

Multimodality

Learning from Text, Speech, and Vision

CMU 11-4/611 Natural Language Processing

Lecture 28

April 14, 2020

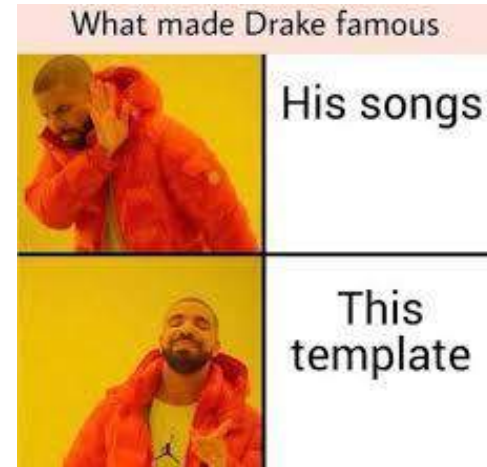
Shruti Palaskar

Outline

- I. What is multimodality?
- II. Types of modalities
- III. Commonly used Models
- IV. Multimodal Fusion and Representation Learning
- V. Multimodal Tasks: Use Cases

I. What is Multimodality?

Human Interaction is Inherently Multimodal



How We Perceive

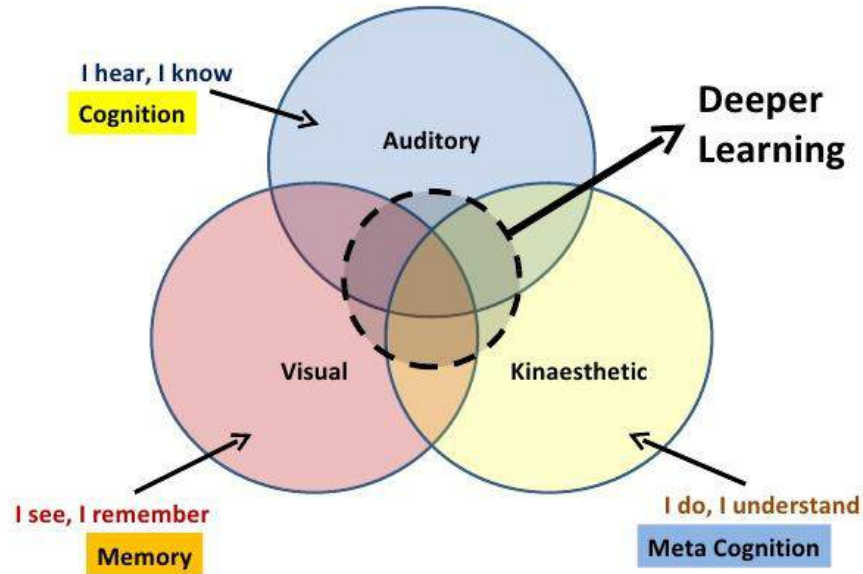


The curse that afflicts abstract painting



How We Perceive

Multi-Modal Learning

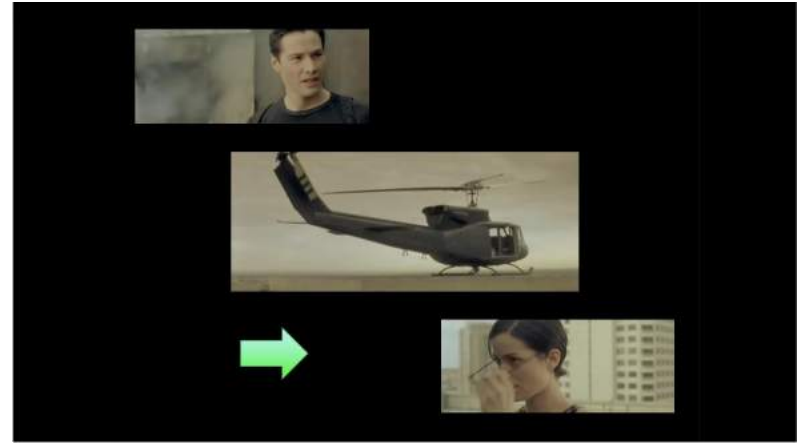


cc Steve Wheeler, University of Plymouth, 2009

The Dream: Sci-Fi Movies



JARVIS



The Matrix

Reality?

Give a caption.



Give a caption.



Human: A Small Dogs Ears Stick Up As It Runs In The Grass.

Model: A Black And White Dog Is Running On Grass With A Frisbee In Its Mouth

Single sentence image description -> Captioning

Give a caption.



Give a caption.

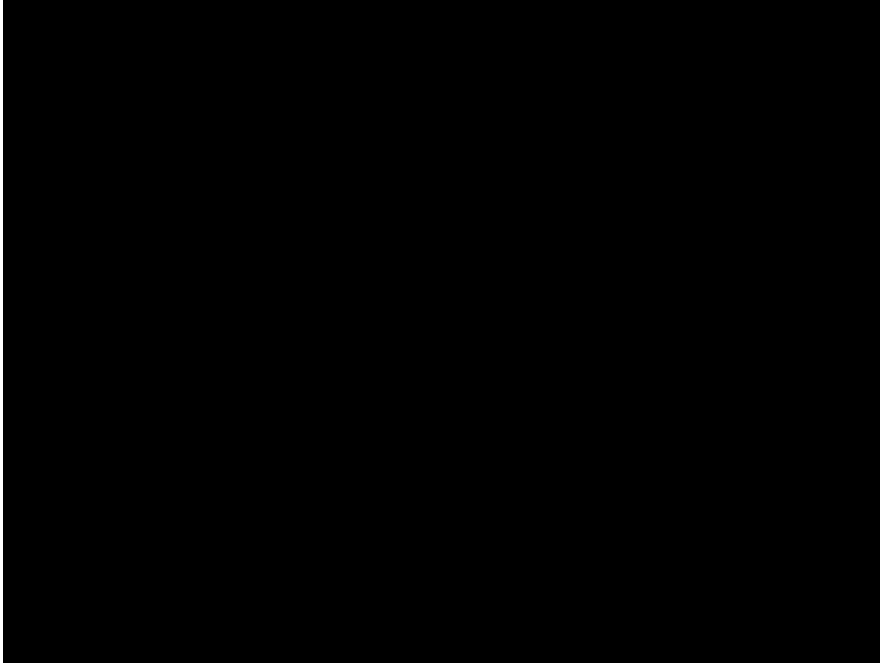


Human: A Young Girl In A White Dress Standing In Front Of A Fence And Fountain.

Model: Two Men Are Standing In Front Of A Fountain

Reality?

Watch the video and answer questions.



QUESTIONS

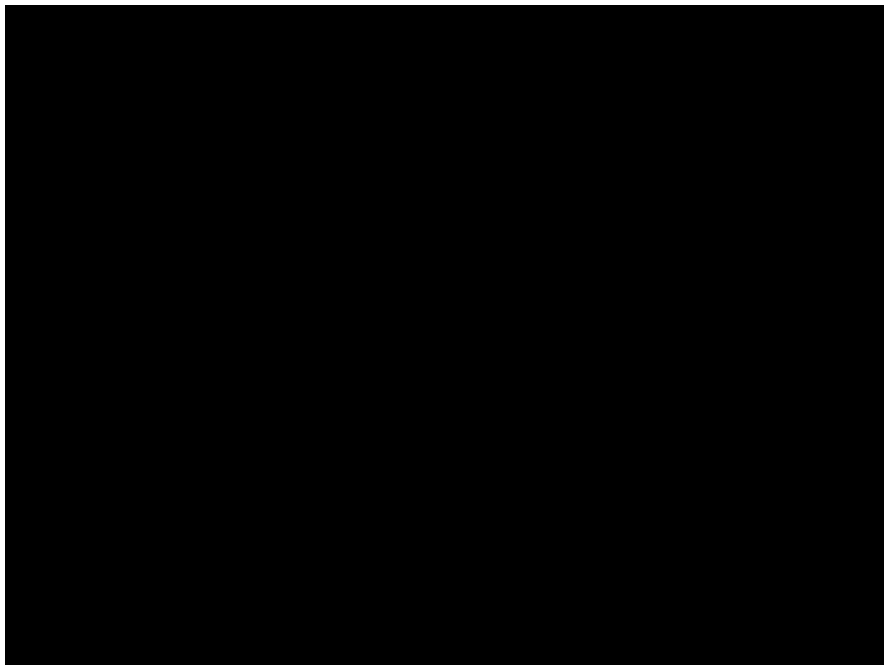
Q. is there only one person ?

Q. does she walk in with a towel around her neck ?

Q. does she interact with the dog ?

Q. does she drop the towel on the floor ?

Watch the video and answer questions.



QUESTIONS

Q. is there only one person ?

A. there is only one person and a dog .

Q. does she walk in with a towel around her neck ?

A. she walks in from outside with the towel around her neck .

Q. does she interact with the dog ?

A. she does not interact with the dog

Q. does she drop the towel on the floor ?

A. she dropped the towel on the floor at the end of the video .

Simple questions, simple answers -> Video
Question Answering

Reality? Baby Steps. Still a long way to go.

...Challenges

Common challenges based on the tasks we just saw

- Training Dataset bias
- Very complicated tasks
- Lack of common sense reasoning within models
- No world knowledge available like humans do
 - Physics, Nature, Memory, Experience

How do we teach machines to perceive?

Outline

- I. What is multimodality?
- II. Types of modalities
- III. Commonly used Models
- IV. Multimodal Fusion and Representation Learning
- V. Multimodal Tasks: Use Cases

II. Types of modalities

Types of Modalities



IMAGE/VIDEO



TEXT

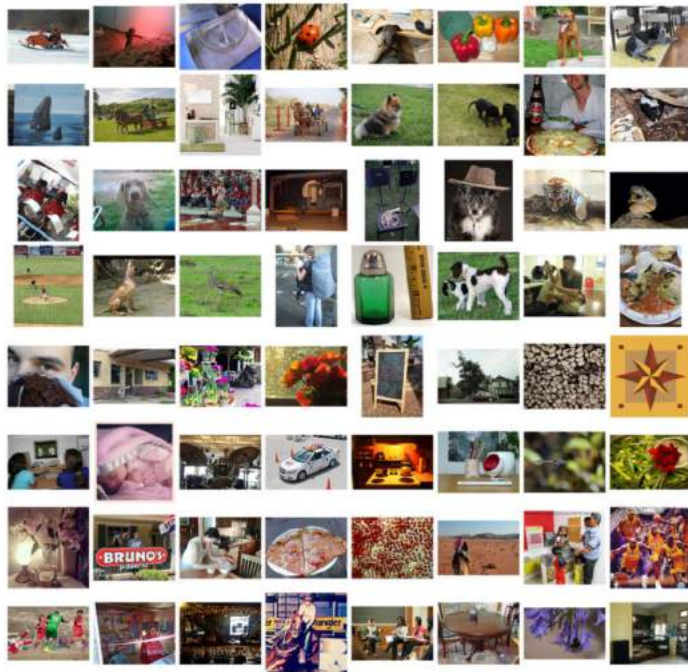


SPEECH/AUDIO



EMOTION/AFFECT
/SENTIMENT

Example Dataset: ImageNet



- Object Recognition
- Image Tagging/Categorization
- ~14M images
- Knowledge Ontology
- Hierarchical Tags
 - Mammal -> Placental -> Carnivore -> Canine -> Dog -> Working Dog -> Husky

Example Dataset: How2 Dataset



I'm very close to the green but I didn't get it on the green so now I'm in this grass bunker.

Eu estou muito perto do green, mas eu não pus a bola no green, então agora estou neste bunker de grama.

In golf, get the body low in order to get underneath the golf ball when chipping out of thick grass from a side hill lie.

- Speech
- Video
- English Transcript

- Portuguese Transcript
- Summary

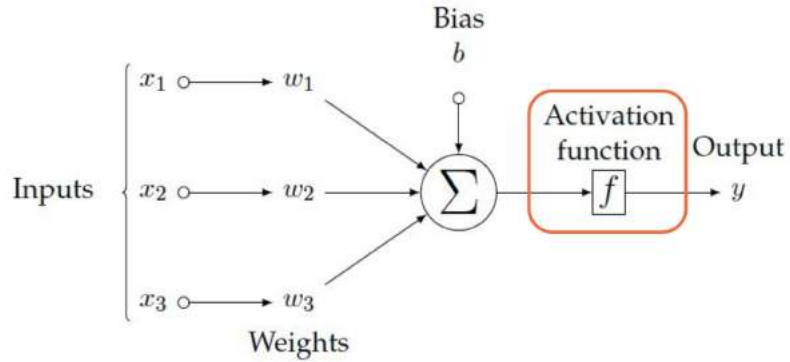
Example Dataset: Open Pose



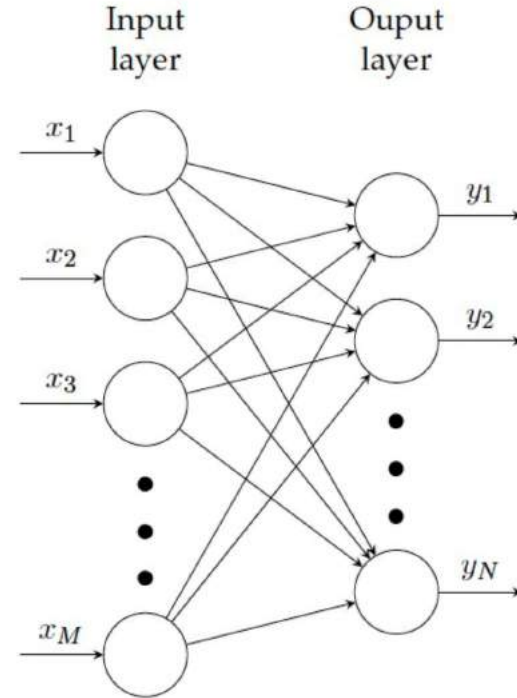
- Action Recognition
- Pose Estimation
- Human Dynamic
- Body Dynamics

III. Commonly Used Models

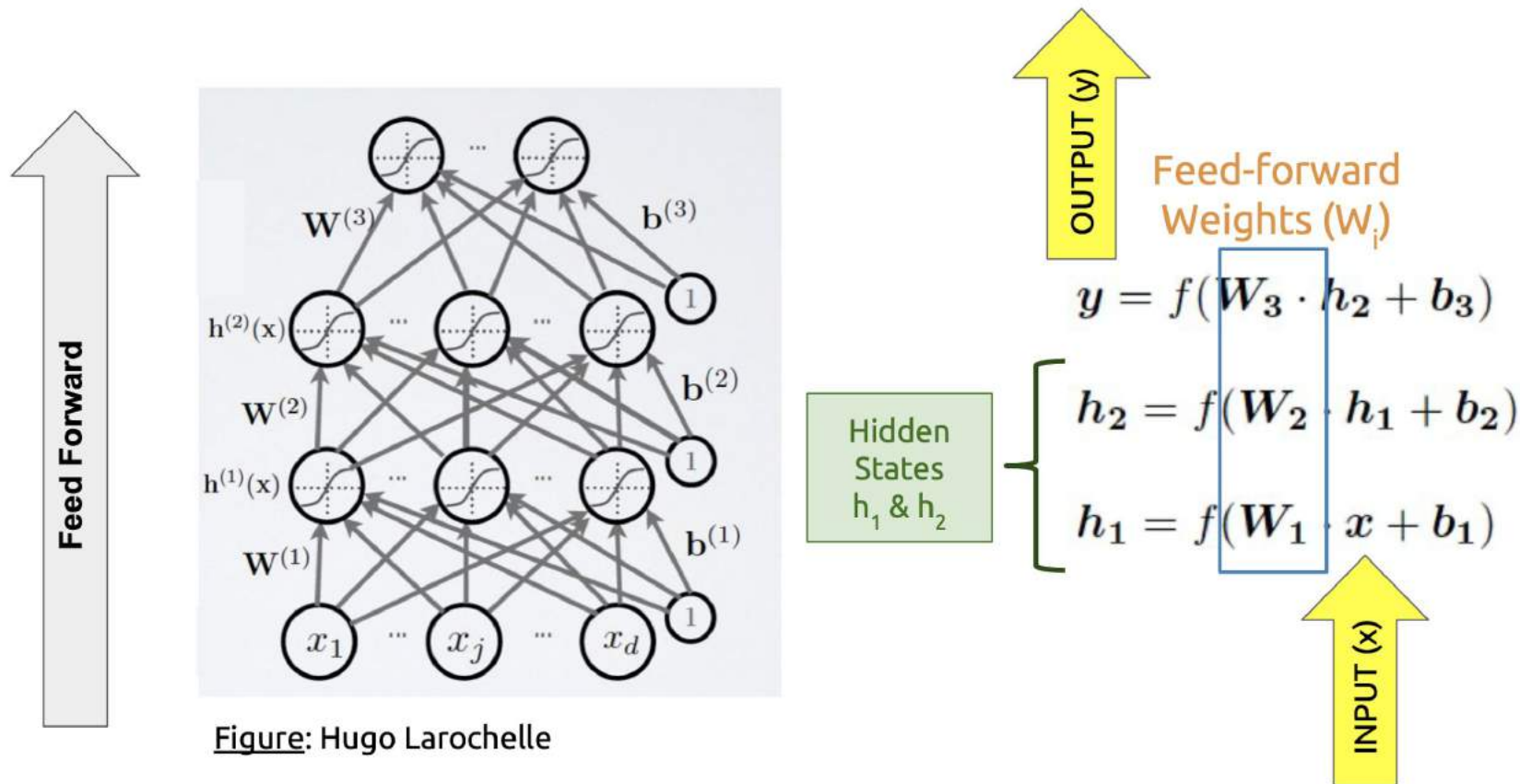
Multilayer Perceptrons



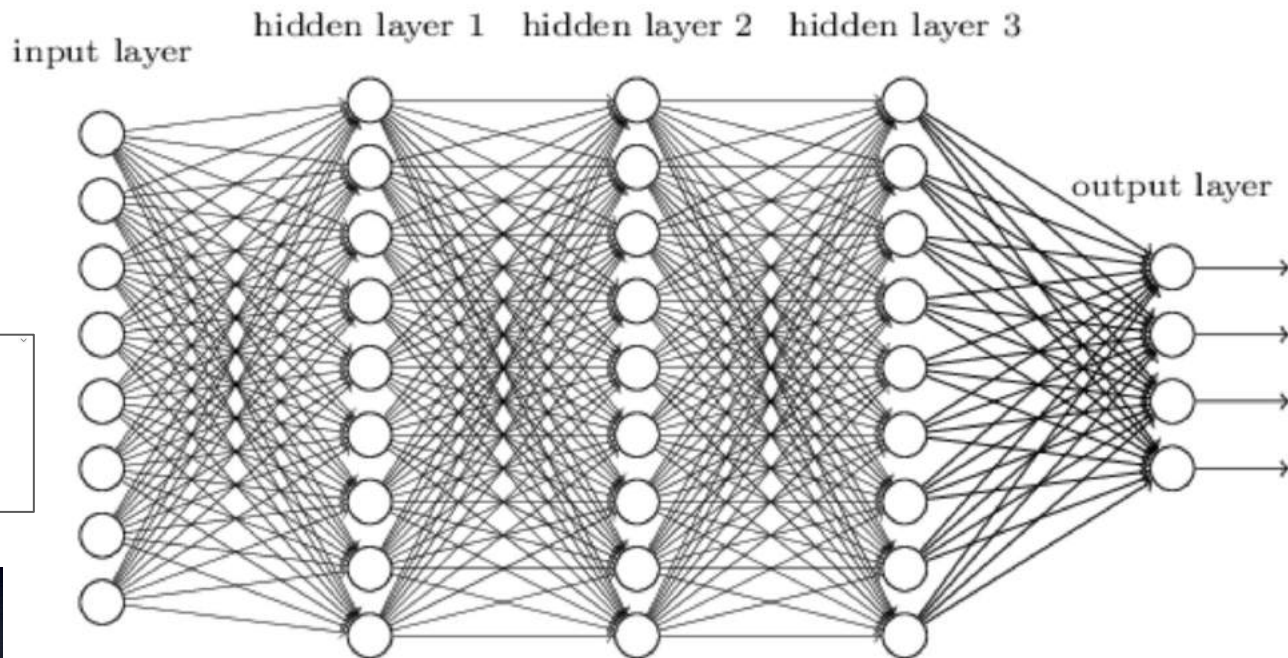
Single Perceptron



Multilayer Perceptrons



Multilayer Perceptrons: Uses in Multimedia



Multilayer Perceptrons: Limitations

Limitation #1

Very large amount of input data samples (x_i), which requires a gigantic amount of model parameters.

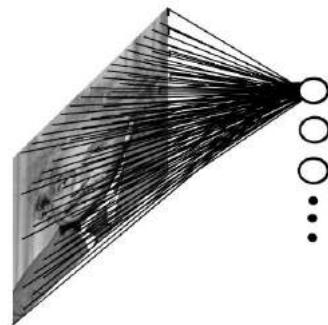
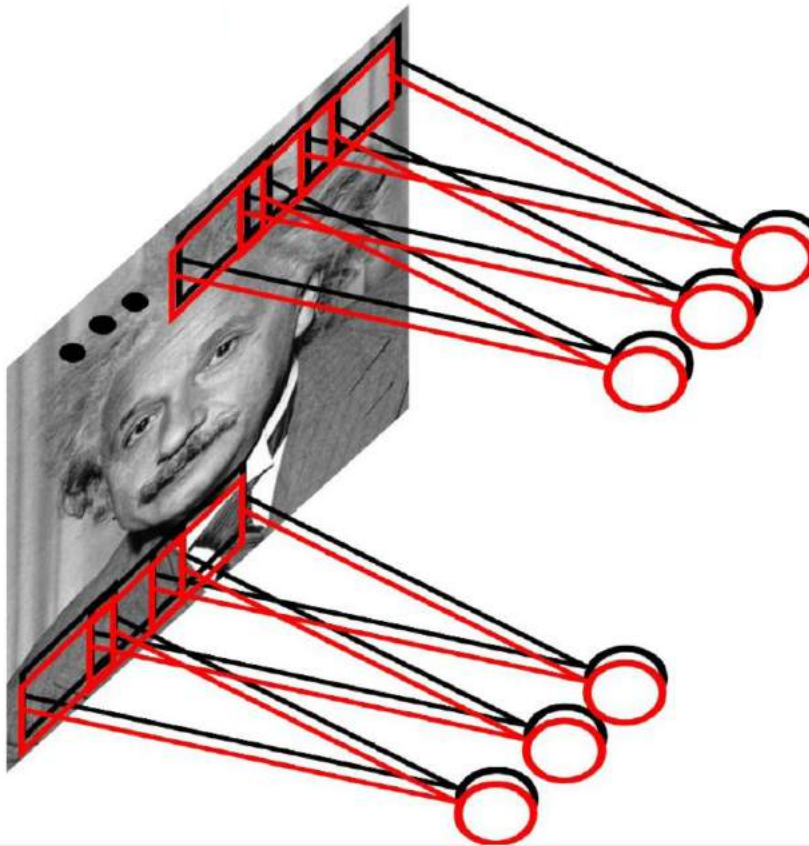


Figure: Ranzatto

Convolutional Neural Networks (CNNs)

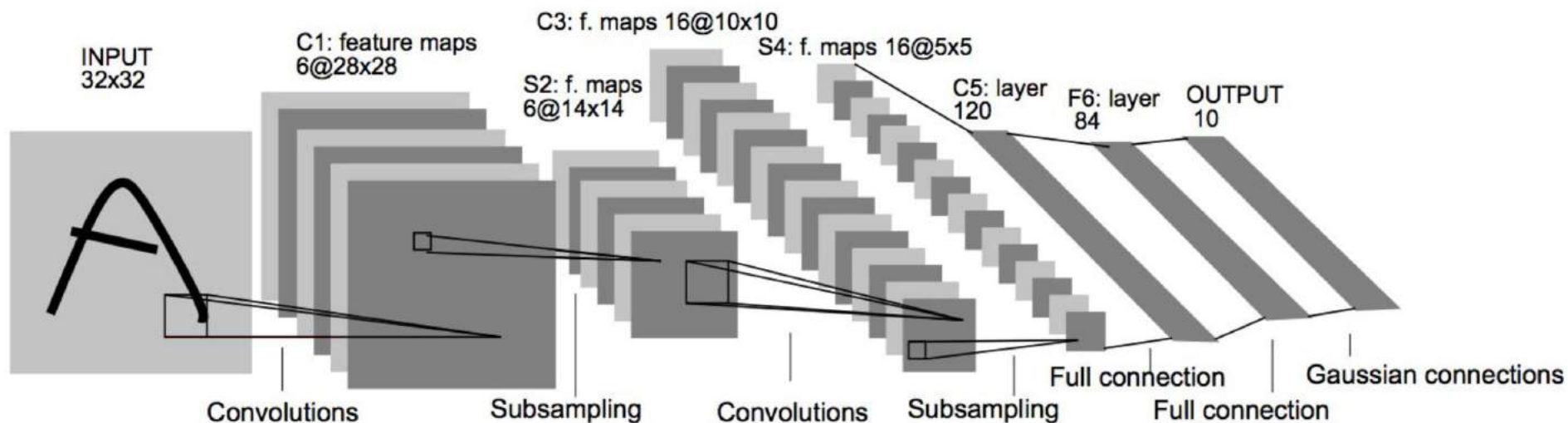


Translation invariance: we can use same parameters to capture a specific “feature” in any area of the image. We can use different sets of parameters to capture different features.

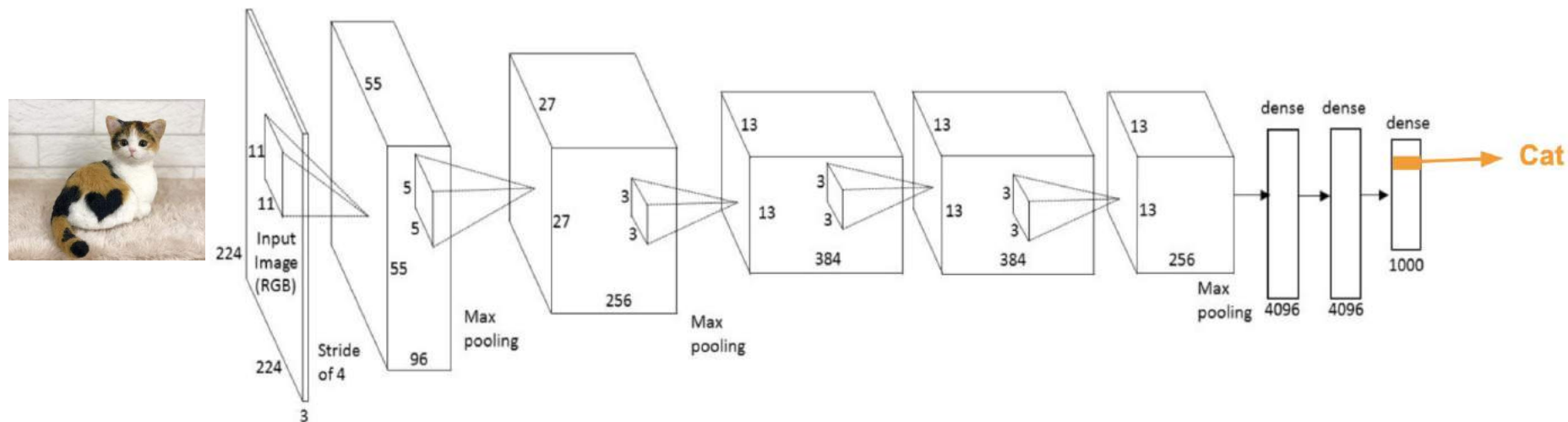
These operations are equivalent to perform **convolutions** with different filters.

Convolutional Neural Networks (CNNs)

LeNet-5



Convolutional Neural Networks (CNNs) for Image Encoding



Multilayer Perceptrons: Limitations

Limitation #1

Very large amount of input data samples (x_i), which requires a gigantic amount of model parameters.

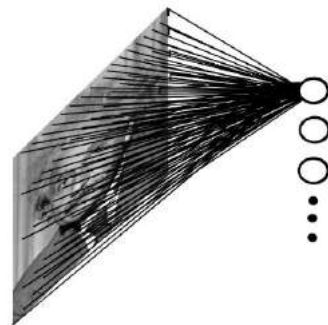


Figure: Ranzatto

Limitation #2

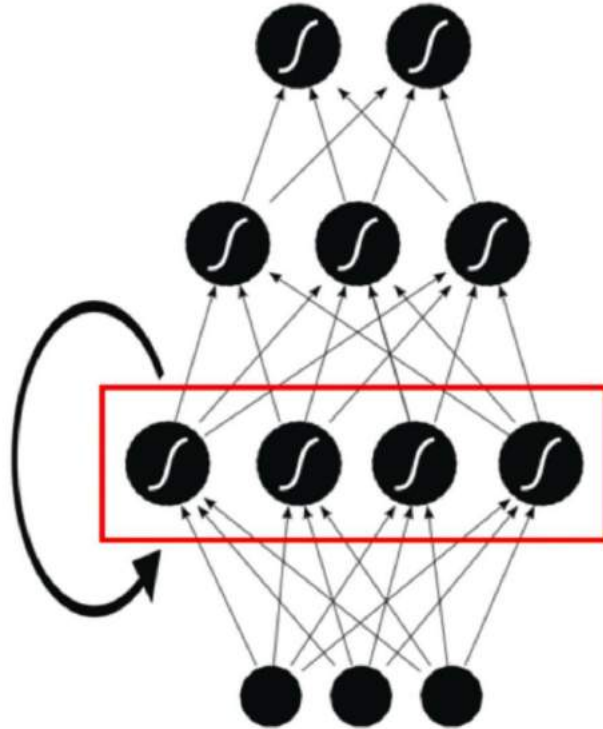
Does not naturally handle input data of variable dimension

(eg. audio/video/word sequences)

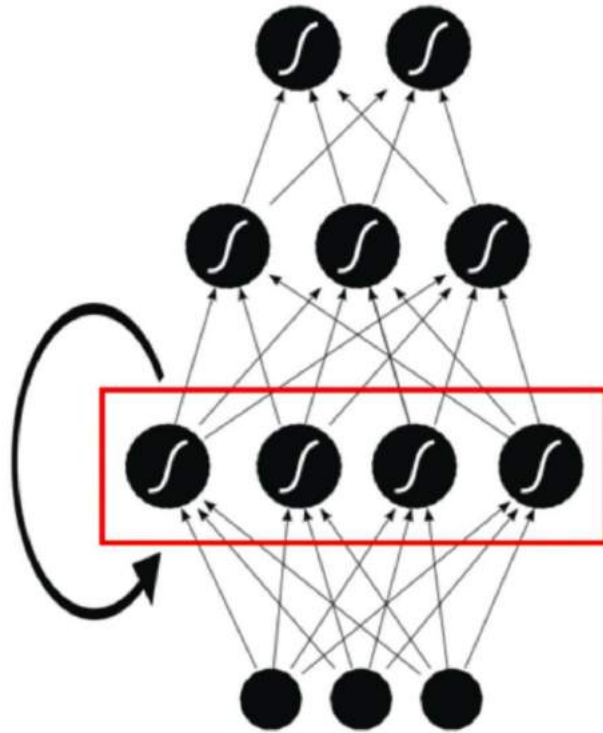
Recurrent Neural Networks

Build specific connections
capturing the temporal
evolution

→ **Shared weights in time**



Recurrent Neural Networks



Feed-forward
Weights (W)

$$h_t = f(\boxed{W} \cdot x_t + \boxed{U} \cdot h_{t-1} + b)$$

Recurrent
Weights (U)

Updated
state

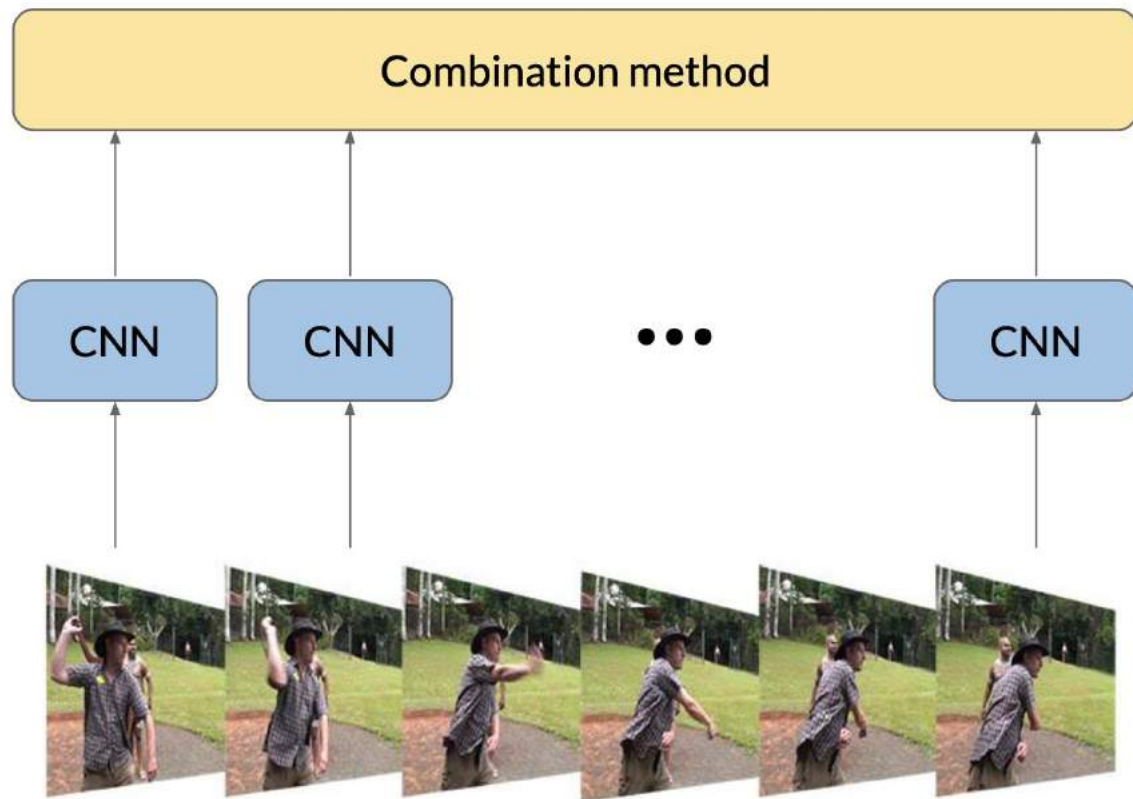
$$h_t = f(W \cdot x_t + U \cdot h_{t-1} + b)$$

Previous
state

$$h_t = f(W \cdot x_t + U \cdot h_{t-1} + b)$$



Recurrent Neural Networks for Video Encoding

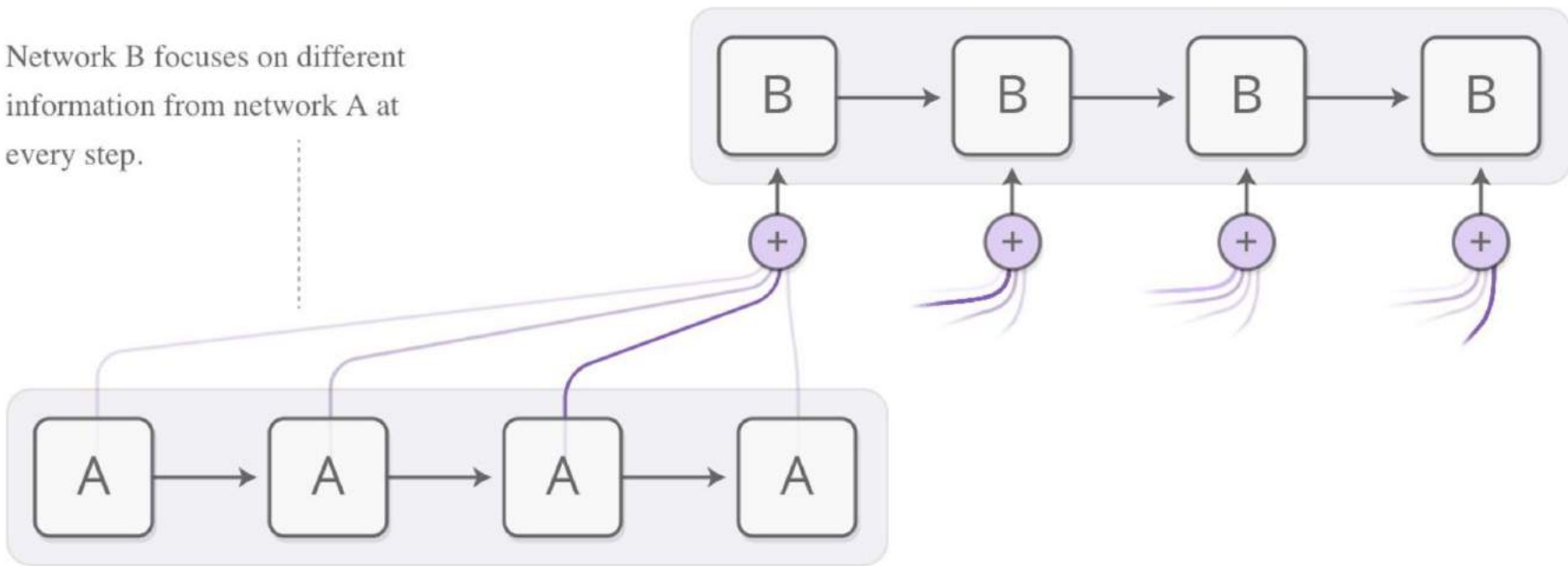


Combination is commonly implemented as a small NN on top of a pooling operation (e.g. max, sum, average).

Recurrent Neural Networks are well suited for processing sequences.

Attention Mechanism

Network B focuses on different information from network A at every step.



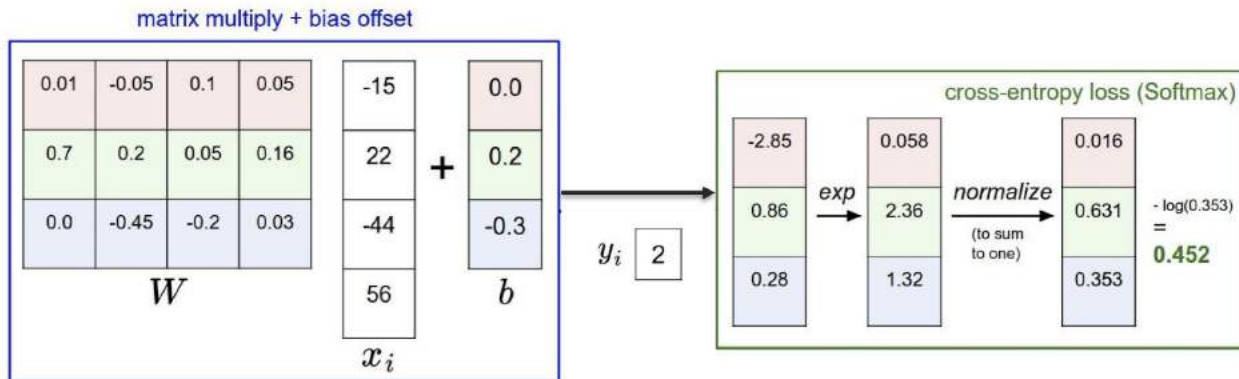
Loss Function: Softmax

Cross-entropy loss:

$$L_i = -\log \left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}} \right)$$

Softmax function

Minimizing the negative log likelihood.

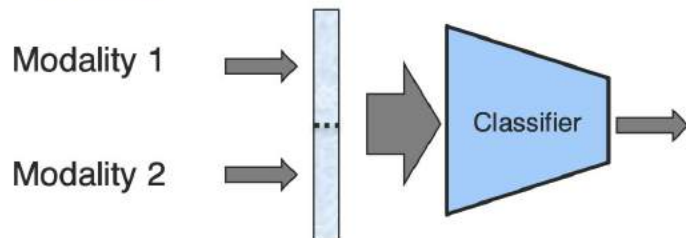


IV. Multimodal Fusion & Representation Learning

Fusion: Model Agnostic

A Model-Agnostic Approaches

1) Early Fusion



2) Late Fusion

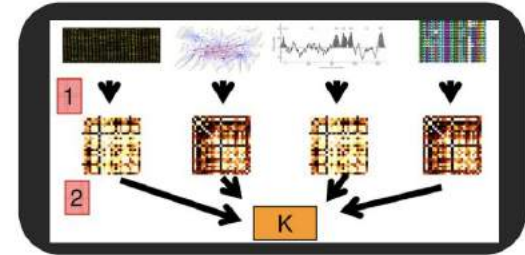


Fusion: Model Based

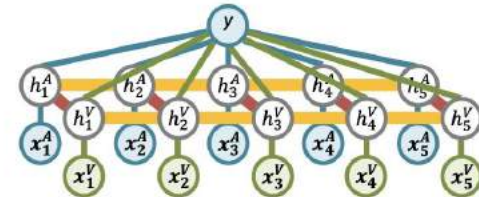
Definition: To join information from two or more modalities to perform a prediction task.

B Model-Based (Intermediate) Approaches

- 1) Deep neural networks
- 2) Kernel-based methods
- 3) Graphical models

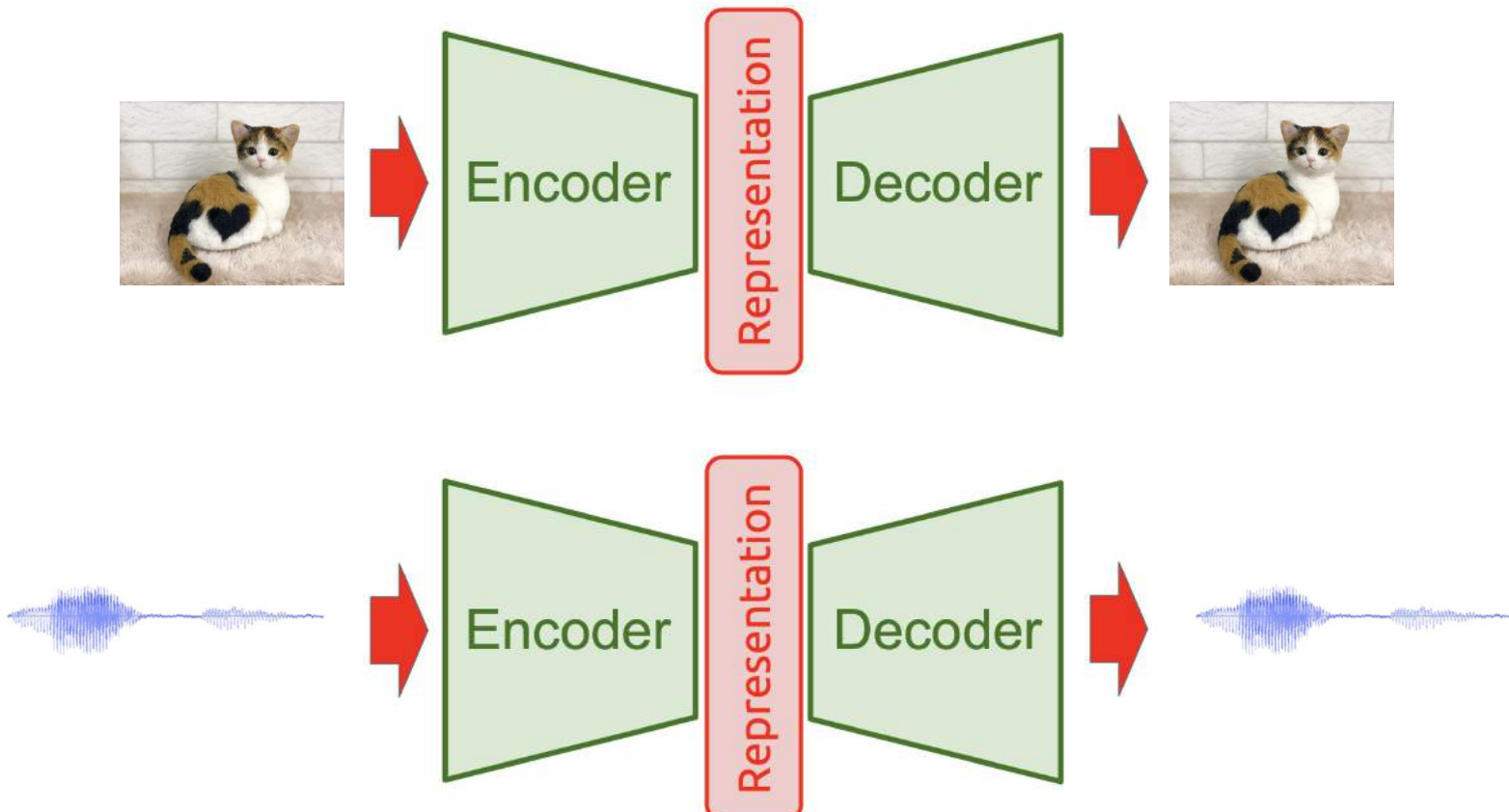


Multiple kernel learning

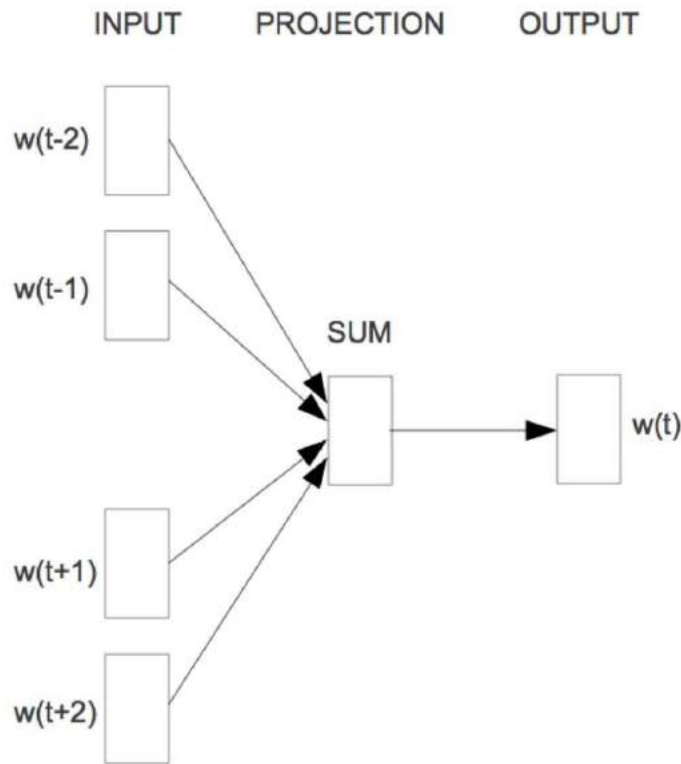


Multi-View Hidden CRF

Representation Learning: Encoder-Decoder



Representation Learning



the cat climbed a tree

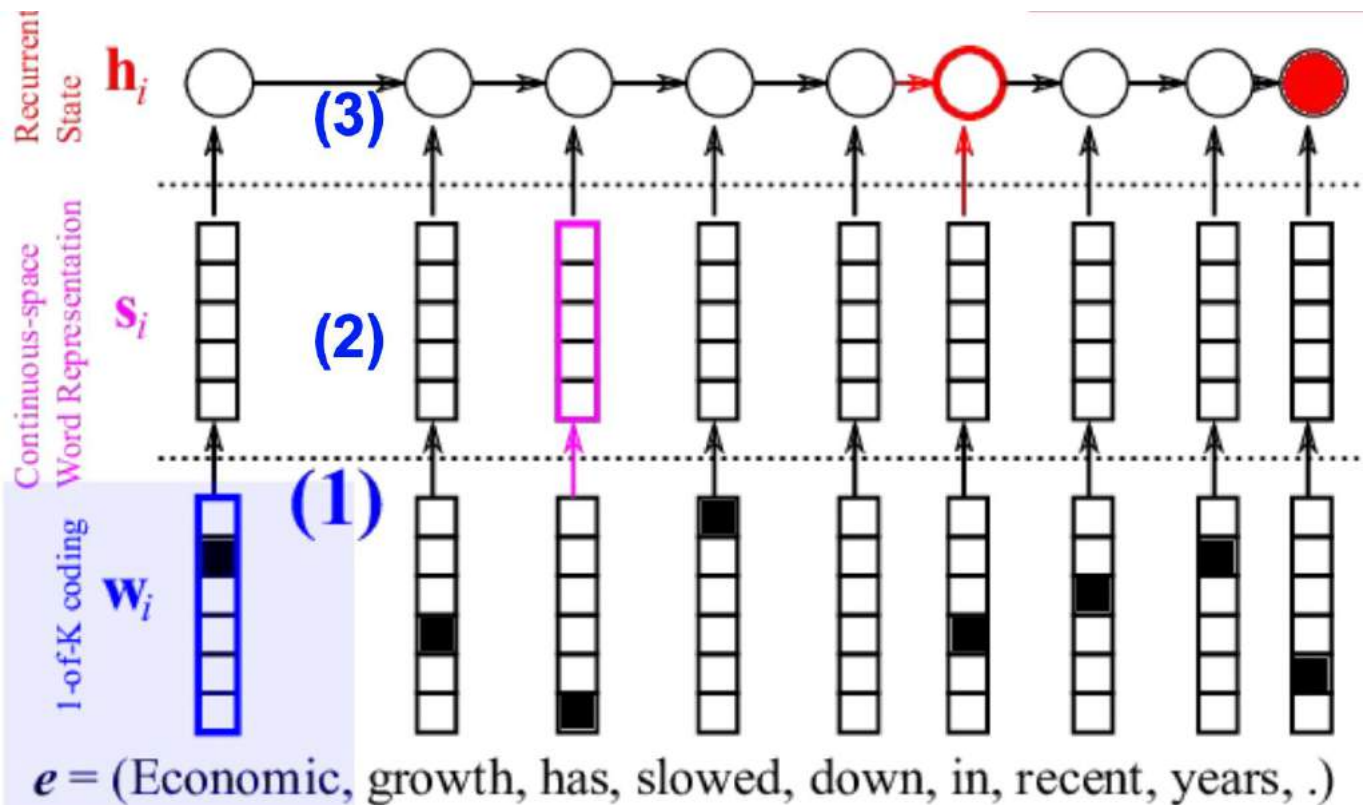
Given context:

a, cat, the, tree

Estimate prob. of
climbed

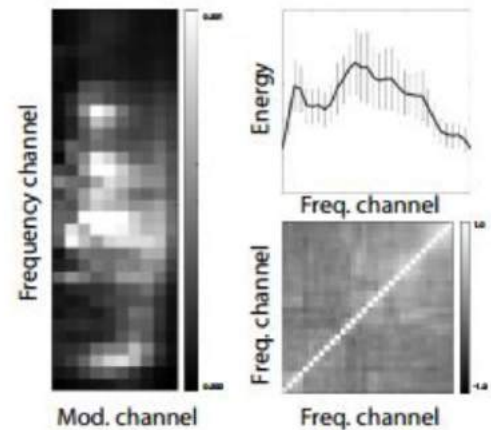
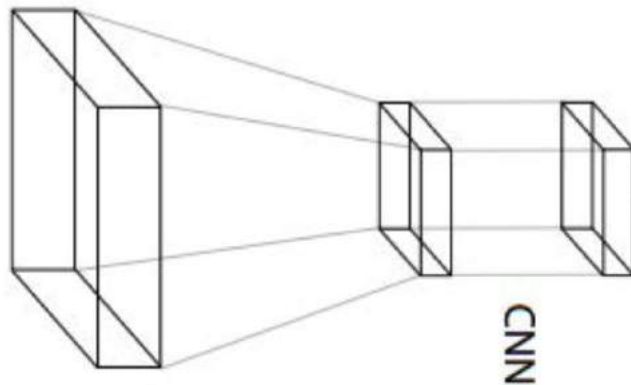
Word2Vec

Representation Learning: RNNs



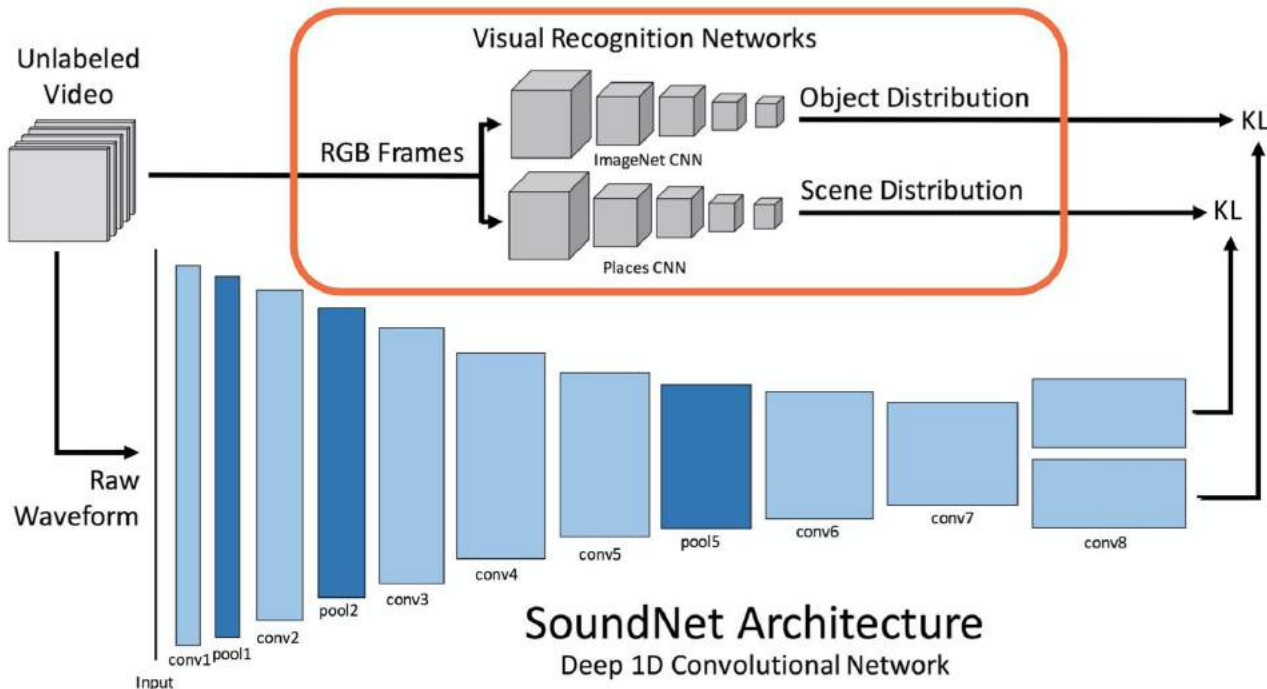
Representation Learning: Self-Supervised

Use videos to train a CNN that predicts the audio statistics of a frame.

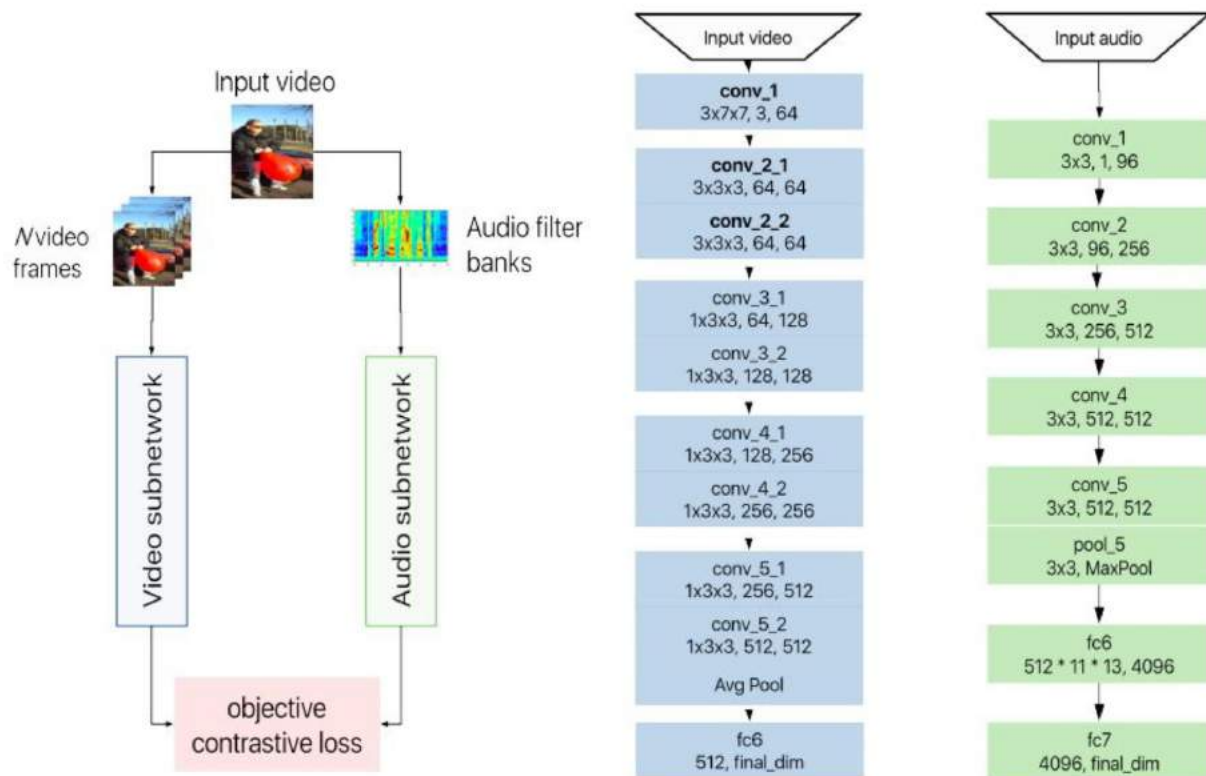


Representation Learning: Transfer Learning

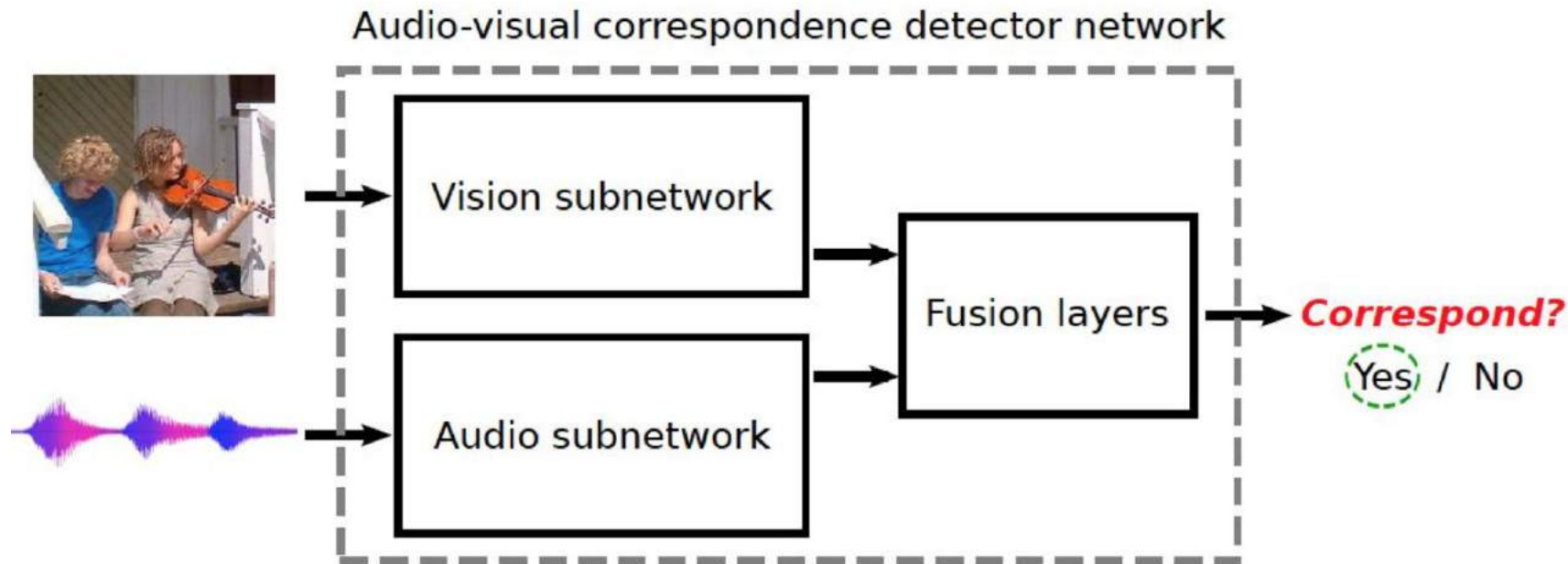
Teacher network: Visual Recognition (object & scenes)



Representation Learning: Joint Learning



Representation Learning: Joint Learning (Similarity)



V. Common Tasks, Use Cases

V. Common Tasks

1. Vision and Language
2. Speech, Vision and Language
3. Multimedia
4. Emotion and Affect

- Image/Video Captioning
- Visual Question Answering
- Visual Dialog
- Video Summarization
- Lip Reading
- Audio Visual Speech Recognition
- Visual Speech Synthesis
- ...

1. Vision and Language Common Tasks

Image Captioning

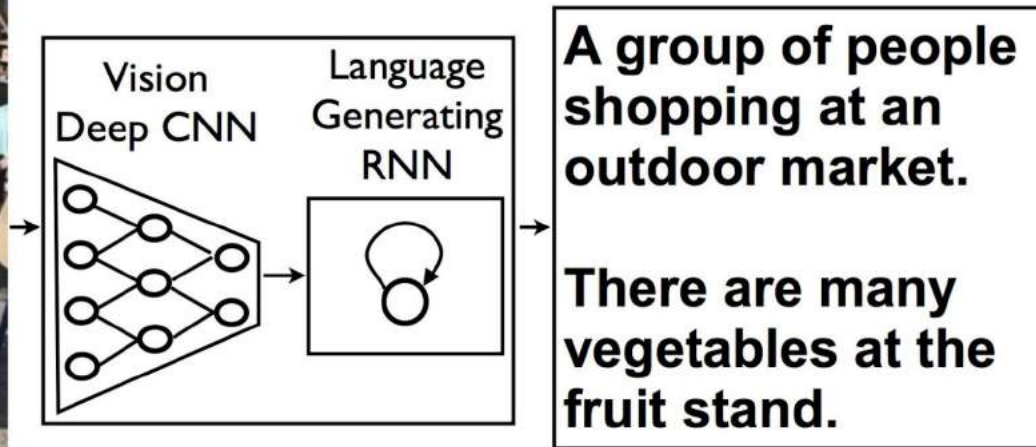
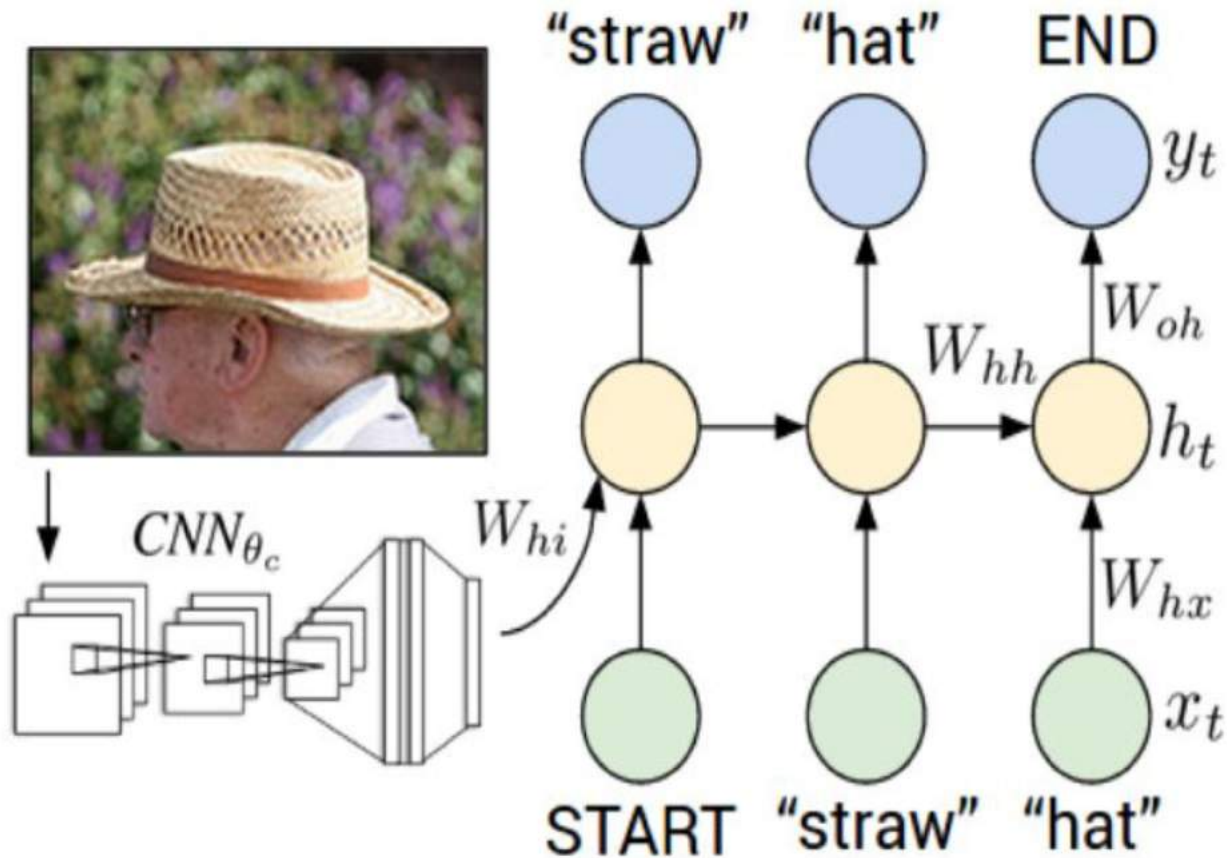


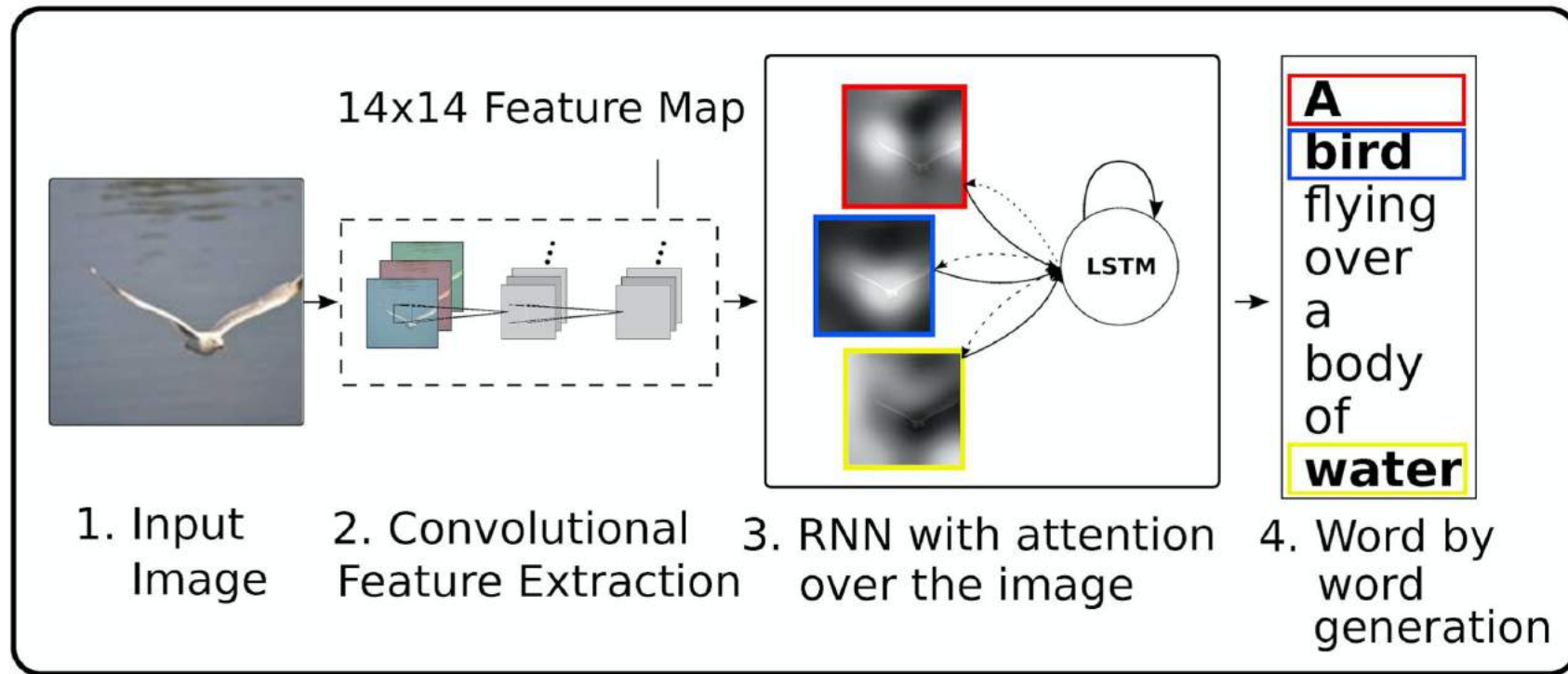
Image Captioning



Karpathy et al. 2015

Slides by Marc Bolaños 54

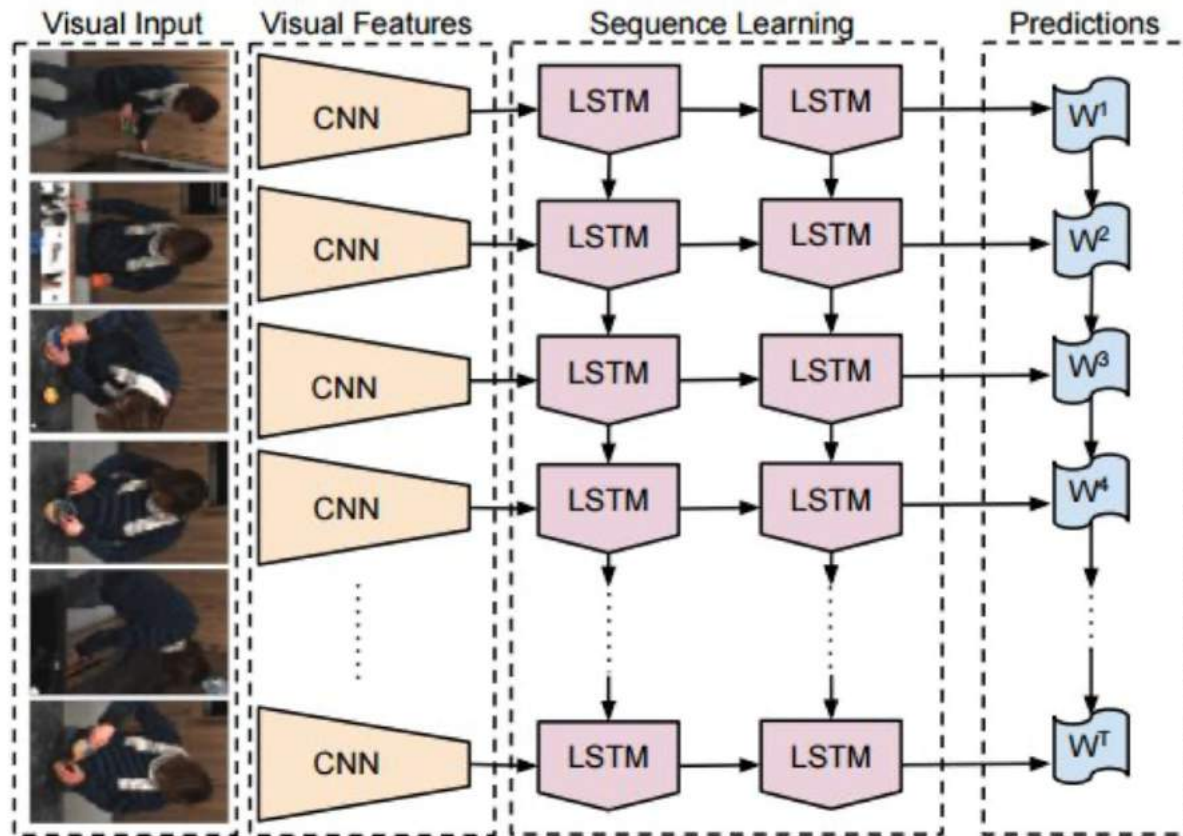
Image Captioning: Show, Attend and Tell



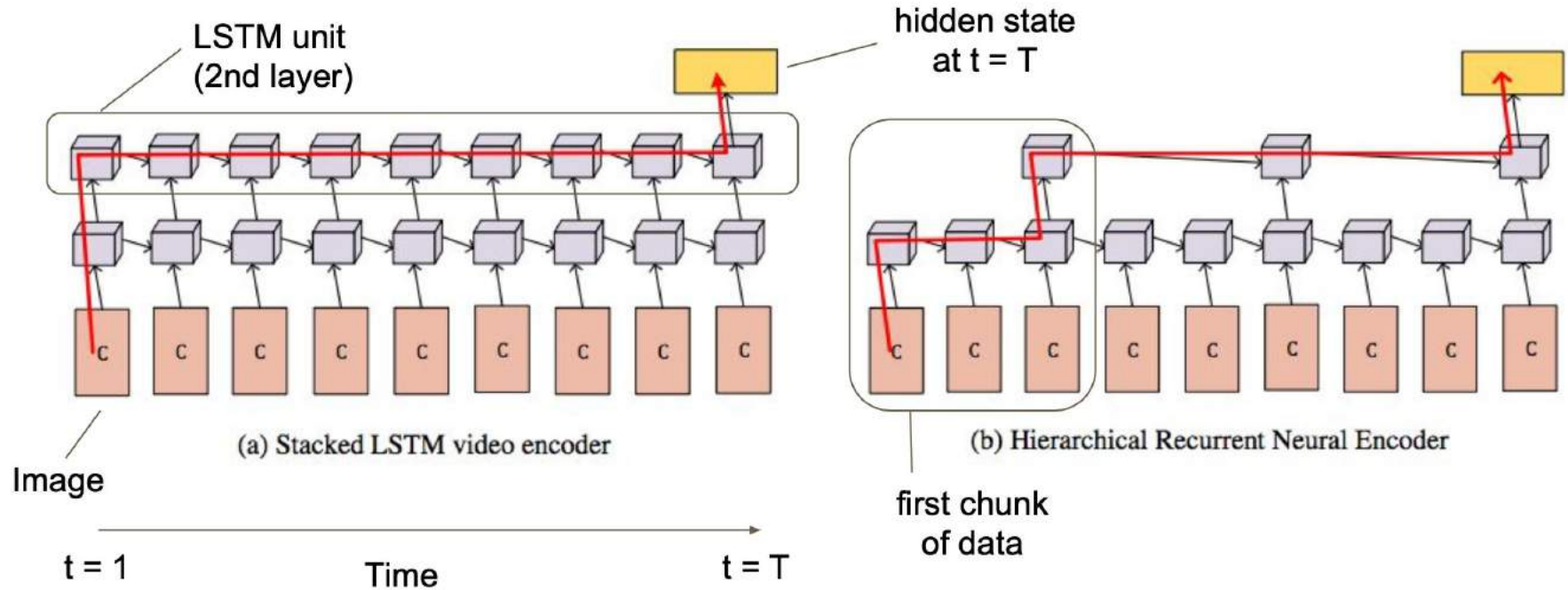
a blue cup on a table
 a plate of food
 a small bowl of sauce
 a bowl of soup
 a plate of food
 a bowl of cheese
 a slice of meat
 a glass of water
 a blue bowl with red sauce
 a cup of coffee
 a silver metal cup
 food on a plate
 yellow and white cheese
 a bowl of orange

a plate of food. food on a plate. a blue cup on a table. a plate of food. a blue bowl with red sauce. a bowl of soup. a cup of coffee. a bowl of chocolate. a glass of water. a plate of food. a silver metal container. a small bowl of sauce. table with food on it. a slice of orange. a table with food on it. a slice of meat. yellow and white cheese.

Video Captioning



Video Captioning



Pan et al. 2016

Visual Question Answering



What is the mustache
made of?

AI System

Visual Question Answering

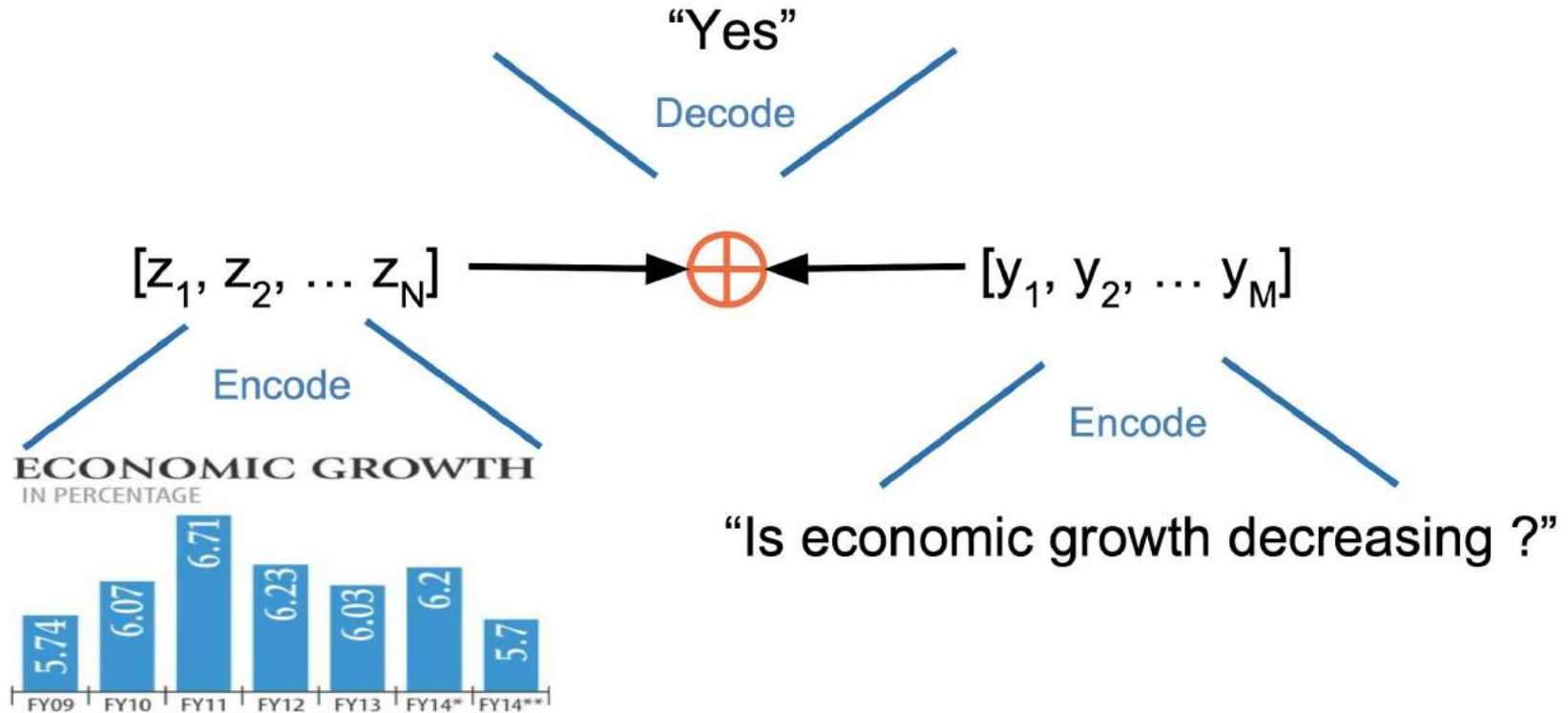


What is the mustache
made of?

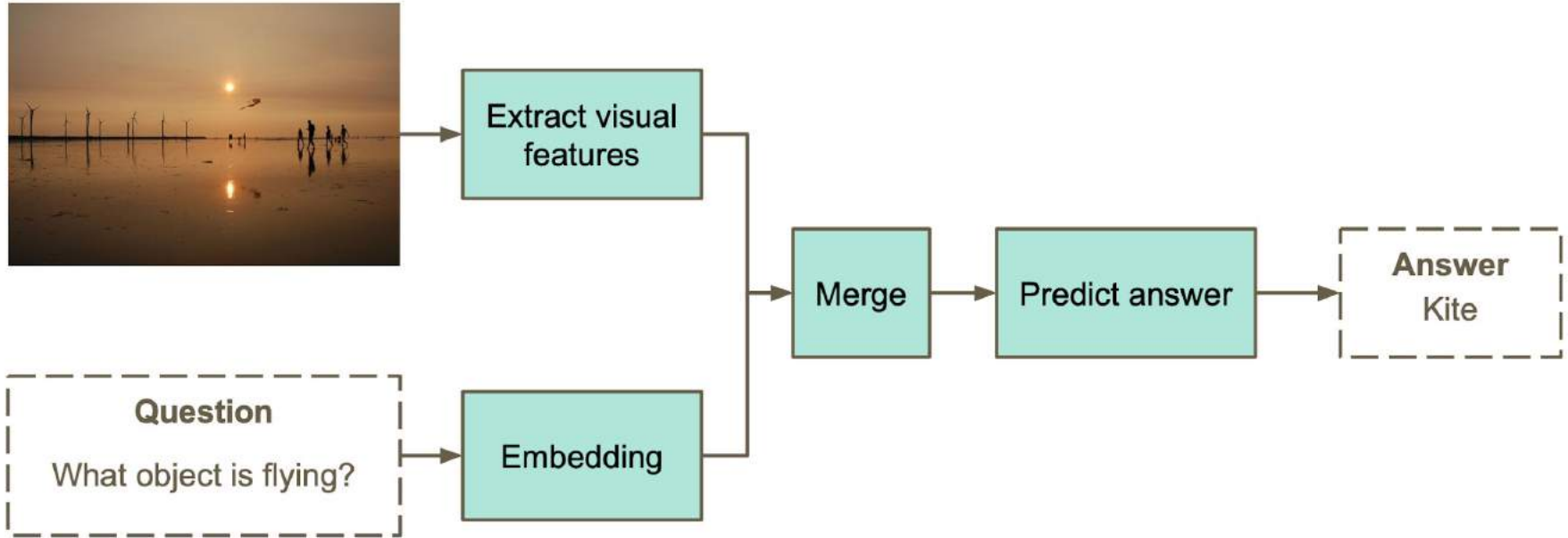
AI System

bananas

Visual Question Answering



Visual Question Answering



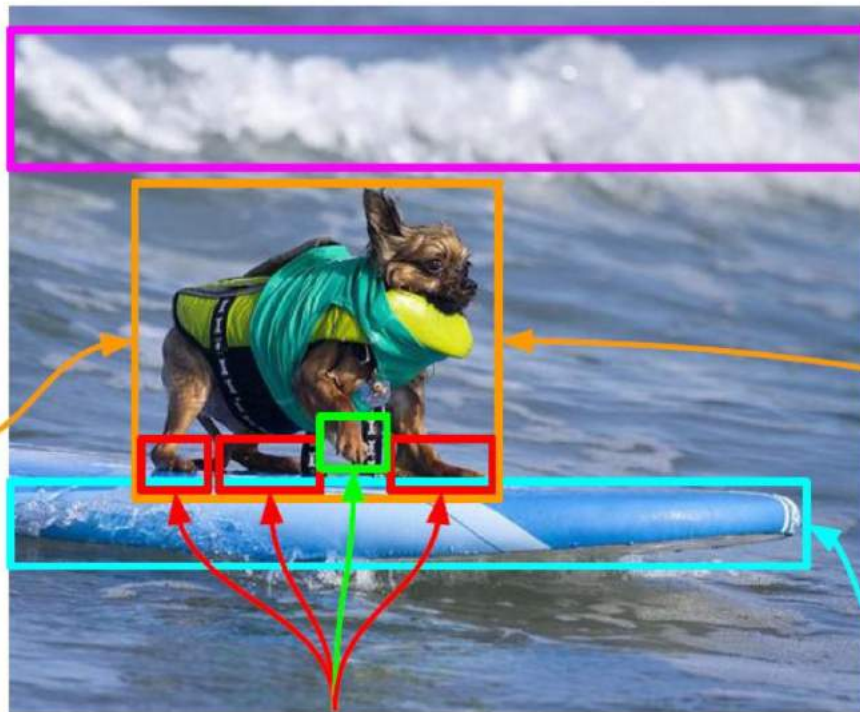
Visual Question Answering

Where does this scene take place?

- A) In the sea. ✓
- B) In the desert.
- C) In the forest.
- D) On a lawn.

What is the dog doing?

- A) Surfing. ✓
- B) Sleeping.
- C) Running.
- D) Eating.



Why is there foam?

- A) Because of a wave. ✓
- B) Because of a boat.
- C) Because of a fire.
- D) Because of a leak.

What is the dog standing on?

- A) On a surfboard. ✓
- B) On a table.
- C) On a garage.
- D) On a ball.

Video Summarization

~1.5 minutes of audio and video

“Teaser” (33 words on avg)

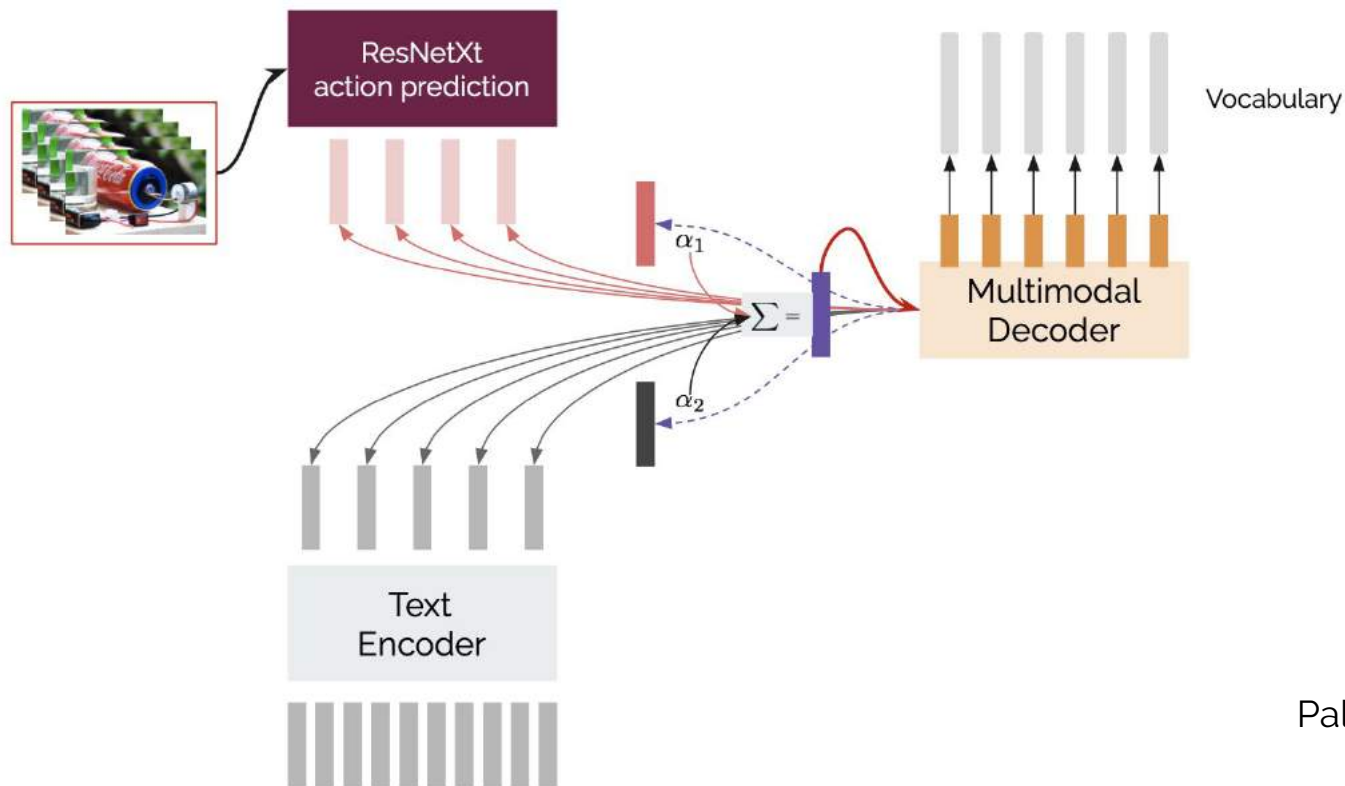
how to cut peppers to make a spanish omelette ; get expert tips and advice on making cuban breakfast recipes in this free cooking video .

Transcript (290 words on avg)

on behalf of expert village my name is lizabeth muller and today we are going to show you how to make spanish omelet . i 'm going to dice a little bit of peppers here . i 'm not going to use a lot , i 'm going to use very very little . a little bit more then this maybe . you can use red peppers if you like to get a little bit color in your omelet . some people do and some people do n't . but i find that some of the people that are mexicans who are friends of mine that have a mexican she like to put red peppers and green peppers and yellow peppers in hers and with a lot of onions . that is the way they make there spanish omelets that is what she says . i loved it , it actually tasted really good . you are going to take the onion also and dice it really small . you do n't want big chunks of onion in there cause it is just pops out of the omelet . so we are going to dice the up also very very small . so we have small pieces of onions and peppers ready to go .



Video Summarization: Hierarchical Model



Action Recognition

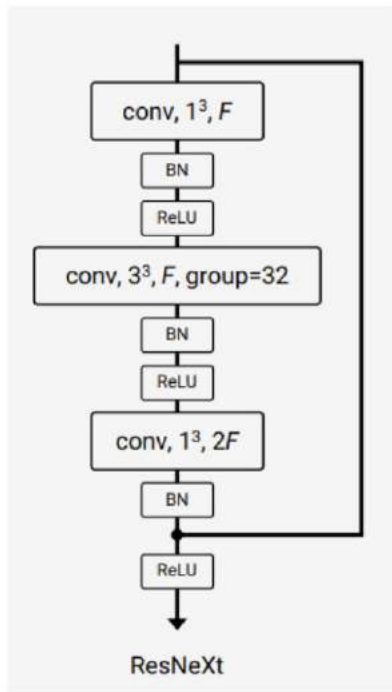
Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?

Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh

National Institute of Advanced Industrial Science and Technology (AIST)

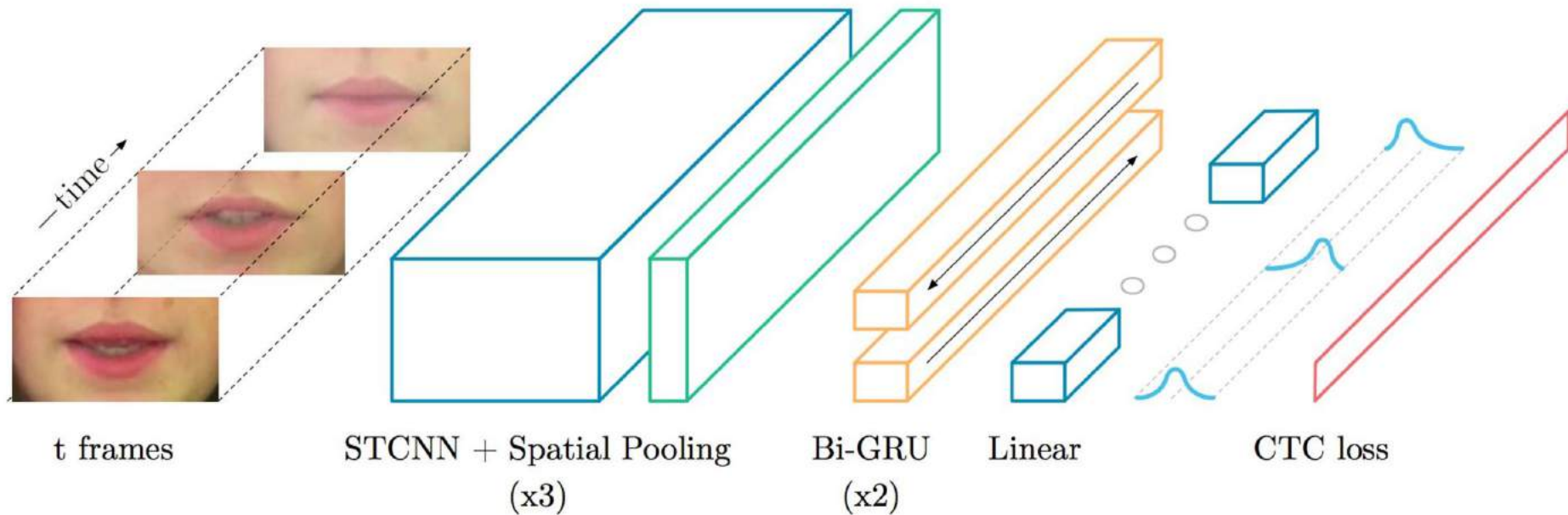
Tsukuba, Ibaraki, Japan

{kensho.hara, hirokatsu.kataoka, yu.satou}@aist.go.jp



2. Speech, Vision and Language Common Tasks

Audio Visual Speech Recognition: Lip Reading



Assael et al. 2016

Lip Reading: Watch, Listen, Attend and Spell

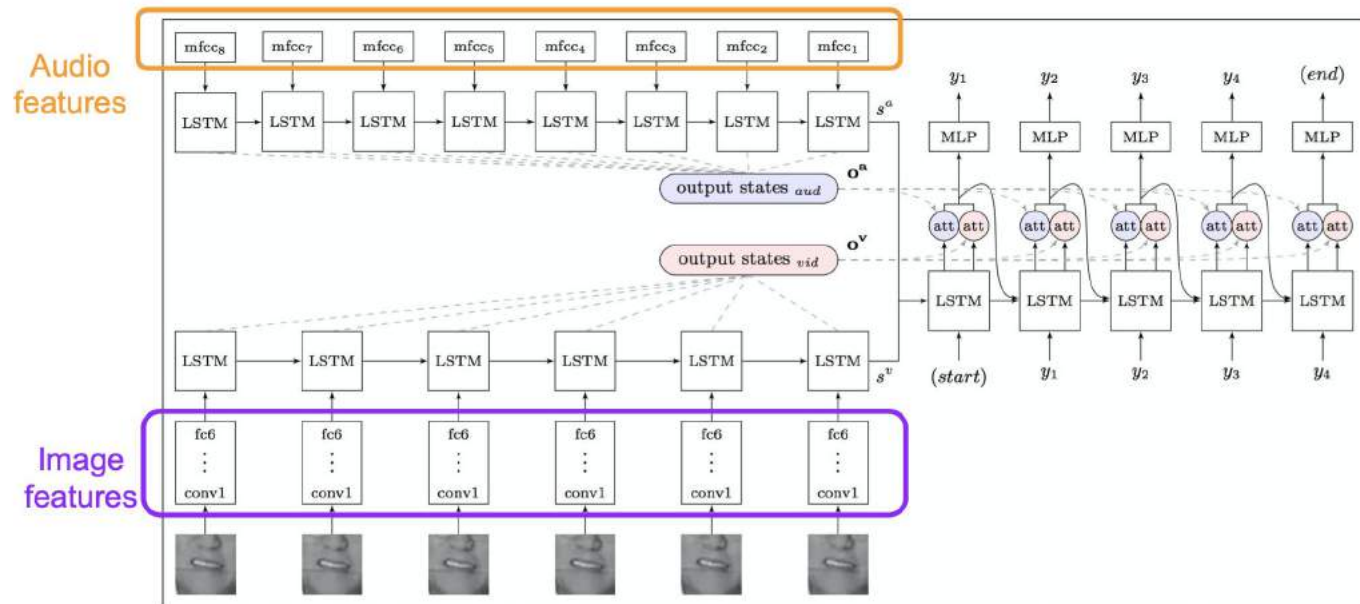


Figure 1. Watch, Listen, Attend and Spell architecture. At each time step, the decoder outputs a character y_i , as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

3. Multimedia Common Tasks

Multimedia Retrieval

Query






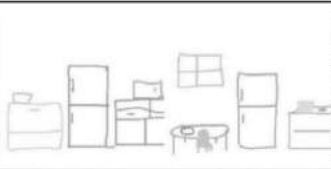










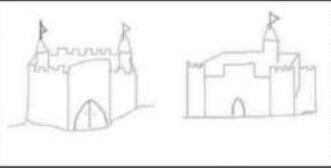






Real

Clip art

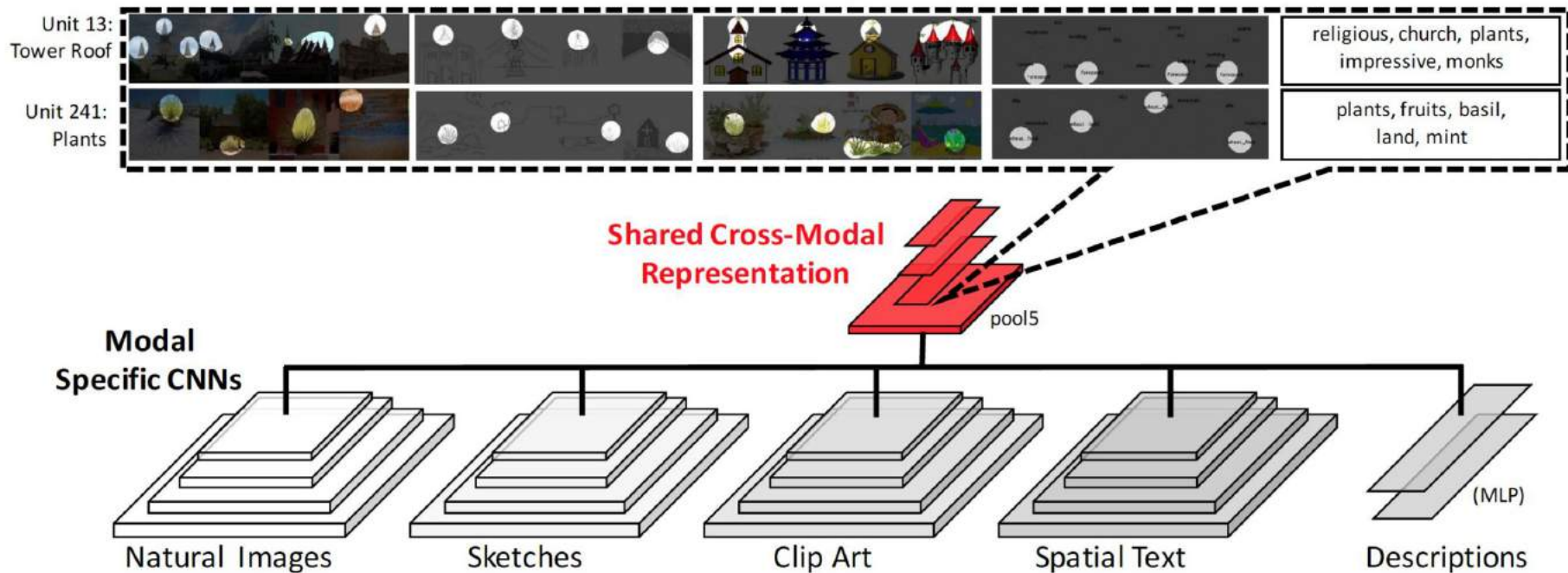
Spatial text

Sketches

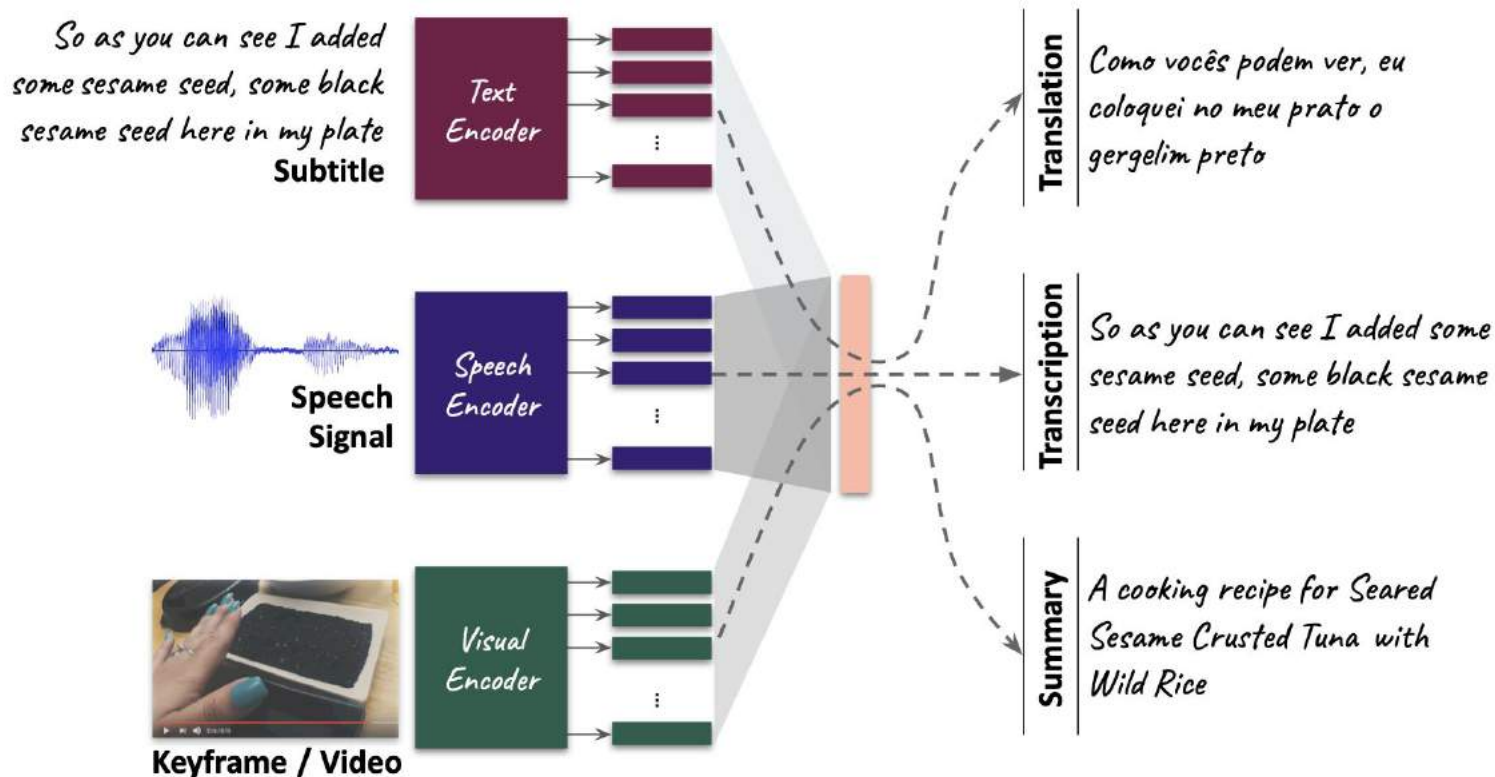
Descriptions

					<p>cabinet door</p> <p>wall</p> <p>wall cabinet sink</p> <p>floor</p>	<p>cabinet door</p> <p>wall</p> <p>wall cabinet sink</p> <p>floor</p>		<p>Everything you could need to make dinner, all in one place. Not quite the size of a full kitchen, but everything is there: microwave, refrigerator, and oven.</p> <p>A very small or compact kitchen. These tiny kitchens typically have all of the regular equipment found in their larger counterparts such as a refrigerator, stove, and microwave, but they are often smaller than full sized appliances. The main purpose of these smaller kitchens</p>
					<p>sky</p> <p>window building</p> <p>window</p> <p>window</p>	<p>sky</p> <p>window building</p> <p>window</p> <p>window</p>		<p>A structure in which people work. It usually has many floors in which the various floors are rented out to different companies. It usually has vending machines on each floor.</p> <p>I had walked inside a very tall building that had many stories in it. I just faced forward and saw the receptionist desk right in front of me. I see several men and woman dressed in suits and their work attire. You could tell this was a serious setting.</p>
<p>sky</p> <p>castle</p> <p>wall</p> <p>road</p>					<p>sky</p> <p>castle</p> <p>wall wall</p> <p>road plants</p>	<p>sky</p> <p>castle</p> <p>wall wall</p> <p>plants road</p>		<p>The building appeared grand from the outside, with its turrets and thick strong walls, but inside the stone air was cold and clammy. The few small windows were all that allowed the sunlight to penetrate the cavernous darkness. There were many old rooms to explore in this ancient</p> <p>This defines the perimeter of an Islamic city with high, fortified walls to keep out intruders. There are often many defenders inside and outside the walls. The residents are relatively safe within the borders of this area.</p>
					<p>sky</p> <p>snowy_mountain crevasse</p>	<p>snowy_mountain</p> <p>sky</p>		<p>A large white covered land mass. It is surrounded by clouds at the top. You can see skiers using trails and ski lifts above the ground. It is winter and many of the cabins at the base of the land mass are occupied. There is a sign that warns be careful of avalanches.</p> <p>Large ice mountain. Usually weather near icebergs is very cold and windy. Huge water bubble sound occurs when the mountain starts melting. Whenever I think of Titanic Ship, I think of ice mountain that caused it.</p>

Multimedia Retrieval



Multimedia Retrieval: Shared Multimodal Representation



Multimedia Retrieval

Ingredients

- 3 lbs salmon
- 1 teaspoon cajun seasoning
- 1 tablespoon olive oil

Cooking Instructions

1. Rinse off salmon and pat dry with paper towel.
2. Drizzle cookie sheet with olive oil.
3. Place salmon (skin side down) on cookie sheet and drizzle more oil on top.
4. Shake Cajun seasoning on salmon to taste.
5. Broil 15-20 minutes or until center of salmon is done.

Images



joint
embedding

4. Emotion and Affect

Affect Recognition:

Emotion, Sentiment, Persuasion, Personality



Happy



Sad



Fear



Anger



Surprise



Disgust



Fearful



Angry



Sad



Happy



Disgusted



Surprised

Outline

- I. What is multimodality?
- II. Types of modalities
- III. Commonly used Models
- IV. Multimodal Fusion and Representation Learning
- V. Multimodal Tasks: Use Cases

Takeaways

- Lots of multimodal data generated everyday
- Need automatic ways to understand it
 - Privacy
 - Security
 - Regulation
 - Storage
- Different models used for different downstream tasks
 - Highly open-ended research!
- Try it out for fun on Kaggle!

Thank you!
spalaska@cs.cmu.edu