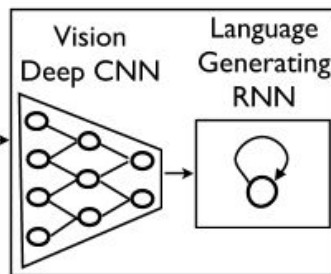


Multimodal Learning

Victoria Dean



A group of people shopping at an outdoor market.

There are many vegetables at the fruit stand.

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?

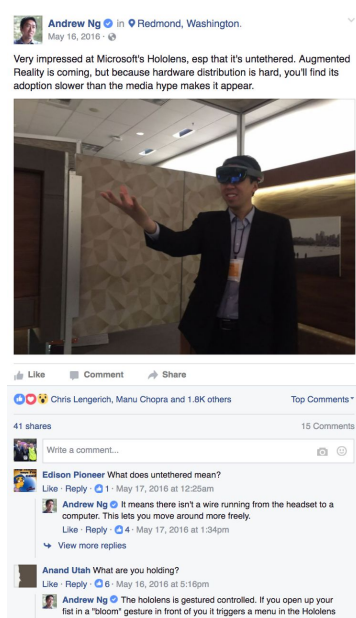
- In general, learning that involves **multiple modalities**
- This can manifest itself in different ways:
 - Input is one modality, output is another
 - Multiple modalities are learned jointly
 - One modality assists in the learning of another
 - ...

Data is usually a collection of modalities

- Multimedia web content



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



You**Tube**

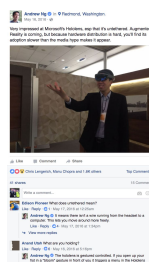


flickr

Google

- Multimedia web content

- Multimedia web content



- Product recommendation systems

Customer Reviews

★★★★☆ 239

4.3 out of 5 stars

| | |
|--------|------------------------|
| 5 star | <div><div></div></div> |
| 4 star | <div><div></div></div> |
| 3 star | <div><div></div></div> |
| 2 star | <div><div></div></div> |
| 1 star | <div><div></div></div> |

Share your thoughts with other customers

Write a customer review

[See all verified purchase reviews.](#)

Top Customer Reviews

★★★★★ **Amazing! 3 Cheers for Greenies Pill Pockets**

By [PhoebeCat](#) on April 21, 2010

Verified Purchase

These Pill Pockets are a total god-send. I have two cats, one of which has to be 'pilled' twice daily. This has meant a horrible process starting with her being wrapped up in a blanket. Now, thank heavens, we no longer have to go through that procedure. I simply pop her tablet into a Pill Pocket and she eats it up. My other cat is beyond fussy; she won't even touch fresh chicken, salmon, cream ... none of the normal cat 'treats' we humans offer. However, she actually BEGS for the Pill Pockets, which I now give her minus any medication at all. I treat it as a reward for her to eat anyone who she can do dogs keeps a pack of these handy. And if your pet is super fussy, they just might like these as treats. In fact, I'm so happy with them, I have already placed another, larger order. The only 'complaint' I have is that they only come in two flavors.

Comment 10 people found this helpful. Was this review helpful to you? [Report abuse](#)

★★★★★ A great invention. why didn't they invent it sooner?!?

By [Joseph D.](#) on April 20, 2012

Verified Purchase

I wish they had invented these ten years ago when my late cat, Jack, needed heart medication three times a day. It would have saved me and the cat much grief getting his pills down.

Customer Images

[See all customer images](#)

Most Recent Customer Reviews

★★★★★ Five Stars

Saved a lot of money (3 cats all on meds) and they were fresh.

Published 1 month ago by A. Duran

★★★★★ Five Stars

Best price I found & so helpful to pill an elderly cat who loves them...

Published 1 month ago by Amazon Customer

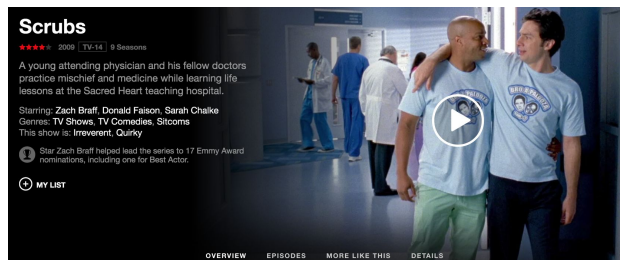
★☆☆☆☆ don't waste your money

★★★★★ **don't waste your money**
don't waste your money. get soft kitty treats and
smash your pill in one of those. we liked these but
they became too expensive when giving our feline a
pill 2x each day.

Published 1 month ago by Robert K. Rutkowski

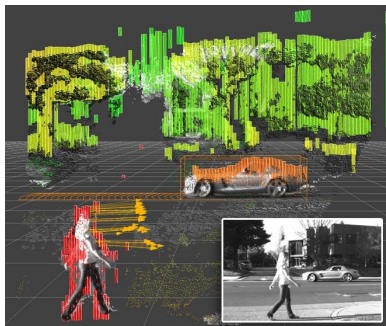
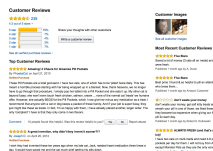
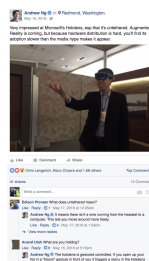
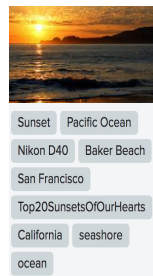
★★★★★ ALWAYS FRESH (and that's saying a

I have two cats on multi-meds and need 4-6 pill pockets per day for them. I will not buy from ANYONE except Monster Pets as they are the only vendor who always sends the [Read more](#)



Data is usually a collection of modalities

- Multimedia web content
- Product recommendation systems
- Robotics



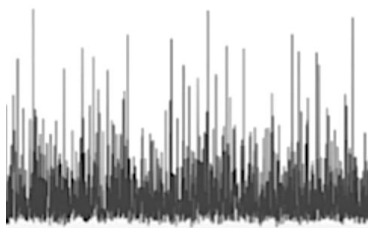
Why is multimodal learning hard?

- Different representations

Images



Real-valued,
Dense



Text

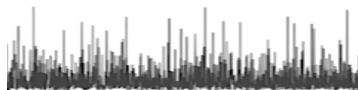


Discrete,
Sparse



Why is multimodal learning hard?

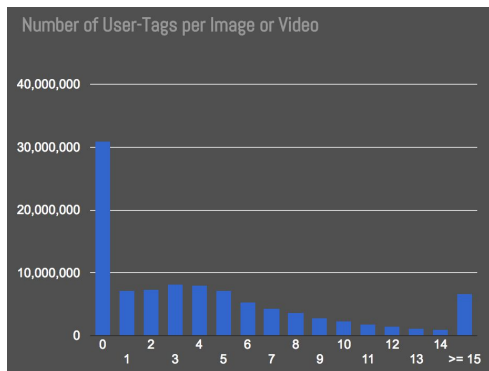
- Different representations



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



- Noisy and missing data



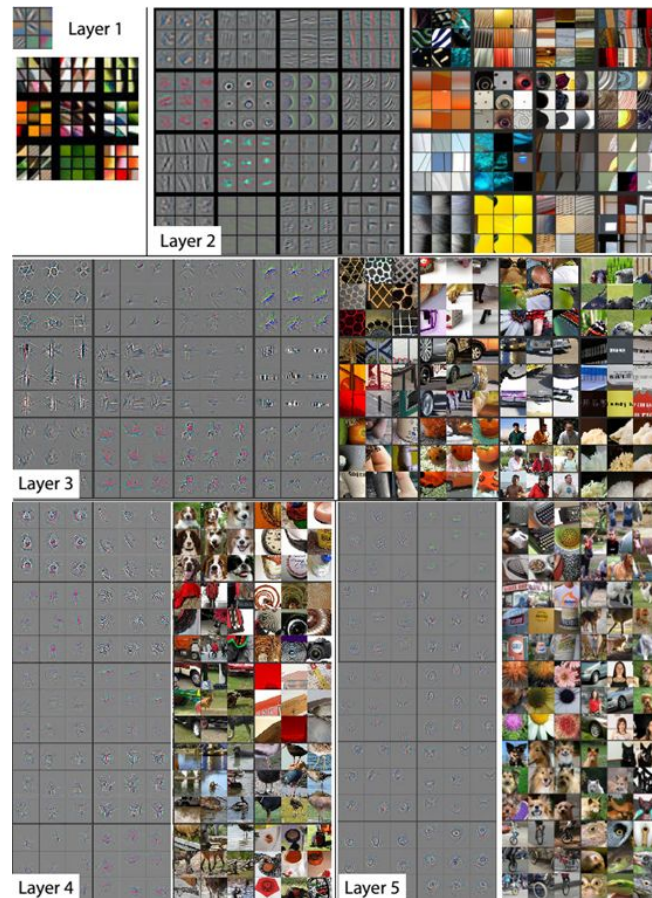
Berlin Pentax *istDL
Pentax-DA 50-200 Cat
Katze Dolores
Backyard smc b/w

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

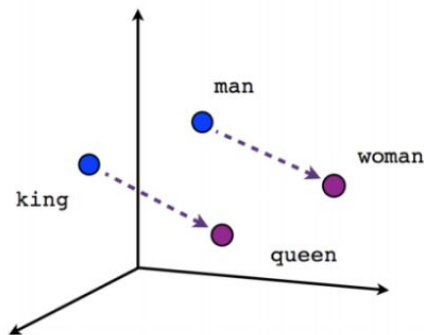


Embeddings

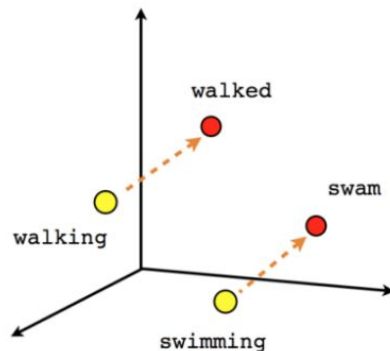
- A way to represent data
- In deep learning, this is usually a high-dimensional vector
- A neural network can take a piece of data and create a corresponding vector in an embedding space
- A neural network can take an embedding vector as an input
- Example: word embeddings

Word embeddings

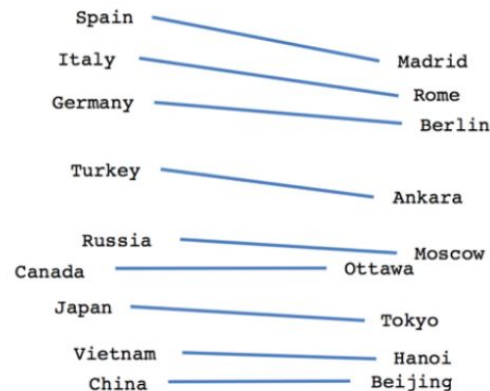
- A word embedding: word \rightarrow high-dimensional vector
- Interesting properties



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

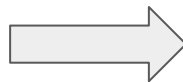


Sunset Pacific Ocean
Nikon D40 Baker Beach
San Francisco
Top20SunsetsOfOurHearts
California seashore
ocean

Text

Flickr tagging: task

Images

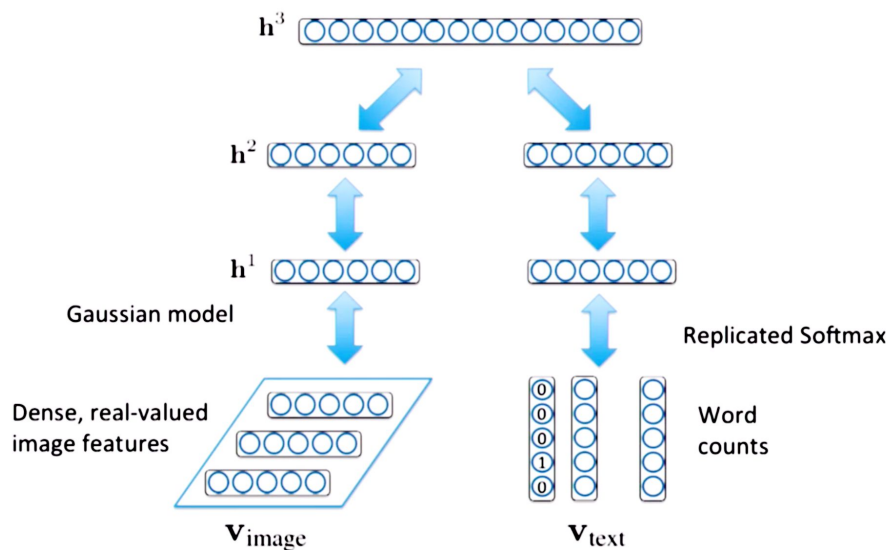


Sunset Pacific Ocean
Nikon D40 Baker Beach
San Francisco
Top20SunsetsOfOurHearts
California seashore
ocean

Text

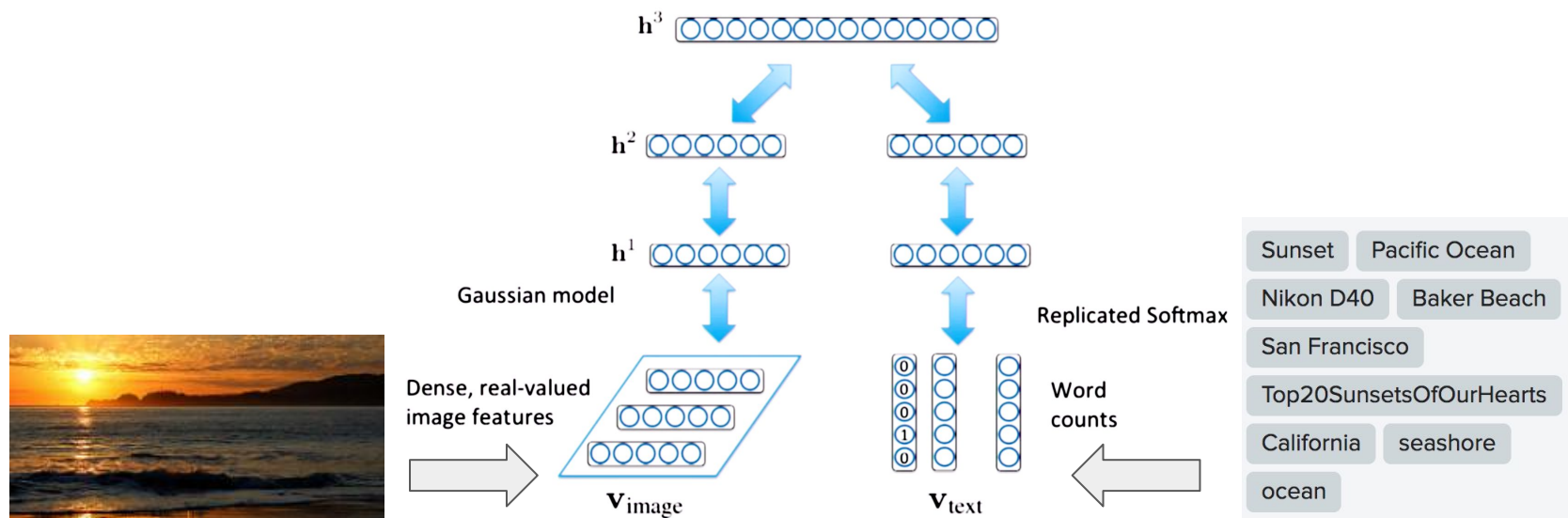
- 1 million images from flickr
- 25,000 have tags

Flickr tagging: model



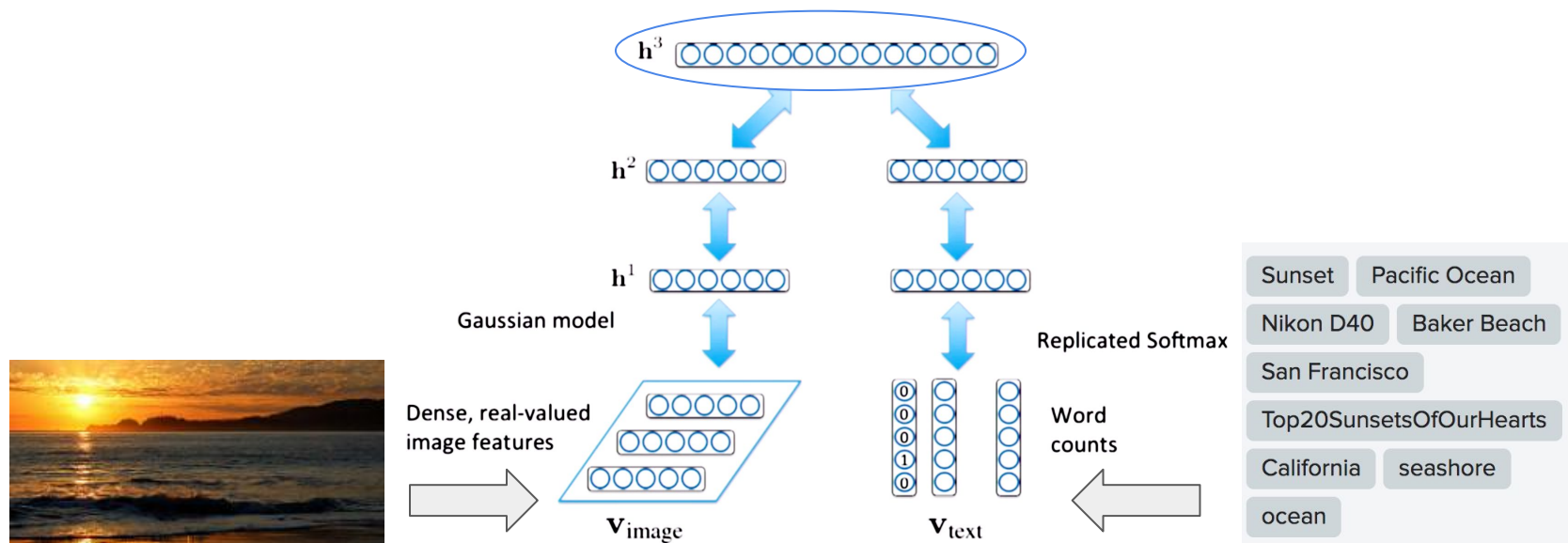
Pretrain unimodal models and combine them at a higher level

Flickr tagging: model



Pretrain unimodal models and combine them at a higher level

Flickr tagging: 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



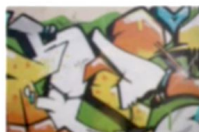
portrait, child, kid,
ritratto, kids, children,
boy, cute, boys, italy

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

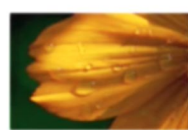
Given

water, red,
sunset

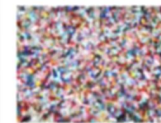
Retrieved



nature, flower,
red, green



blue, green,
yellow, colors



chocolate, cake



Flickr tagging



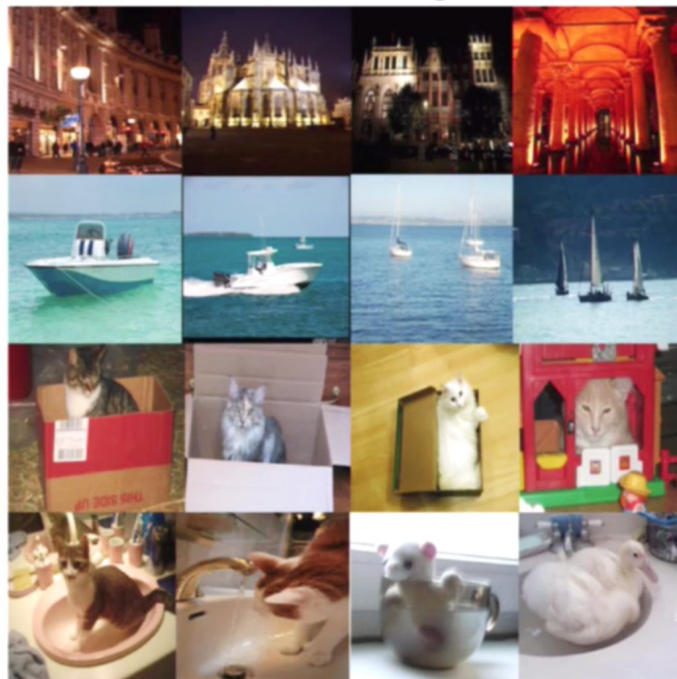
- day + night =

- flying + sailing =

- bowl + box =

- box + bowl =

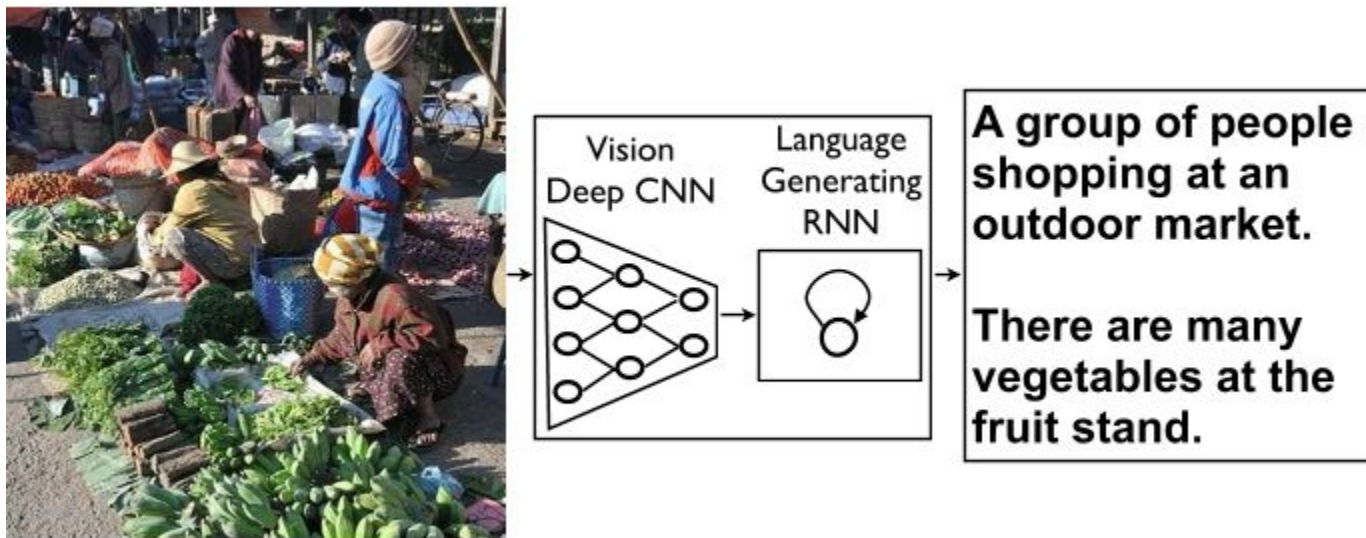
Nearest Images



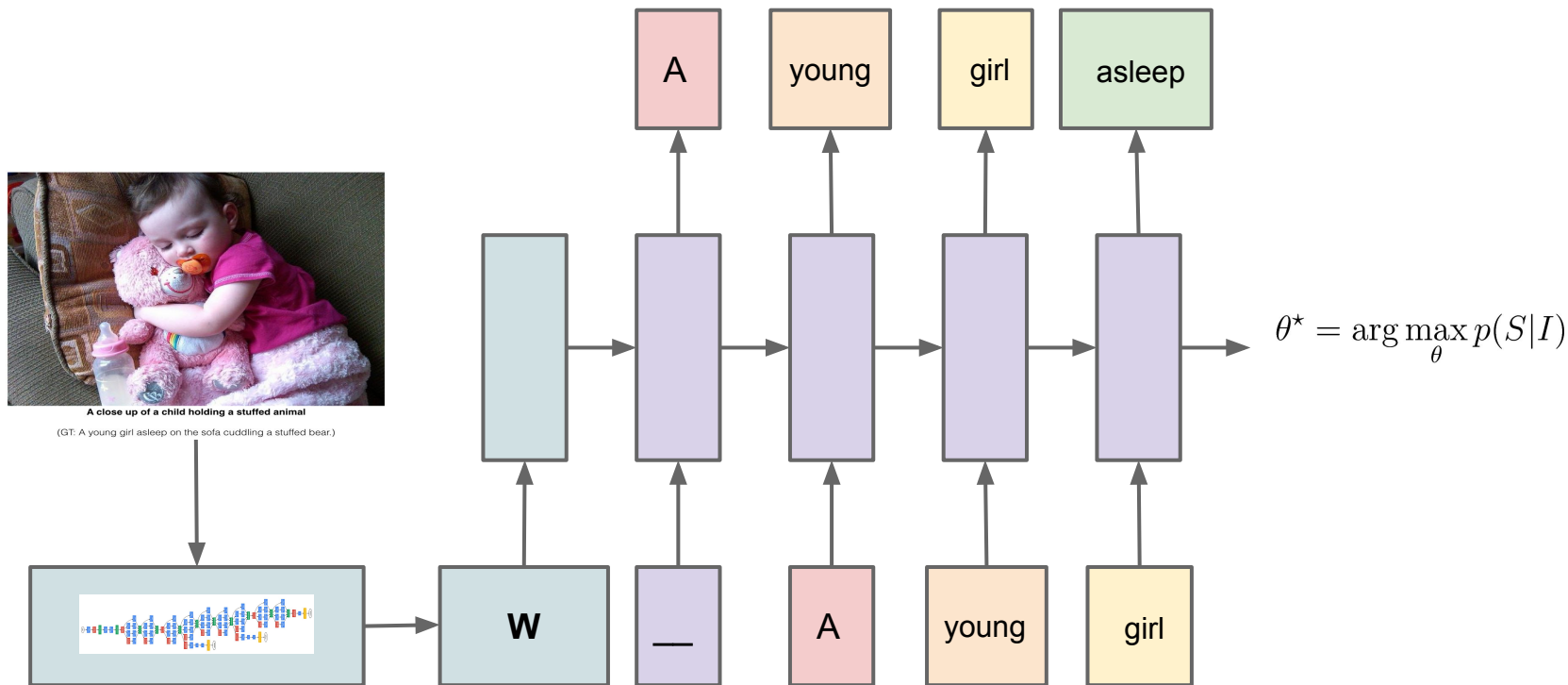
Talk outline

- 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



Example: image captioning





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

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

Computer: A baby is asleep next to a teddy bear.



Human: A close up of two bananas with bottles in the background.

Computer: A bunch of bananas and a bottle of wine.



*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, github.com/ryankiros/neural-storyteller

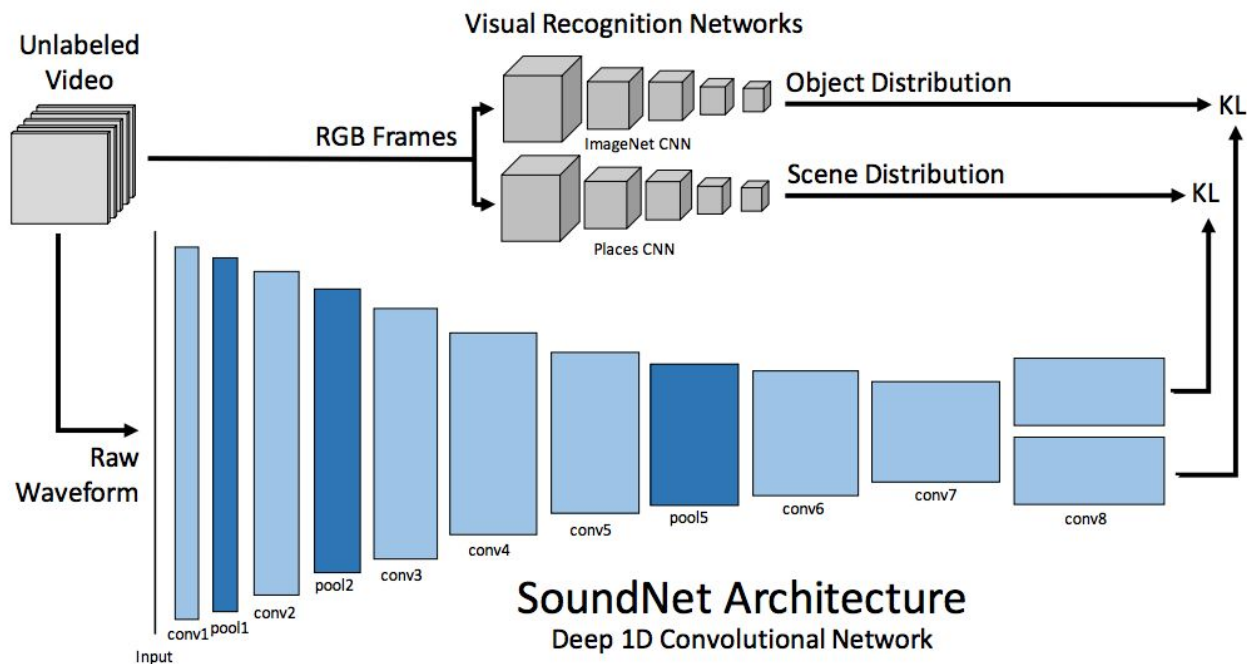
Talk outline

- 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

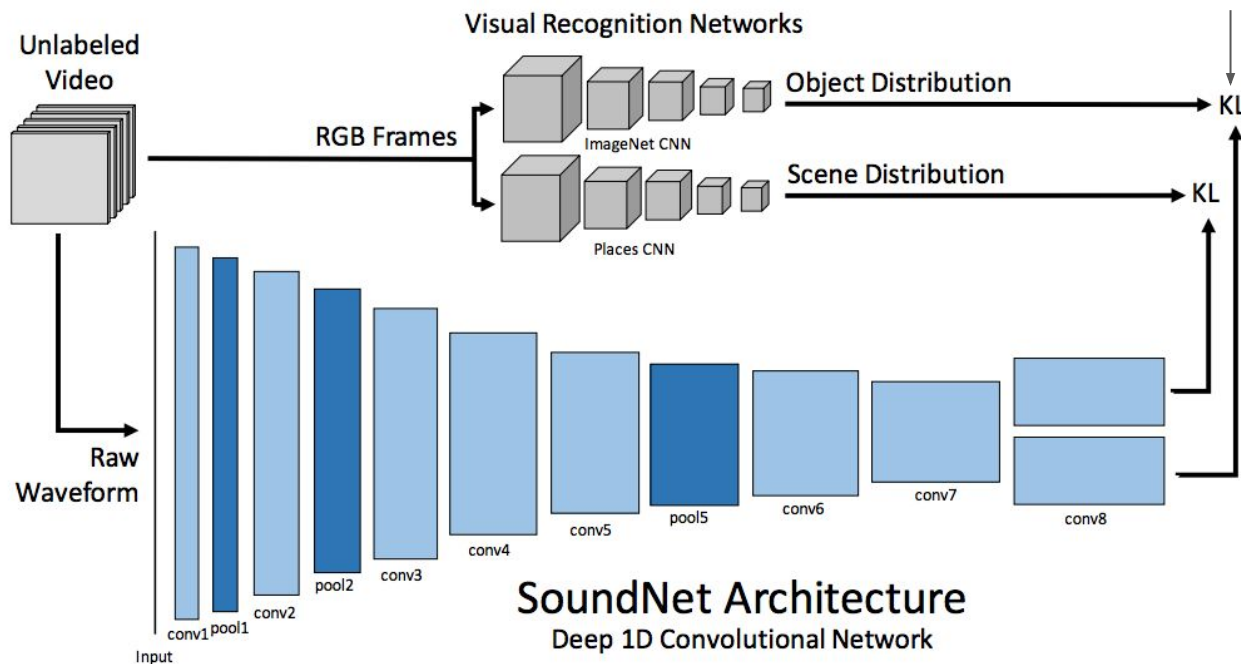
SoundNet training



SoundNet training

Loss for the sound CNN:

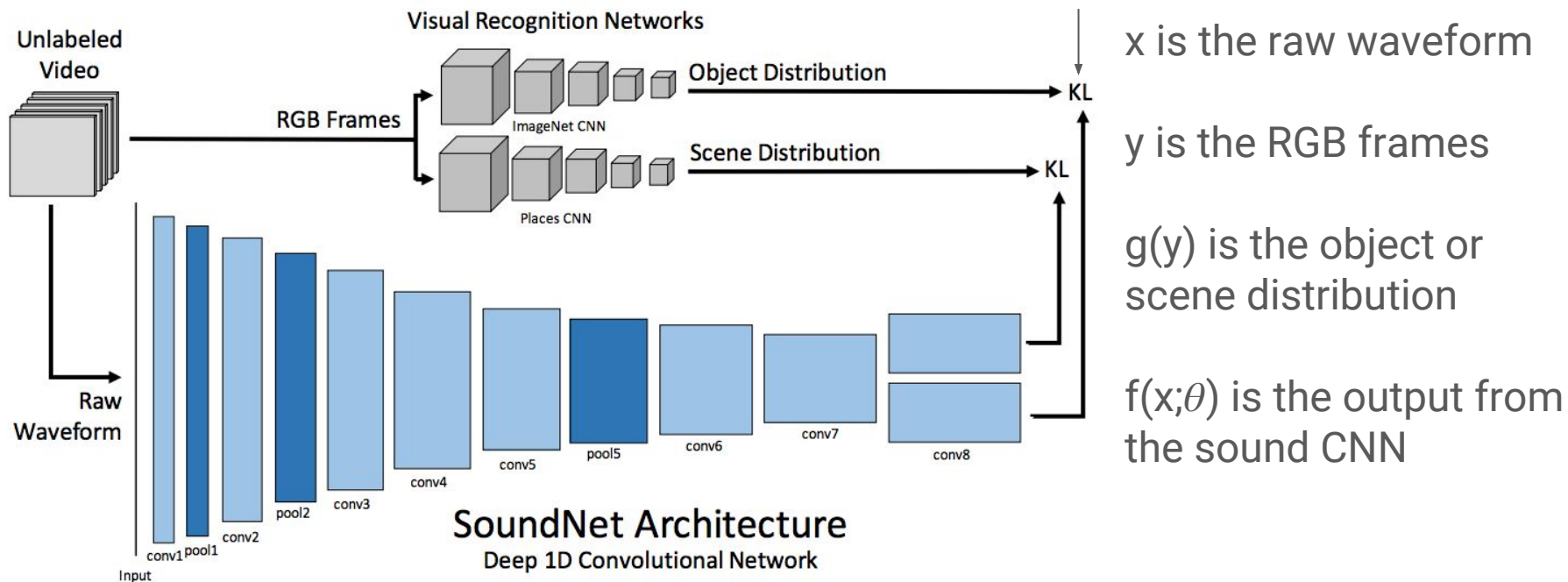
$$D_{KL}(g(y) \parallel f(x; \theta))$$



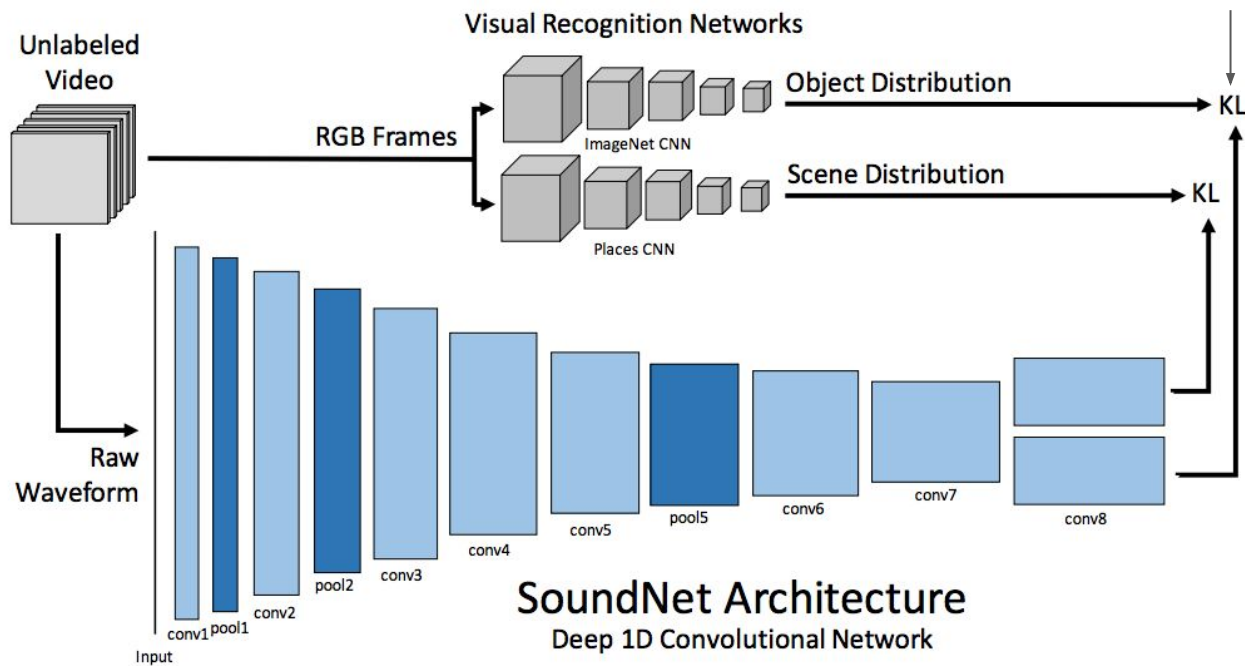
SoundNet training

Loss for the sound CNN:

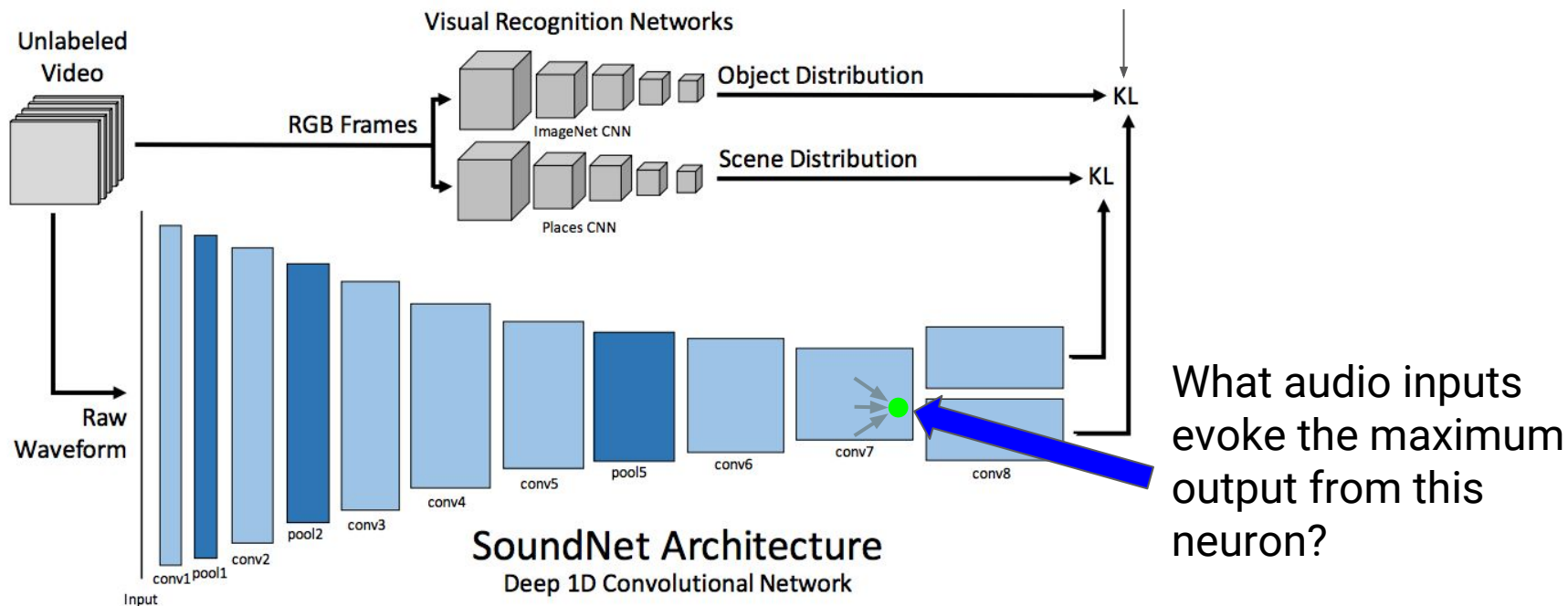
$$D_{KL}(g(y) \parallel f(x; \theta))$$



SoundNet visualization



SoundNet visualization



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)