# Language and Vision

Jim Glass / MIT 6.806-6.864 / Spring 2021

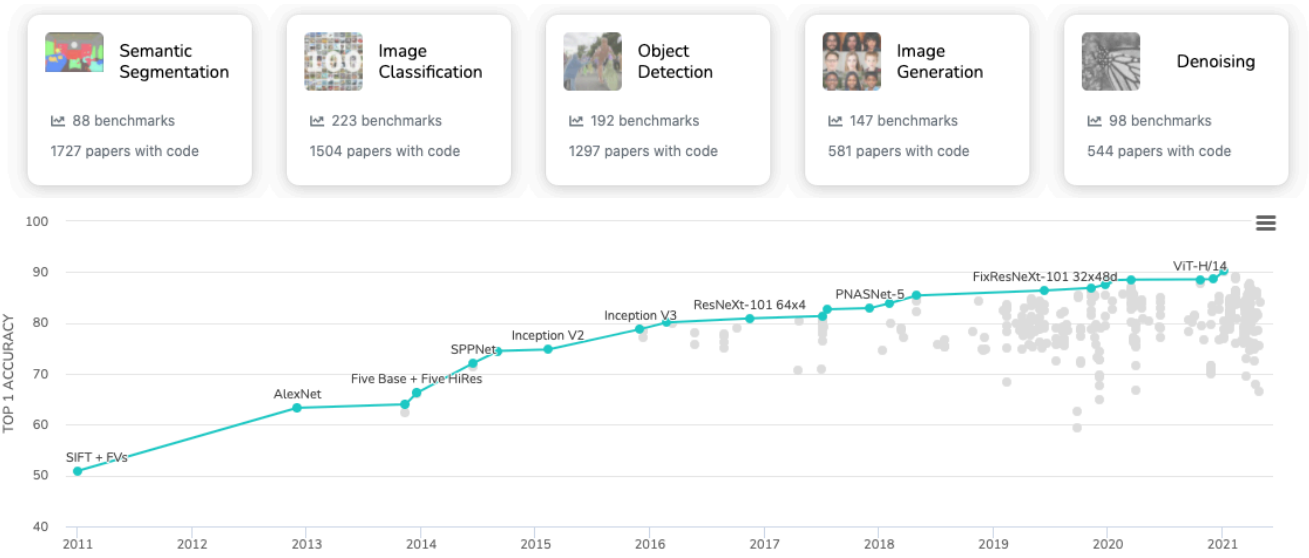
1

## Language and Vision

- NLP and vision
  - Image captioning
  - Visual question-answering
- Speech and vision

2

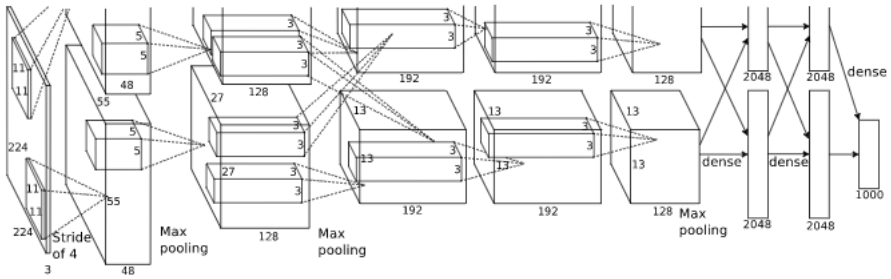
# Computer Vision: Image Classification

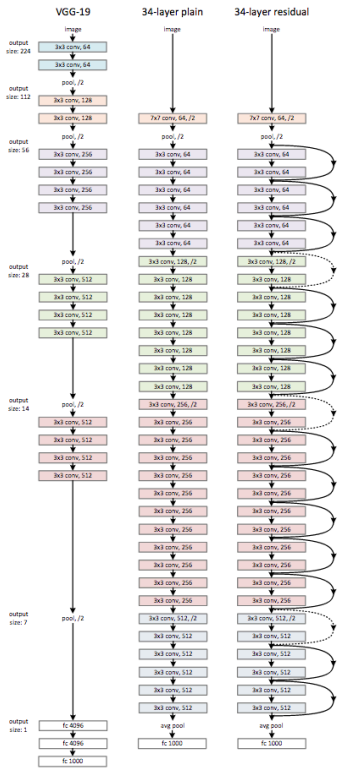



3

# Image Classification

- A wide variety of deep learning architectures have been explored for computer vision tasks
  - Typically include many layers of convolutional neural networks (CNNs) with gradual subsampling
  - For image classification a final label layer (e.g., 1K)



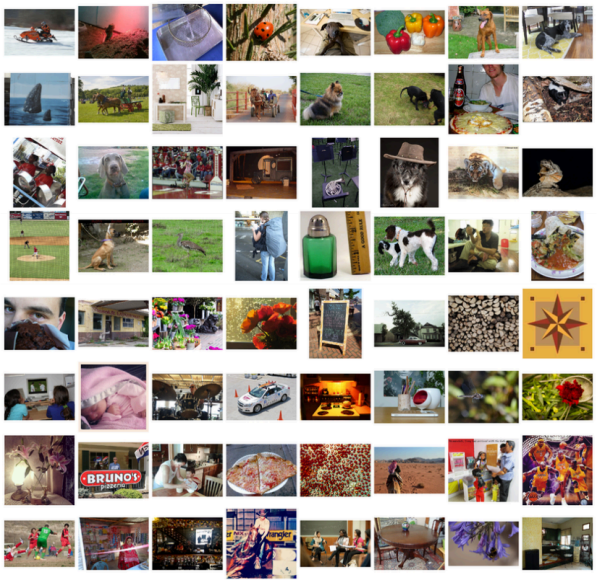

4

# ImageNet Object Classification Dataset

**ImageNet Large Scale Visual Recognition Challenge**

Olga Russakovsky* · Jia Deng* · Hao Su · Jonathan Krause · Sanjeev Satheesh · Sean Ma · Zhiheng Huang · Andrej Karpathy · Aditya Khosla · Michael Bernstein · Alexander C. Berg · Li Fei-Fei

- ImageNet annotated image corpus
  - Basis for visual recognition challenge
  - ~14M images covering ~1k objects
  - Annotations crowdsourced via AMT
  - Includes image classification and multi-label and segmentation tasks
- Many other image datasets, e.g.,
  - COCO, LabelMe



5

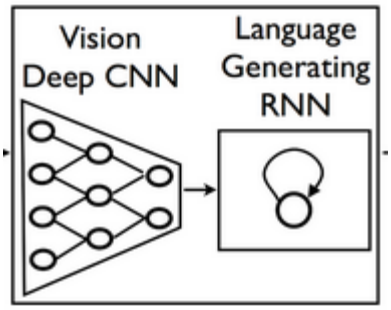
# Why Language and Vision?

- Language and perception and fundamental human capabilities
  - Pictures are everywhere
  - Words are how humans communicate
  - Vision and language combinations create new opportunities
- Many potential applications, e.g.,:
  - Multi-modal human-computer interaction
  - Interact with, organize, navigate and summarize visual data
- Measuring joint (vision + NLP) AI technology capabilities
  - Image captioning
  - Visual question answering, visual dialogue etc.
- Recent progress in deep learning enabled cross-modal research

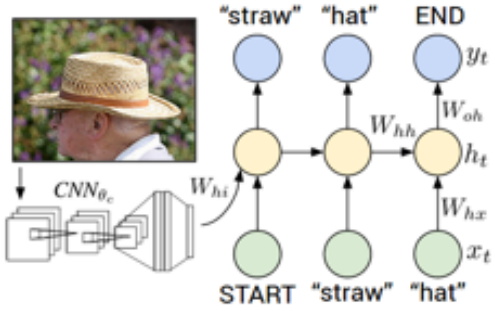
8

# Image Captioning

- The use of deep learning embeddings for visual and linguistic information enabled the combination of the two modalities
- In a cross-modal formulation of an encoder-decoder model:
  - A deep CNN-based image model would generate an embedding
  - An RNN-based language model would use it to generate a caption




Vinyals et al., 2015




Karpathy et al., 2015

9

# Image Captioning Datasets

- A number of image datasets were captioned via crowdsourcing
  - Typically multiple (e.g., 5) captions were generated for each image
  - E.g., Flickr8k, Flickr30k, COCO, …

"A cow is standing in the field"

"This is a close up of a brown cow"

"There is a brown cow with long hair and two horns"

"There are two trees and a cloud in the background of a field with a large cow in it"

10

Man in blue wetsuit is surfing on wave

Karpathy et al., 2015

Vinyals et al., 2015

A group of young people playing a game of Frisbee

A car is parked in the middle of nowhere

Kiros et al., 2015

Fang et al., 2015

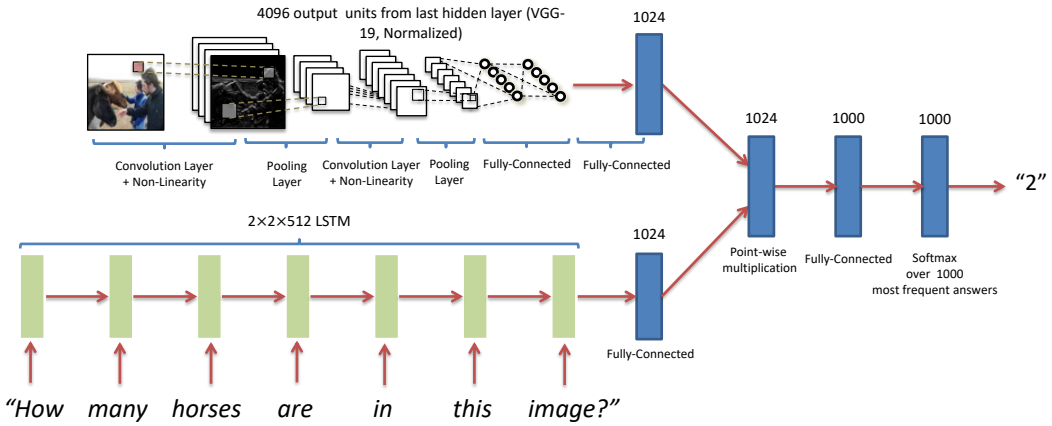
A pot of broccoli on a stove

[Devi Parikh]

11

# Visual Question-Answering

- Visual question-answering (VQA) an alternative to passive image captioning that allows for interaction, fine-grained understanding
  - Combines visual input and text question as context to generate an answer




[Antol et al., 2015]

12

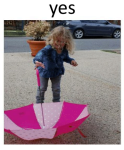
# VQA Dataset Examples

- VQA contains crowdsourced questions about images
  - ~265k images from COCO
  - ~5 questions/image
  - ~10 answers/question

Who is wearing glasses?
man woman


Where is the child sitting?
fridge arms

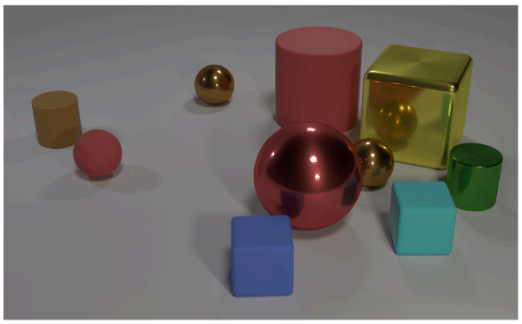

Is the umbrella upside down?
yes no




How many children are in the bed?
2 1


https://visualqa.org/

- CLEVR tests compositional language and visual reasoning
  - ~100k synthetic images
  - ~10 questions/image



**Q:** Are there an equal number of large things and metal spheres?
**Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
**Q:** How many objects are either small cylinders or metal things?

13

# Speech and Vision

- Visually-grounded speech recognition
  - Use image context as input to neural language model
- Learning spoken language through vision
  - Weakly supervised acquisition of spoken language

14

## A Fundamental Challenge for AI

- Many successful machine learning tasks rely on large quantities of *annotated* training data
  - Annotated data comes in {Input, Output} pairs

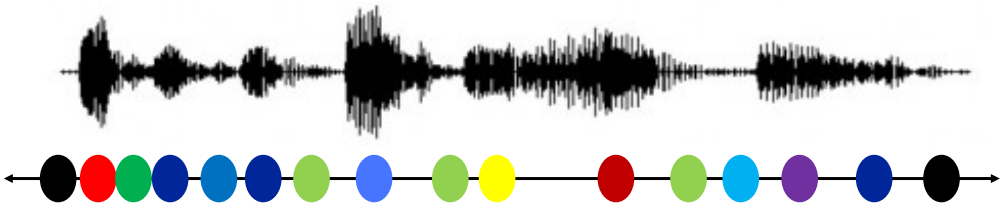
**Input** ⇨ **Model** ⇦ **Output**

- Issues:
  - Training data should match the "testing" conditions
  - Annotating large corpora is time-consuming and expensive
- Challenge:
  - There is far more raw data in the world than annotated data
  - Can we build models that learn with much less supervision?

15

## The Automatic Speech Recognition Learning Paradigm

- The training paradigm for speech recognition is >40 years old
  - {Speech, words} pairs enable alignment at phone/character level
  - Training becomes an exercise in aligning "beads on a string"

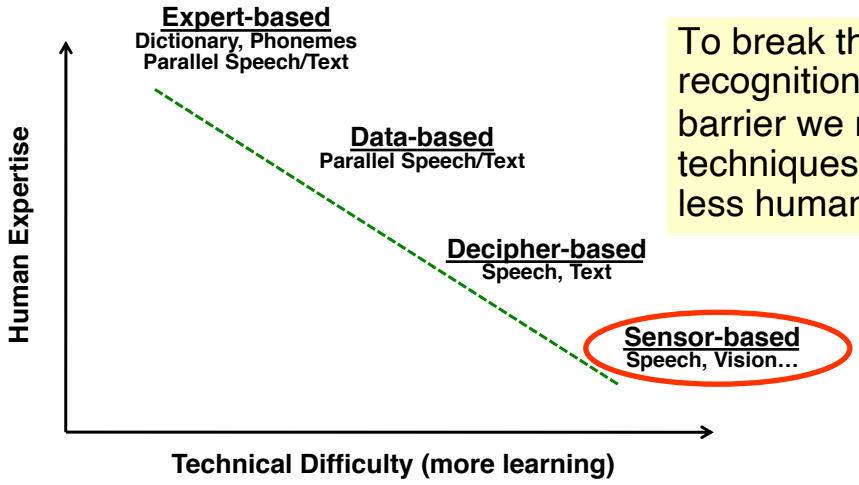


  - *This is not how humans learn speech!*
- Cost of annotations limits ASR to major languages of the world
- An ability to learn 1) with weakly constrained inputs from 2) freely available data, will be a major paradigm shift for low-resource speech tasks

16

# The State of Spoken Language Processing

Most (~98%) of the worlds languages have not been addressed by resource and expert intensive supervised speech recognition training methods




To break the speech recognition language barrier we need to develop techniques that require less human annotations!

17

# Learning Spoken Language through Vision

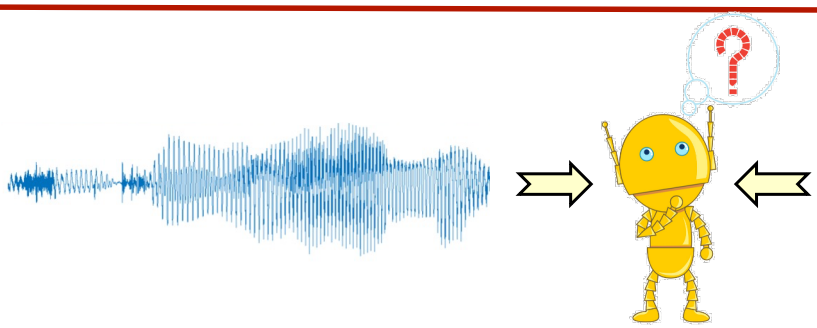



**David Harwath** **Wei-Ning Hsu** **Adrià Recasens** **Antonio Torralba**

18


19

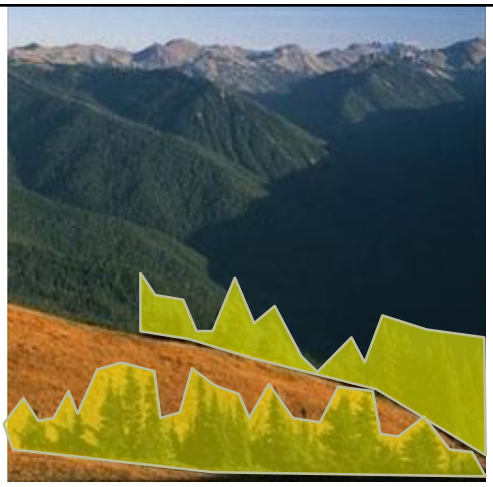
## A New Learning Paradigm

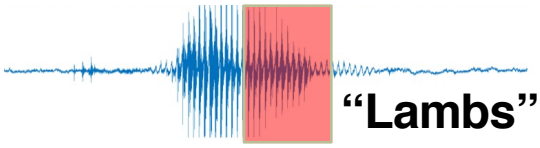


If you can associate the speech waveform that you hear with the things that you see, then you must have implicitly learned to:

1) Recognize words in speech
2) Recognize visual objects
3) Ground words to objects

20





“Lambs”

“Trees”

~~Two white lambs in a meadow.~~

~~A mountain chain with forests and trees.~~

22

# Crowdsourcing Spoken Image Captions

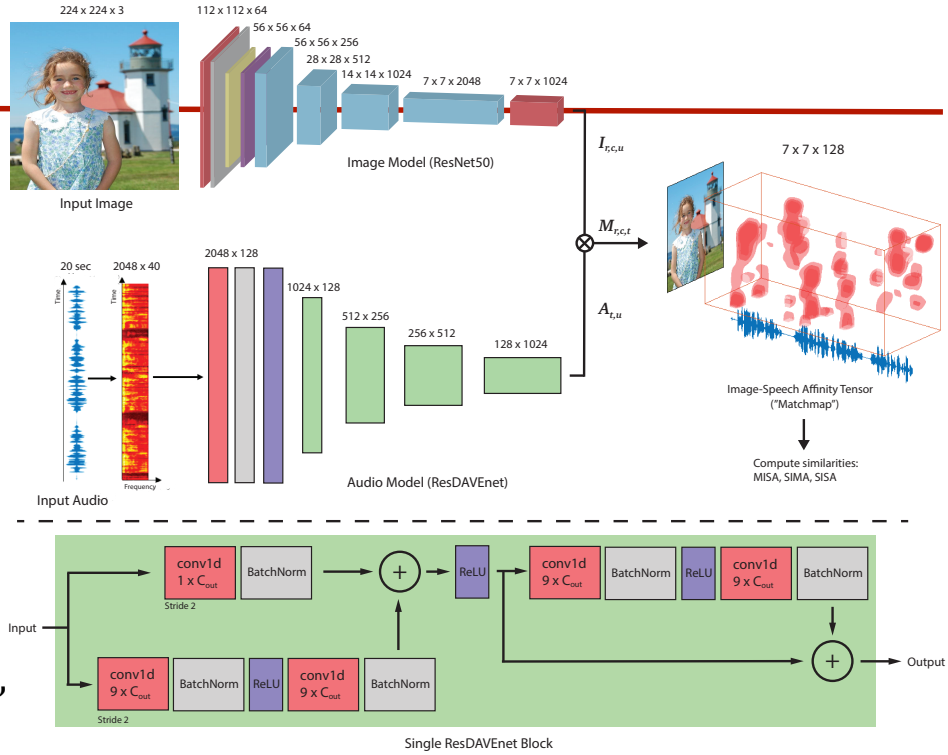

23

# ResDAVEnet

- Image branch based on ResNet50
- Audio branch uses a novel architecture based on 1-D convolutions with residual connections
- Matchmap similarity

$$M_{r,c,t} = sim(I_{r,c,:}, A_{t,:})$$

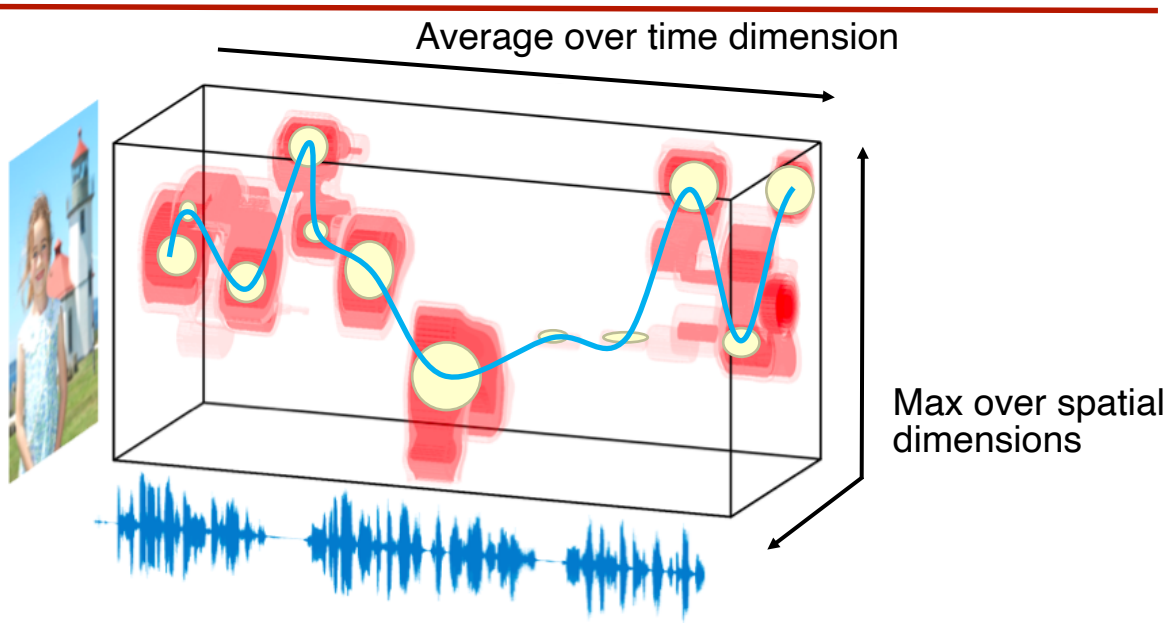
$sim(x, y) =$ dot product, cosine similarity, etc.




[Harwath et al., IJCV 2019]

25

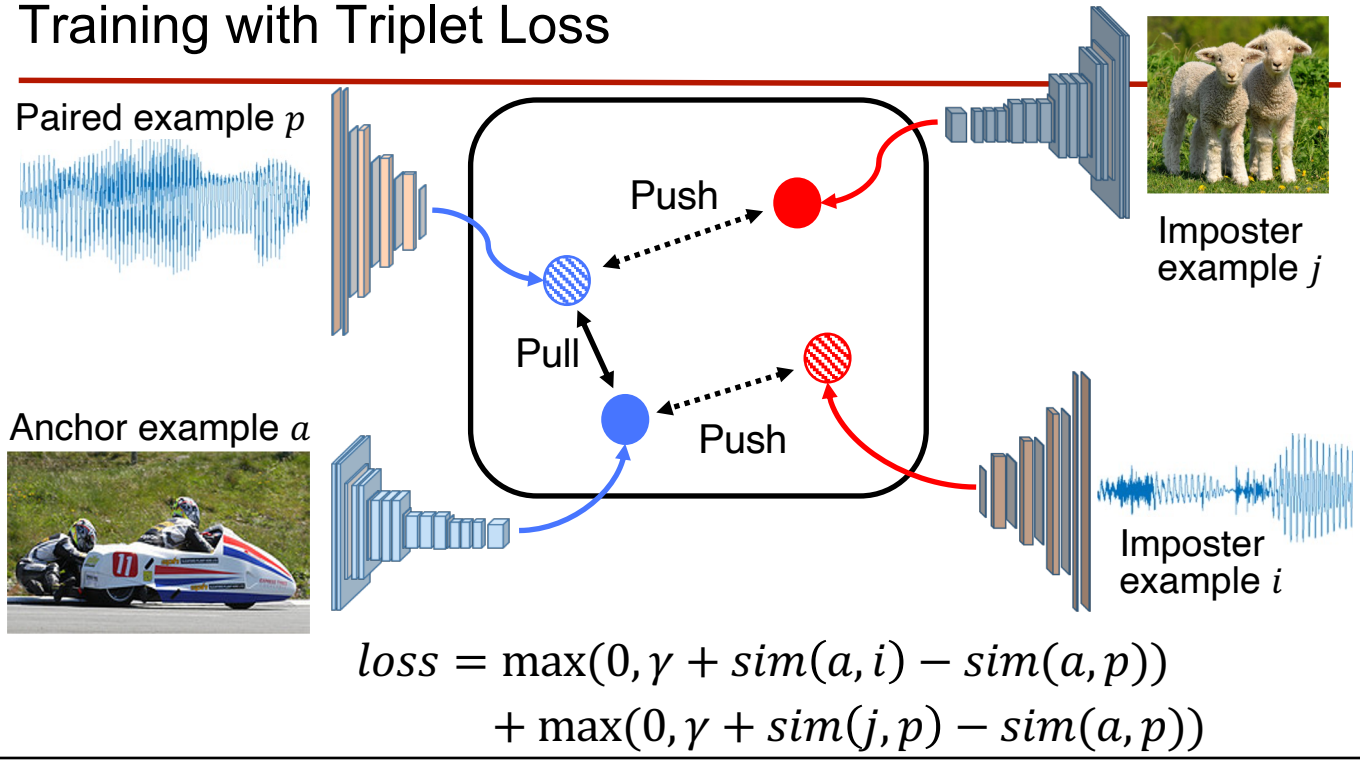
# Computing Image-Caption Similarity




(Average pooling works well too but is slightly worse)

26

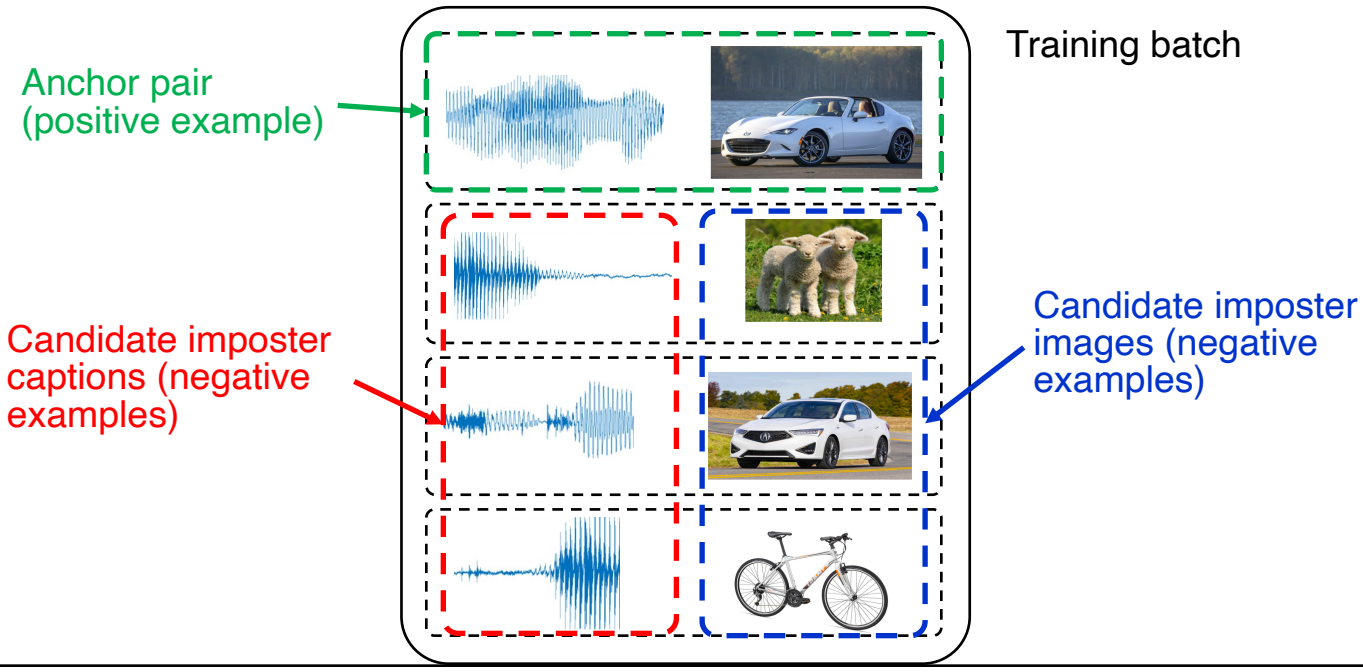
## Training with Triplet Loss




$$loss = \max(0, \gamma + sim(a, i) - sim(a, p)) + \max(0, \gamma + sim(j, p) - sim(a, p))$$

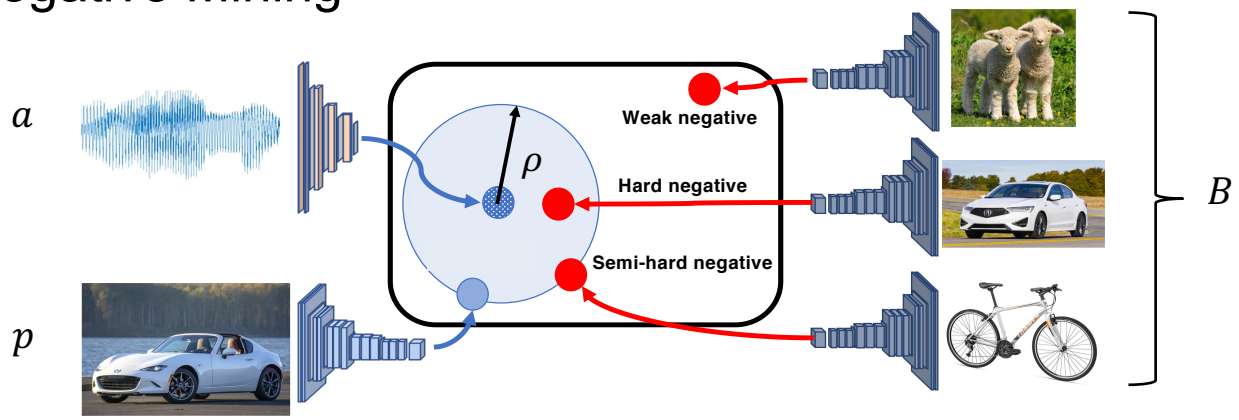
27

## Selecting Triplet Loss Examples




28

# Negative Mining




Recall the loss for a single positive pair: $\max(0, \gamma + sim(a,i) - sim(a,p))$

Given a batch of candidate imposter examples $B$:

- Hard negative mining:

$$i = argmax_{x \in B}(sim(a,x))$$

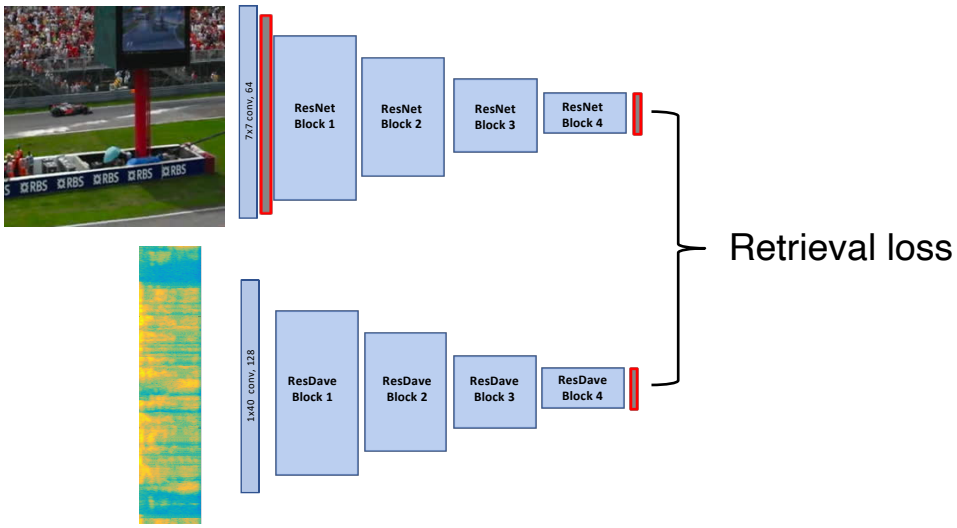
- Semi-hard negative mining

$i = argmax_{x \in B \setminus E}(sim(a,x))$ where $E = \{x \in B \mid sim(a,x) < sim(a,p)\}$

29

# Unsupervised Pre-Training

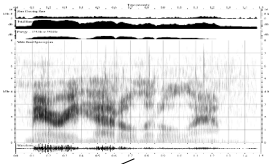
- Use videos containing natural sounds (waterfalls, racecars, etc.) from Flickr
- Pre-train the model from a random initialization, fine tune on speech captions

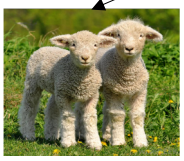



31

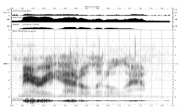
# Evaluation: Image and Caption Retrieval

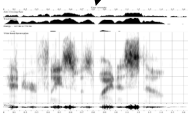
Image Retrieval:
Given caption, find image

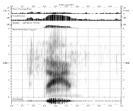




Caption Retrieval:
Given image, find caption



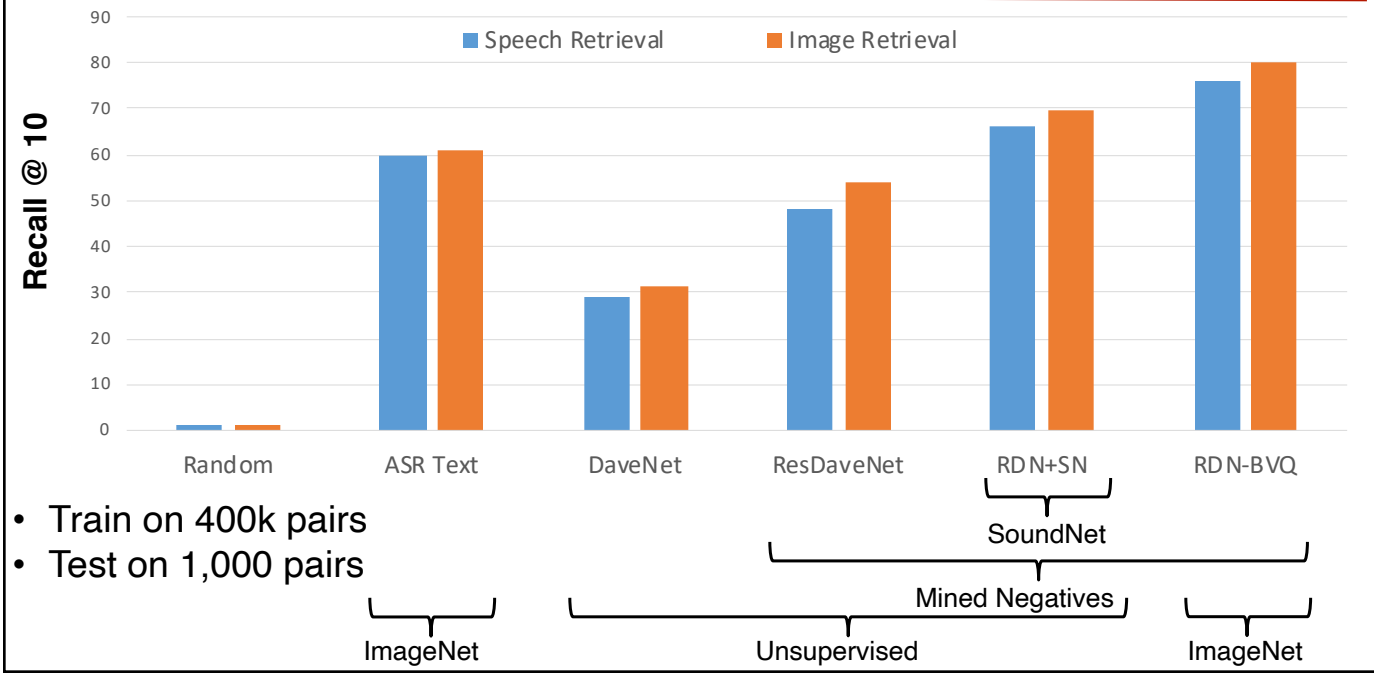




Evaluation metric: $P$(correct result is in top 10 retrieved examples)
(Recall @ 10)

32

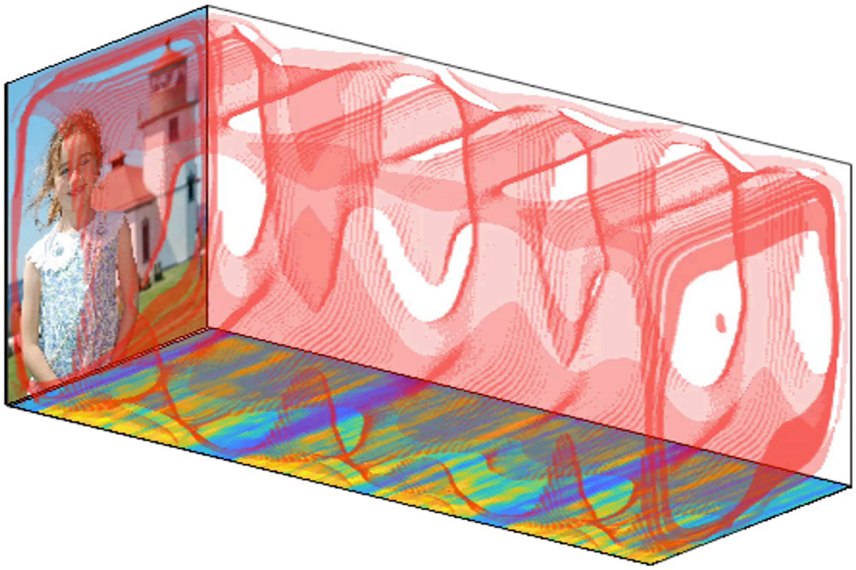
# Image and Caption Retrieval on MIT Places Corpus




- Train on 400k pairs
- Test on 1,000 pairs

33

# Matchmap Convergence Illustration



*D. Harwath et al., "Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input," Proc. ECCV 2018*

35

# Matchmap Visualization

36

## Matchmap Visualization

37

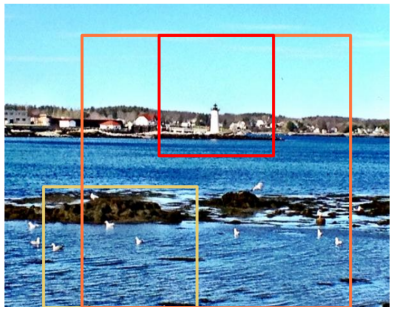
## Matchmap Visualization

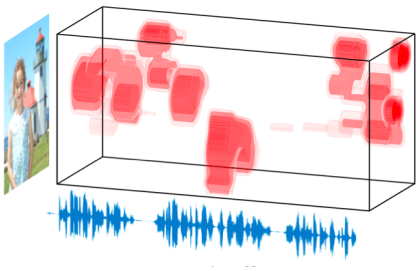
38

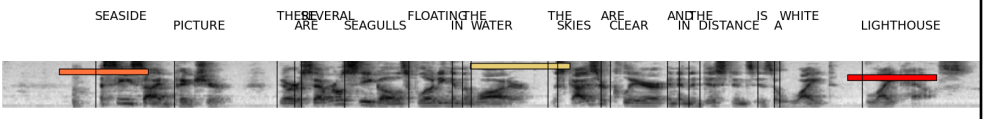
# Matchmap Visualization

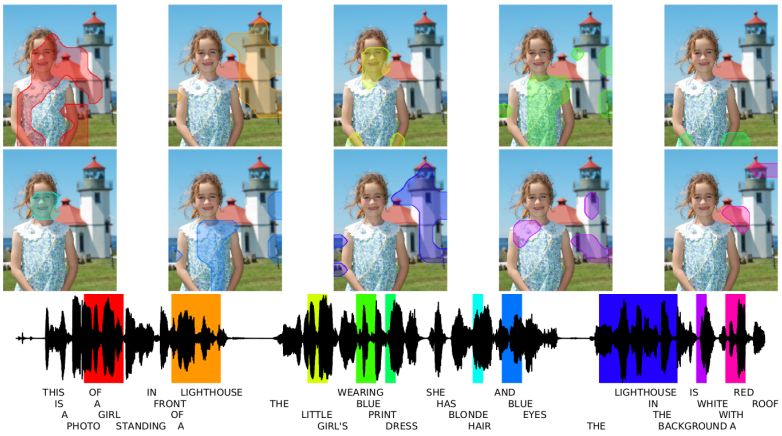

39

# Semantic Co-Segmentation



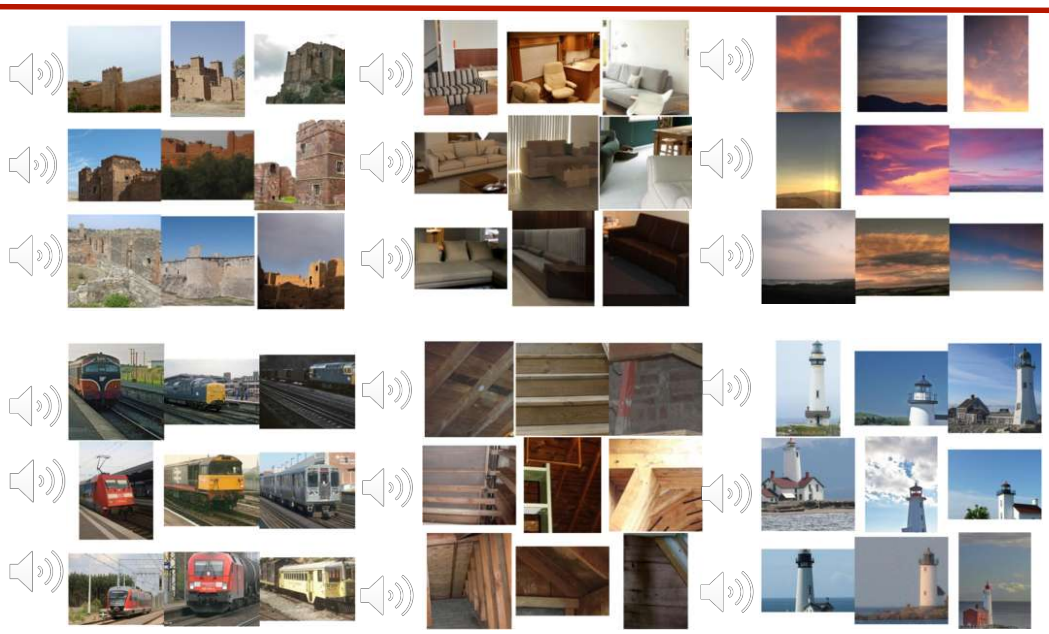




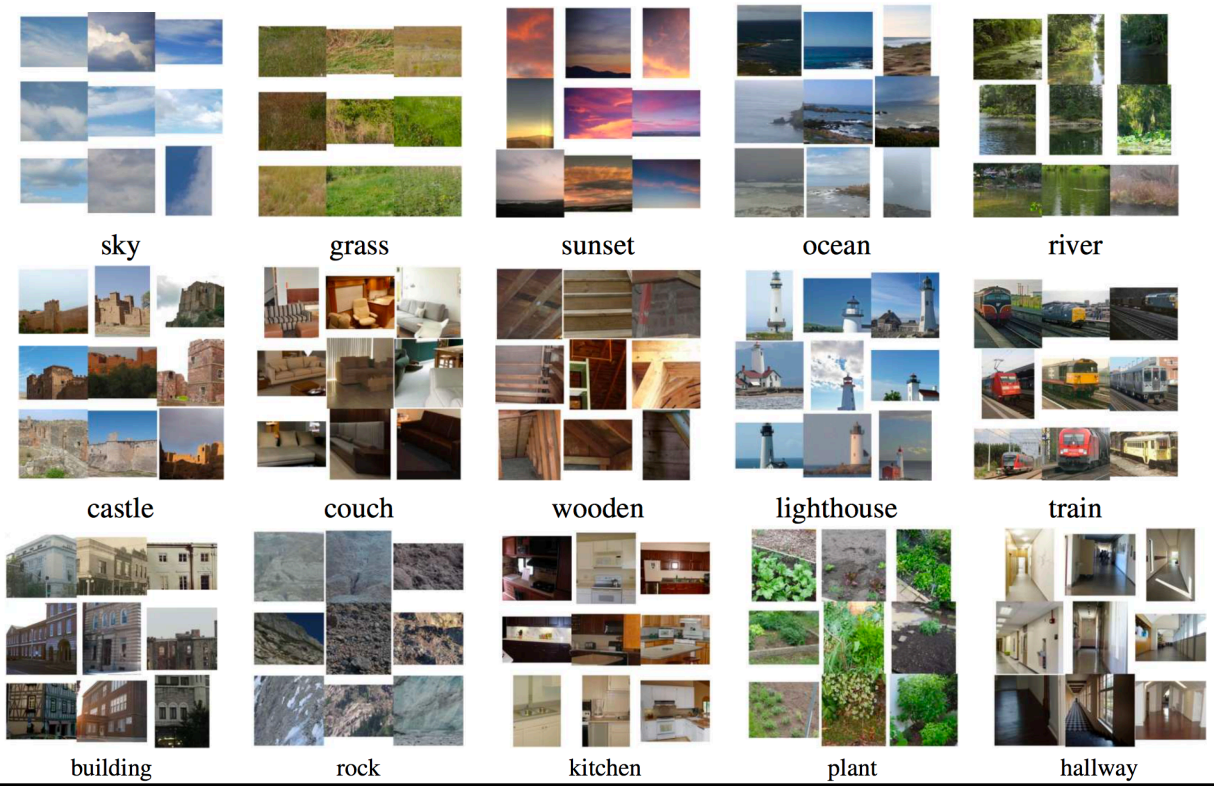




40

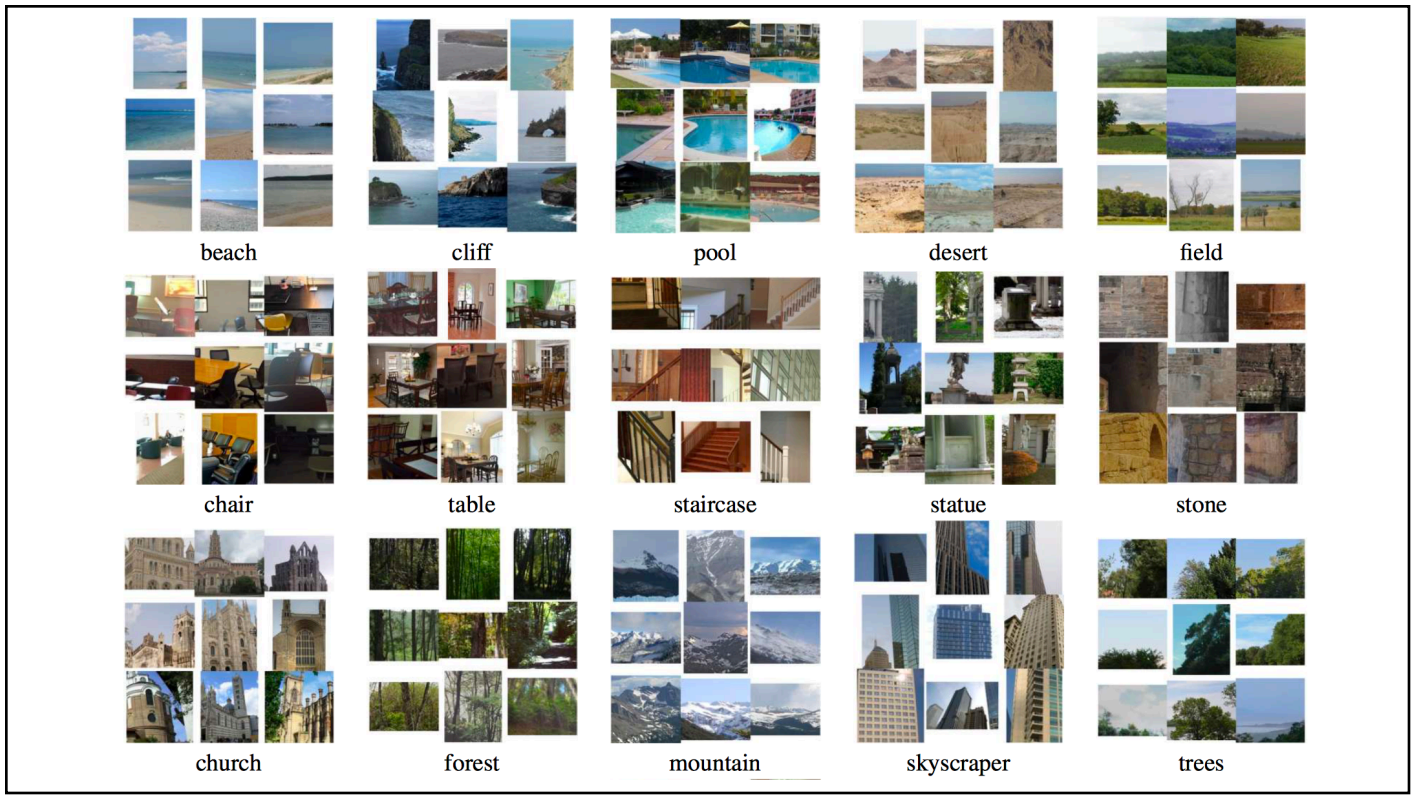
# Examples of audio-visual clusters



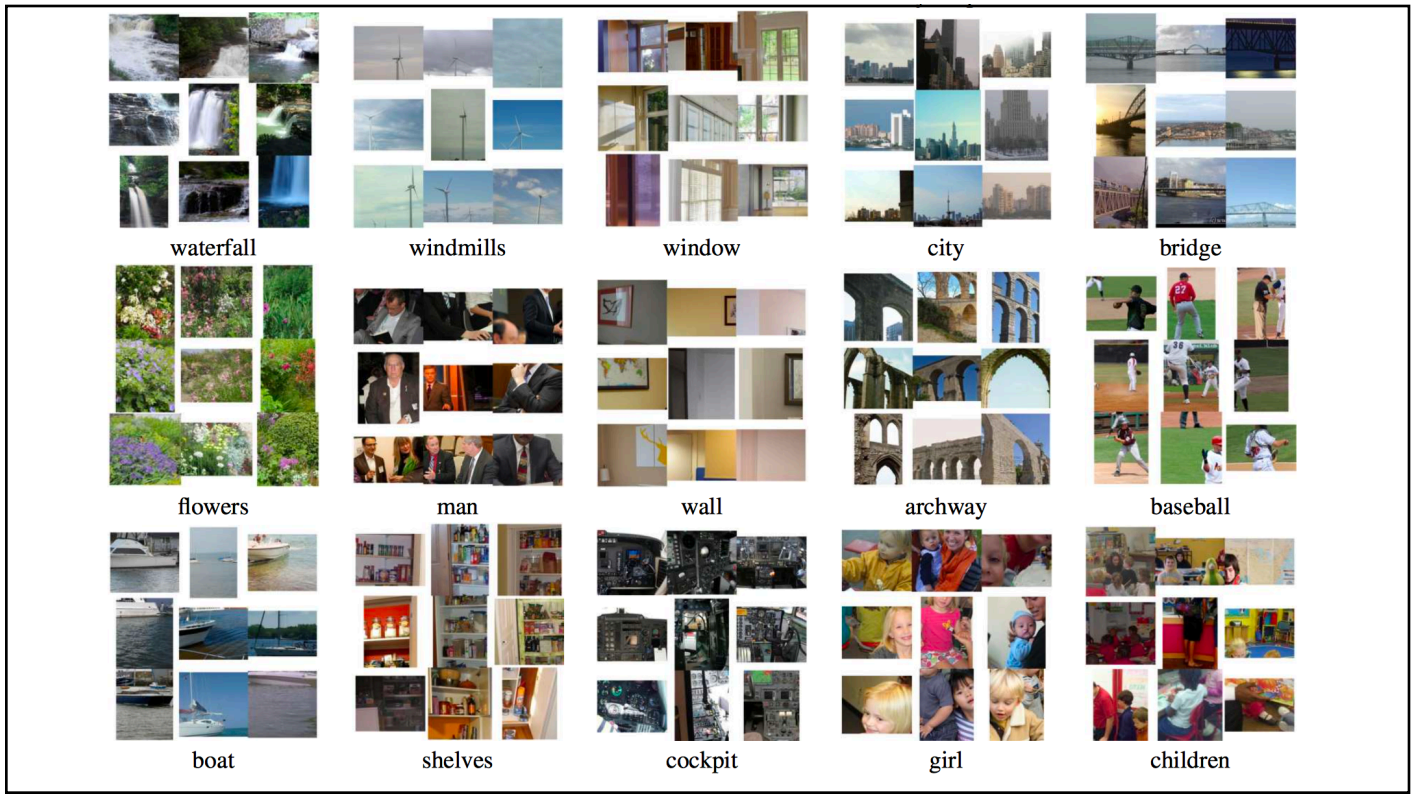
41




42

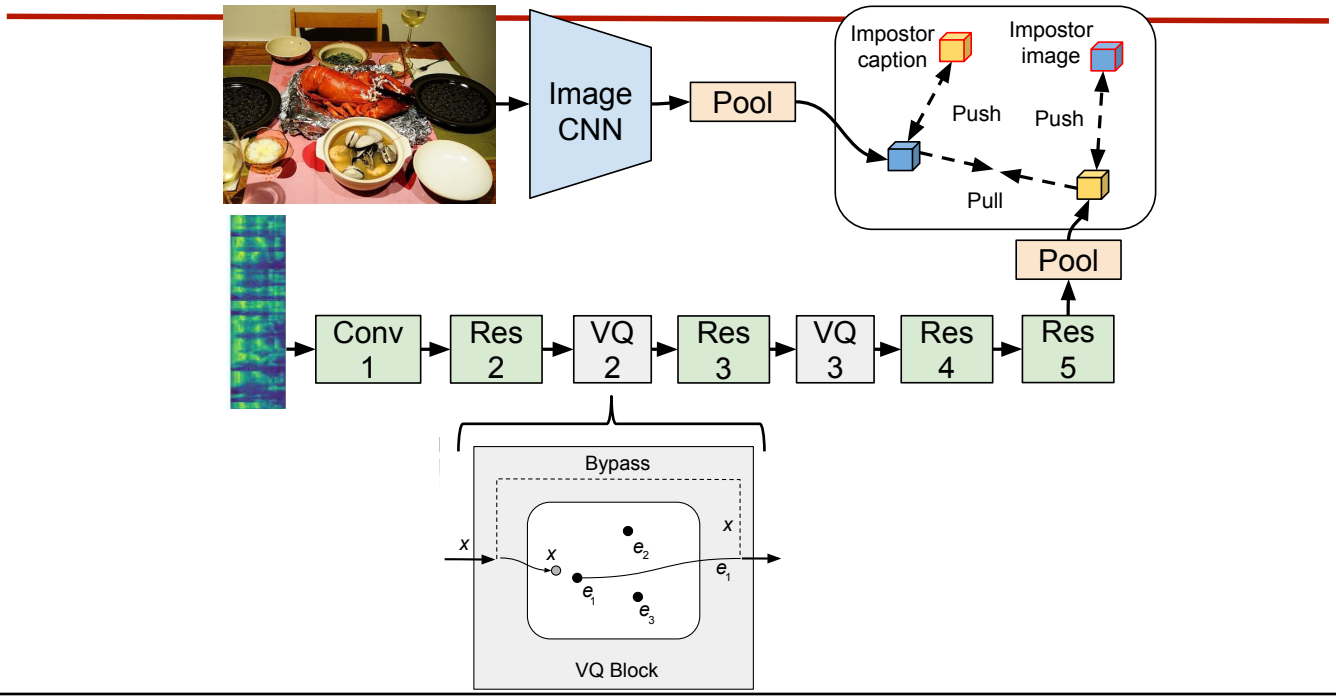

beach
cliff
pool
desert
field
chair
table
staircase
statue
stone
church
forest
mountain
skyscraper
trees

43


waterfall
windmills
window
city
bridge
flowers
man
wall
archway
baseball
boat
shelves
cockpit
girl
children

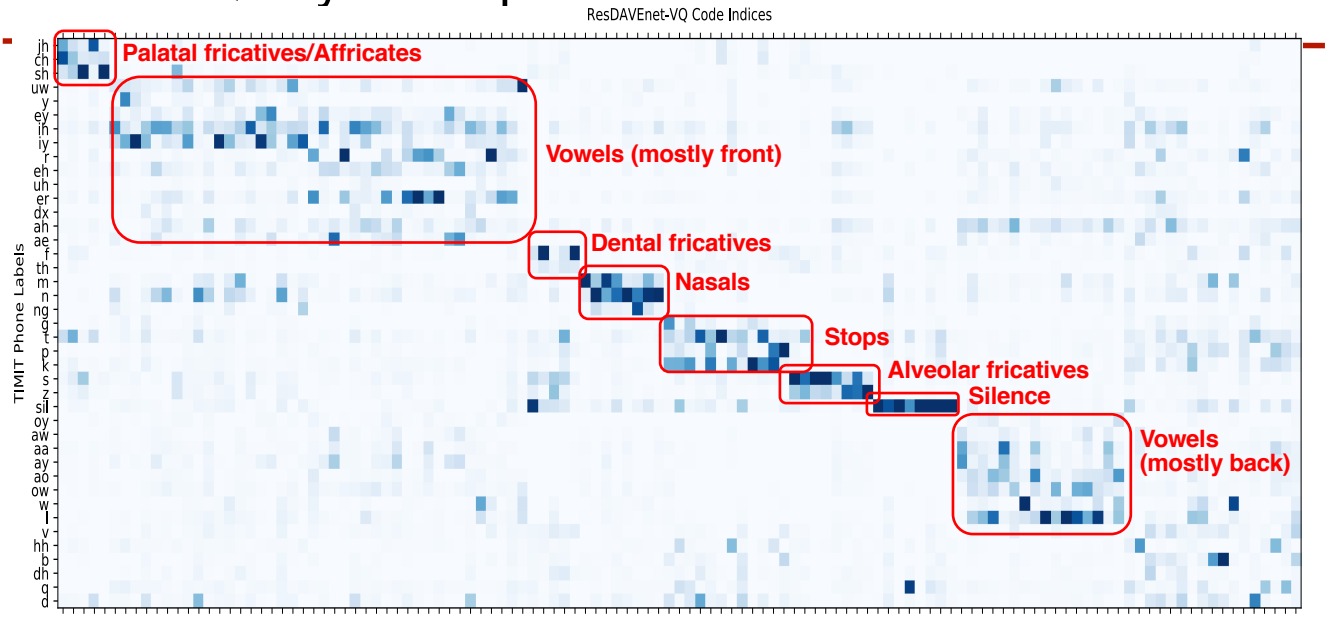
44

# Explicitly Modeling Discrete Structure




50

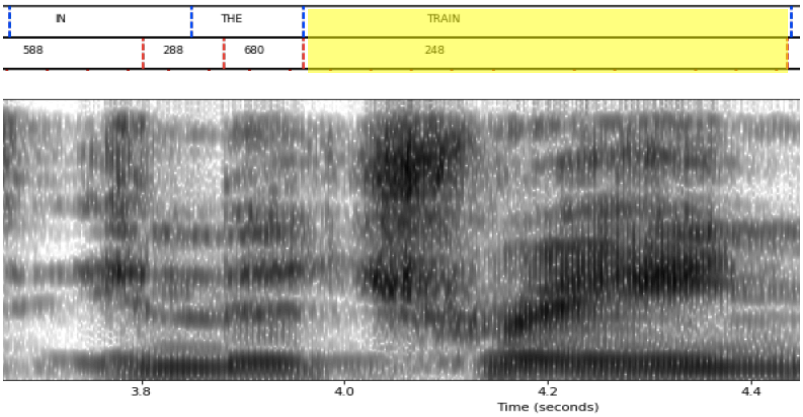
# Lower VQ Layers Capture Phones




Frame-level $P(phone|unit)$ computed across TIMIT
(Darker color is higher conditional probability)

54

# Evaluate Word-Like Unit Learning




$$Recall(word, code) = \frac{\#\textbf{co-occurrences of } (word, code)}{\#\text{occurrences } \textbf{of } word} \qquad Precision(word, code) = \frac{\#\textbf{co-occurrences of } (word, code)}{\#\text{occurrences } \textbf{of } code}$$

$$F1(word, code) = 2\frac{precision(word, code) \times recall(word, code)}{precision(word, code) + recall(word, code)}$$

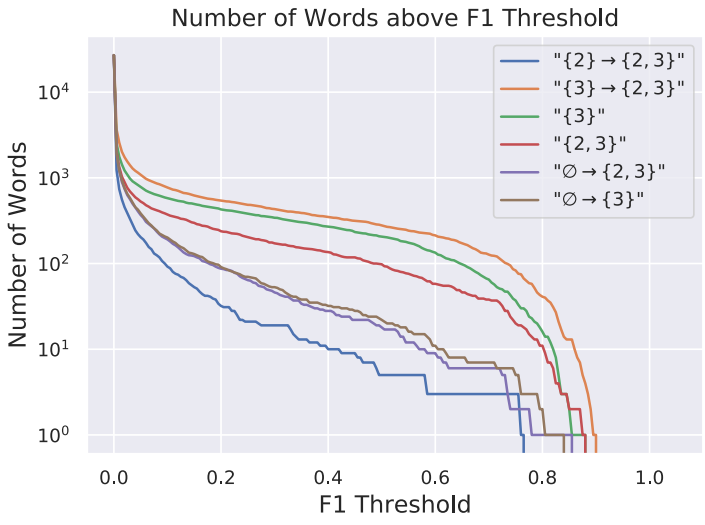
56

# Higher VQ Layers Capture Words

Finding "word detectors"

For each VQ code vector:

1. Compute $F1(word, code)$ for each word that appears in the test data
2. Assign a score to the code vector equal to $max_{w'} F1(w', code)$



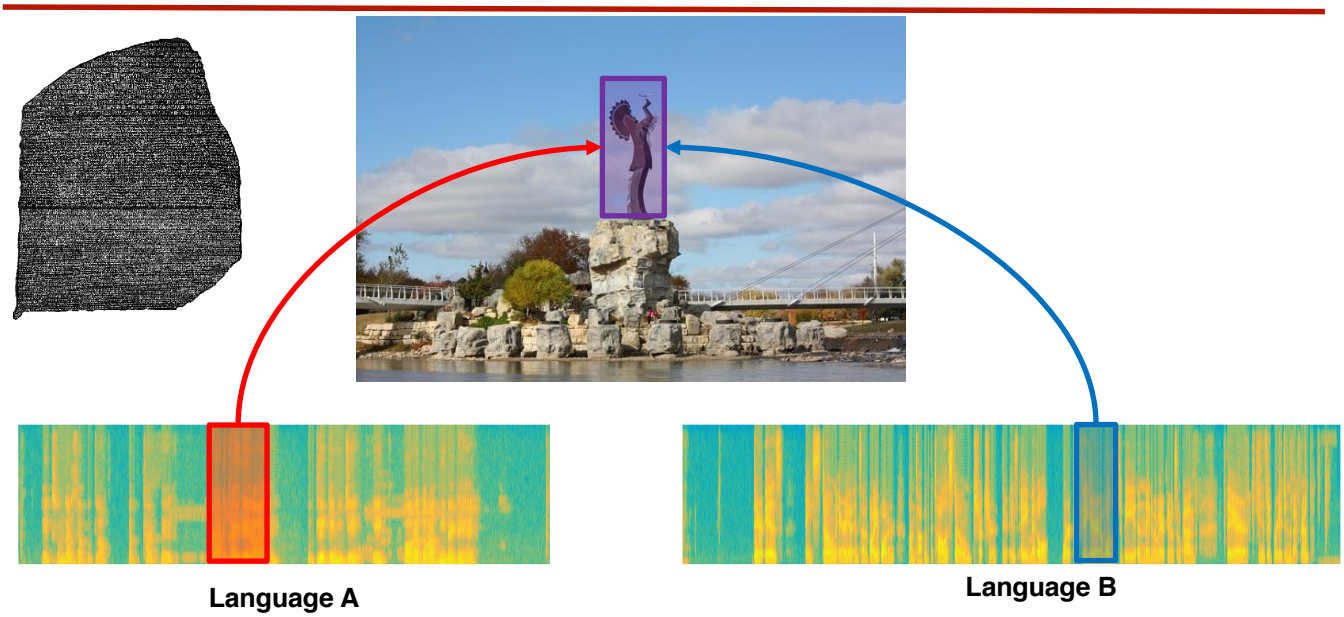

57

# Ranked Word Detectors

| rank | code | Top Hypotheses word | F1 | P | R | occ | Second Hypotheses word | F1 | P | R | occ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 58 | baseball | 90.09 | 89.45 | 90.75 | 3266 | player | 2.82 | 1.60 | 12.18 | 139 |
| 2 | 706 | background | 89.26 | 92.36 | 86.36 | 17740 | backgrounds | 0.71 | 0.36 | 68.75 | 66 |
| 3 | 88 | classroom | 88.78 | 87.39 | 90.21 | 1512 | class | 4.65 | 2.79 | 13.99 | 82 |
| 4 | 116 | construction | 88.16 | 90.19 | 86.22 | 2484 | constructed | 1.98 | 1.03 | 26.29 | 56 |
| 5 | 584 | playground | 87.84 | 85.68 | 90.12 | 930 | play | 2.70 | 1.81 | 5.33 | 51 |
| 6 | 596 | kitchen | 87.26 | 86.42 | 88.11 | 5313 | kitchenette | 1.68 | 0.85 | 86.44 | 51 |
| 7 | 48 | desert | 87.17 | 87.85 | 86.50 | 3319 | doesn't | 1.59 | 0.89 | 7.41 | 28 |
| 8 | 625 | background | 86.79 | 94.12 | 80.52 | 16541 | back | 1.45 | 0.95 | 3.01 | 225 |
| 9 | 557 | concrete | 86.61 | 91.17 | 82.49 | 1917 | country | 0.64 | 0.49 | 0.92 | 13 |
| 10 | 5 | airport | 86.46 | 89.04 | 84.02 | 962 | escalator | 2.29 | 1.25 | 13.86 | 23 |
| 11 | 534 | background | 86.09 | 80.24 | 92.86 | 19076 | back | 8.44 | 5.67 | 16.52 | 1233 |
| 12 | 274 | subway | 86.01 | 86.62 | 85.41 | 1264 | station | 1.82 | 1.61 | 2.09 | 99 |
| 13 | 44 | patio | 85.93 | 89.94 | 82.27 | 2056 | patios | 2.02 | 1.02 | 77.78 | 28 |
| 14 | 310 | rocky | 85.90 | 86.19 | 85.62 | 2375 | rock | 3.52 | 2.72 | 4.97 | 266 |
| 15 | 18 | driveway | 85.68 | 91.60 | 80.49 | 1159 | sidewalk | 1.58 | 1.13 | 2.63 | 47 |
| 16 | 560 | hospital | 84.35 | 90.16 | 79.24 | 1191 | hot | 0.92 | 0.62 | 1.81 | 17 |
| 17 | 598 | palm | 83.68 | 82.26 | 85.16 | 2071 | concrete | 1.67 | 1.28 | 2.41 | 56 |
| 18 | 769 | bamboo | 83.68 | 85.57 | 81.88 | 1265 | abandoned | 5.38 | 3.31 | 14.44 | 109 |
| 19 | 892 | walking | 83.56 | 85.02 | 82.15 | 7747 | walk | 7.97 | 4.51 | 34.35 | 519 |
| 20 | 124 | stage | 83.43 | 84.65 | 82.24 | 2103 | concert | 1.60 | 0.87 | 10.74 | 51 |
| 186 | 280 | station | 68.23 | 54.23 | 91.95 | 4366 | gas | 28.72 | 17.35 | 83.40 | 1773 |
| 187 | 829 | shirt | 68.21 | 70.83 | 65.78 | 6592 | shirts | 15.44 | 8.70 | 68.57 | 757 |
| 188 | 548 | night | 68.15 | 64.95 | 71.69 | 2277 | nighttime | 13.14 | 7.15 | 81.42 | 241 |
| 189 | 554 | computer | 68.14 | 71.66 | 64.95 | 1360 | computers | 22.99 | 13.78 | 69.21 | 254 |
| 190 | 2 | empty | 67.96 | 76.93 | 60.87 | 2805 | terminal | 4.13 | 2.17 | 42.21 | 176 |
| 191 | 993 | ruins | 67.93 | 59.53 | 79.08 | 586 | ruin | 21.58 | 12.90 | 66.05 | 142 |
| 192 | 820 | coffee | 67.88 | 63.95 | 72.32 | 1335 | cream | 7.04 | 4.32 | 19.09 | 357 |
| 193 | 164 | man | 67.80 | 73.79 | 62.71 | 16622 | men | 13.85 | 9.77 | 23.79 | 1390 |
| 194 | 461 | baby | 67.71 | 64.00 | 71.88 | 1204 | baby's | 24.10 | 13.92 | 89.89 | 249 |
| 195 | 803 | train | 67.62 | 85.63 | 55.87 | 4720 | trains | 10.06 | 5.52 | 56.98 | 306 |
| 196 | 446 | lake | 67.56 | 78.98 | 59.02 | 2294 | late | 2.85 | 1.58 | 14.25 | 53 |
| 197 | 225 | house | 67.40 | 58.83 | 78.89 | 10776 | houses | 18.45 | 10.44 | 78.82 | 1868 |
| 198 | 234 | alleyway | 67.26 | 56.62 | 82.81 | 944 | alley | 37.25 | 30.06 | 48.98 | 863 |
| 199 | 1000 | orange | 67.25 | 89.09 | 54.00 | 2987 | oranges | 3.30 | 1.71 | 48.84 | 63 |
| 200 | 842 | pink | 67.19 | 78.12 | 58.94 | 2723 | paint | 5.05 | 3.07 | 14.26 | 140 |

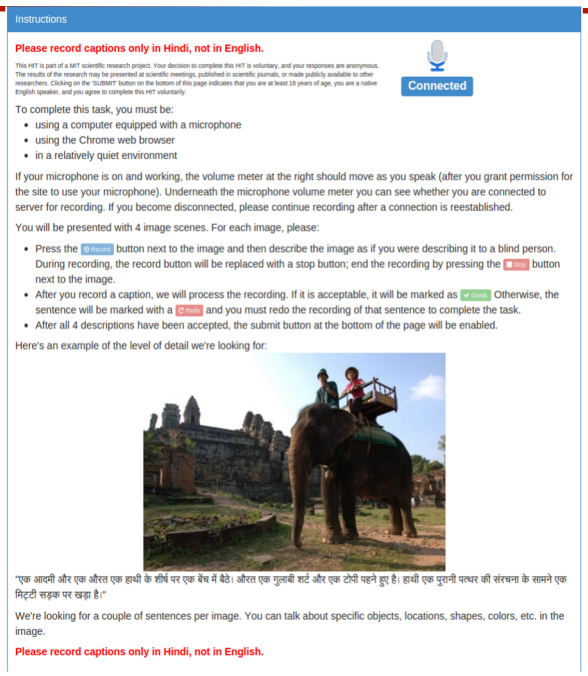
58

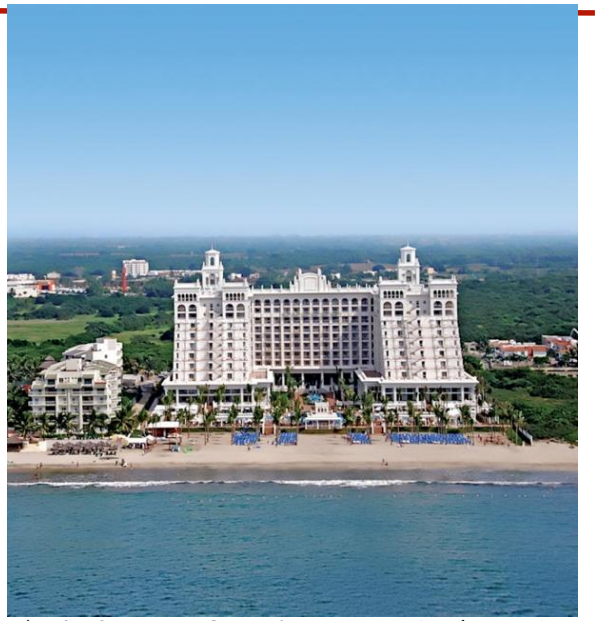
# Using Images as a "Rosetta Stone"




61

# Collecting Hindi Spoken Captions

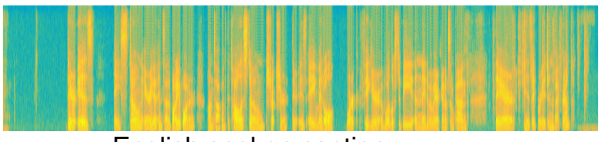


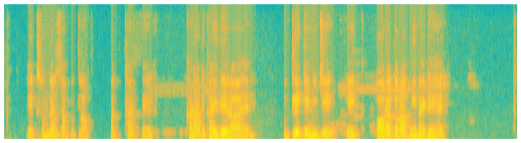
सफेद रंग की बढ़िया मार था जिसके पीछे हरे भरे पेड़ पौधे और कुछ छोटे लगाना है और सामने समुंदर का किनारा

62

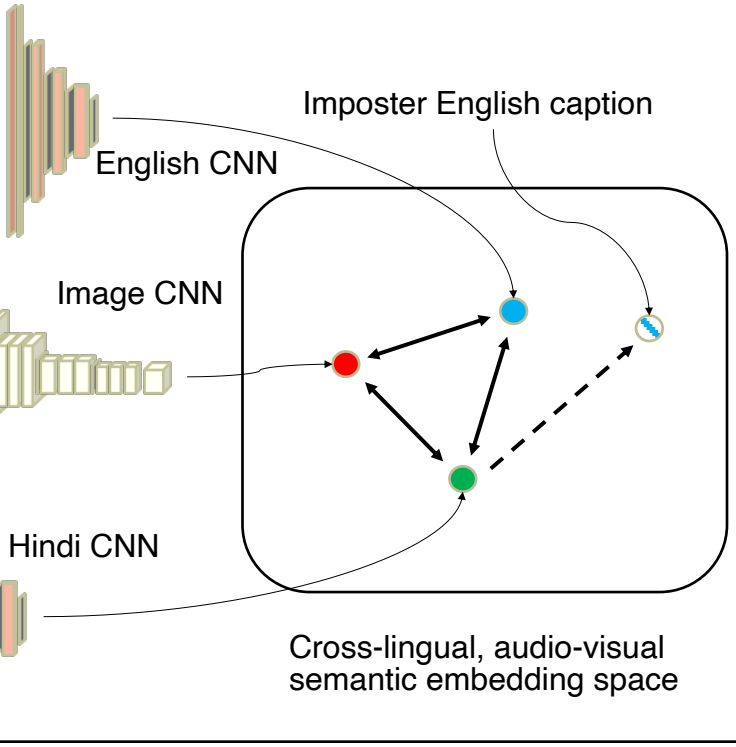
# Multi-Lingual Grounding



English spoken caption

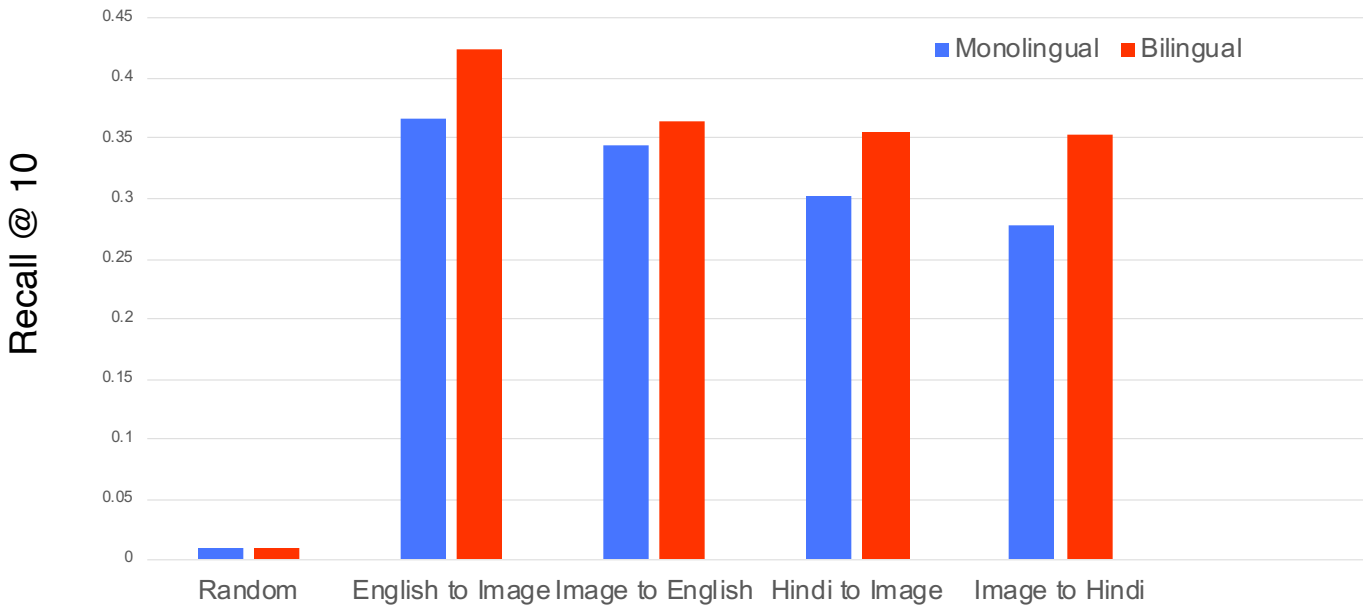


Hindi spoken caption




63

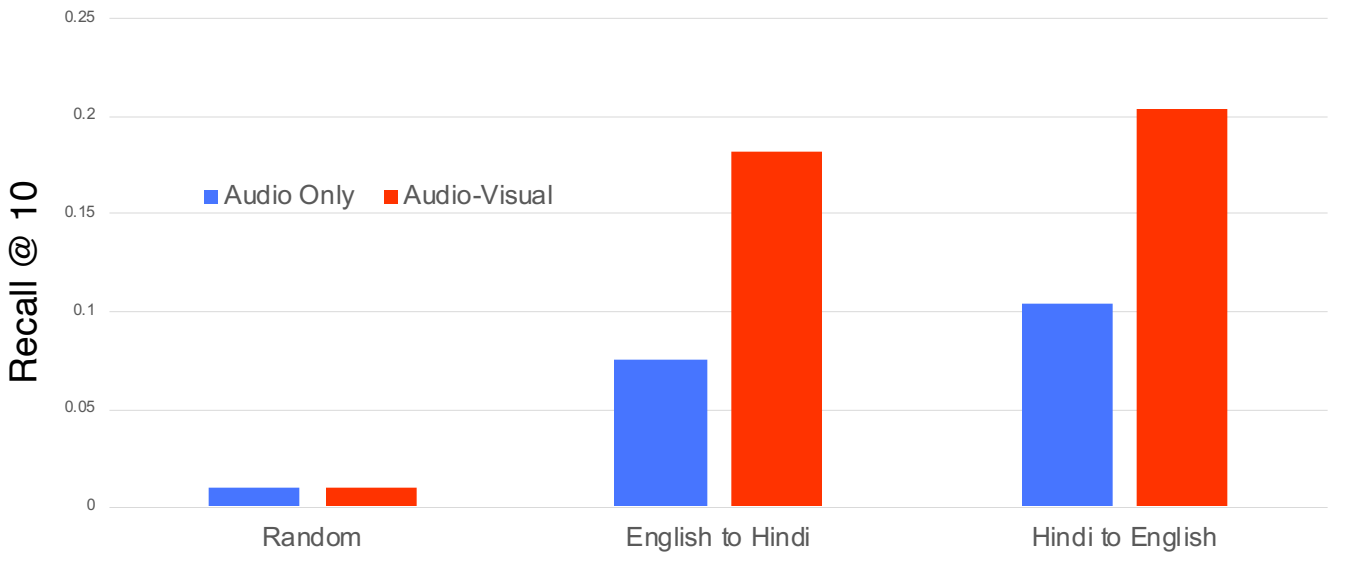
## Multi-Lingual Speech and Image Retrieval




83k training triplets, 1k testing triplets

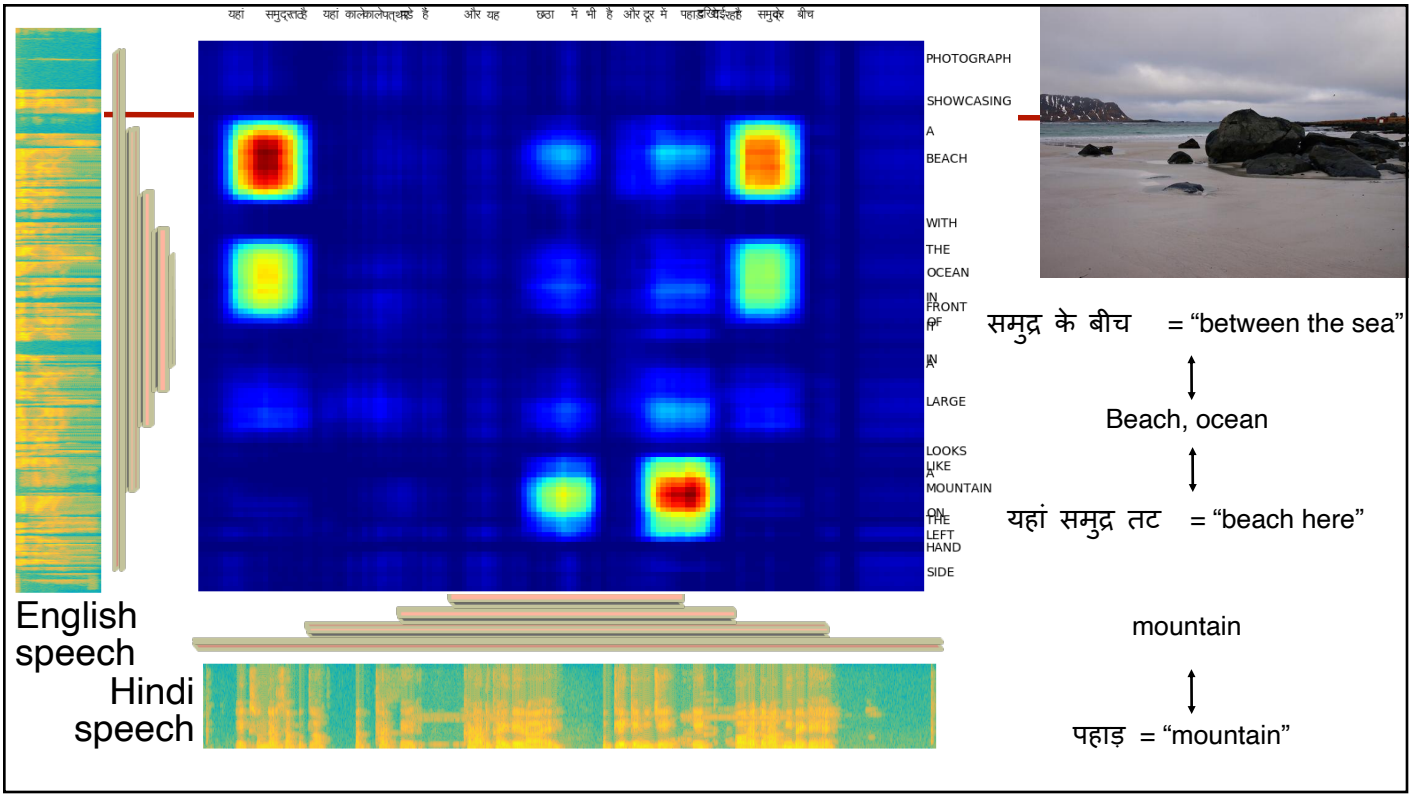
64

## Cross-Lingual Speech to Speech Retrieval




83k training triplets, 1k testing triplets

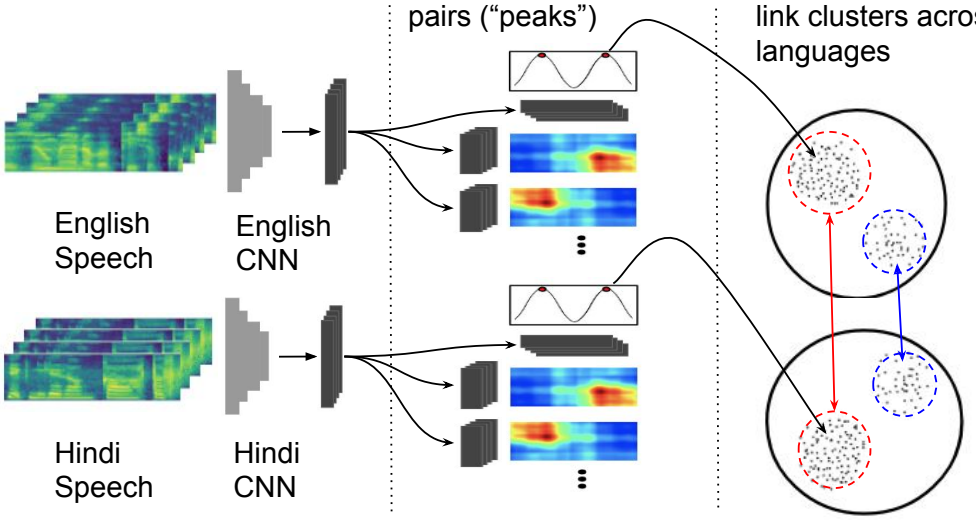
65




66

# Bilingual Speech Lexicon Discovery

1. Extract embeddings for each spoken caption
2. Find locations of maximum similarity between utterance pairs (“peaks”)
3. Cluster the peaks into pseudo-word categories in each language, then link clusters across languages




67

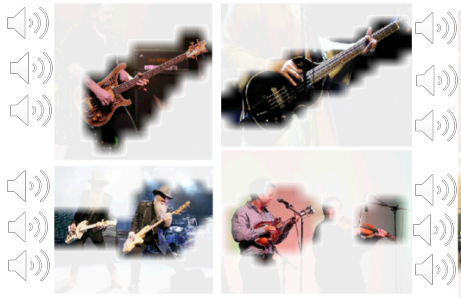
*people (0.84);* लोग*:the people (0.56)*



*trees (0.83);* पेड़*:the trees (0.75)*

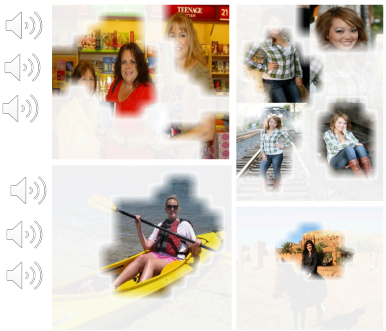


*water (0.78);* दुकान*:water (0.51)*



*guitar (0.79);* गिटार*:guitars (0.55)*

*kitchen (0.82);* रसोईघर*:kitchen (0.62)*



*woman (0.74);* औरत*:woman(0.384)*

68

# Other Ongoing Research

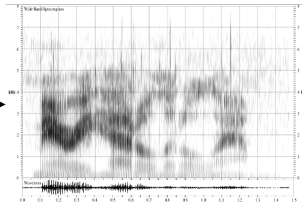
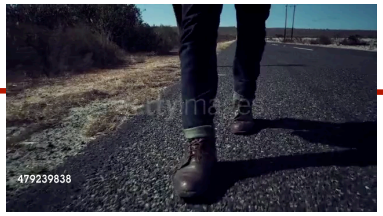


Image to speech synthesis



Learning from spoken video captions (Moments in Time)

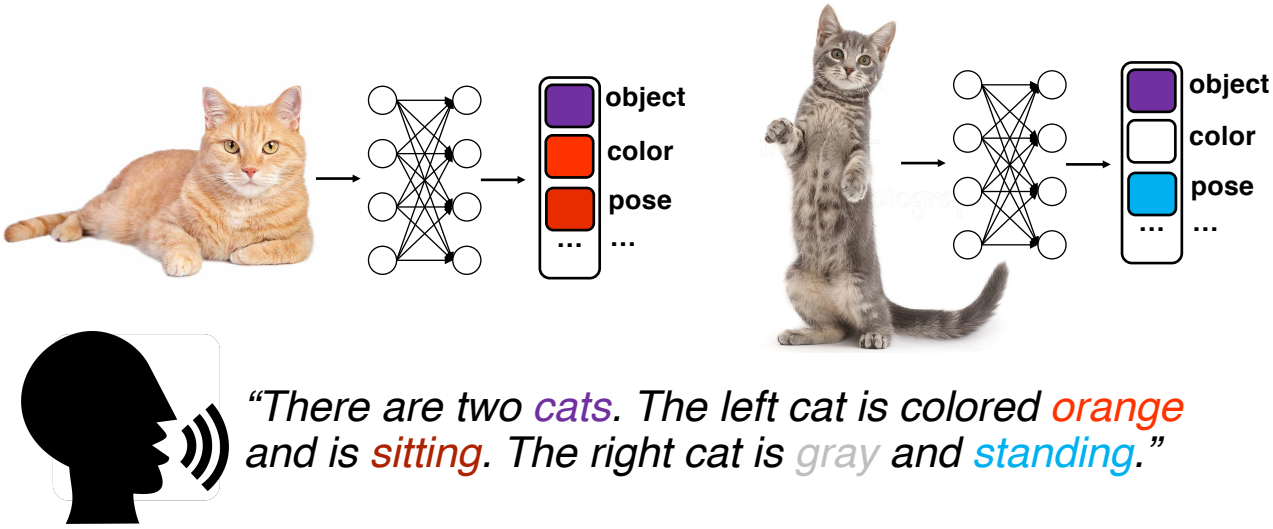
Many-lingual spoken caption datasets

Learning from videos “in the wild” e.g. YouTube

69

## Future: Language Acquisition

Using speech and language as a prior for learning disentangled representations, and using disentanglement to help learn language






*"There are two cats. The left cat is colored orange and is sitting. The right cat is gray and standing."*

70

## References


- Harwath et al., "Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input," *IJCV*, 2019
- Harwath et al., "Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech," Proc. ICLR, 2020.
- Harwath and Glass, "Towards Visually Grounded Sub-Word Speech Unit Discovery," *Proc. ICASSP*, 2019.
- Azuh et al., "Towards Bilingual Lexicon Discovery From Visually Grounded Speech Audio," *Proc. Interspeech*, 2019.


Papers, data and code:
http://people.csail.mit.edu/dharwath/

72