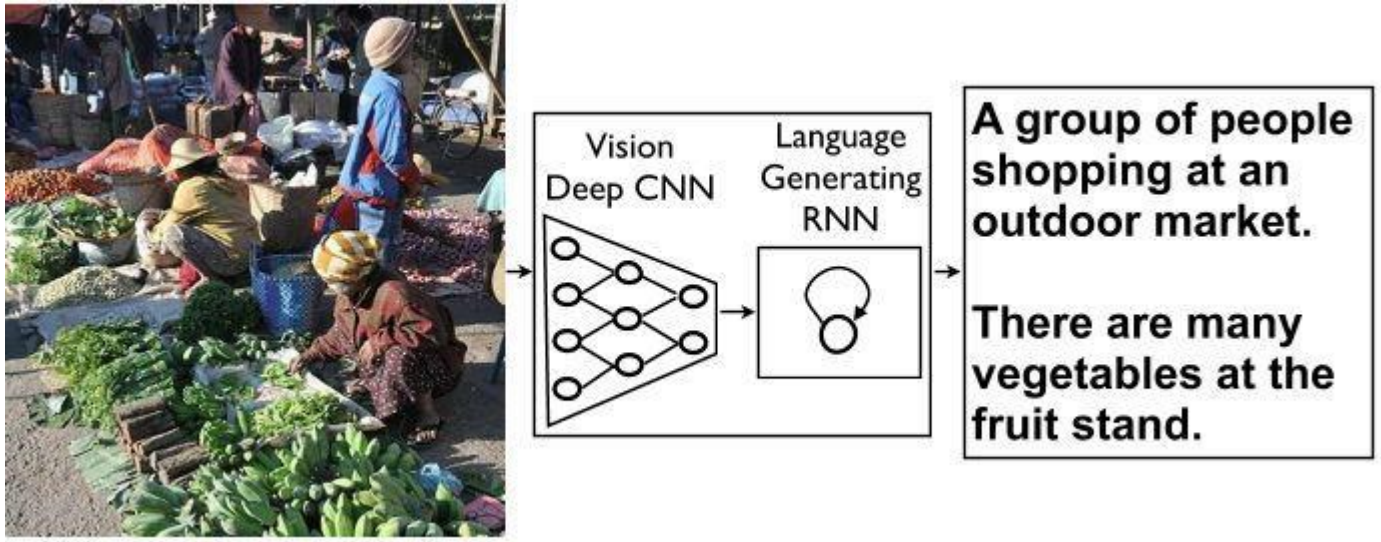
Multimodal Deep Learning

Ahmed Abdelkader

Design & Innovation Lab, ADAPT Centre



Talk outline

- What is multimodal learning and what are the challenges?
- Flickr example: joint learning of images and tags
- Image captioning: generating sentences from images
- SoundNet: learning sound representation from videos

Talk outline

- **What is multimodal learning and what are the challenges?**
- Flickr example: joint learning of images and tags
- Image captioning: generating sentences from images
- SoundNet: learning sound representation from videos

Deep learning success in single modalities

Deep learning success in single modalities

Deep learning success in single modalities

Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

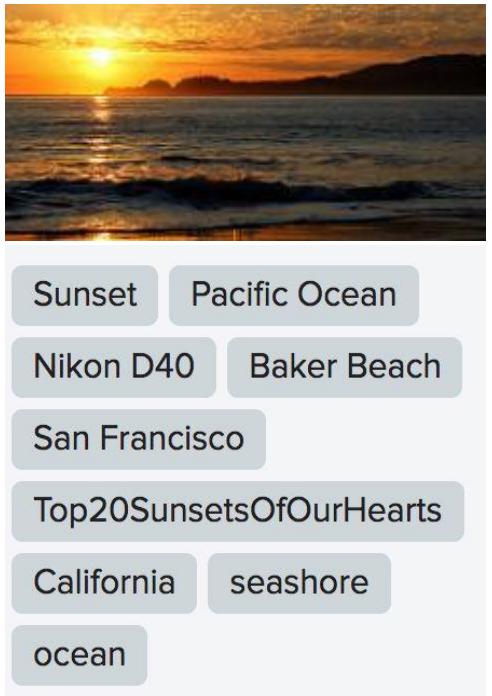
Super Bowl 50 decided the NFL champion for what season?
Ground Truth Answers: 2015 | the 2015 season | 2015
Prediction: 2015

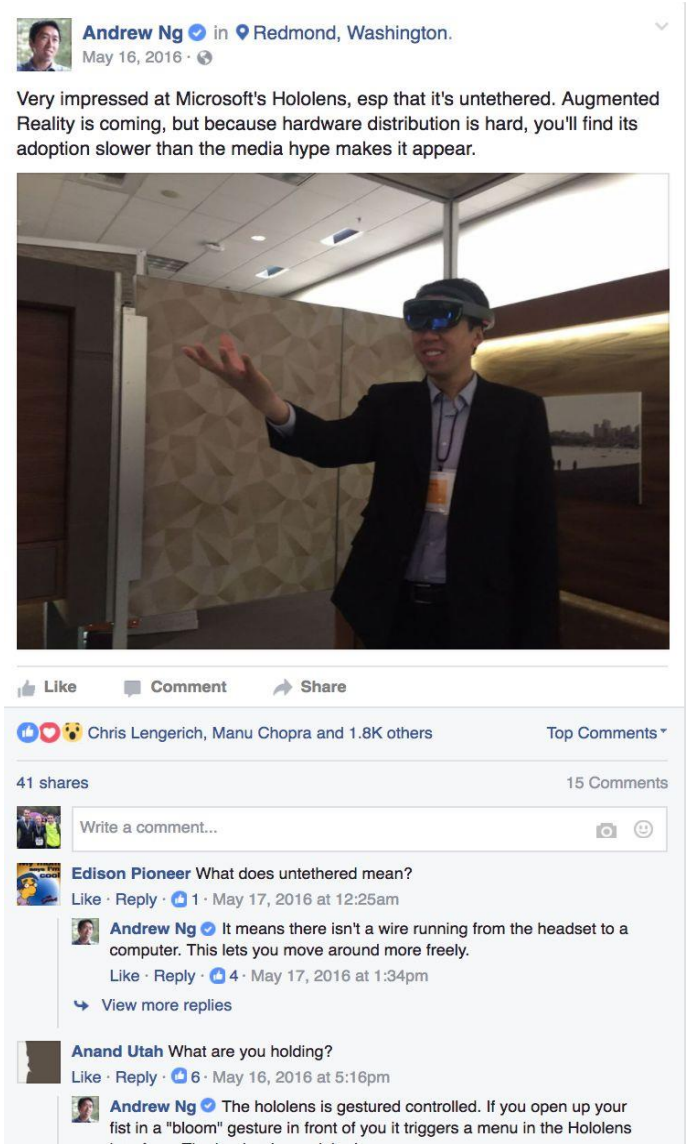
What is multimodal learning?

- In general, learning that involves **multiple modalities**
- This can manifest itself in different ways:
  - Input is one modality, output is another
  - Multiple modalities are learned jointly
  - One modality assists in the learning of another
  - …

Data is usually a collection of modalities

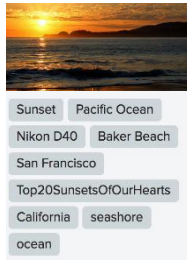
- Multimedia web content





Data is usually a collection of modalities

- Multimedia web content

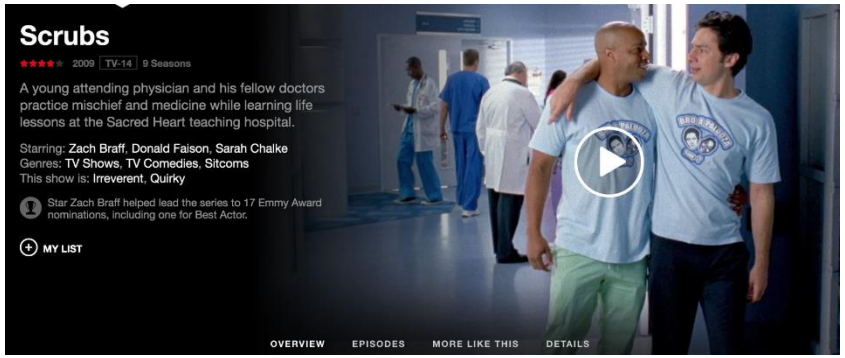


flickr

Google

YouTube

- Product recommendation systems



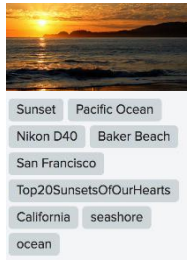
NETFLIX

ebay

amazon

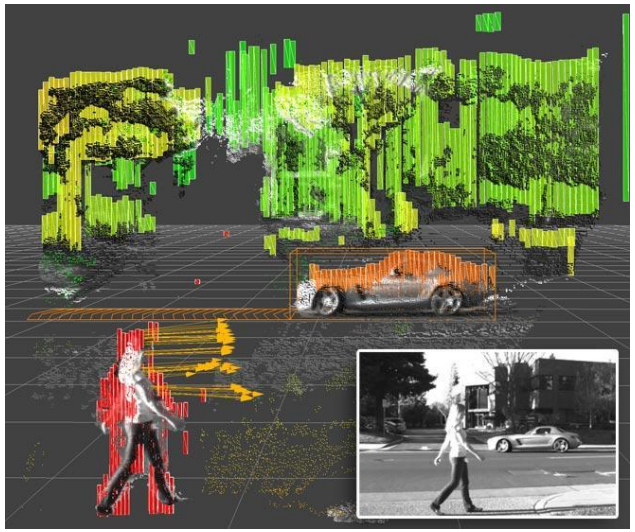
Data is usually a collection of modalities

- Multimedia web content



- Product recommendation systems

- Robotics



Why is multimodal learning hard?

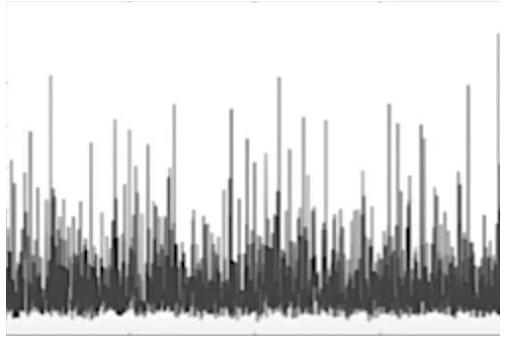
- Different representations

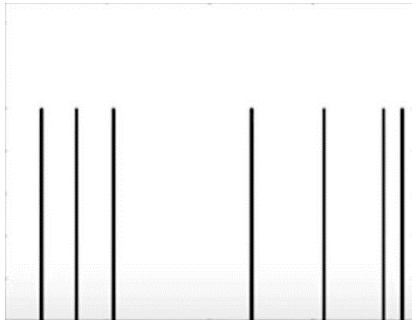
Images

Sunset
Pacific Ocean
Nikon D40
Baker Beach
San Francisco
Top20SunsetsOfOurHearts
California
seashore
ocean

Text

Real-valued
Dense

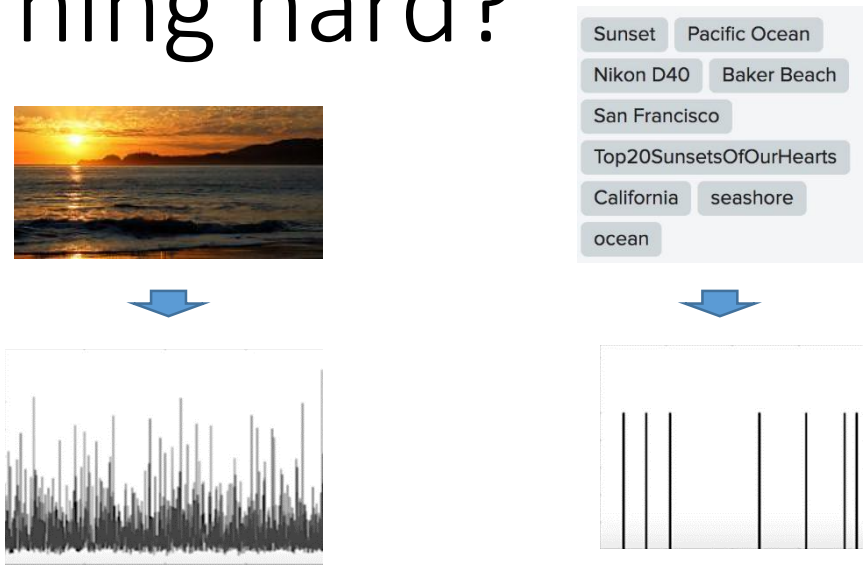




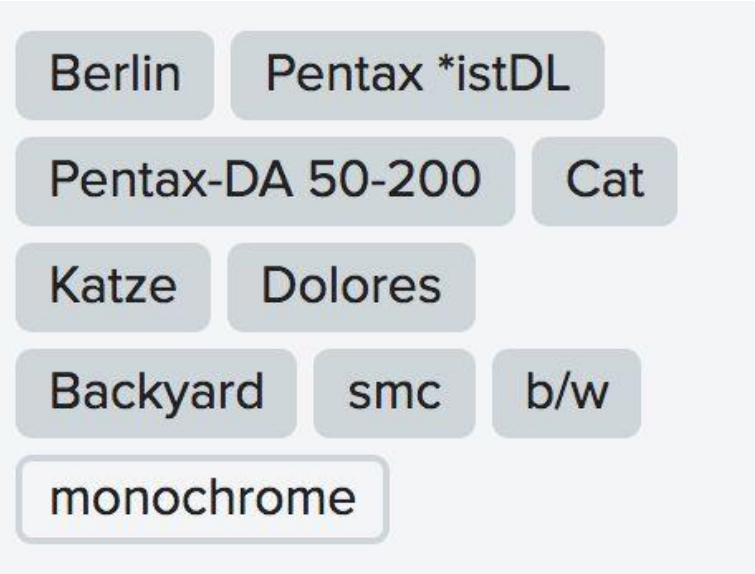
Discrete,
Sparse

Why is multimodal learning hard?

- Different representations



- Noisy and missing data



How can we solve these problems?

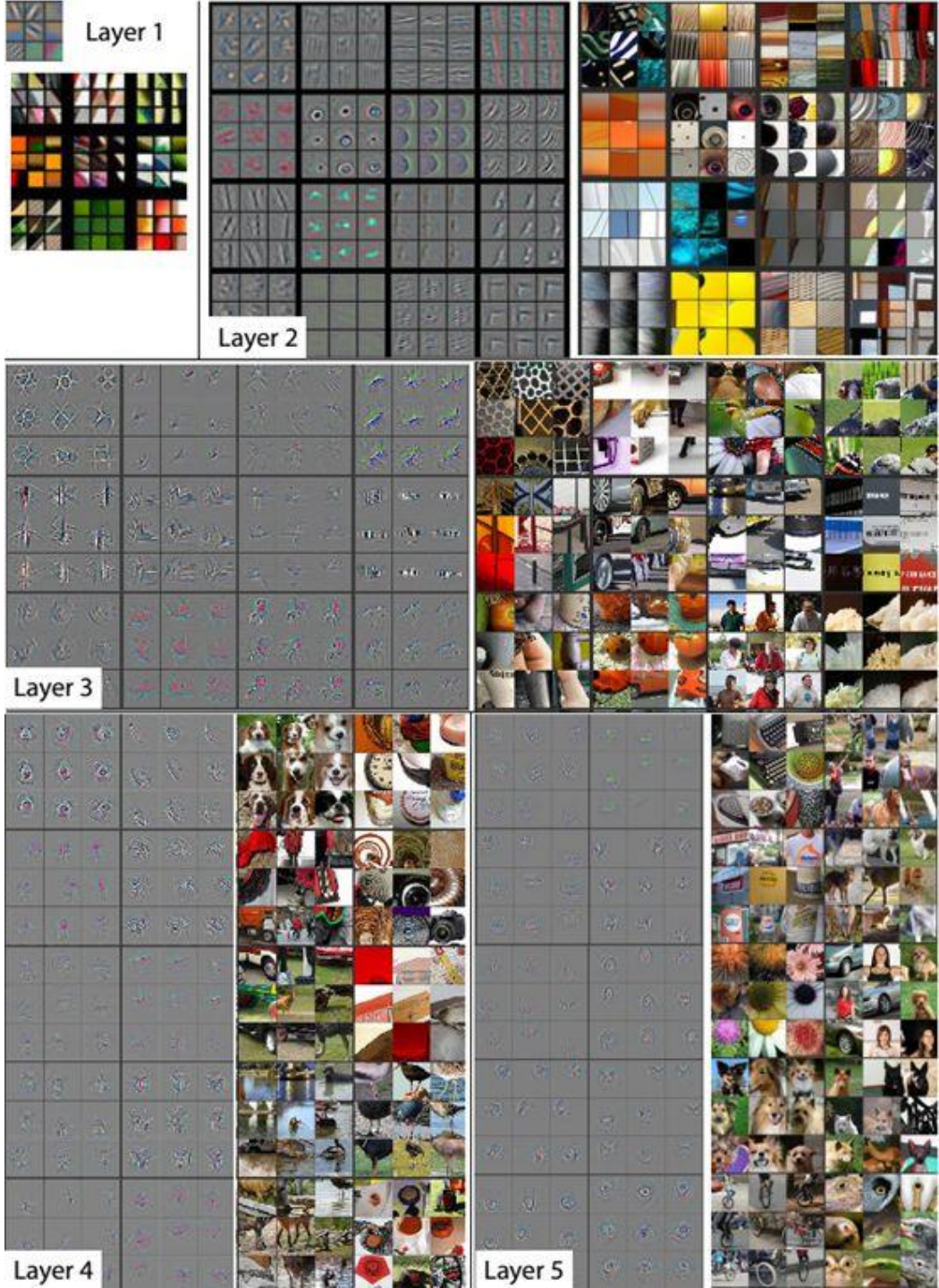
- **Combine** separate models for single modalities at a higher level
- **Pre-train** models on single-modality data
- How do we combine these models? **Embeddings**!

Pretraining

- Initialize with the weights from another network (instead of random)

- Even if the task is different, low-level features will still be useful, such as edge and shape filters for images

- Example: take the first 5 convolutional layers from a network trained on the ImageNet classification task

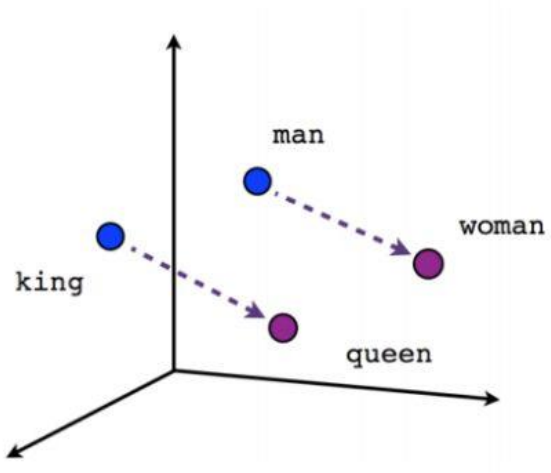


Embeddings

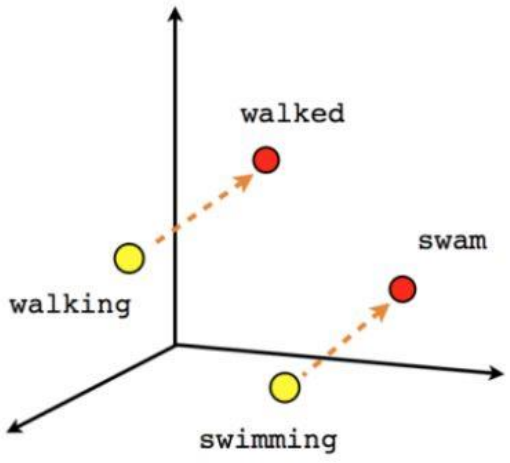
- A way to represent data
- In deep learning, this is usually a high-dimensional vector
- A neural network can take a piece of data and create a corresponding vector in an embedding space
- A neural network can take a embedding vector as an input
- Example: word embeddings

Word embeddings

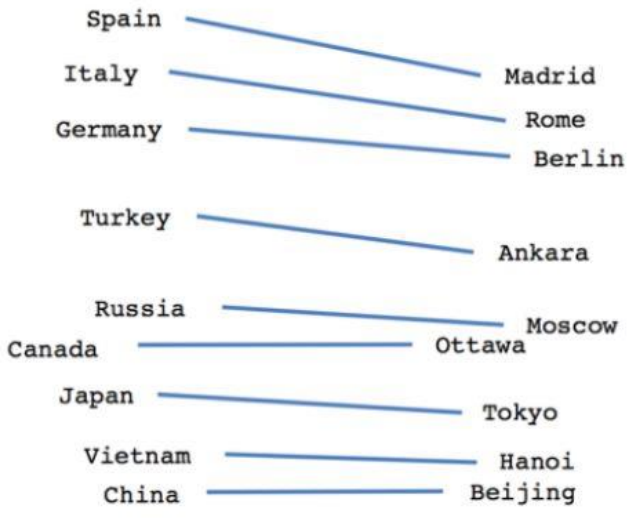
- A word embedding: word ⟶ high-dimensional vector In deep
- Interesting properties



Male-Female



Verb tense



Country-Capital

Embeddings

- We can use embeddings to switch between modalities!
- In sequence modeling, we saw a sentence embedding to switch between languages for translation
- Similarly, we can have embeddings for images, sound, etc. that allow us to transfer meaning and concepts across modalities

Talk outline

- What is multimodal learning and what are the challenges?
- **Flickr example: joint learning of images and tags**
- Image captioning: generating sentences from images
- SoundNet: learning sound representation from videos

Flickr tagging: task

Images

Sunset
Pacific Ocean
Nikon D40
Baker Beach
San Francisco
Top20SunsetsOfOurHearts
California
seashore
ocean

Text

Flickr tagging: task

Images

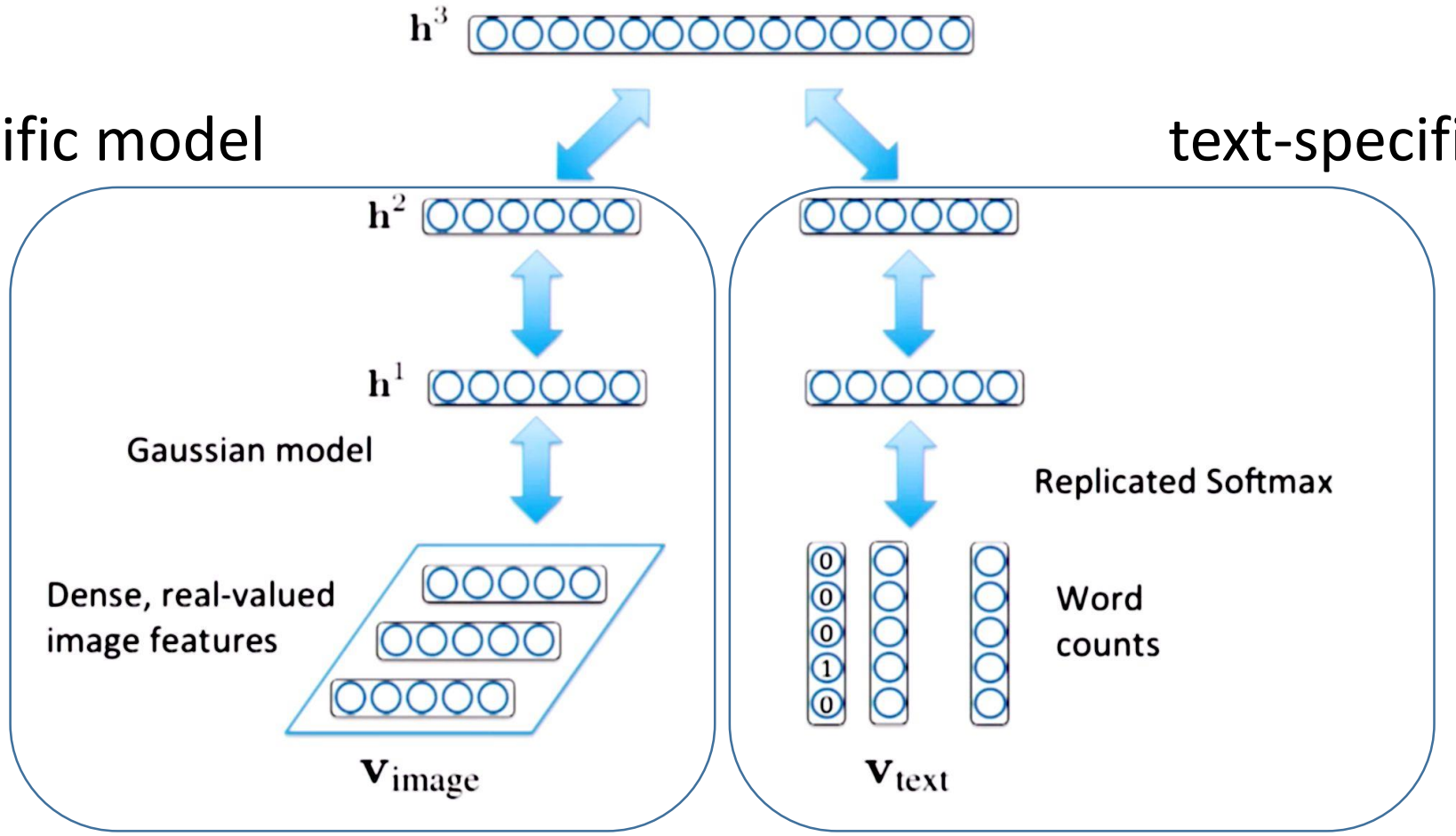
Text

- 1 million images from flickr
- 25,000 have tags
- Goal: create a joint representation of images and text
- Useful for Flickr photo search

Flickr tagging: model

Image-specific model

text-specific model

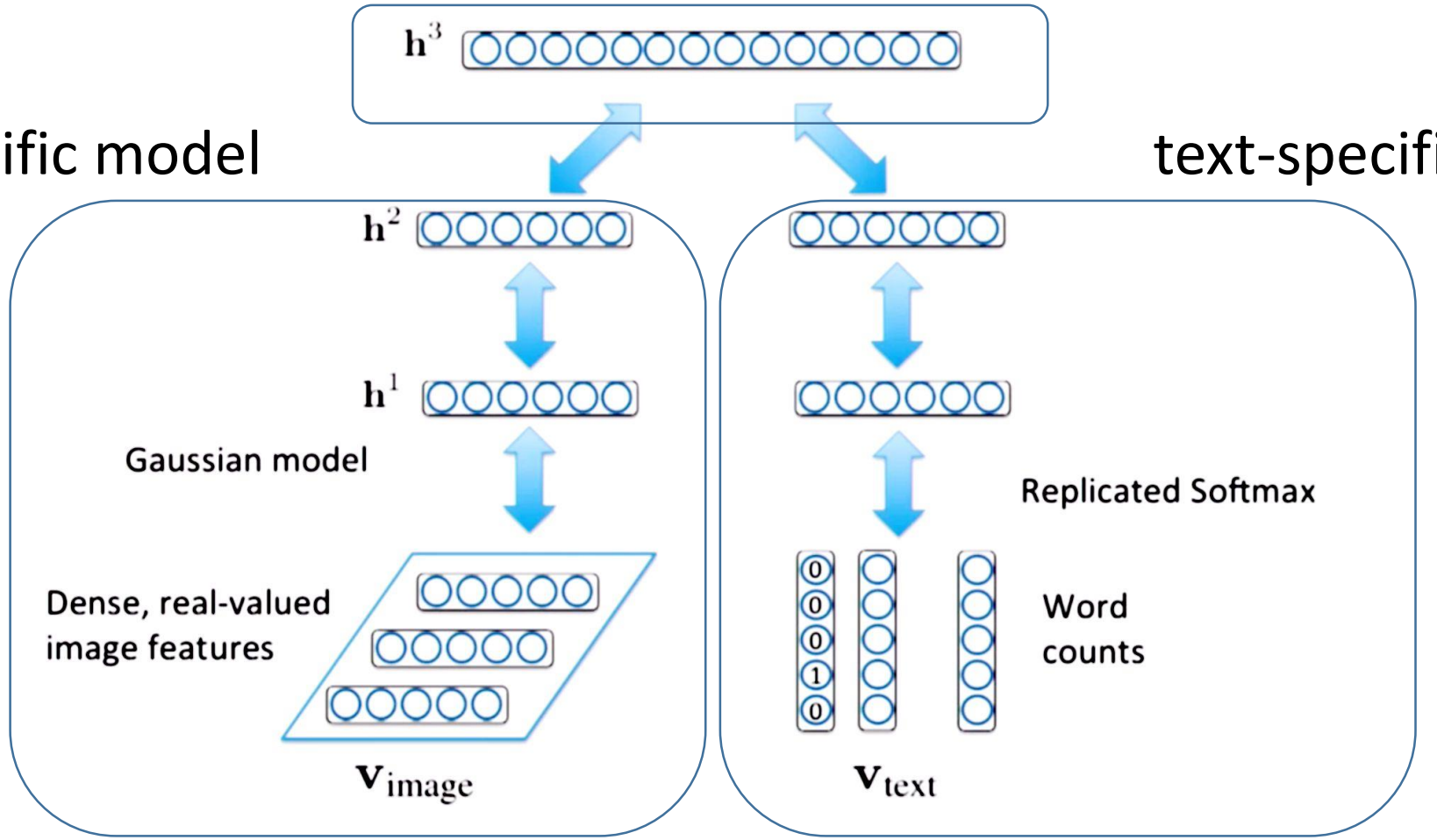


Pretrain unimodal models and combine them at a higher level

Flickr tagging: model

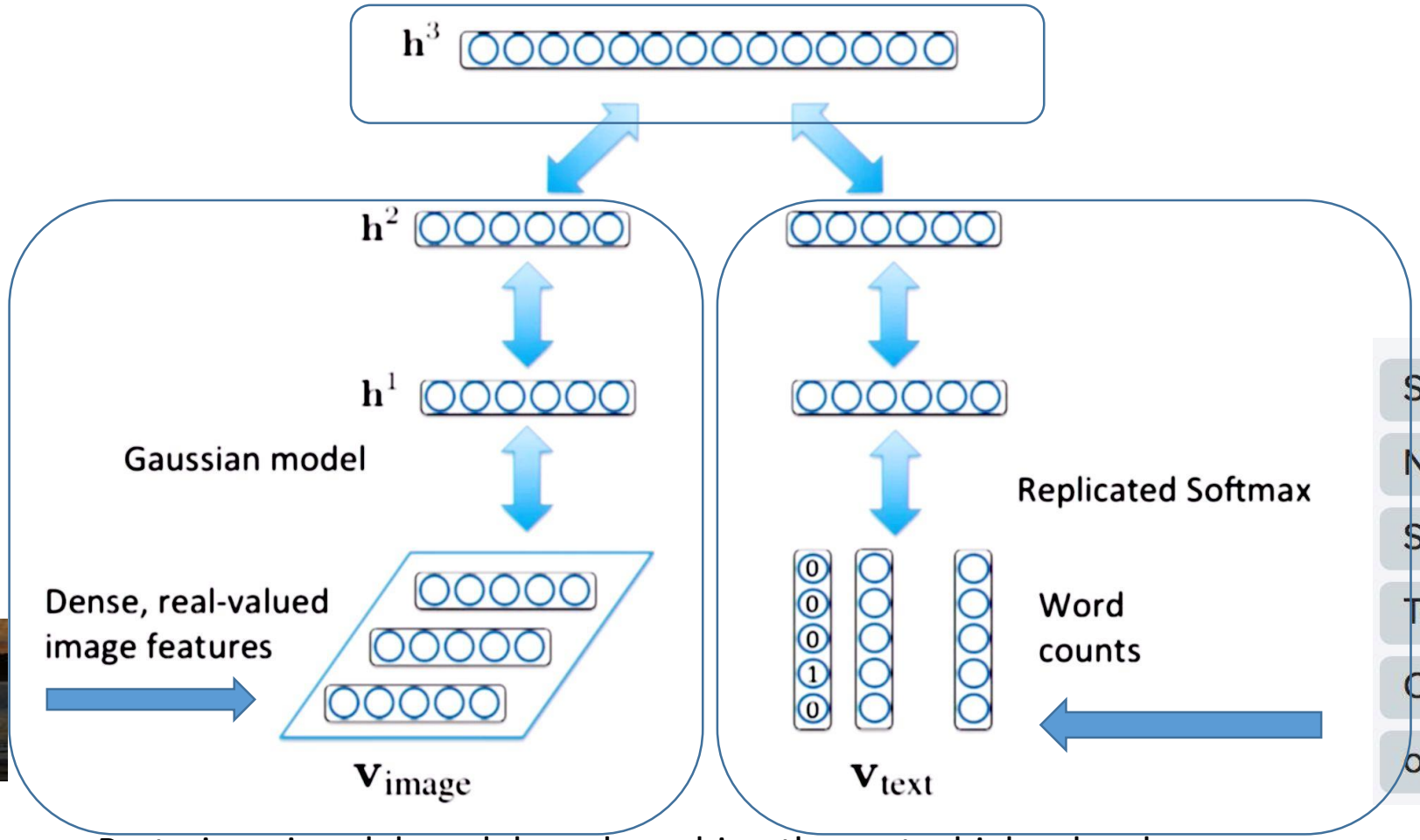
Image-specific model

text-specific model



Pretrain unimodal models and combine them at a higher level

Flickr tagging: model



Pretrain unimodal models and combine them at a higher level

Flickr tagging: example outputs

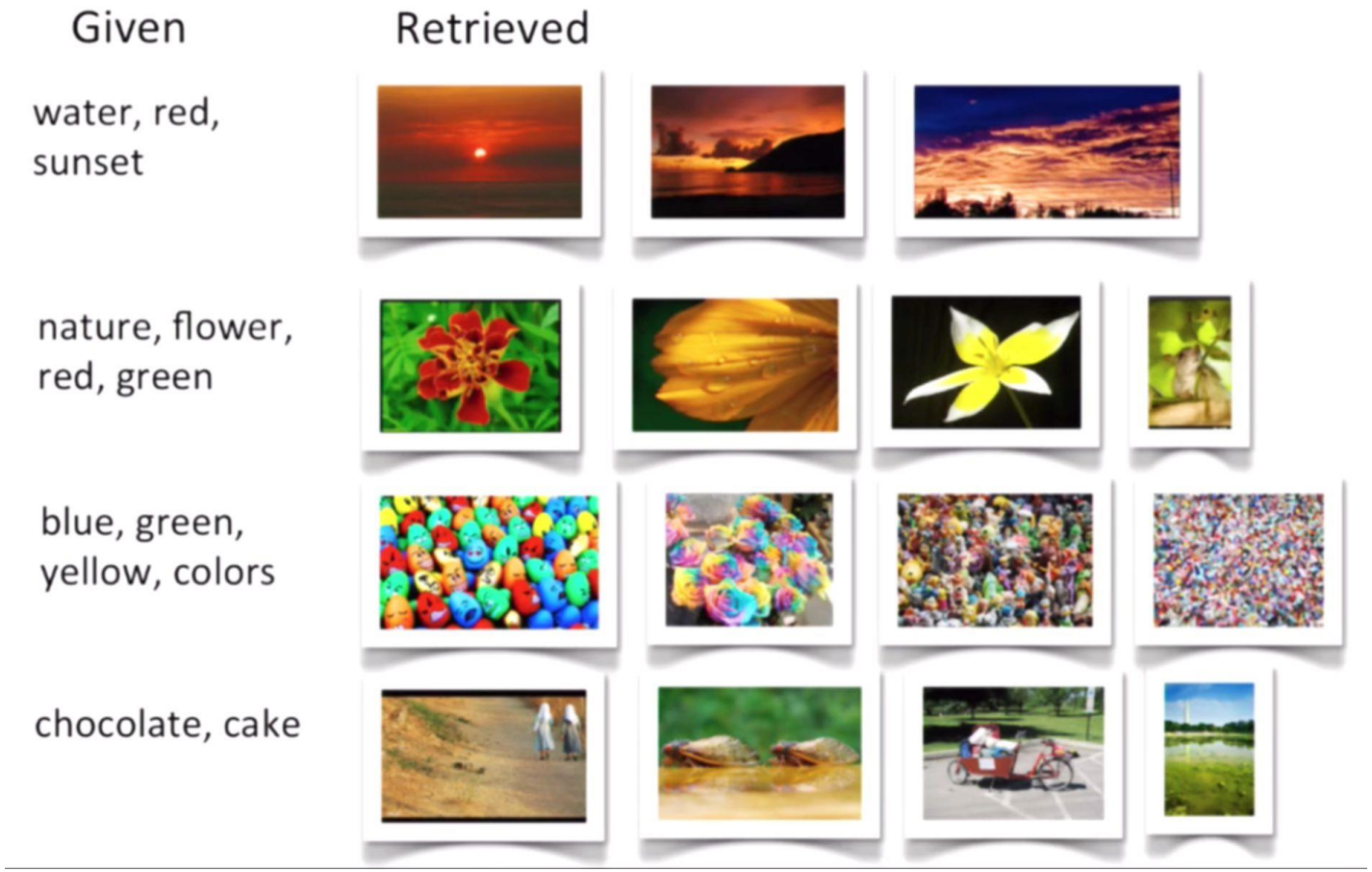
Given	Generated
	dog, cat, pet, kitten, puppy, ginger, tongue, kitty, dogs, furry
	sea, france, boat, mer, beach, river, bretagne, plage, brittany
	portrait, child, kid, ritratto, kids, children, boy, cute, boys, italy

Given	Generated
	insect, butterfly, insects, bug, butterflies, lepidoptera
	graffiti, streetart, stencil, sticker, urbanart, graff, sanfrancisco
	canada, nature, sunrise, ontario, fog, mist, bc, morning

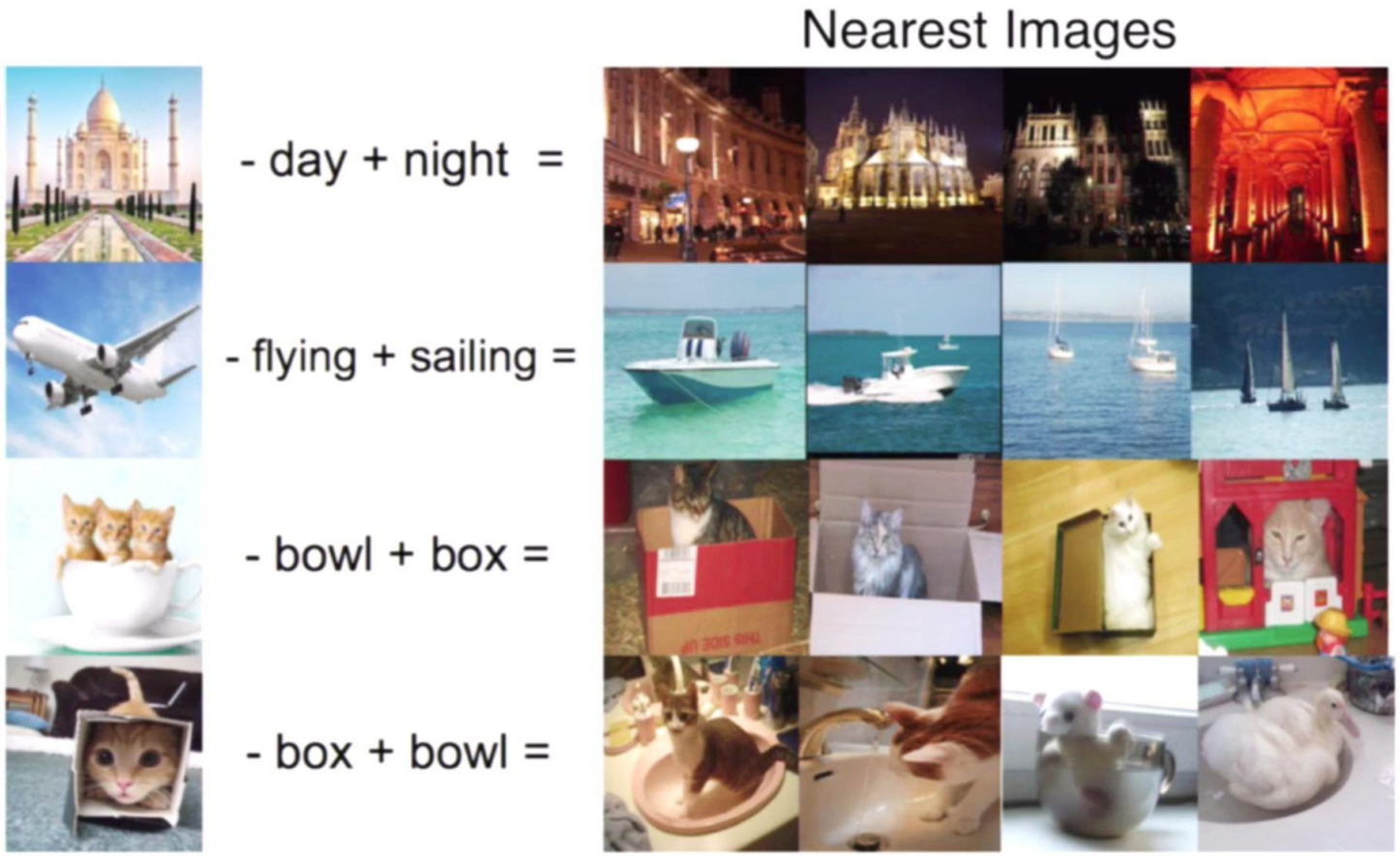
Flickr tagging: example outputs

Given	Generated
	portrait, women, army, soldier, mother, postcard, soldiers
	obama, barackobama, election, politics, president, hope, change, sanfrancisco, convention, rally
	water, glass, beer, bottle, drink, wine, bubbles, splash, drops, drop

Flickr tagging: visualization



Flickr tagging: multimodal arithmetic

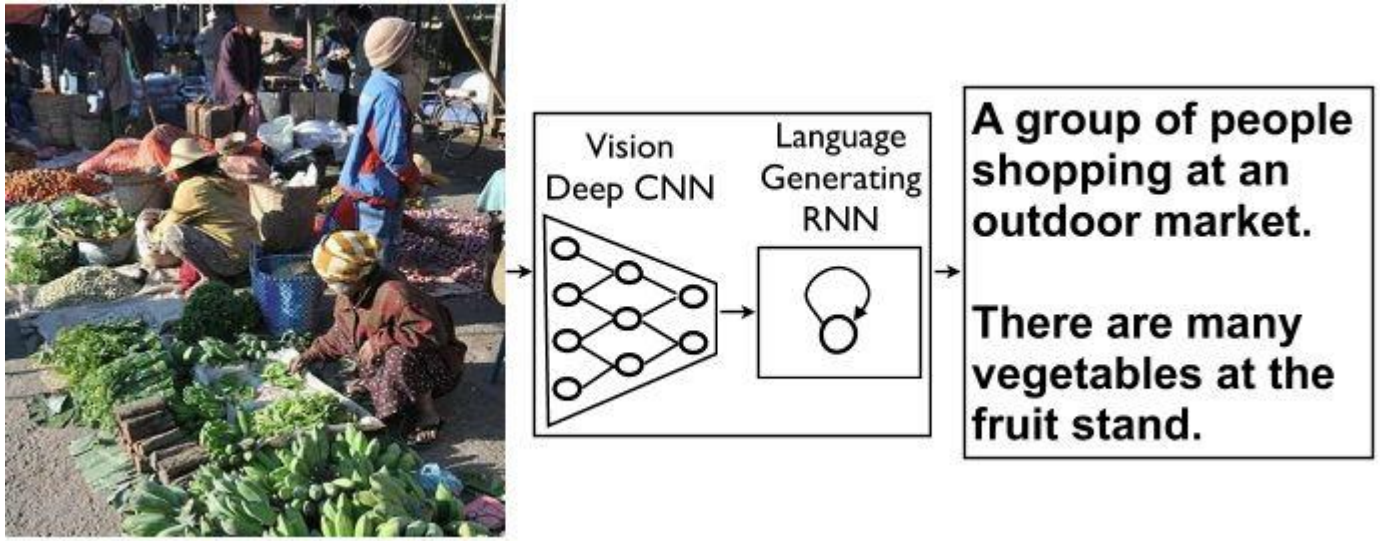


Kiros, Salakhutdinov, Zemel 2015

Talk outline

- What is multimodal learning and what are the challenges?
- Flickr example: joint learning of images and tags
- **Image captioning: generating sentences from images**
- SoundNet: learning sound representation from videos

Example: image captioning



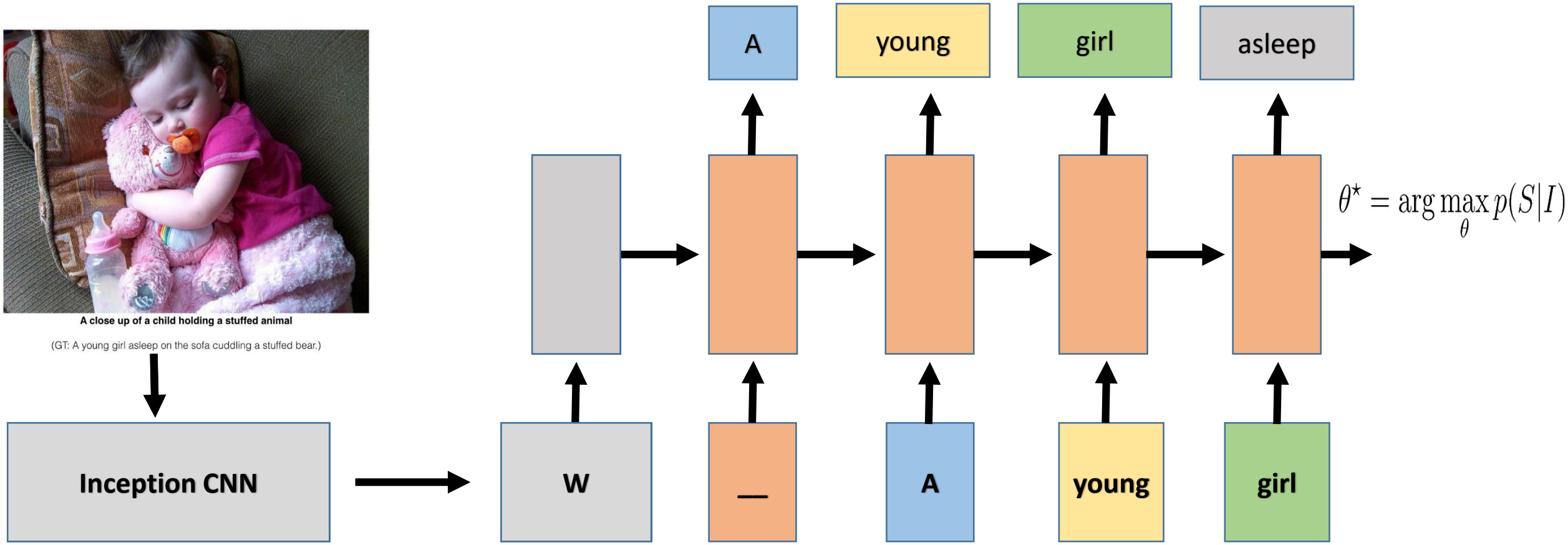
Show and Tell: A Neural Image Caption Generator

Vinyals et al. 2014

Example: image captioning

A close up of a child holding a stuffed animal

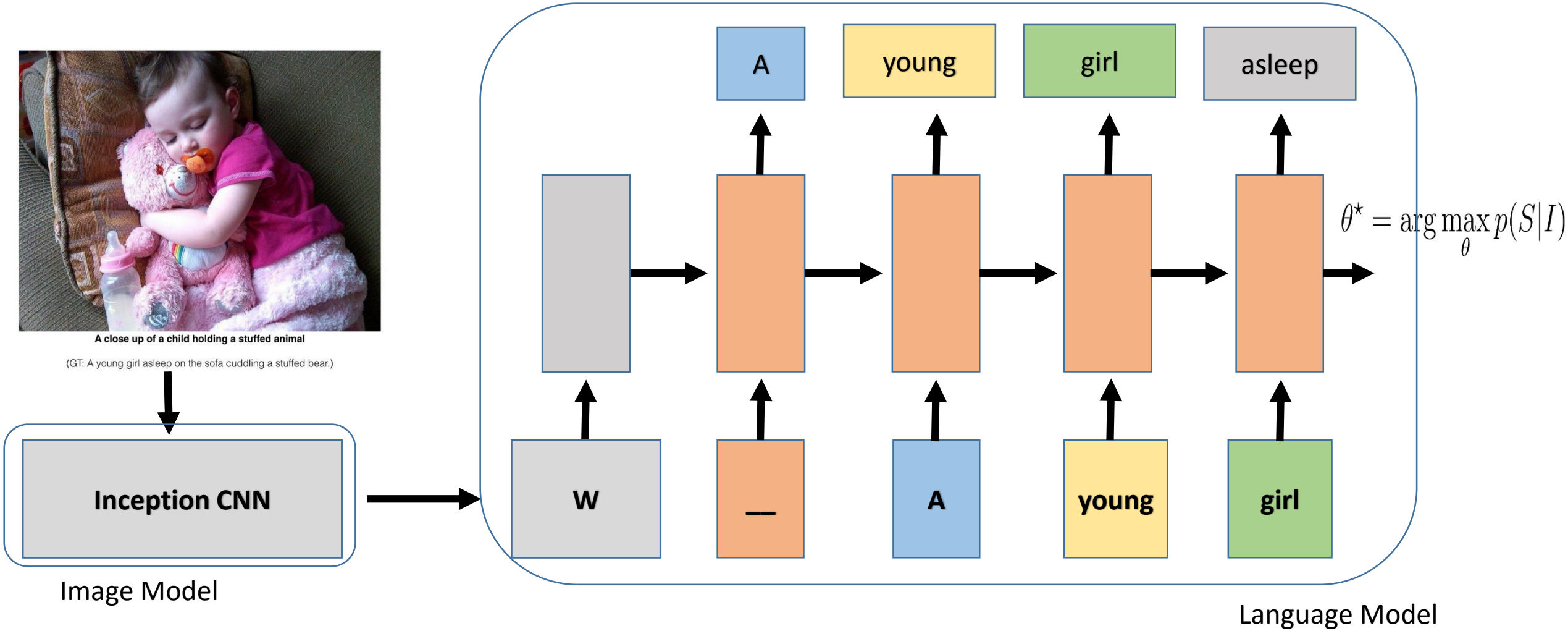
(GT: A young girl asleep on the sofa cuddling a stuffed bear.)



Example: image captioning

A close up of a child holding a stuffed animal

(GT: A young girl asleep on the sofa cuddling a stuffed bear.)



Human: A young girl asleep on the sofa cuddling a stuffed bear

Computer: A close up of a child holding a stuffed animal.

Human: A view of inside of a car where a cat is laying down.

Computer: A cat sitting on top of a black car.

Human: A green monster kite soaring in a sunny sky.

Computer: A man flying through the air while riding a snowboard.

Caption model for neural storytelling

We were barely able to catch the breeze at the beach, and it felt as if someone stepped out of my mind. She was in love with him for the first time in months, so she had no intention of escaping. The sun had risen from the ocean, making her feel more alive than normal. She's beautiful, but the truth is that I don't know what to do. The sun was just starting to fade away, leaving people scattered around the Atlantic Ocean. I'd seen the men in his life, who guided me at the beach once more.

Jamie Kiros, www.github.com/ryankiros/neural-storyteller

Talk outline

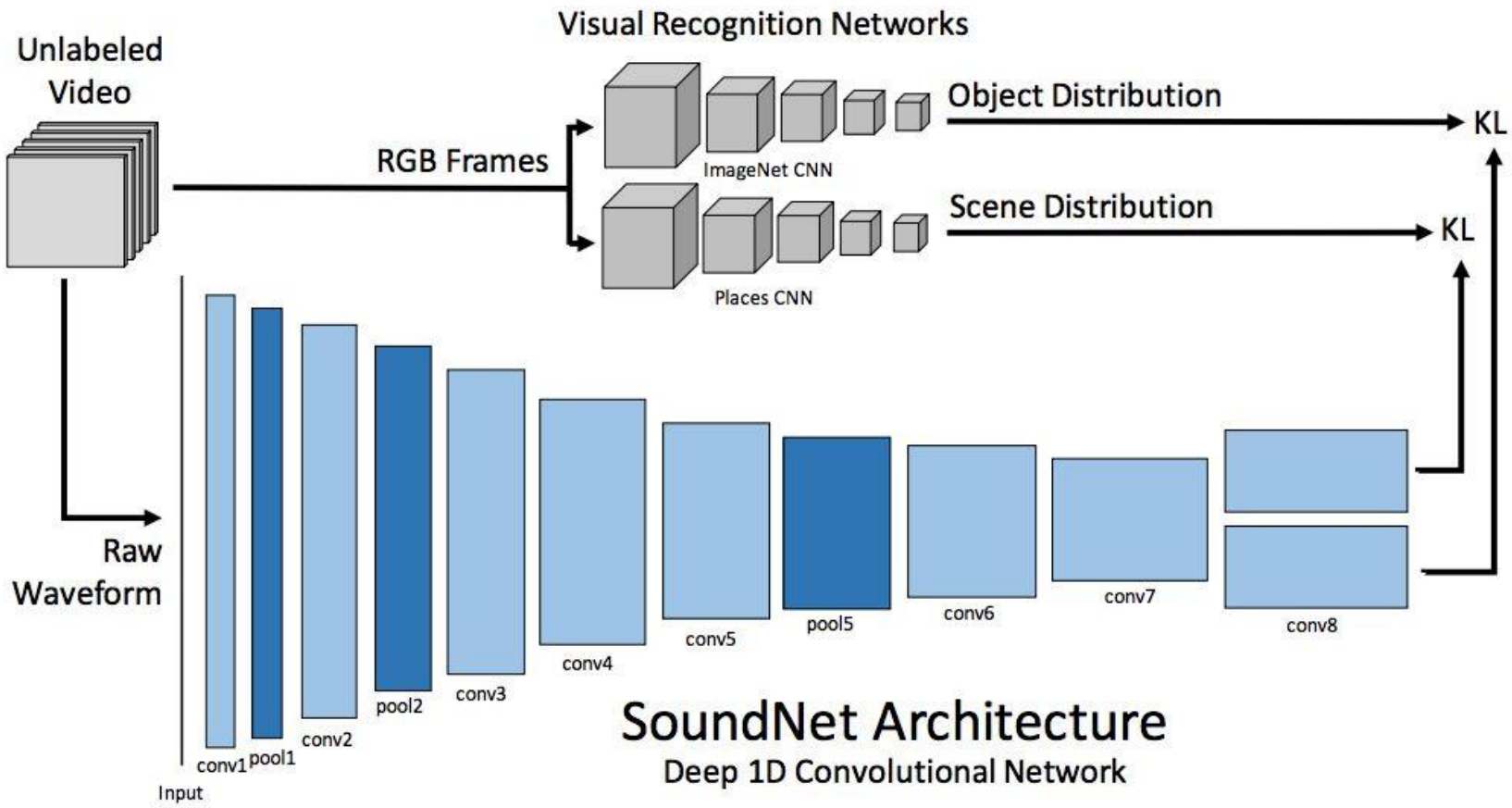
- What is multimodal learning and what are the challenges?
- Flickr example: joint learning of images and tags
- Image captioning: generating sentences from images
- **SoundNet: learning sound representation from videos**

SoundNet

- Idea: learn a sound representation from unlabeled video
- We have good vision models that can provide information about unlabeled videos
- Can we train a network that takes sound as an input and learns object and scene information?
- This sound representation could then be used for sound classification tasks

Aytar, Vondrick, Torralba. NIPS 2016

SoundNet training

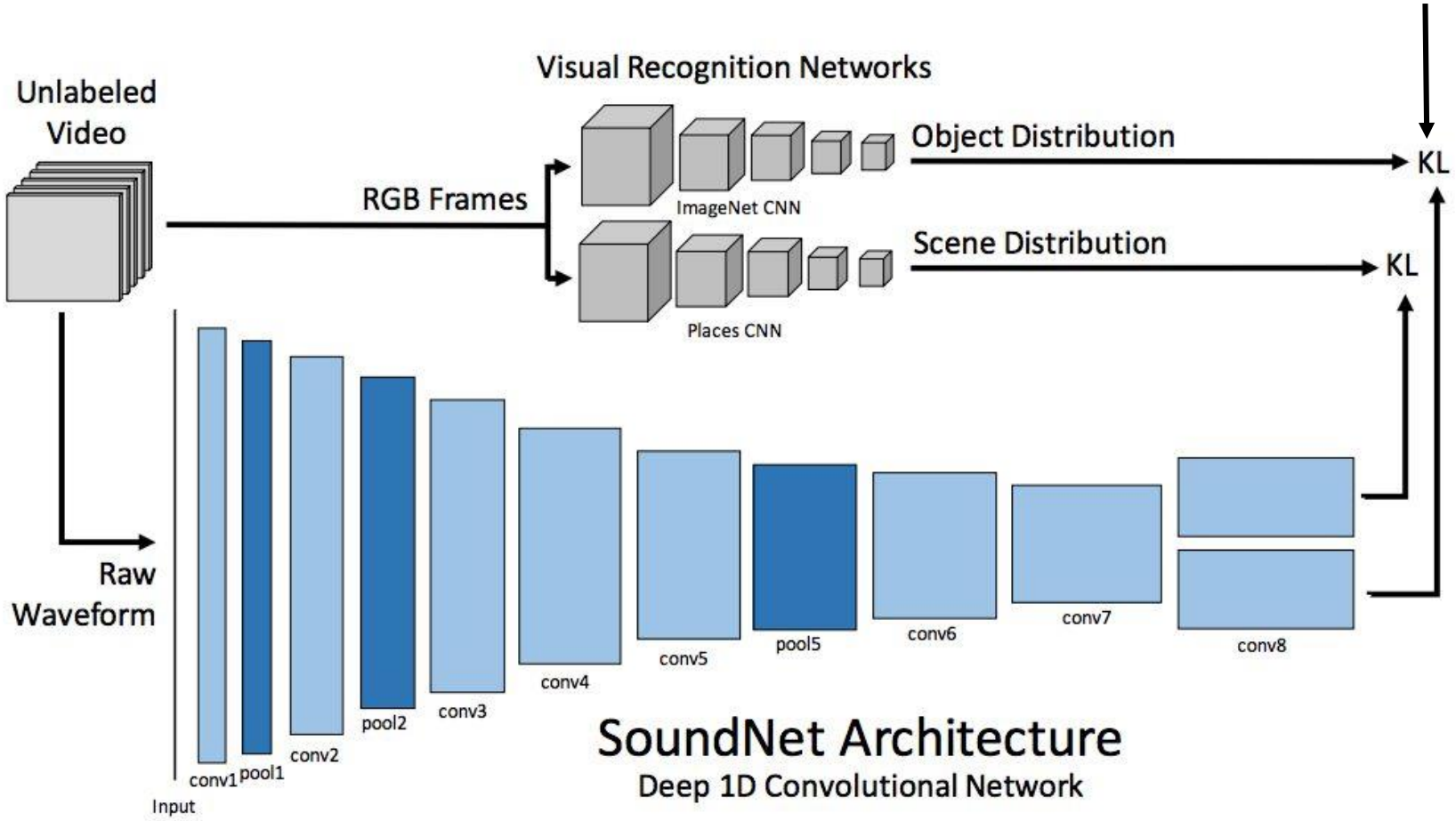


Aytar, Vondrick, Torralba. NIPS 2016

SoundNet **training**

Loss for the sound CNN:

$$\mathrm{D}_{KL}(g(y) \parallel f(x;\theta))$$

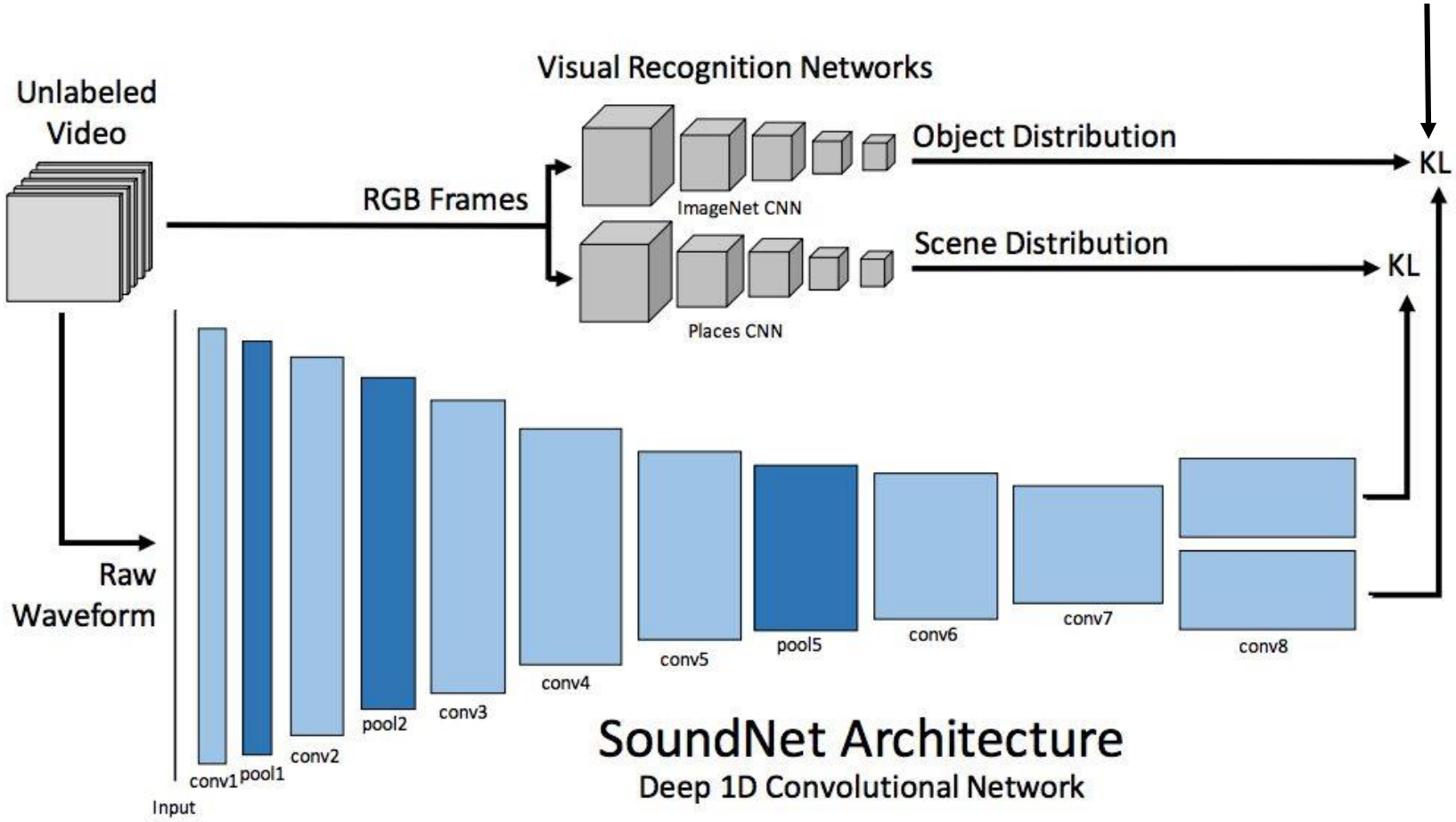


Aytar, Vondrick, Torralba. NIPS 2016

SoundNet **training**

Loss for the sound CNN:

$$\mathrm{D}_{KL}(g(y) \parallel f(x;\theta))$$



x is the raw waveform

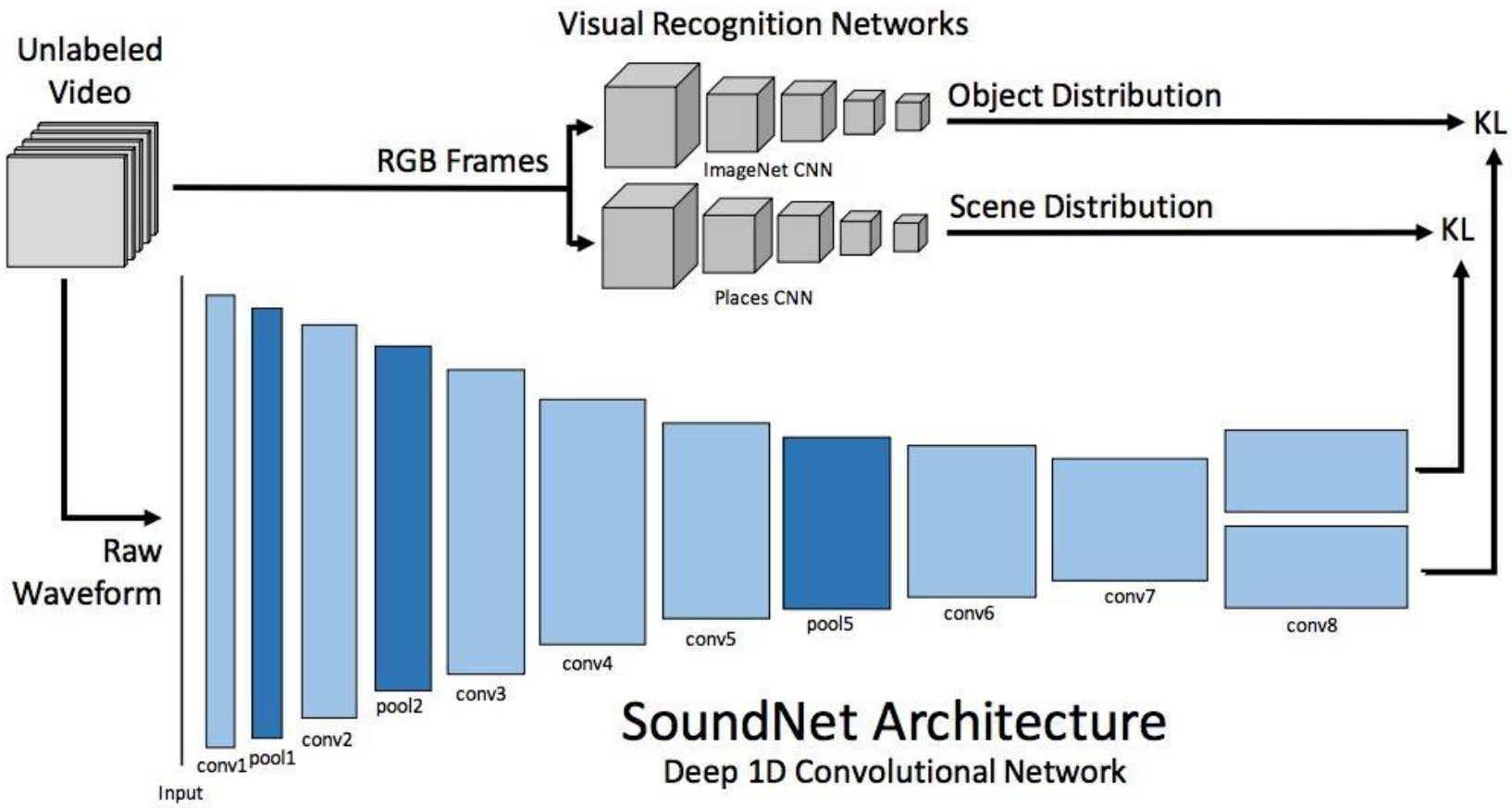
y is the RGB frames

$g(y)$ is the object or scene distribution

$f(x;\theta)$ is the output from the sound CNN

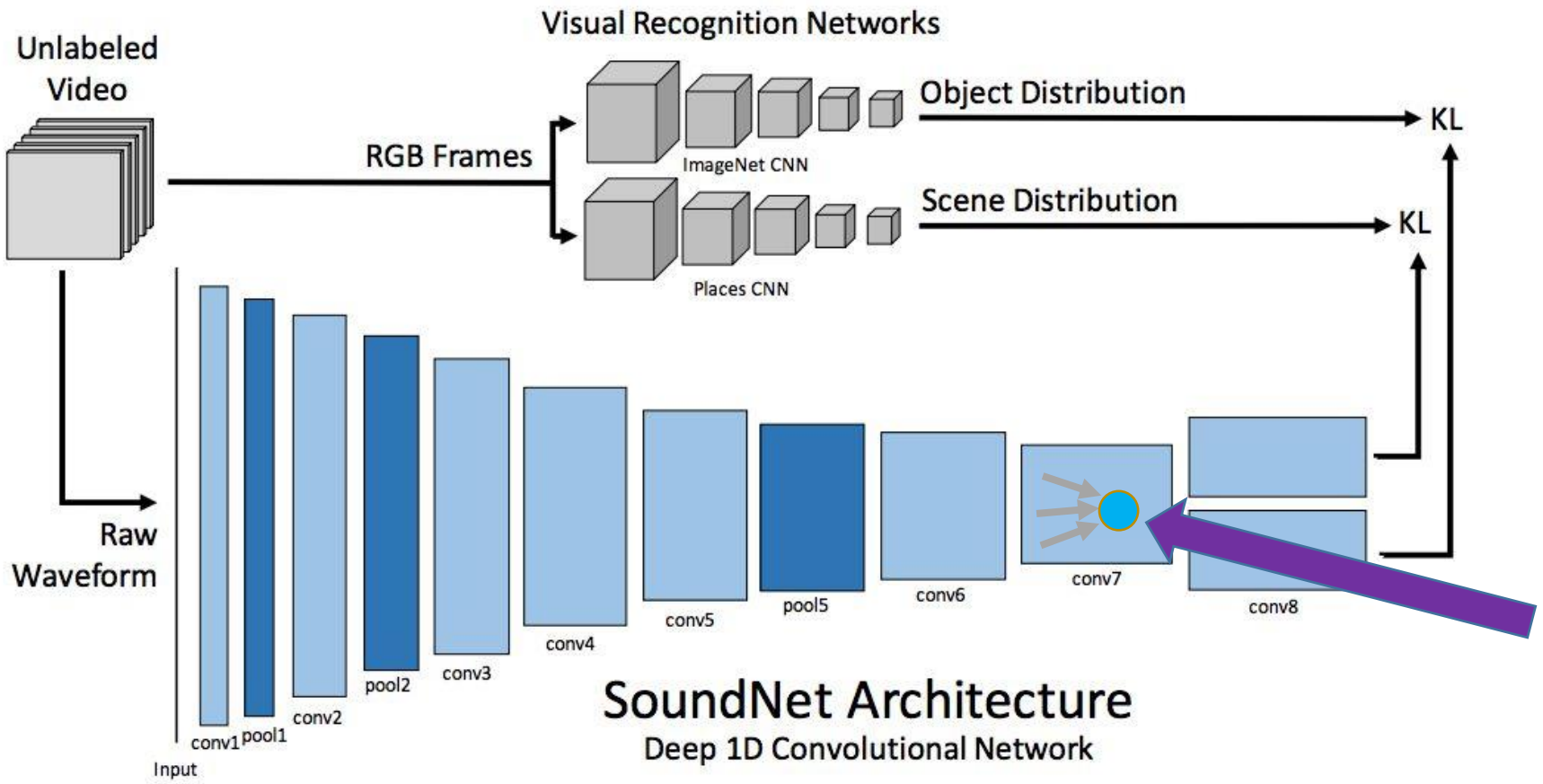
Aytar, Vondrick, Torralba. NIPS 2016

SoundNet **visualization**



Aytar, Vondrick, Torralba. NIPS 2016

SoundNet **visualization**



What audio inputs evoke the maximum output from this neuron?

Aytar, Vondrick, Torralba. NIPS 2016

SoundNet: visualization of hidden units

https://projects.csail.mit.edu/soundnet/

Baby Talk | Bubbles | Cheering | Bird Chirps

Conclusion

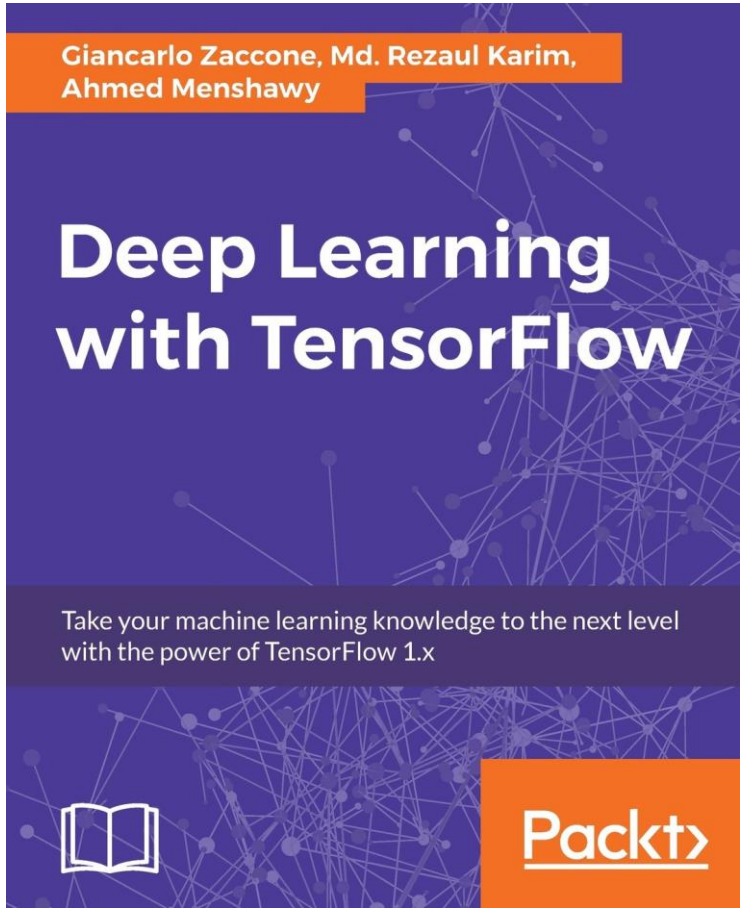
- **Multimodal tasks are hard**
  - Differences in data representation
  - Noisy and missing data

Conclusion

- **Multimodal tasks are hard**
  - Differences in data representation
  - Noisy and missing data
- **What types of models work well?**
  - Composition of unimodal models
  - Pretraining unimodally

Conclusion

- **Multimodal tasks are hard**
  - Differences in data representation
  - Noisy and missing data
- **What types of models work well?**
  - Composition of unimodal models
  - Pretraining unimodally
- **Examples of multimodal tasks**
  - Model two modalities jointly (Flickr tagging)
  - Generate one modality from another (image captioning)
  - Use one modality as labels for the other (SoundNet)



https://www.amazon.co.uk/Deep-Learning-TensorFlow-Giancarlo-Zaccone/dp/1786469782

Questions?