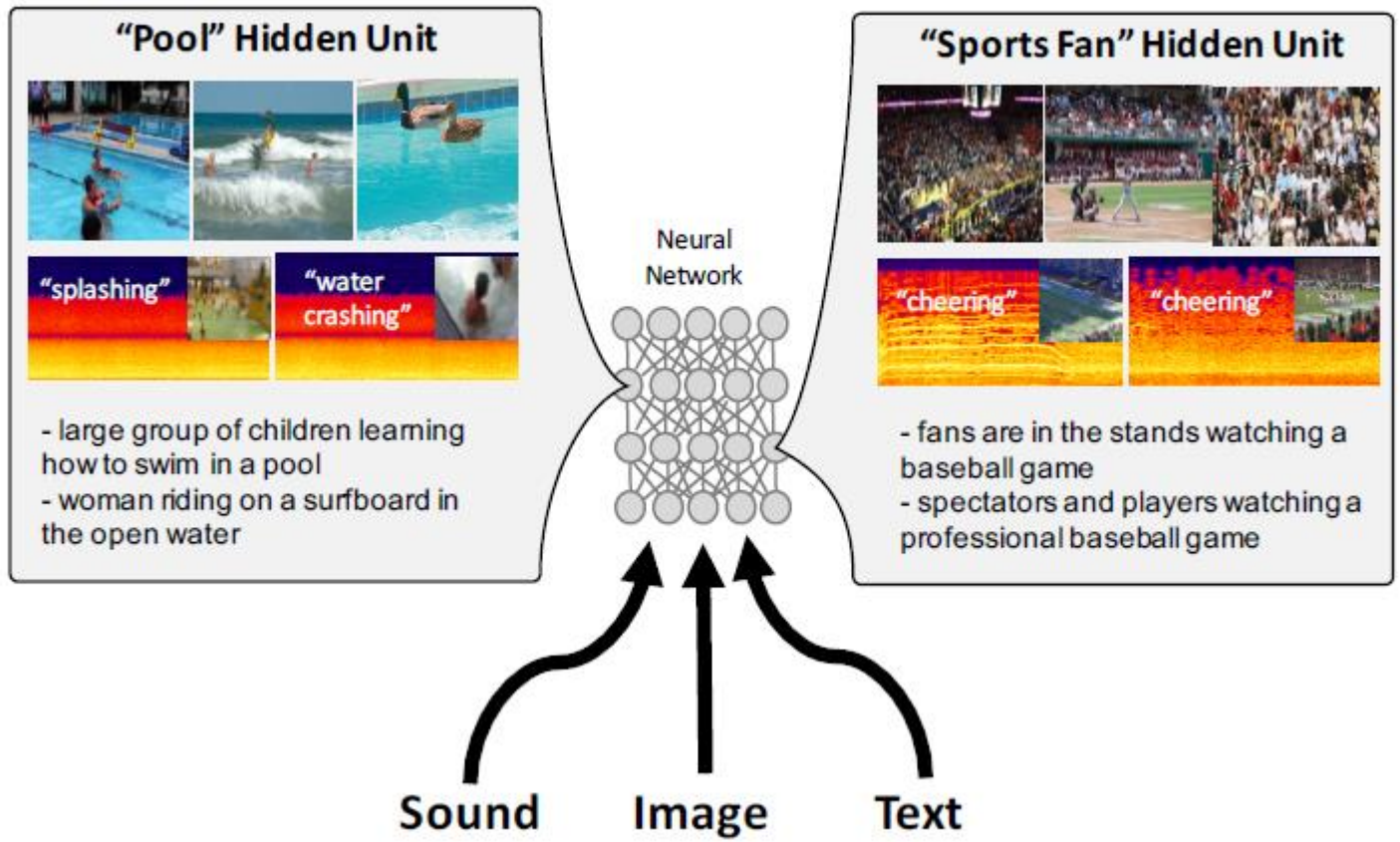


Sound of splashing

The Pool image, the splashing sound, and the above sentence should have similar representations.

= Aligned cross-modal representations.

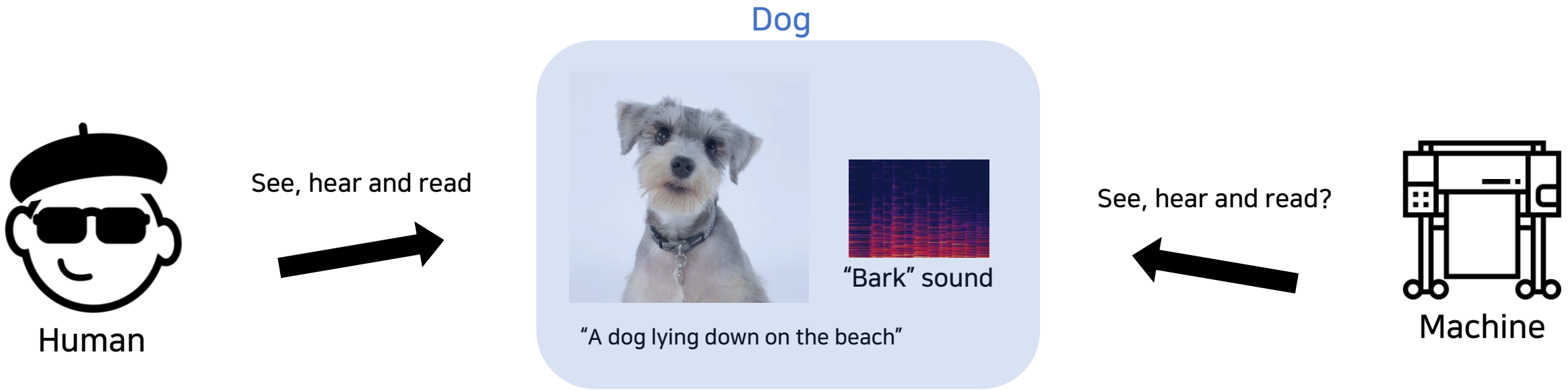
Aligned Representations



Representation aligned across three senses: seeing, hearing, and reading.

Introduction

- Aligned cross-modal representations are fundamental components for machine perception to understand relationships between modalities.
- Many practical applications in recognition and graphics, such as
  - Transferring learned knowledge between modalities.
- Primary contribution
  - Showing how to leverage massive amounts of synchronized data to learn a deep, aligned cross-modal representation.



Related Work

- Canonical-Correlation Analysis (CCA)
  - A way of inferring information from cross-covariance matrices.
- RBM auto-encoders between vision and sound
- Passive-aggressive model for content-based audio retrieval from text queries
- Pioneering work explore image-captioning as a retrieval task

Related Work

- “Cross-modal correlation learning for clustering on image-audio dataset.”
  - Applied CCA between visual and auditory features, and used common subspace features for aiding clustering in image-audio datasets.
- “Multimodal deep learning.”
  - Investigates RBM auto-encoders between vision and sound.
- “Large-scale content-based audio retrieval from text queries.”
  - Applies a passive-aggressive model for content-based audio retrieval from text queries.
- “Every picture tells a story: Generating sentences from images.”
  - Pioneering work explore image-captioning as a retrieval task.

DataSet

- Sound
  - Flickr: Over 750,000 videos from Flickr which provides over a year (377 days) of continuous audio.
  - Extract the spectrogram from the video files and subtract the mean
- Language
  - COCO: 400,000 sentences and 80,000 images
  - Visual Genome: 4,200,000 descriptions and 100,000 images
  - Preprocess the sentences and embedding each word with word2vec
- Image
  - Frames from above sound dataset
  - Images from above language dataset
- Synchronization
  - Pairs of images and sound(from videos)
  - Pairs of images and text(from caption datasets)

Implementation

- Cross-modal retrieval
  - Alignment by Model Transfer
  - Alignment by Ranking
- Transferring classifier
  - Sound and Text Transfer
  - Zero Shot Classifier Transfer
- Visualization of our representation

Cross-Modal Networks

- Let x_i be a sample from modality x, and y_i be the corresponding sample from modality y.
- $f_x(x_i)$ to be the representation in modality x, $f_y(y_i)$ to be the representation in modality y
- Can accept as input either an image, a sound, or a sentence, and produces a common representation shared across modalities.
- Desire the representation to be both aligned and discriminative.

Two Approaches

- Alignment by Model Transfer
  - Student-teacher model
  - KL-divergence
- Alignment by Ranking
  - Ranking loss function

Alignment by Model Transfer

$$\sum_{i}^{N} D_{\mathrm{KL}}\left(g(x_i)||f_y(y_i)\right), \text{ where } D_{\mathrm{KL}(P||Q)} = \sum_j P_j \log \frac{P_j}{Q_j}$$

- Take advantage of discriminative visual models to teach a student model to have an alignment.
- Teacher model could be any image classification model, such as AlexNet.

Alignment by Ranking

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

- ψ is a similarity function
- Δ is a margin hyper-parameter
- j iterates over negative examples
- This loss seeks to push paired examples close together in representation space, and mismatched pairs further apart.
- We use cosine similarity in representation space.

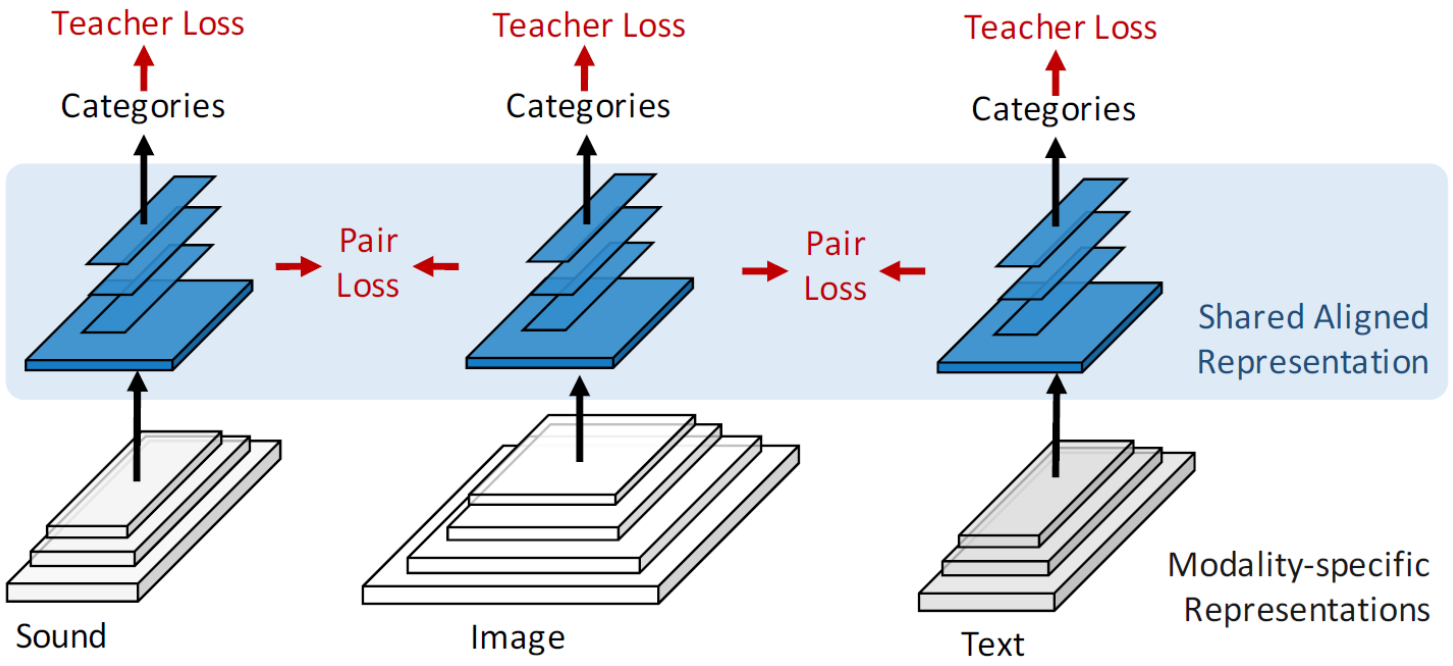
$$\psi(x, y) = \cos(f_x(x), f_y(y))$$

Learning

- Model transfer
  - Train student models for sound, vision, and text to predict class probabilities from a teacher ImageNet model.

- Ranking
  - Apply the ranking loss for alignment between
    - Vision to Text
    - Text to Vision
    - Vision to Sound
    - Sound to Vision
  - We do not have large amounts of sound/text pairs

Network Architecture



- a network that accepts as input either an image, a sound, or a text.
- Modality-specific layers are convolutional, and the shared layers are fully connected.

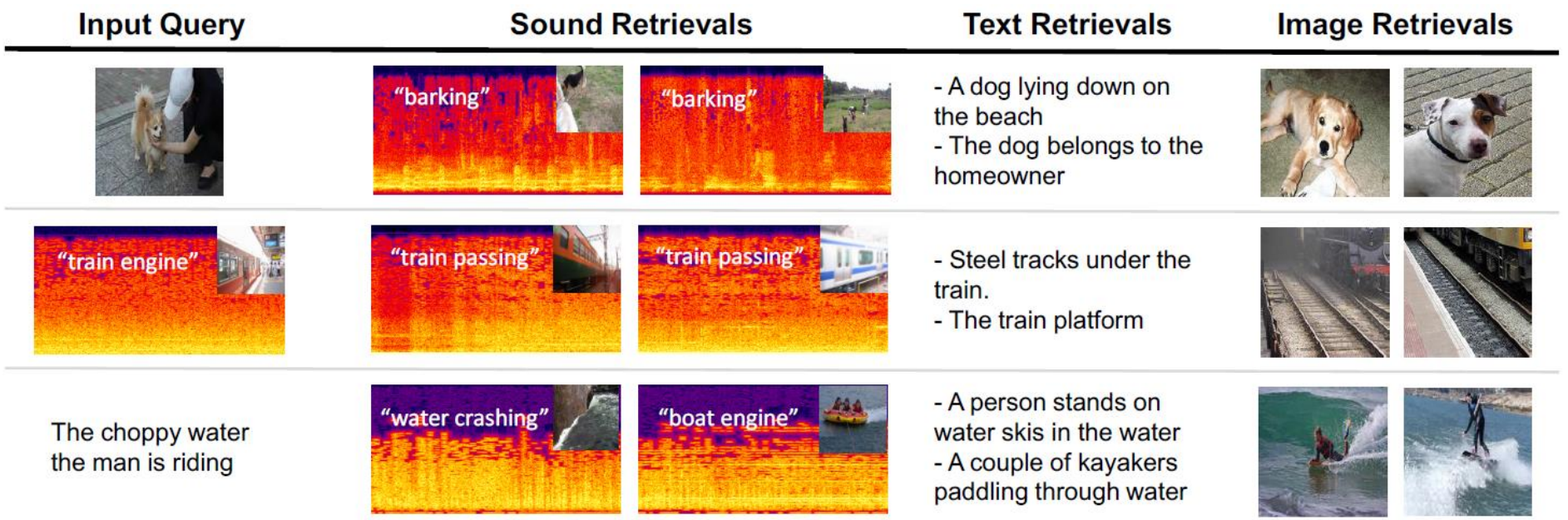
Network Architecture

- Sound Network
  - Input: spectrograms
  - Four-layer one-dimensional convolutional network
- Text Network
  - Input: sentences (where each word is embedded into a word representation using word2vec)
  - Four-layer network
  - Convolutions with fixed kernel sizes
- Vision Network
  - Standard Krizhevsky architecture (AlexNet)
- Shared Network
  - Input: fixed length vectors with the same dimensionality(output from sound, text, vision networks)
  - The weights in the upper layers are shared across all modalities.
  - Two fully connected layers of dimensionality 4096 with rectified linear activations
  - Output: 1000 dimensional with a softmax activation function

Experiments

- Setup
  - For both validation and testing
    - 5,000 video/sound pairs
    - 5,000 image/text pairs
  - Rest is used for training.
- Cross Modal Retrieval
- Sound and Text Transfer
- Zero Shot Classifier Transfer
  - Annotate held-out videos into 42 categories using Amazon Mechanical Turk.

Cross Modal Retrieval



Method	IMG ↓ SND	SND ↓ IMG	IMG ↓ TXT	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**

Lower is better.

Sound and Text Transfer

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**

Cross Modal Retrieval for Sound and Text

- Network was trained using only image/sound and image/text pairs.
- 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.

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
  - 42 categories consisting of objects and scenes
  - Used Amazon Mechanical Turk
  - Training set: 2,799 videos
  - Testing set: 1,050 videos
  - Annotated each video with a short text description

Zero Shot Classifier Transfer

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**

- This shows that the representation is both aligned and discriminative.
- The most challenging source modality for training is sound, which makes sense as vision and text are very rich modalities.
- Our approach still learns to align sound with vision and text.

Conclusion

- Invariant representations enable computer vision systems to operate in unconstrained, real-world environments.
- 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 empirically suggest it has learned an alignment between them, possibly by using images as a bridge internally.

References

- A New Approach to Cross-Modal Multimedia Retrieval
- Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books
- Convolutional Neural Networks for Sentence Classification
- Cross-modal Correlation Learning for Clustering on Image-Audio Dataset
- Distributed Representations of Words and Phrases and their Compositionality
- Every Picture Tells a Story: Generating Sentences from Images
- Large-Scale Content-Based Audio Retrieval from Text Queries
- Multimodal Deep Learning
- Objects that Sound
- SoundNet: Learning Sound Representations from Unlabeled Video
- Visually Indicated Sounds

Thank You !