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

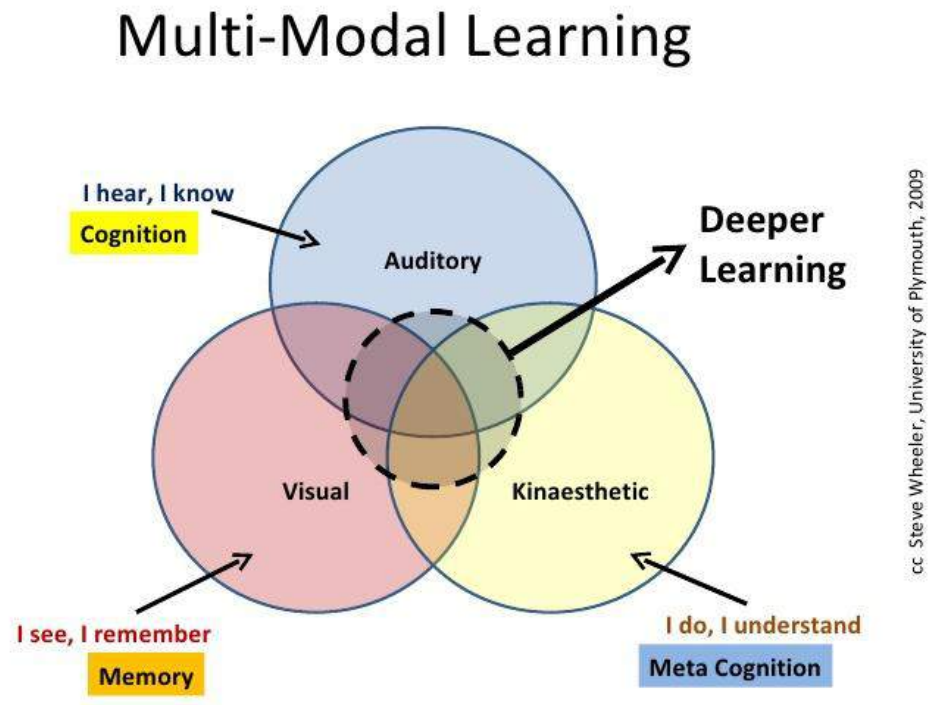
I. What is Multimodality?

Human Interaction is Inherently Multimodal

How We Perceive

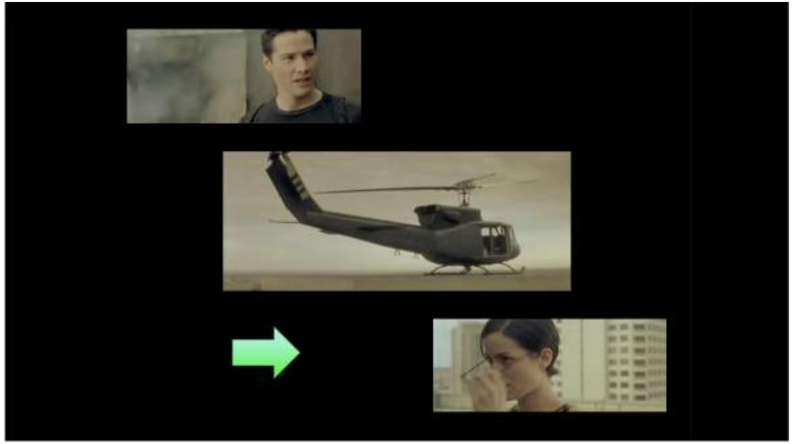
The curse that afflicts abstract painting

How We Perceive



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.

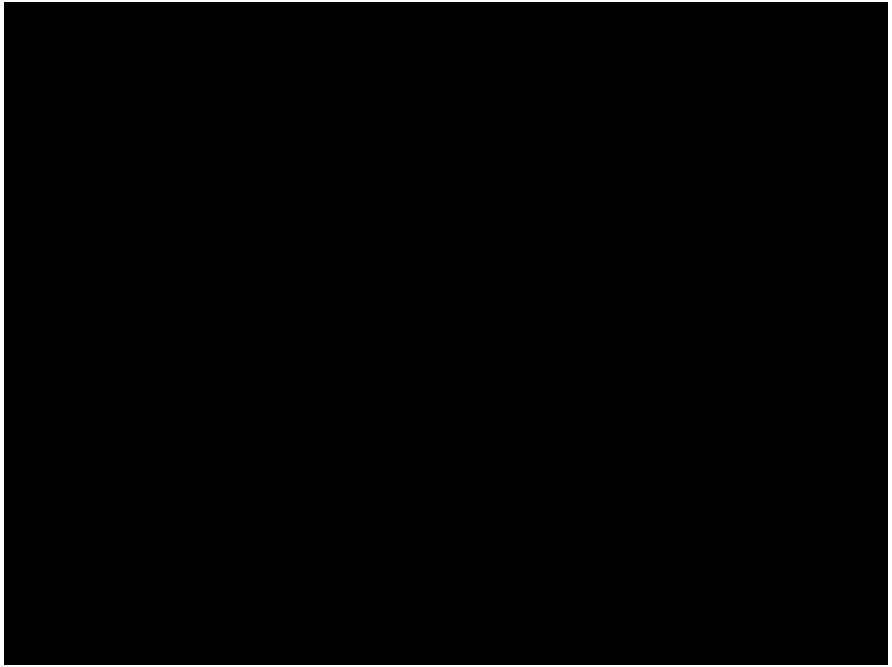
Give a caption.

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

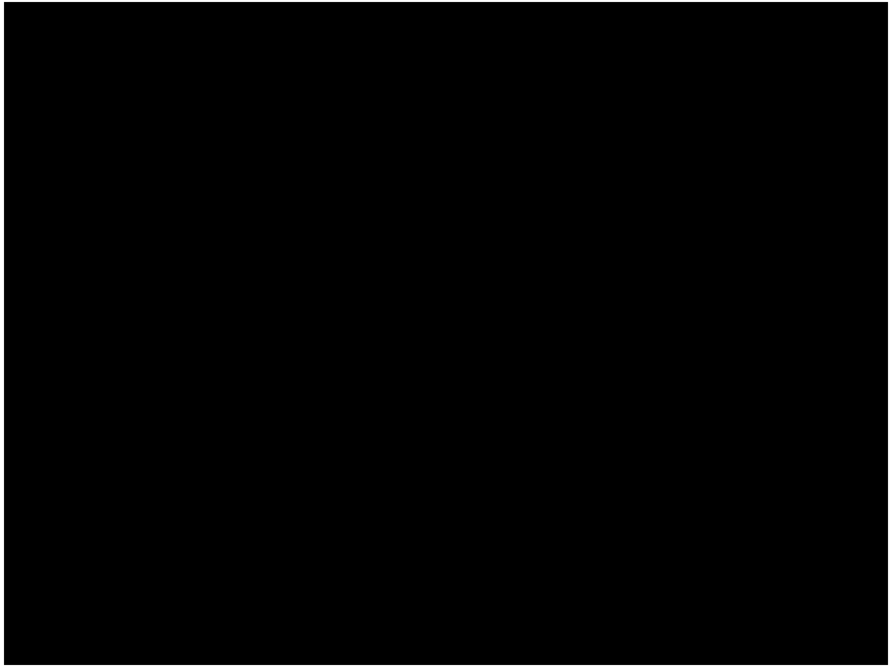
Q. is there only one person ?

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

II. Types of modalities

Types of Modalities

IMAGE/VIDEO

The Dad
@thedad

me: it's bedtime now

kid: please let me do just ONE thing

me: ok

kid: *starts watching one movie*

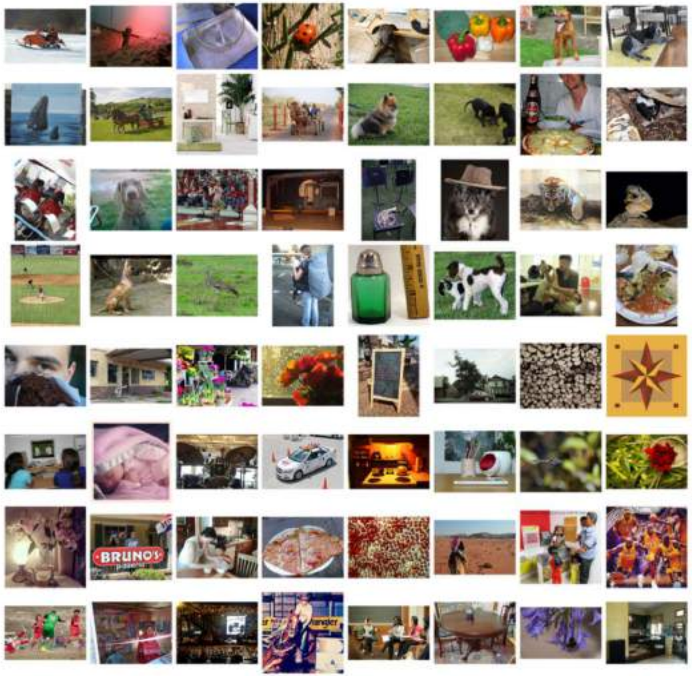
5:59 PM · Oct 25, 2019 · Buffer

TEXT

SPEECH/AUDIO

EMOTION/AFFECT
/SENTIMENT

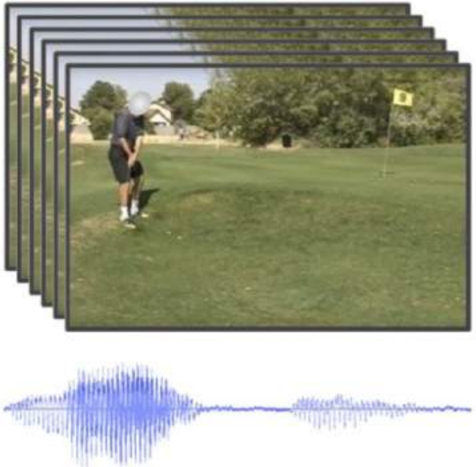
Example Dataset: ImageNet



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

Deng et al. 2009

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.

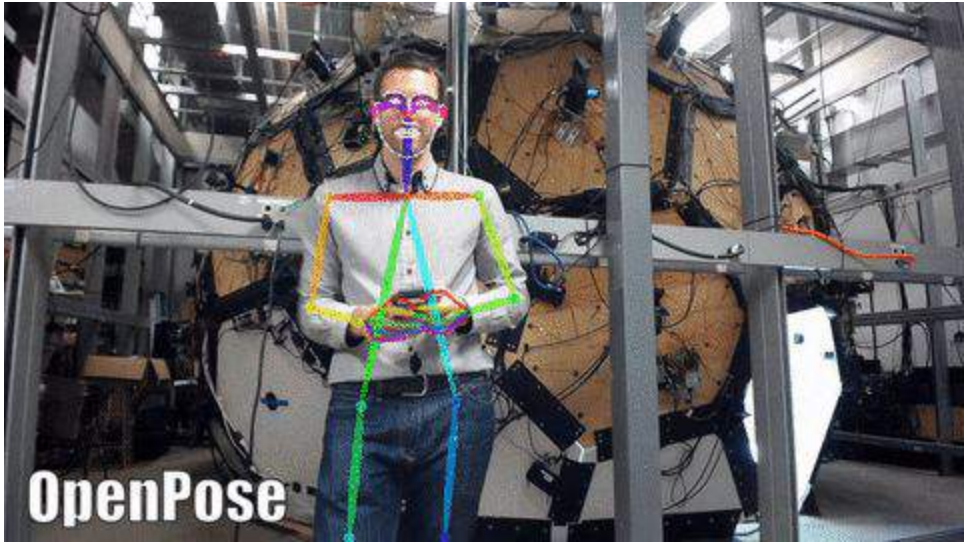
Eu estou muito perto do green, mas eu não pus a bola no green, então agora estou neste bunker de grama.

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

Sanabria et al. 2018

Example Dataset: Open Pose

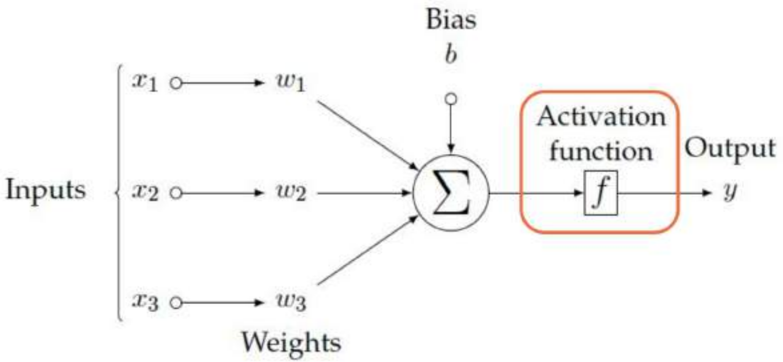


- Action Recognition
- Pose Estimation
- Human Dynamic
- Body Dynamics

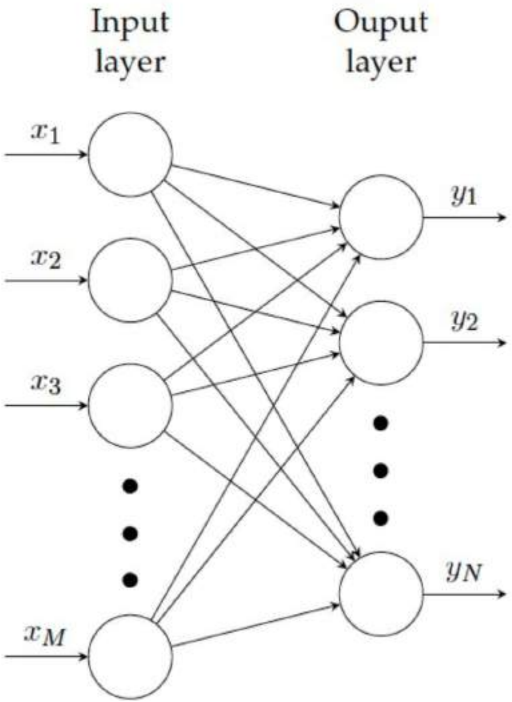
Wei et al. 2016

III. Commonly Used Models

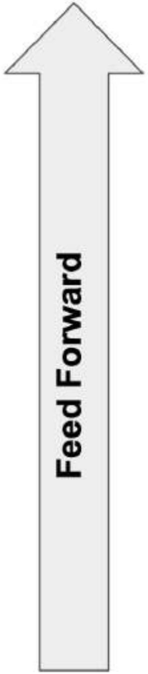
Multilayer Perceptrons



Single Perceptron



Multilayer Perceptrons



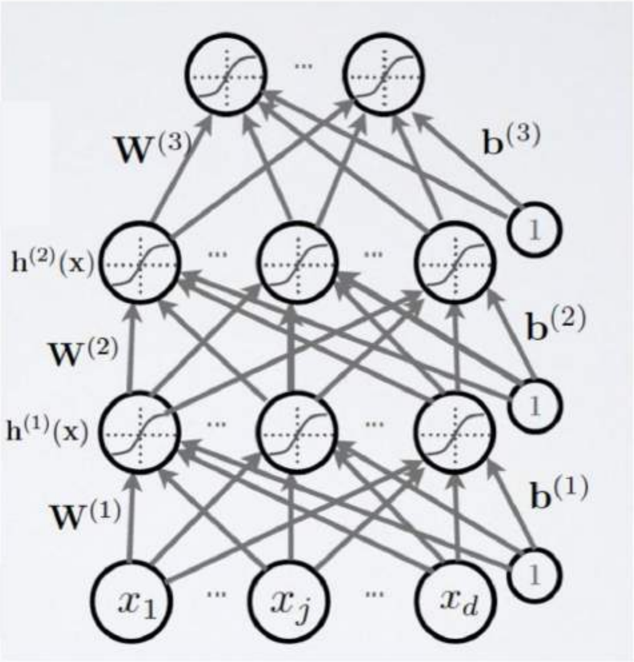


Figure: Hugo Larochelle

OUTPUT (y)

Feed-forward Weights (W_i)

$$y = f(\boldsymbol{W_3} \cdot h_2 + b_3)$$

Hidden States h_1 & h_2

$$h_2 = f(\boldsymbol{W_2} \cdot h_1 + b_2)$$

$$h_1 = f(\boldsymbol{W_1} \cdot x + b_1)$$

INPUT (x)

Multilayer Perceptrons: Uses in Multimedia


The Dad
@thedad

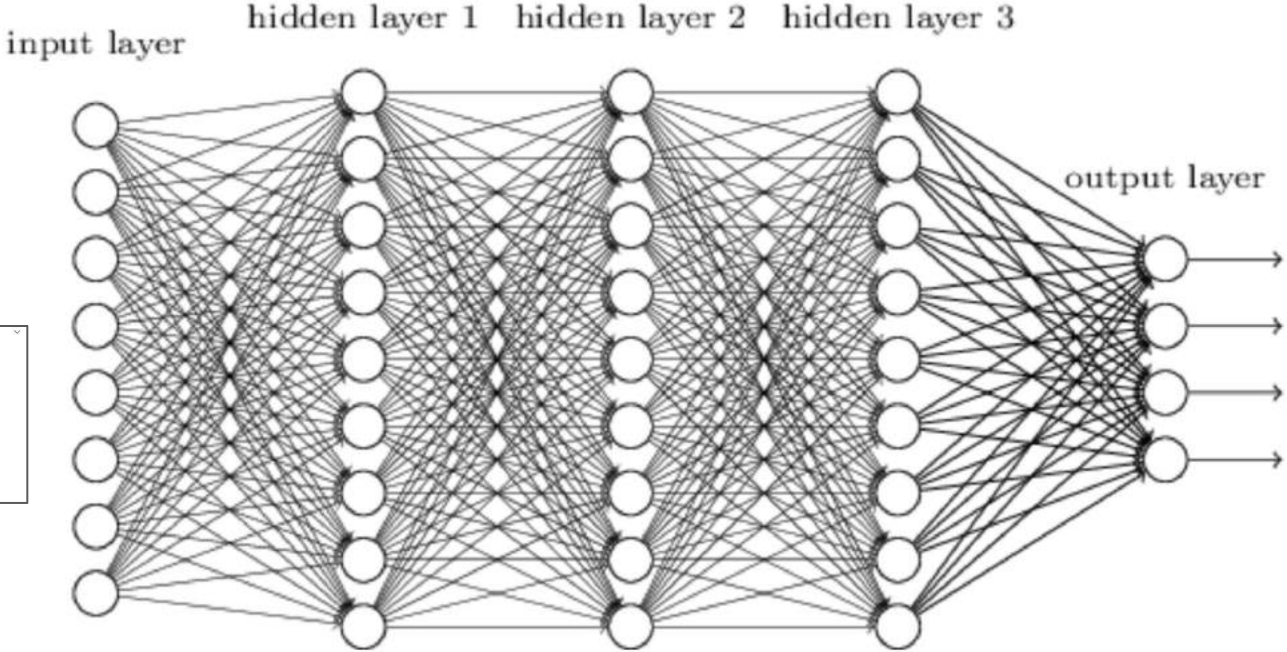
me: it's bedtime now

kid: please let me do just ONE thing

me: ok

kid: *starts watching one movie*

5:59 PM · Oct 25, 2019 · Buffer



Multilayer Perceptrons: Limitations

Limitation #1

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

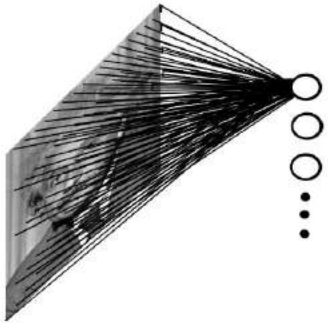
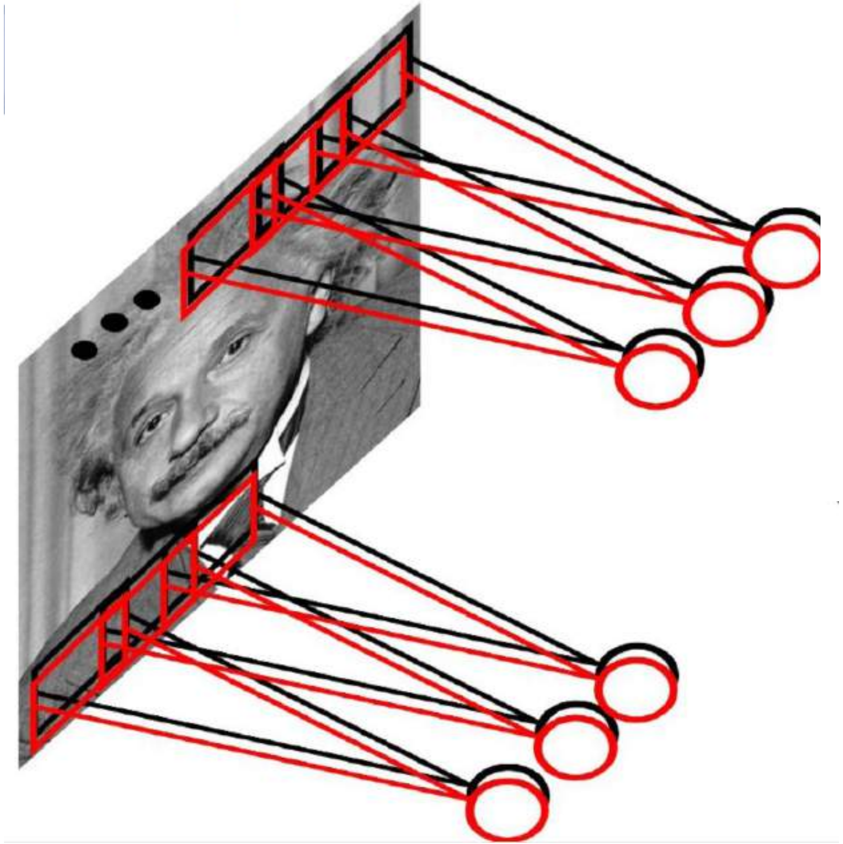


Figure: Ranzatto

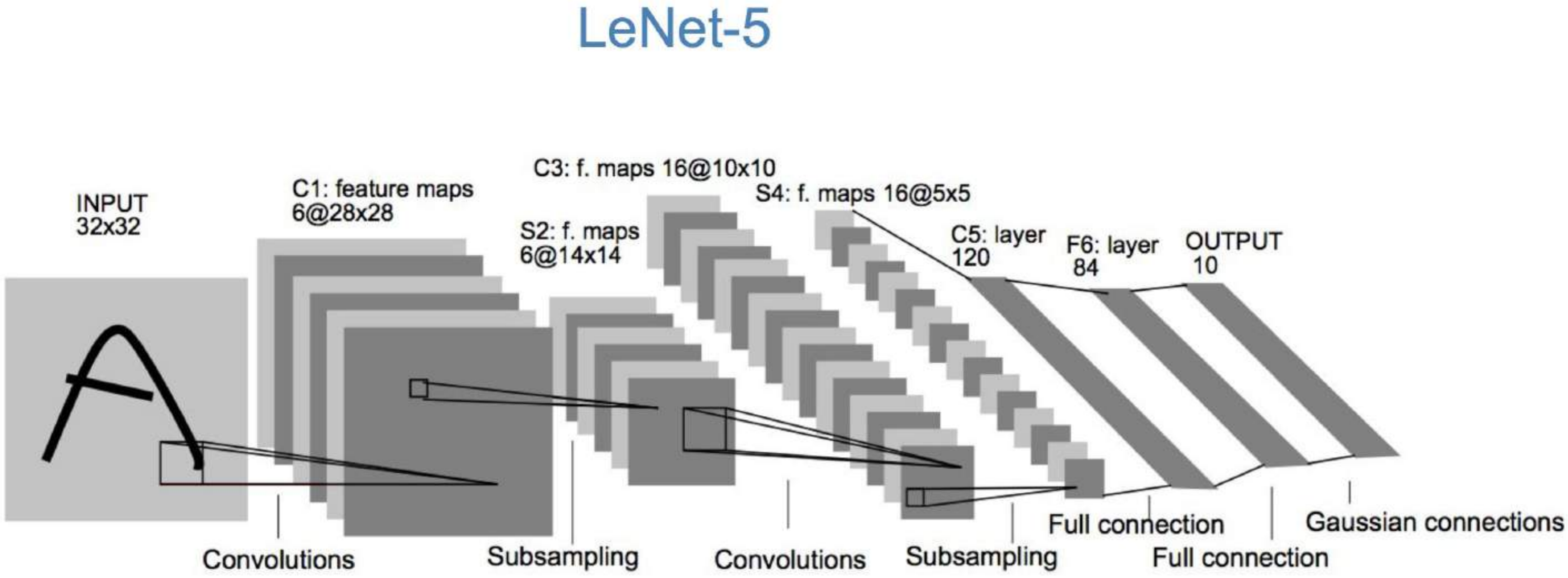
Convolutional Neural Networks (CNNs)



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.

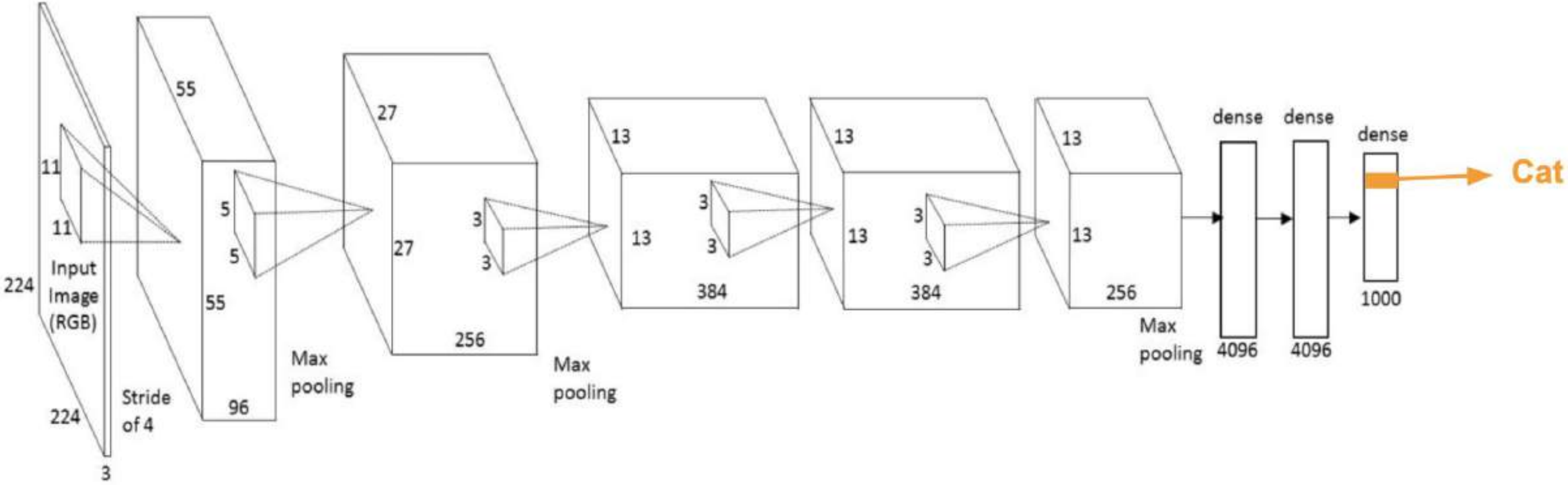
These operations are equivalent to perform **convolutions** with different filters.

Convolutional Neural Networks (CNNs)



LeCun et al. 1998

Convolutional Neural Networks (CNNs) for Image Encoding



Krizhevsky et al. 2012

Multilayer Perceptrons: Limitations

Limitation #1

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

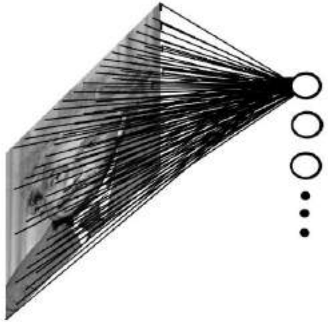


Figure: Ranzatto

Limitation #2

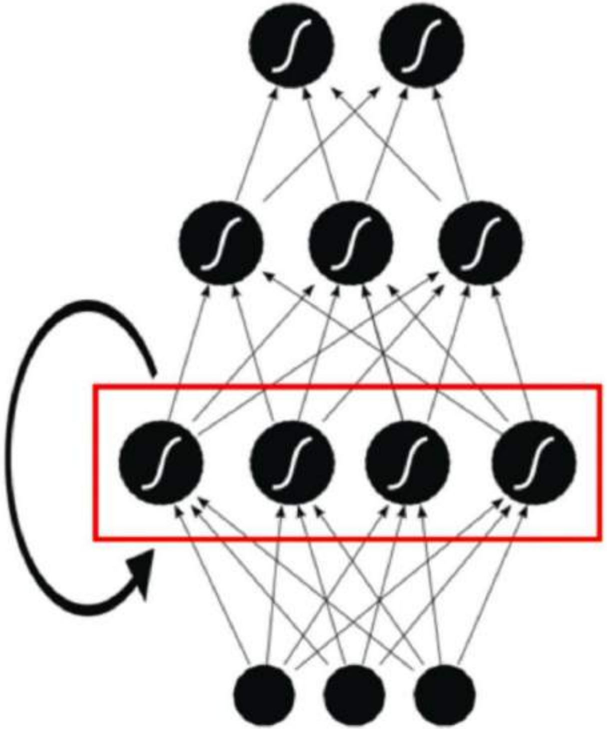
Does not naturally handle input data of variable dimension

(eg. audio/video/word sequences)

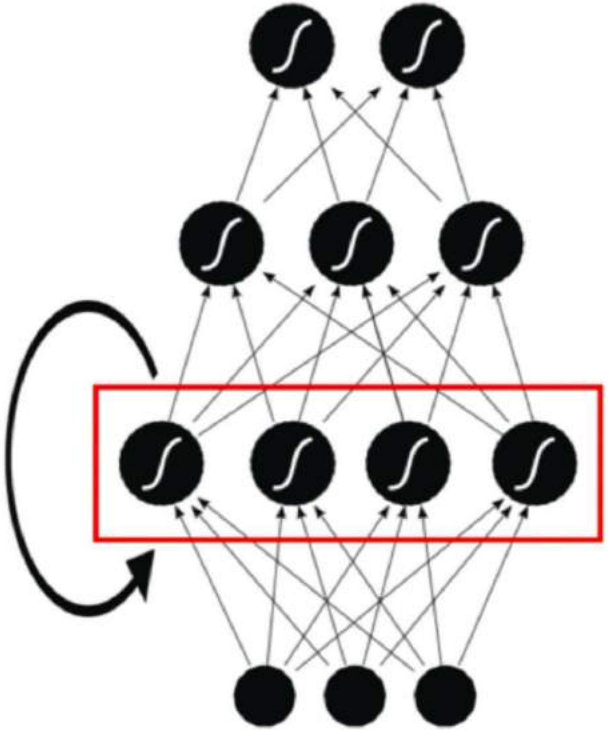
Recurrent Neural Networks

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)

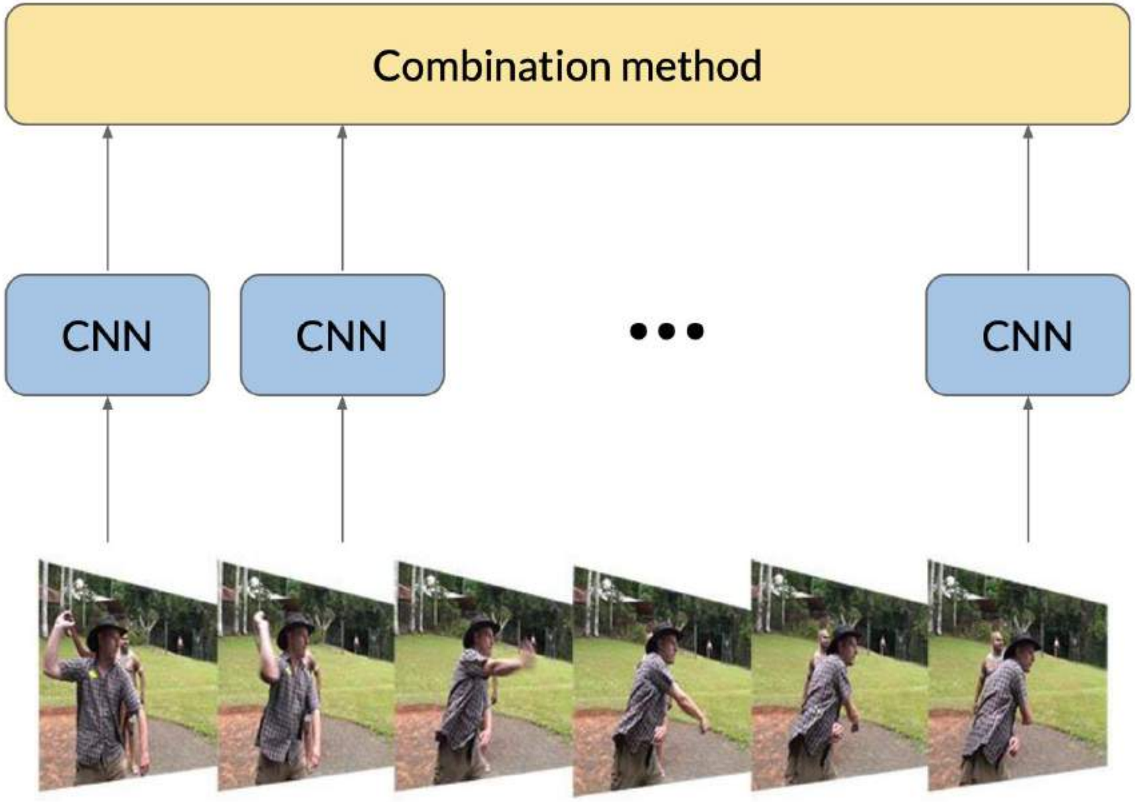
Updated state

Previous state

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

INPUT (x)

Recurrent Neural Networks for Video Encoding

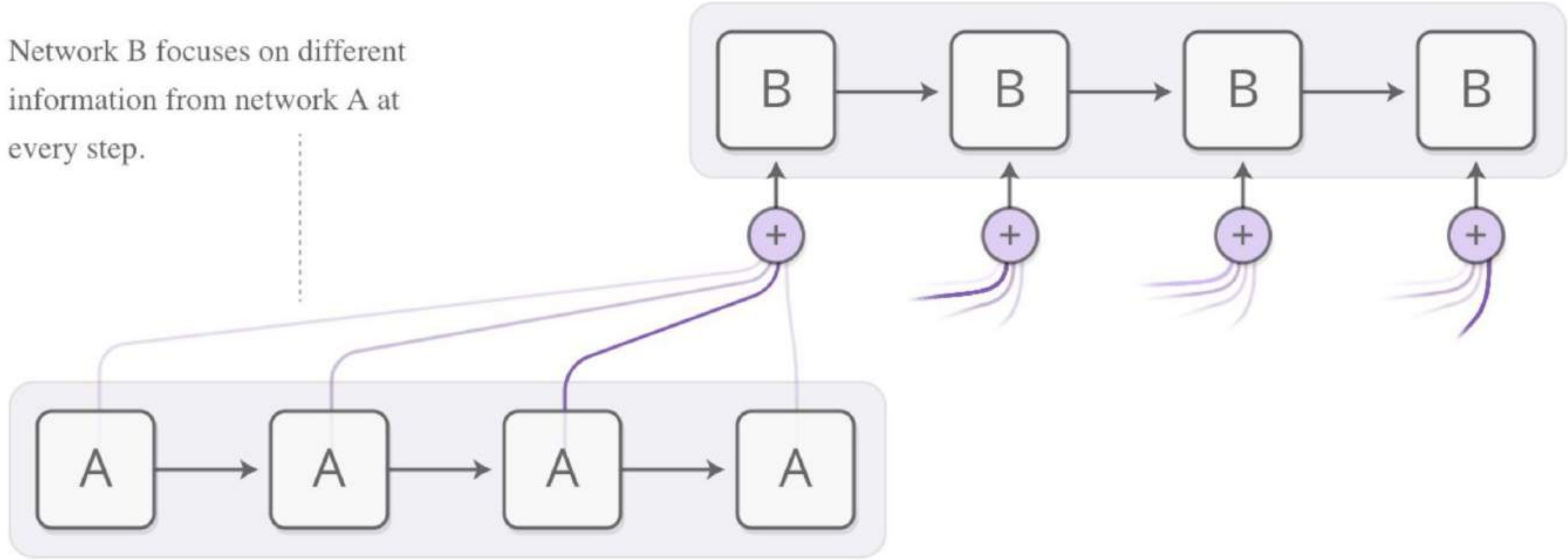


Combination is commonly implemented as a small NN on top of a pooling operation (e.g. max, sum, average).

Recurrent Neural Networks are well suited for processing sequences.

Donahue et al. 2015

Attention Mechanism



Olah and Cate 2016

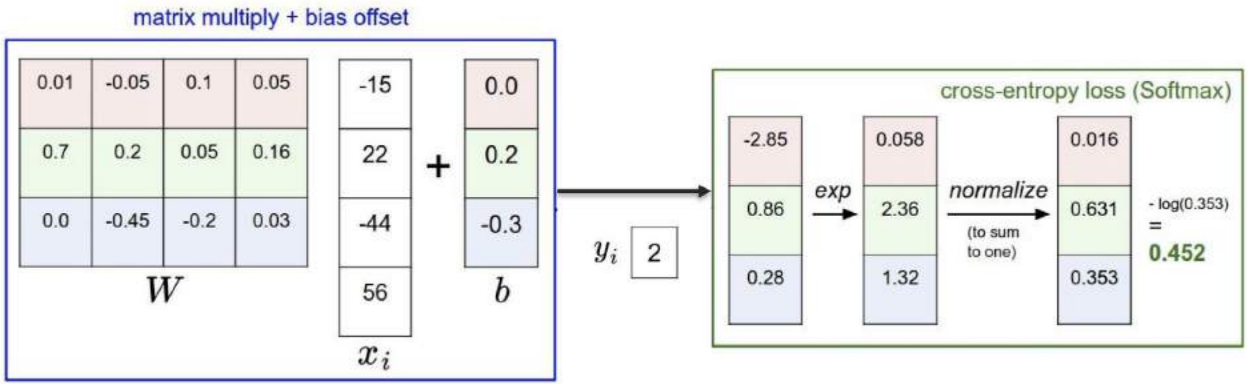
Loss Function: Softmax

Cross-entropy loss:

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

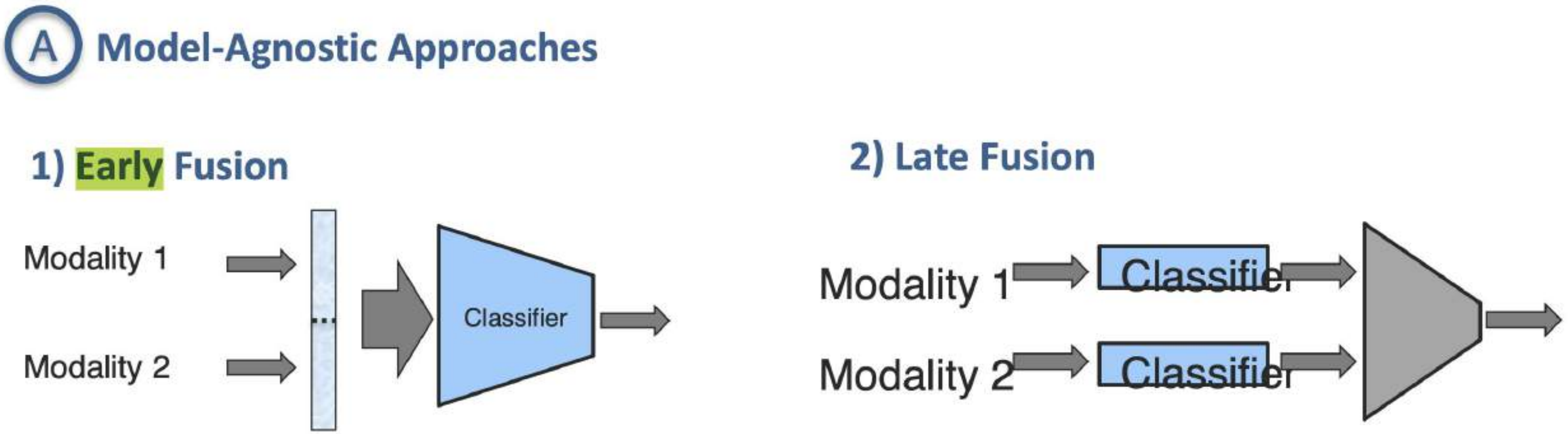
Softmax function

Minimizing the negative log likelihood.



IV. Multimodal Fusion & Representation Learning

Fusion: Model Agnostic

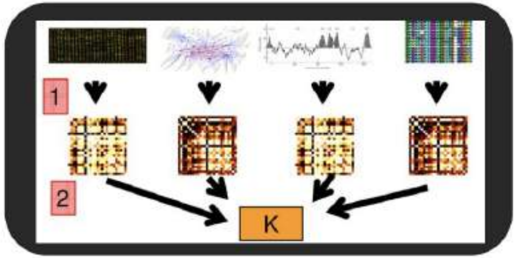


Fusion: Model Based

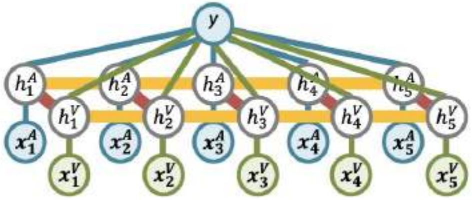
Definition: To join information from two or more modalities to perform a prediction task.

(B) Model-Based (Intermediate) Approaches

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



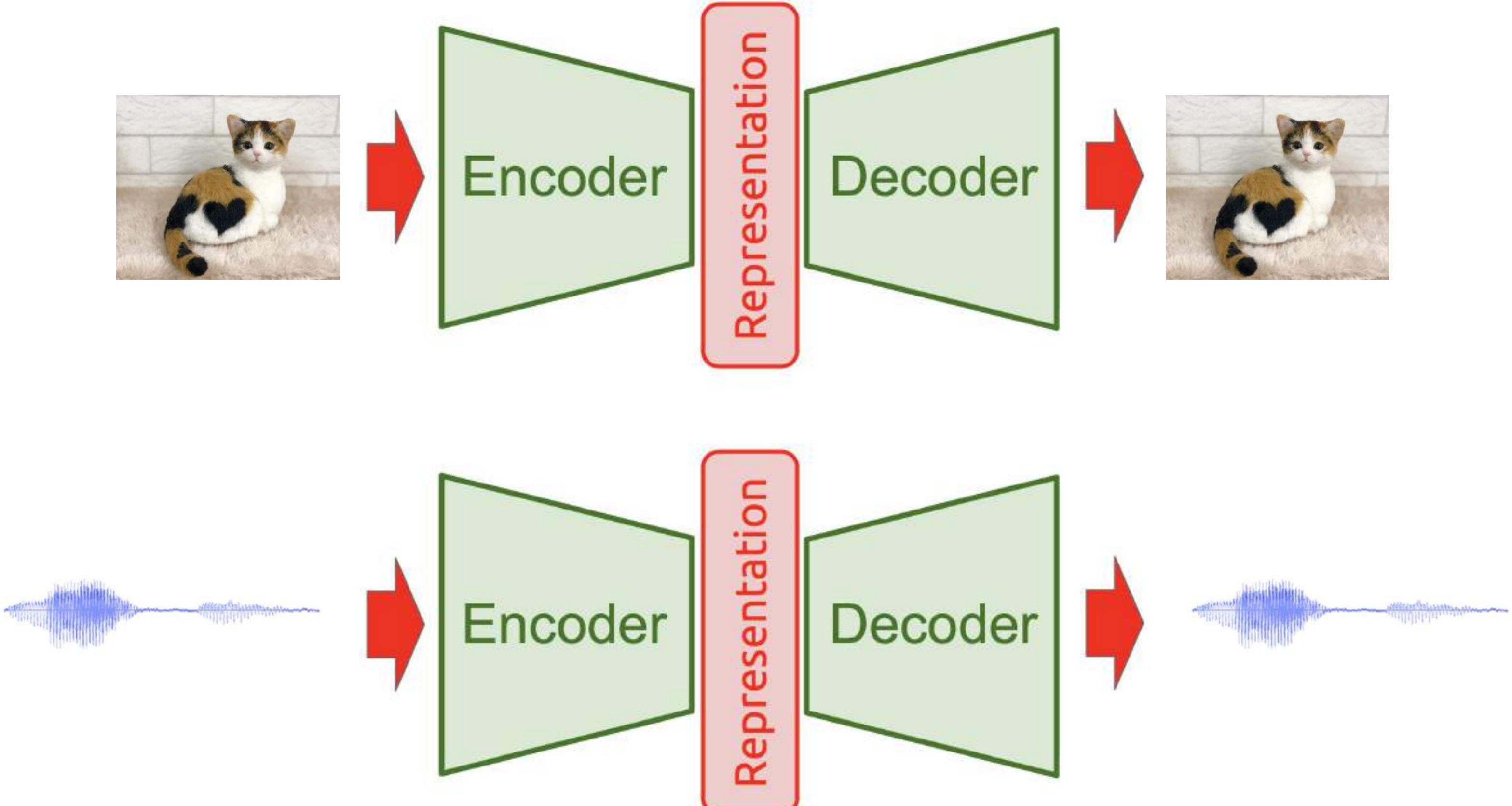
Multiple kernel learning



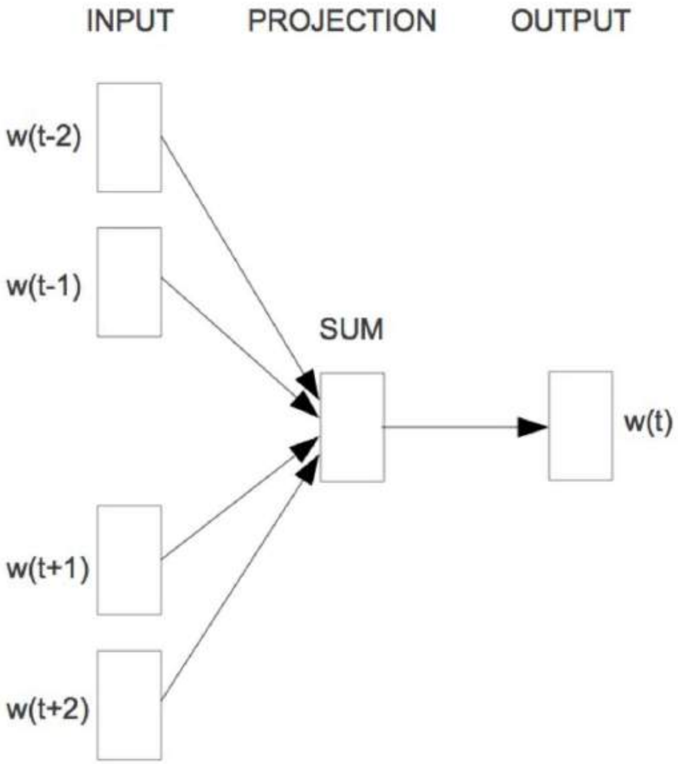
Multi-View Hidden CRF

Slide by LP Morency

Representation Learning: Encoder-Decoder



Representation Learning



`the cat climbed a tree`

Given context:

`a, cat, the, tree`

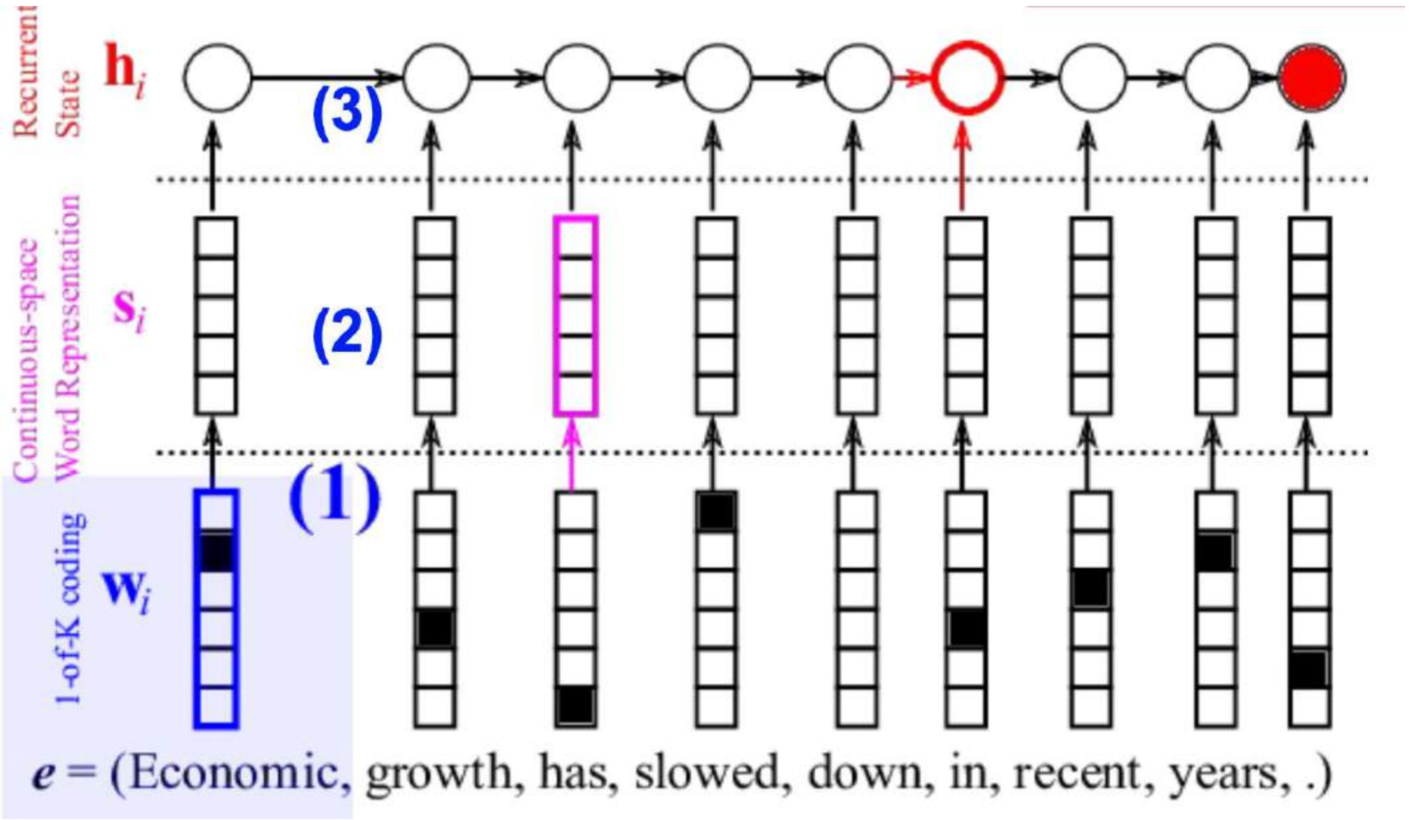
Estimate prob. of

`climbed`

Word2Vec

Mikolov et al. 2013

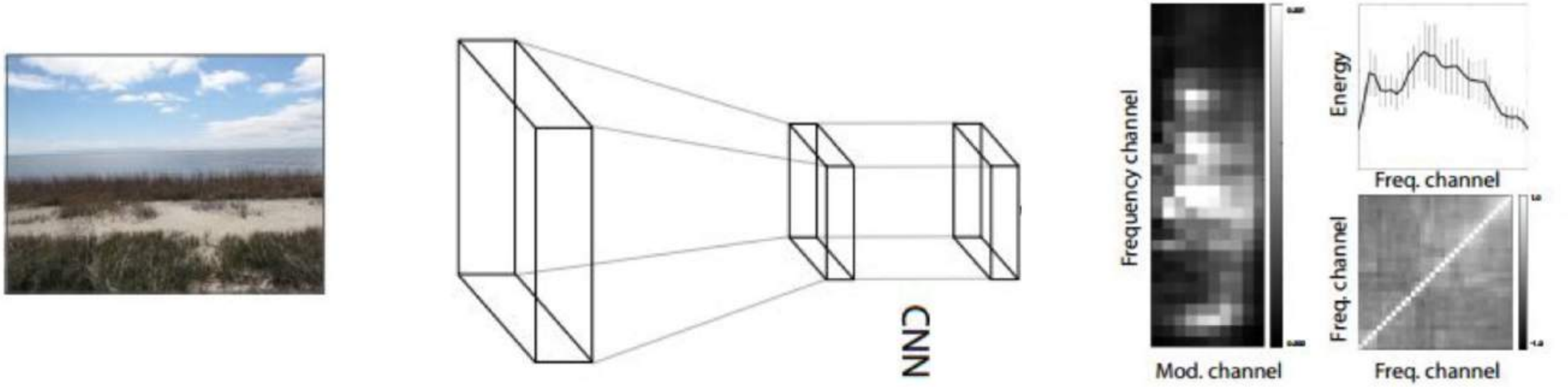
Representation Learning: RNNs



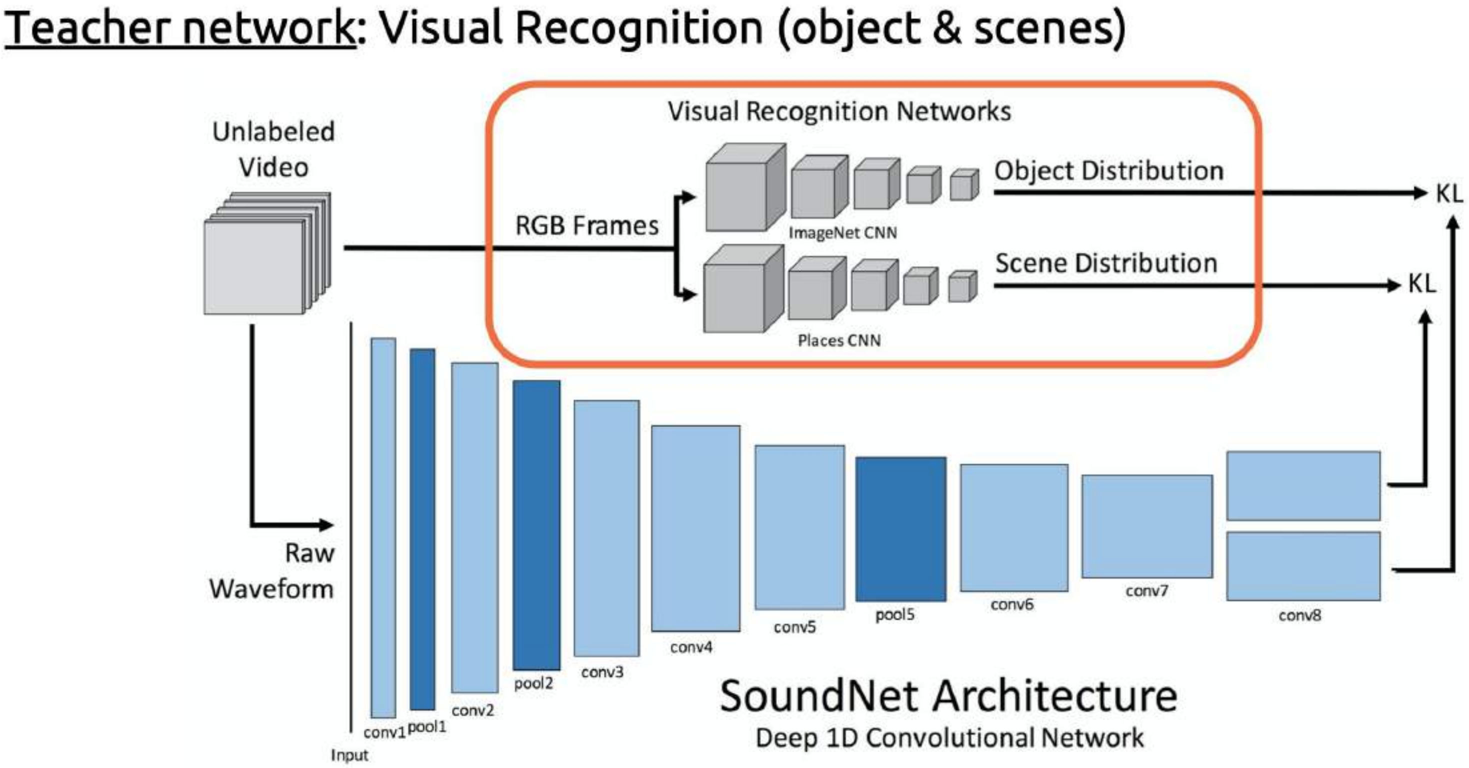
Cho et al. 2014

Representation Learning: Self-Supervised

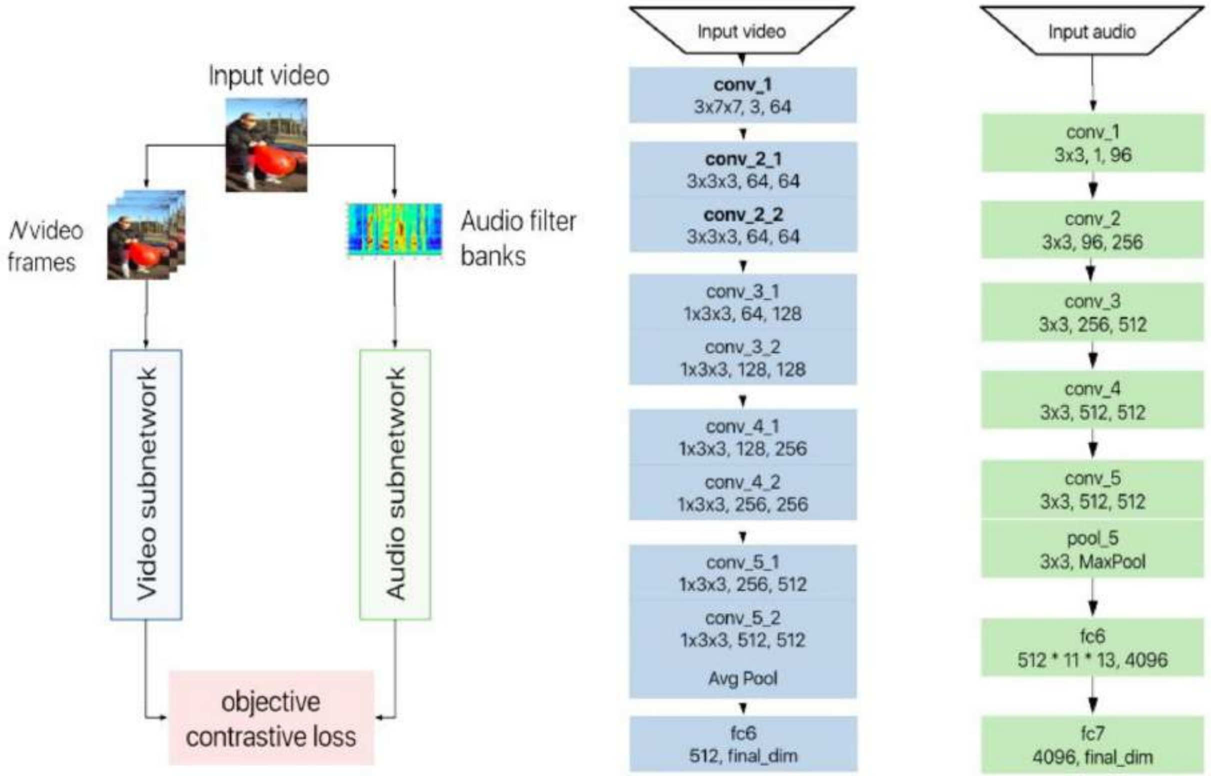
Use videos to train a CNN that predicts the audio statistics of a frame.



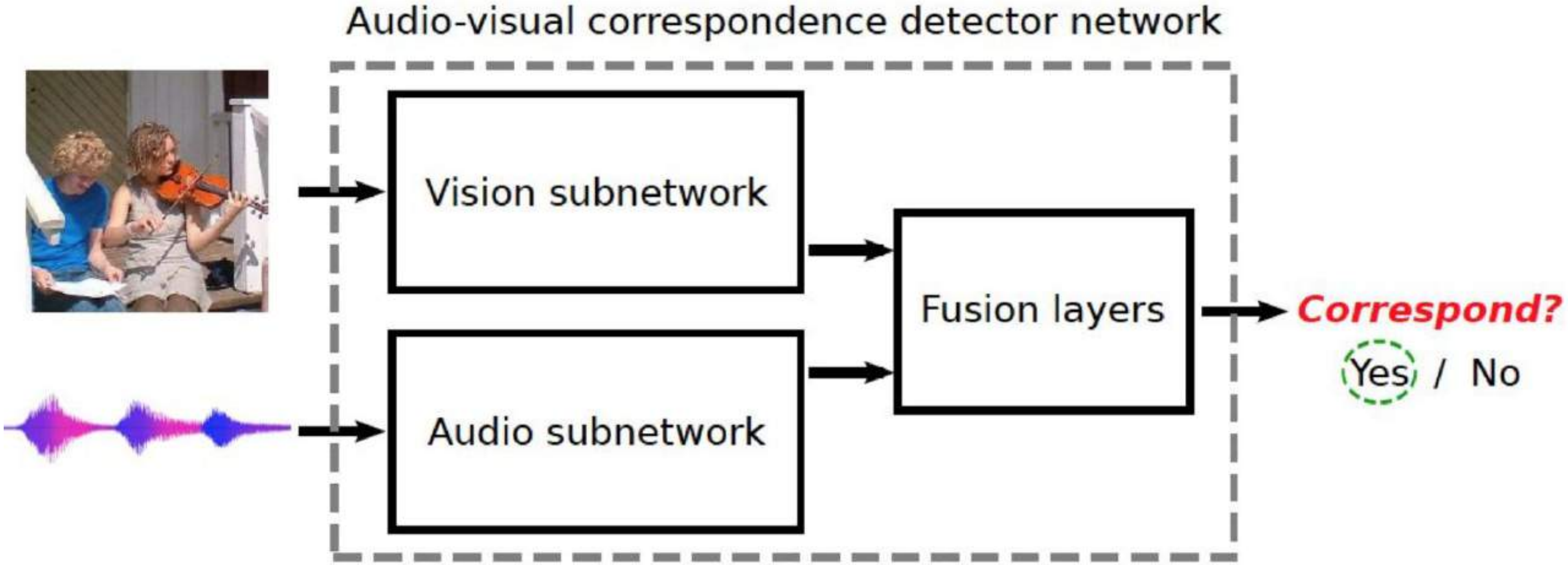
Representation Learning: Transfer Learning



Representation Learning: Joint Learning



Representation Learning: Joint Learning (Similarity)



V. Common Tasks, Use Cases

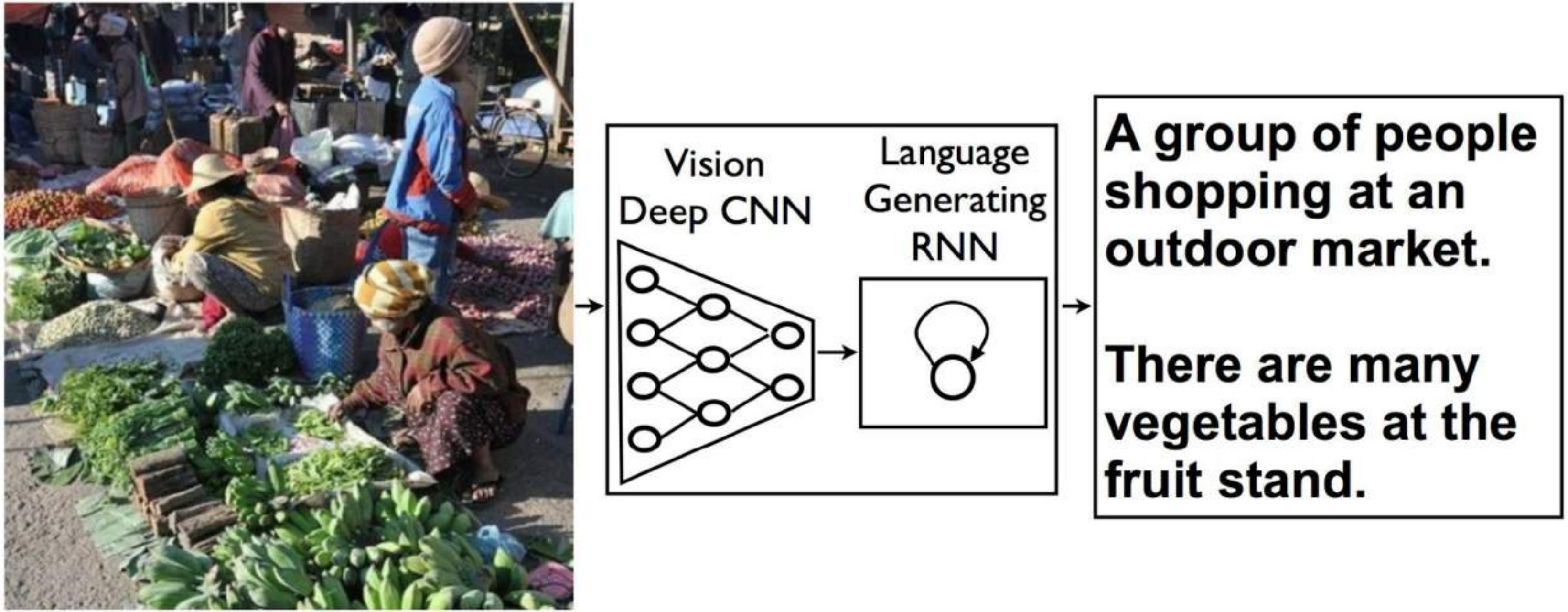
V. Common Tasks

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
- …

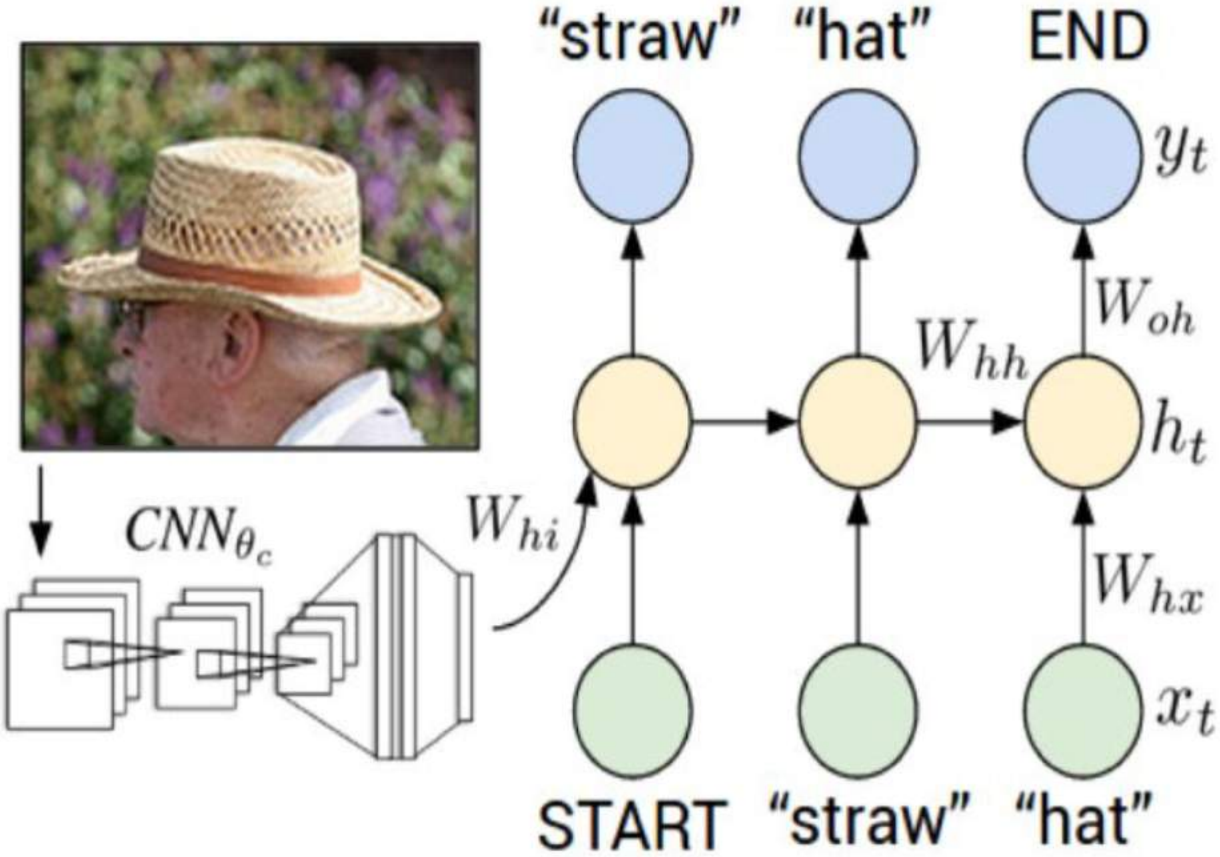
1. Vision and Language Common Tasks

Image Captioning



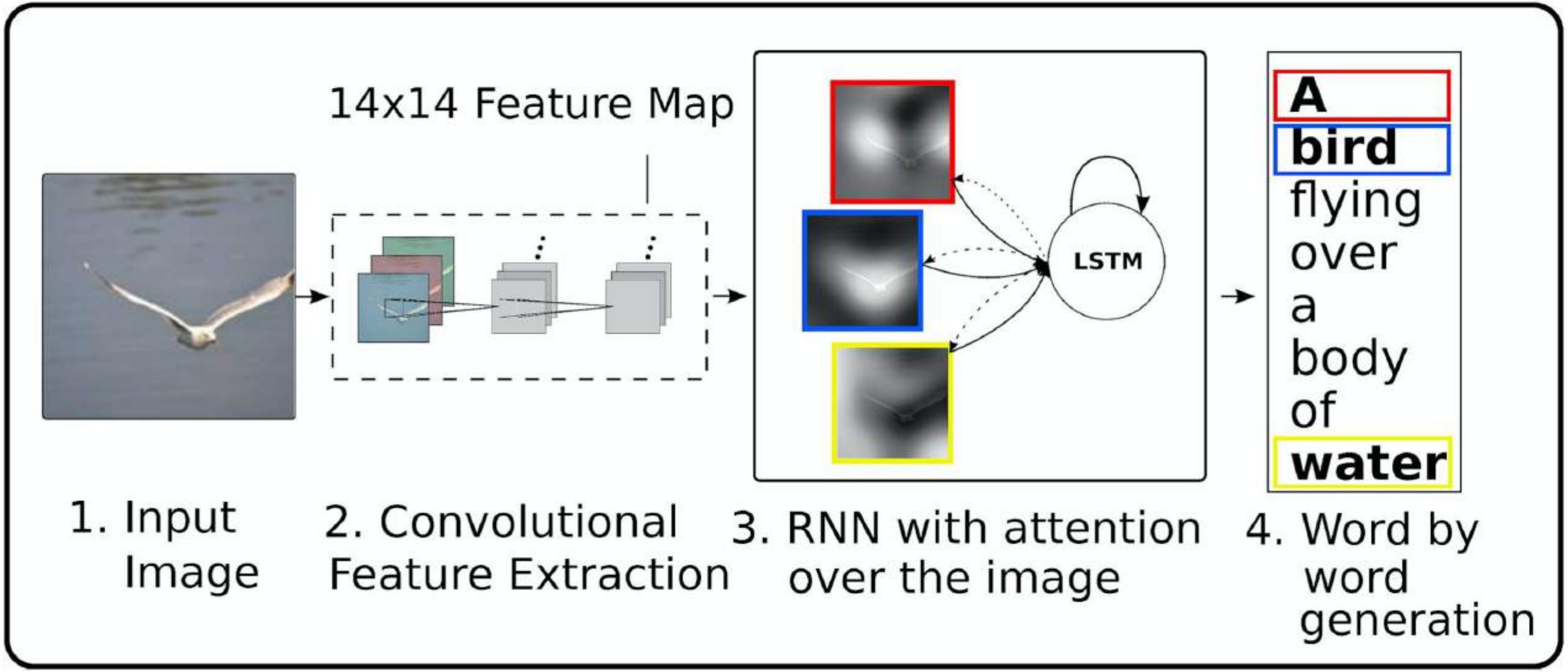
Vinyals et al. 2015

Image Captioning



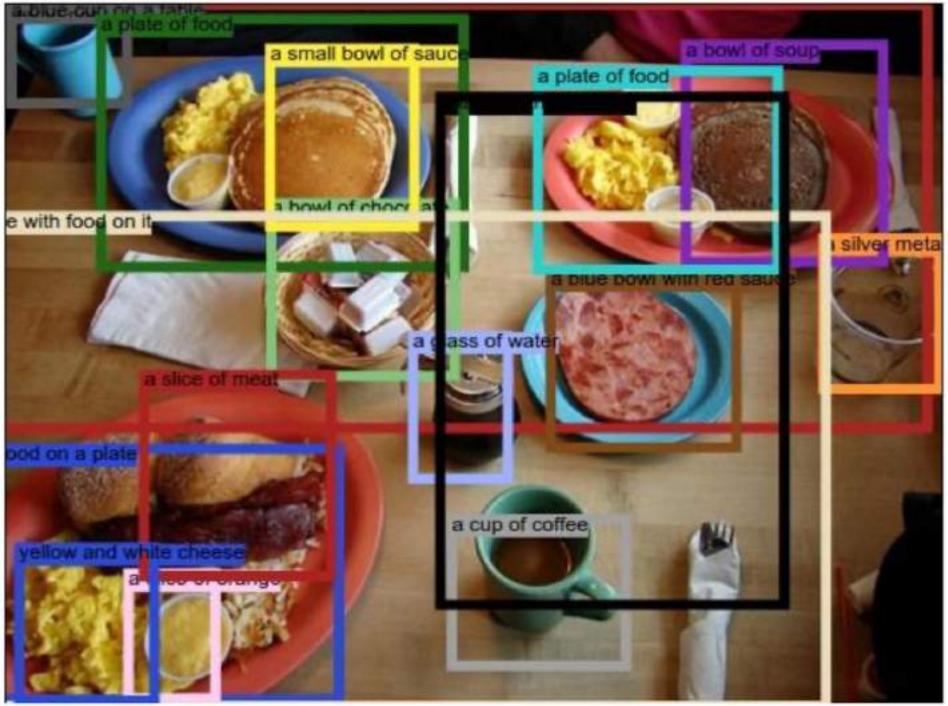
Karpathy et al. 2015

Image Captioning: Show, Attend and Tell



Xu et al. 2015

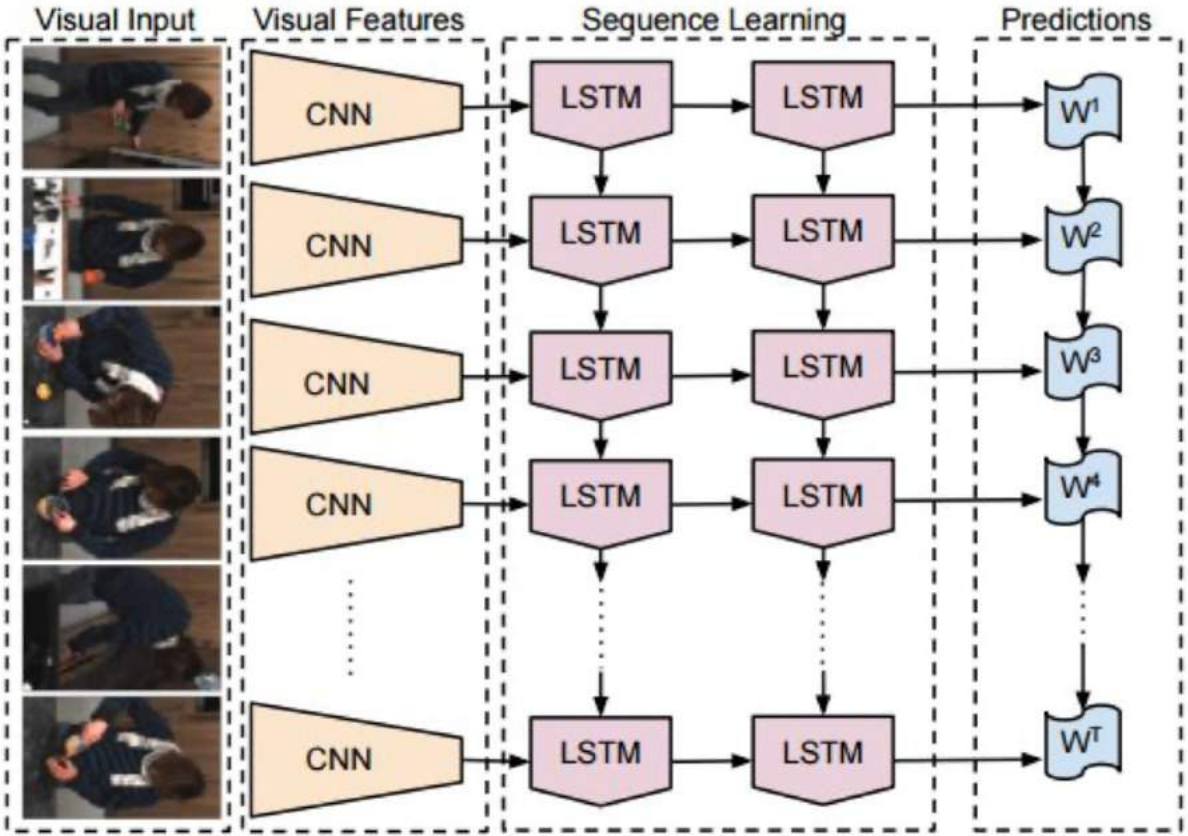
Image Captioning and Detection



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.

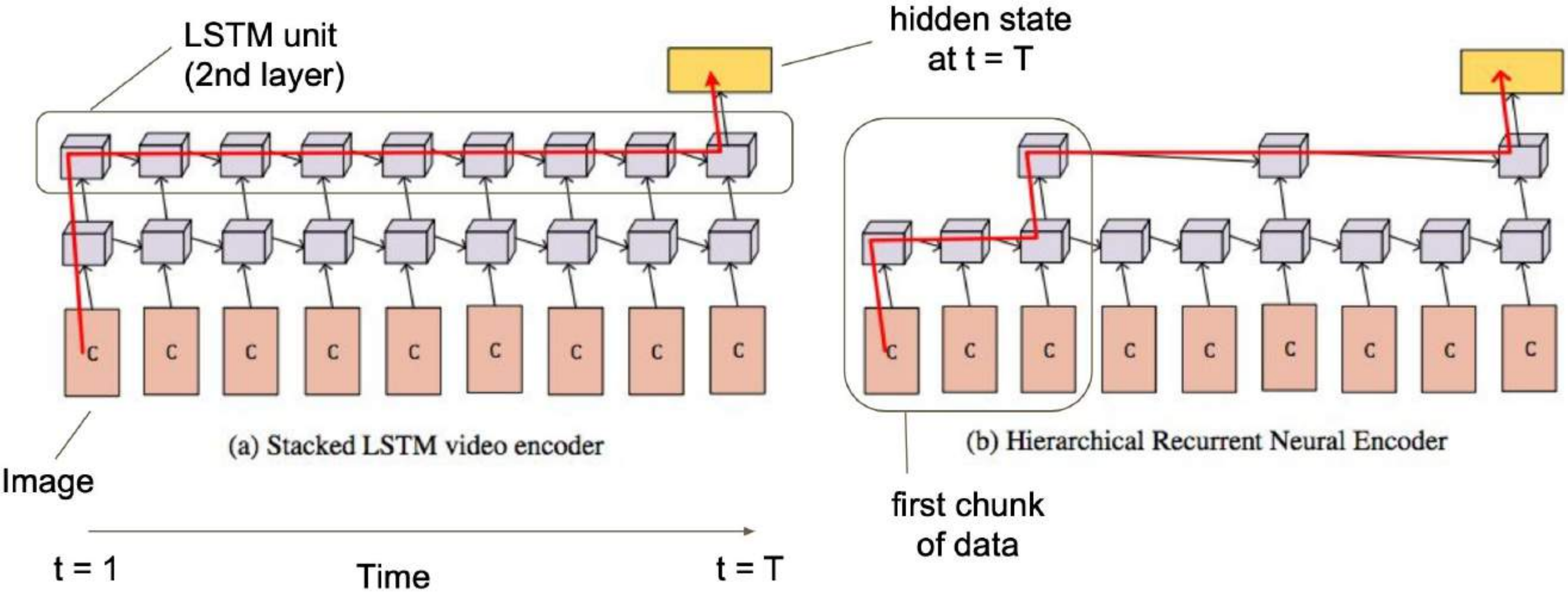
Johnson et al. 2016

Video Captioning



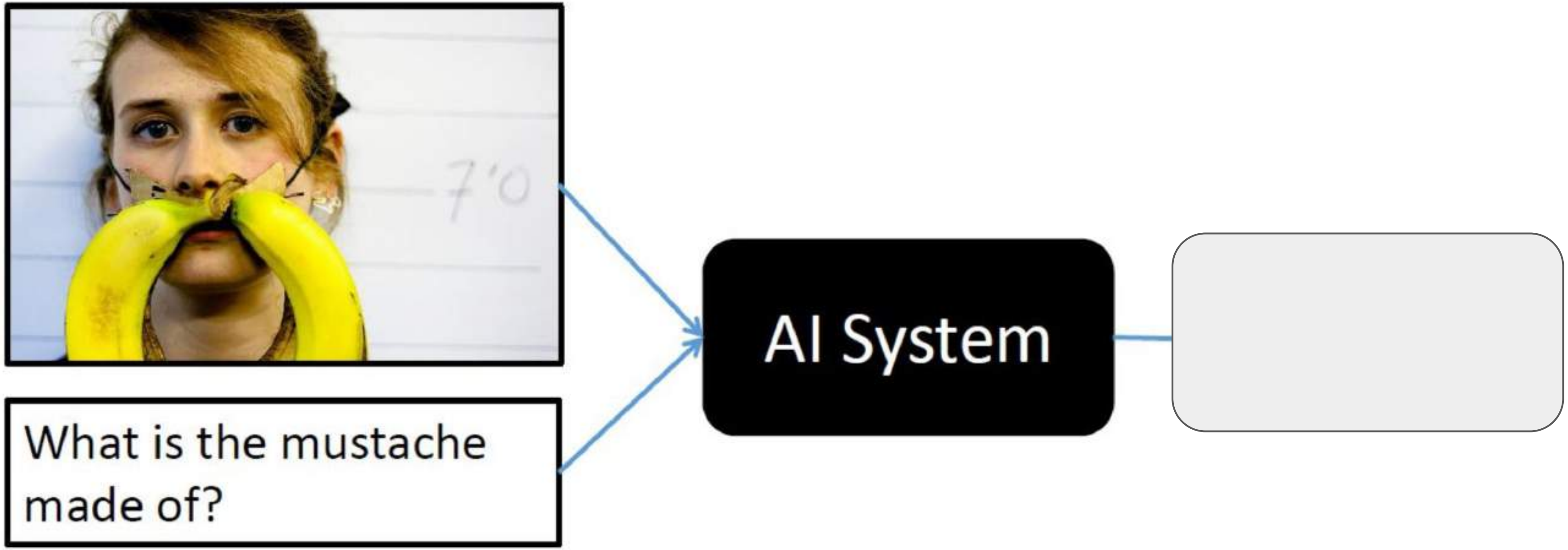
Donahue et al. 2015

Video Captioning

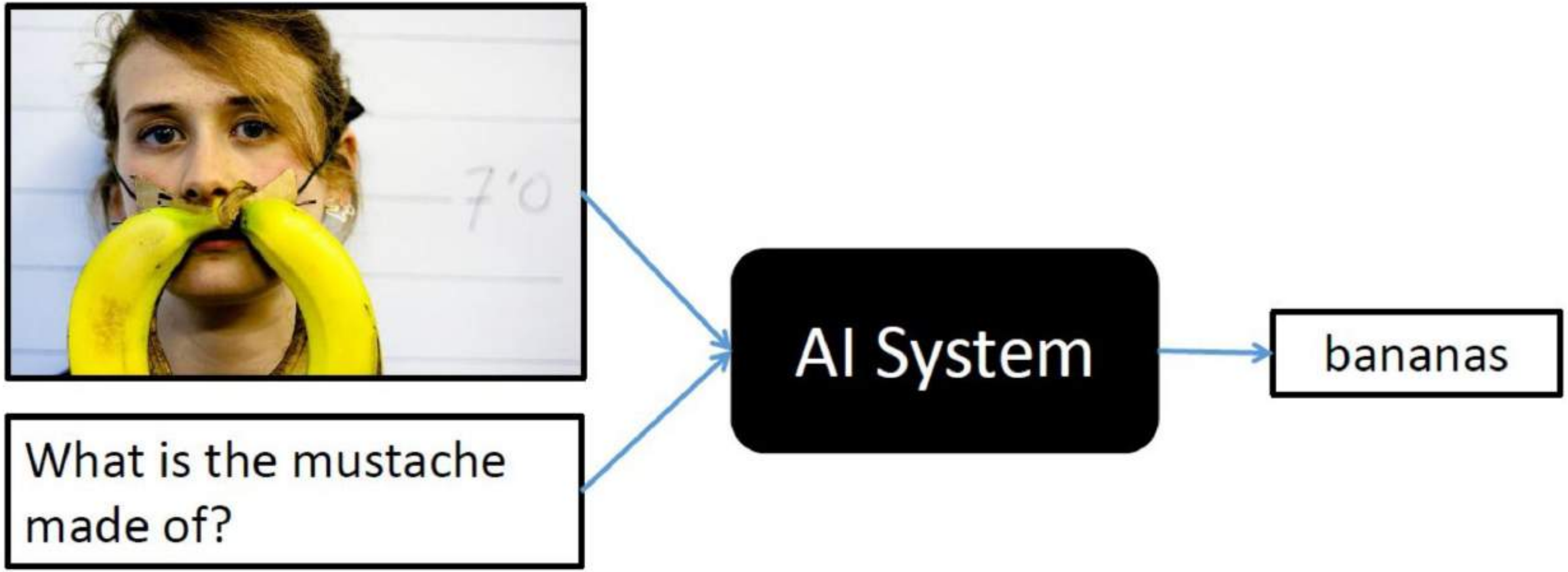


Pan et al. 2016

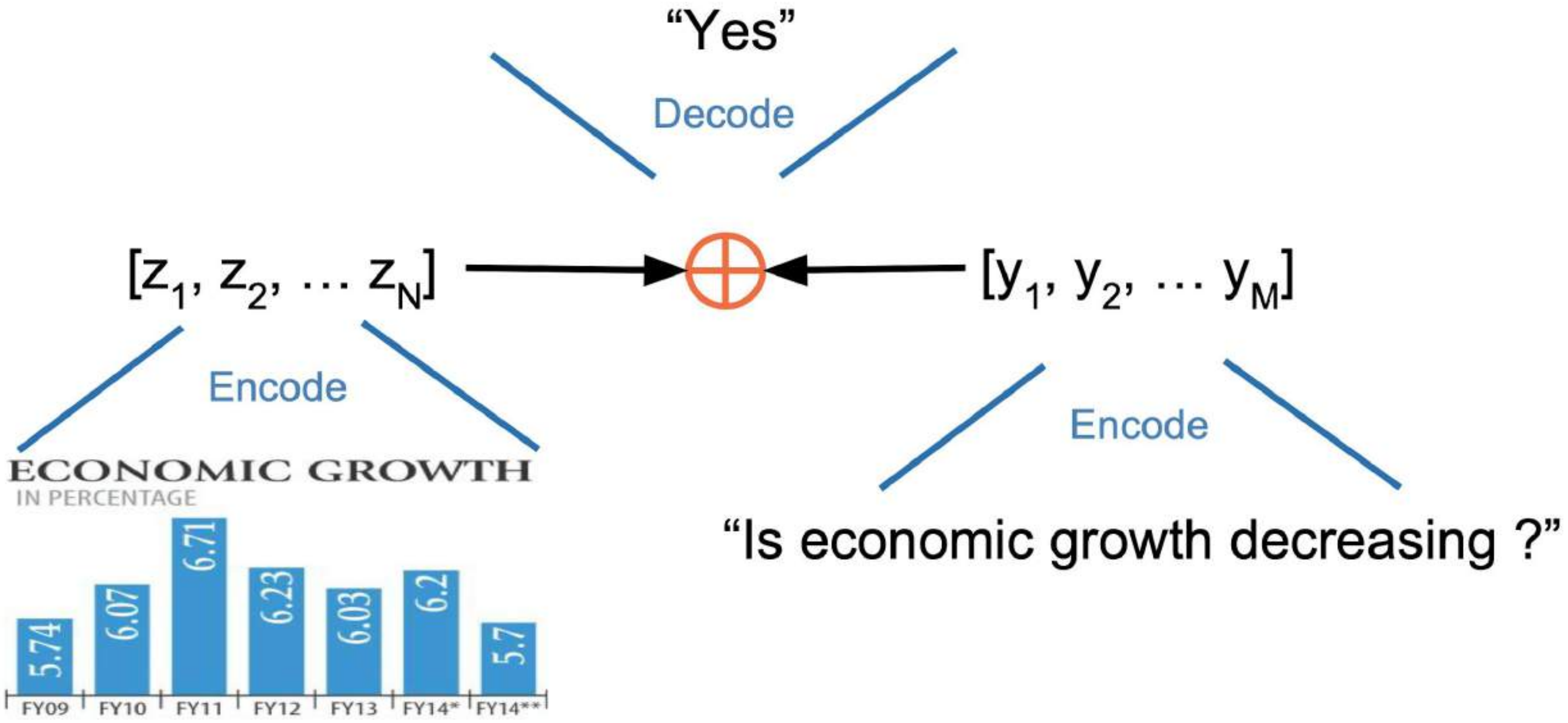
Visual Question Answering



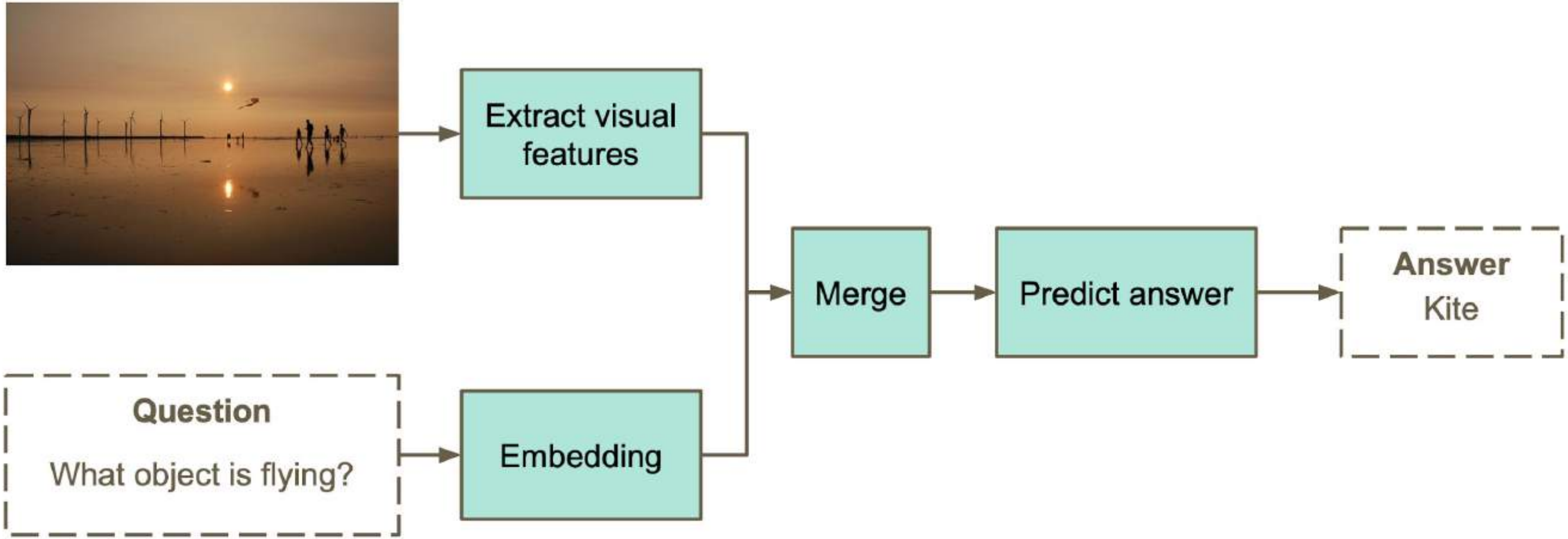
Visual Question Answering



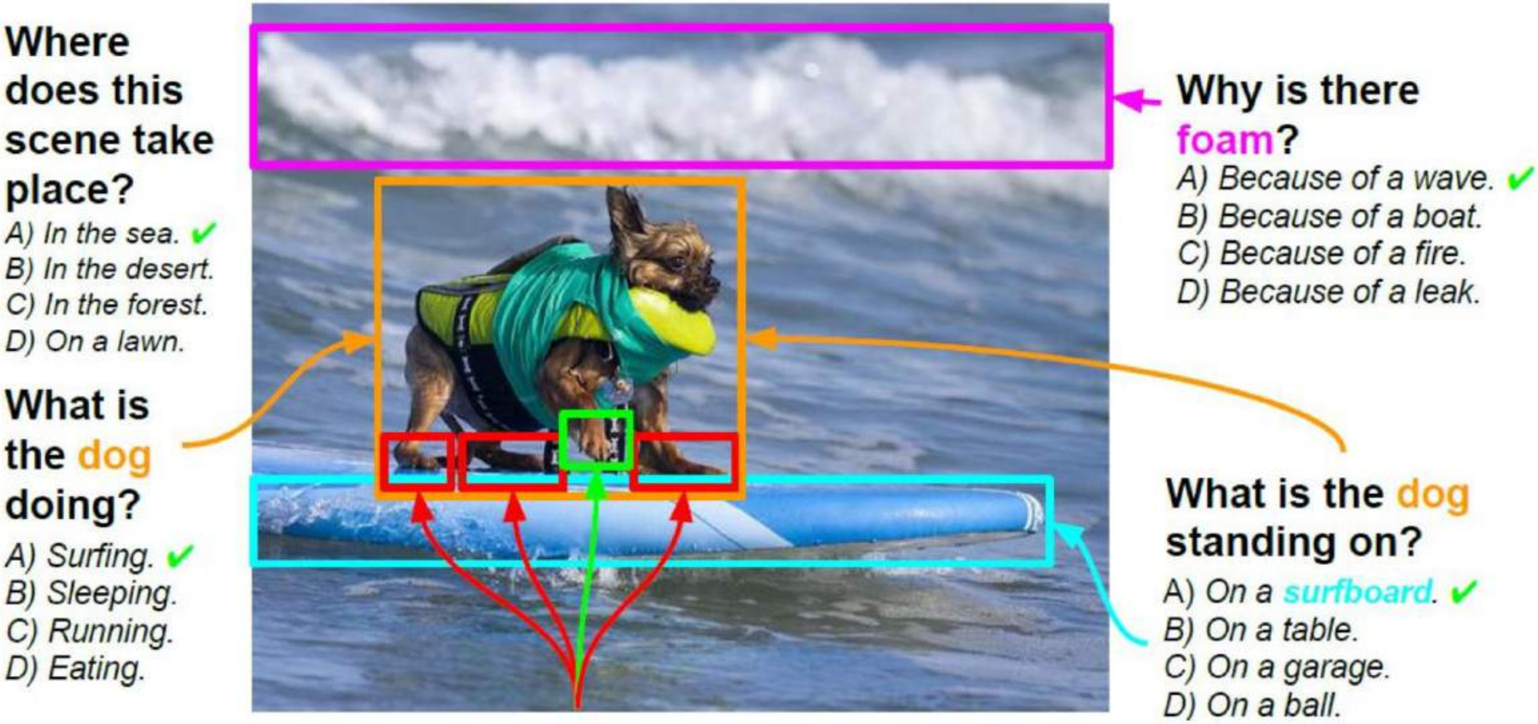
Visual Question Answering



Visual Question Answering



Visual Question Answering



Video Summarization

~1.5 minutes of audio and video

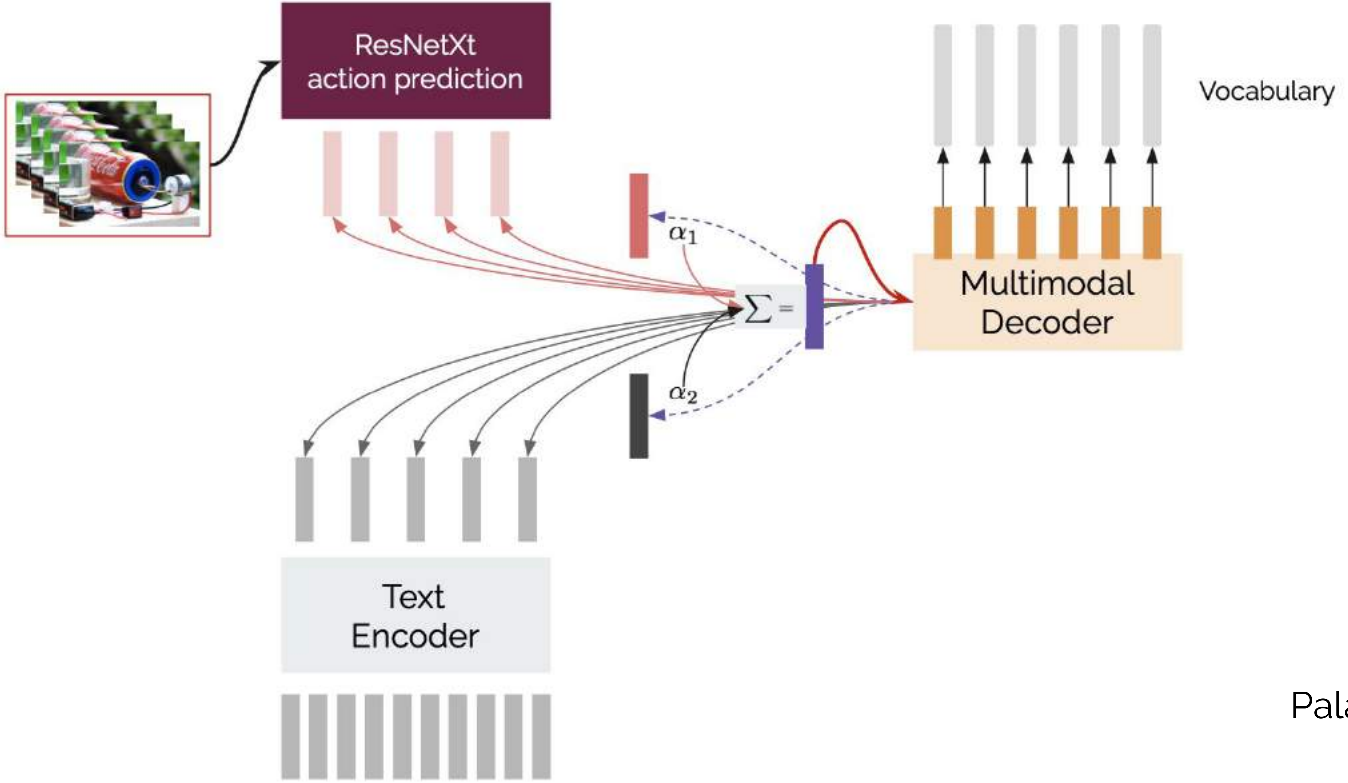
“Teaser” (33 words on avg)

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)

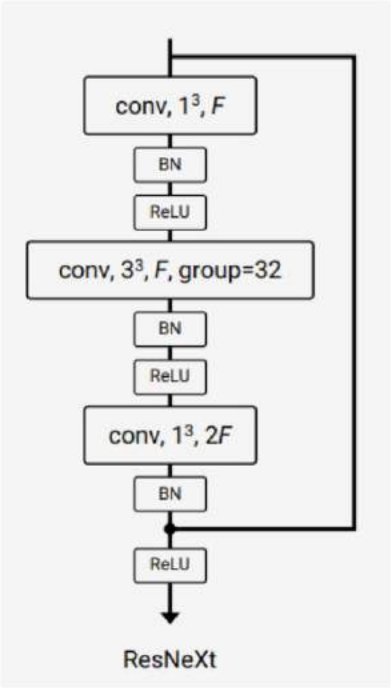
on behalf of expert village my name is lizbeth muller and today we are going to show you how to make spanish omelet . i 'm going to dice a little bit of peppers here . i 'm not going to use a lot , i 'm going to use very very little . a little bit more then this maybe . you can use red peppers if you 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 .

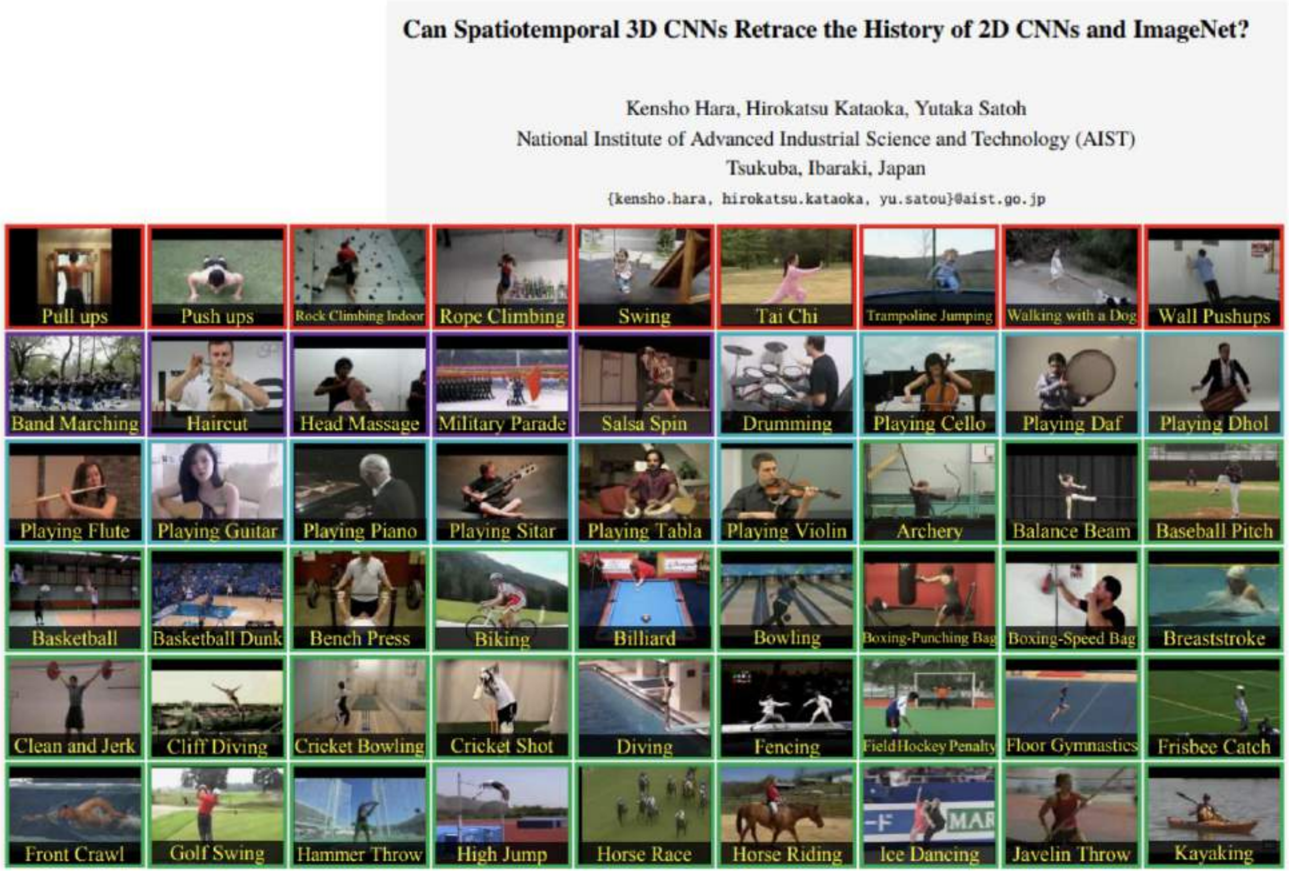
Video Summarization: Hierarchical Model



Palaskar et al. 2019

Action Recognition



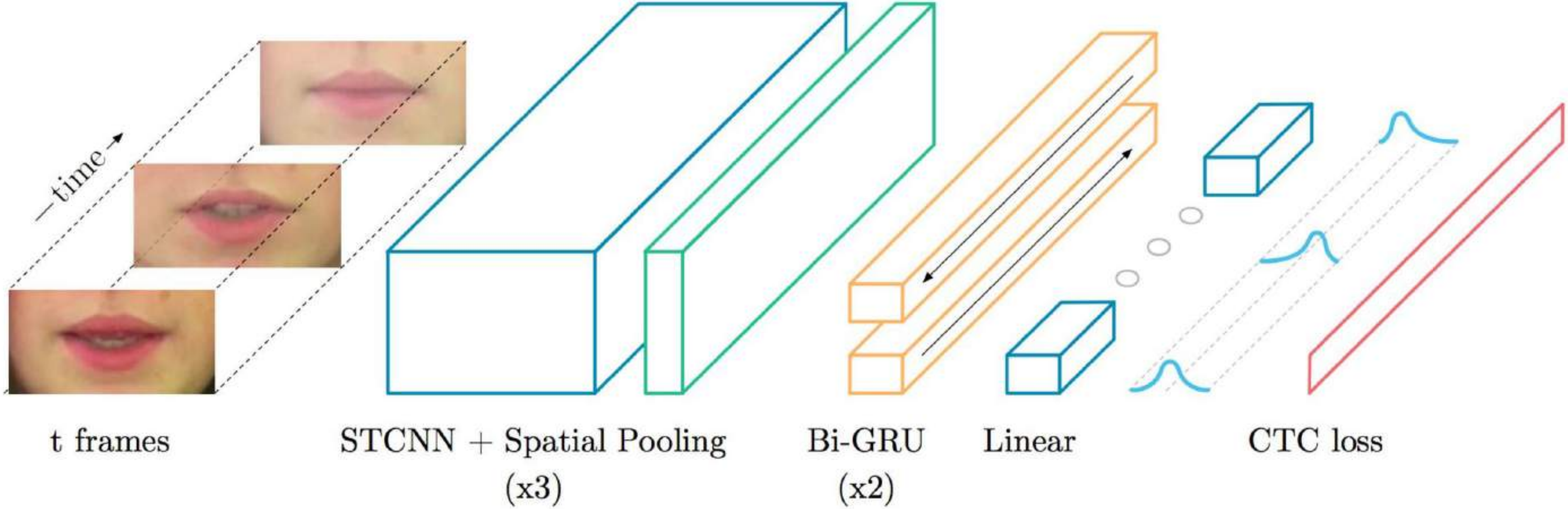


Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?

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

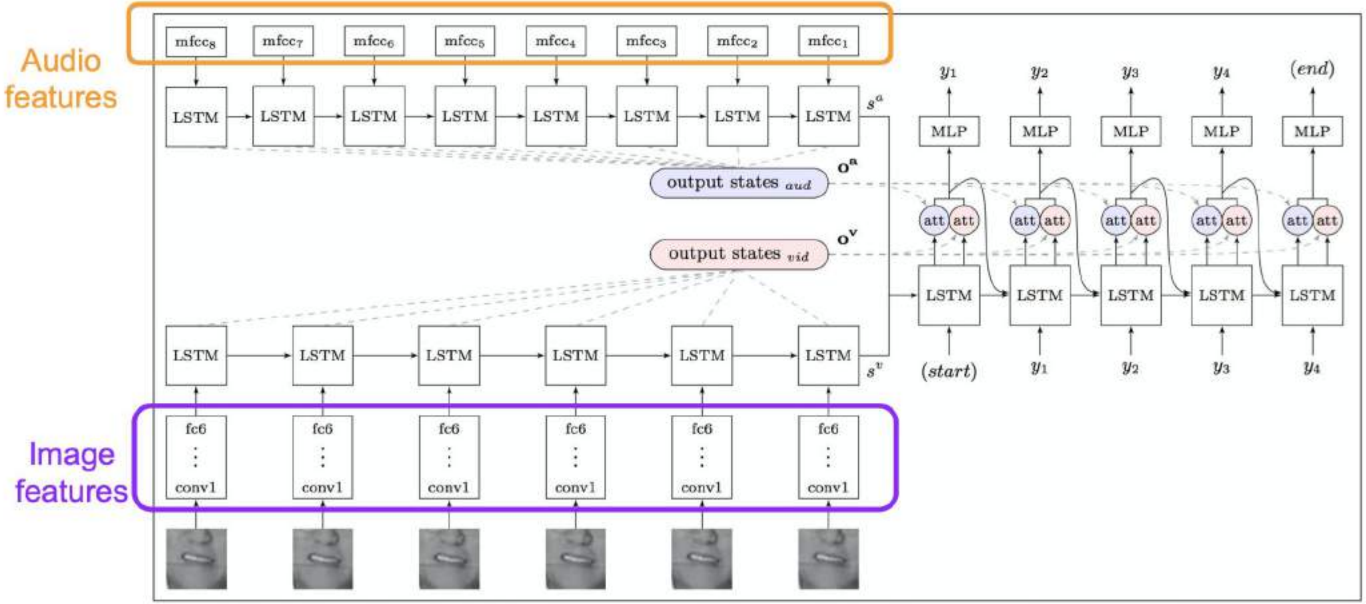
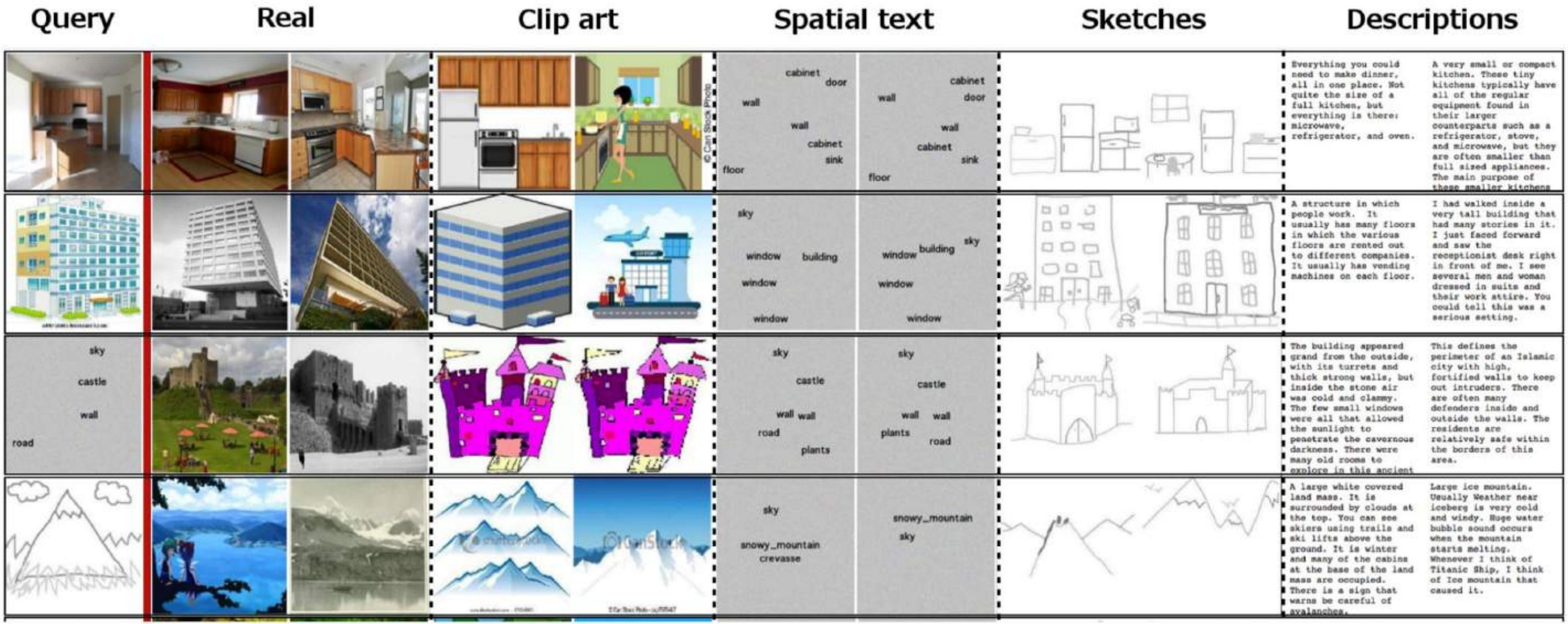


Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character y_i, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

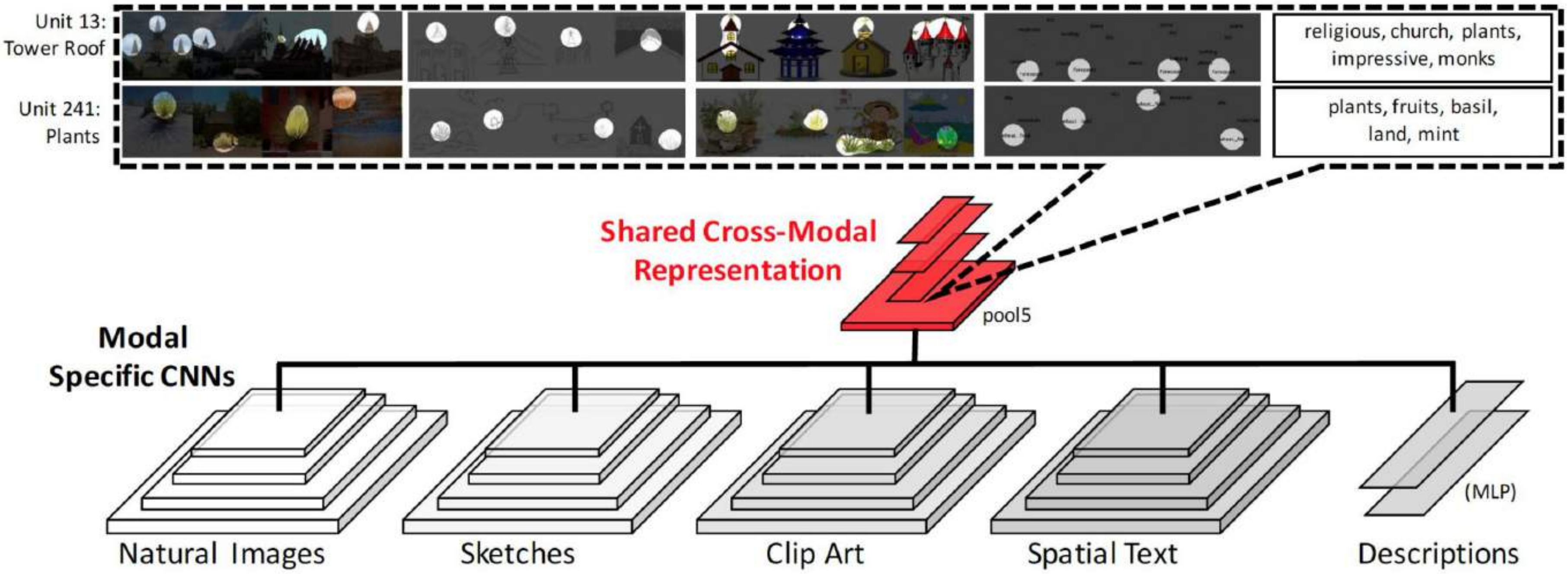
Chung et al. 2017

3. Multimedia Common Tasks

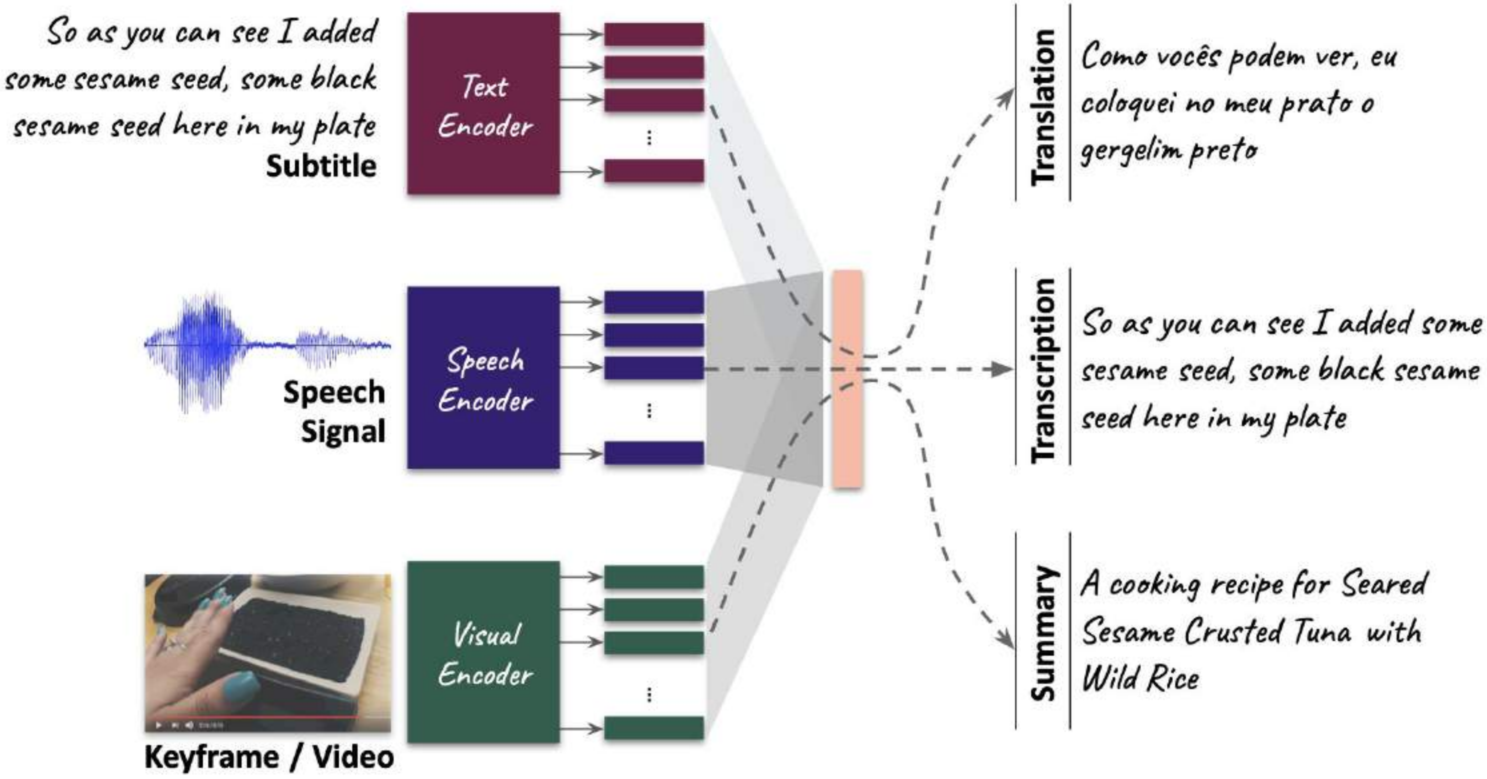
Multimedia Retrieval



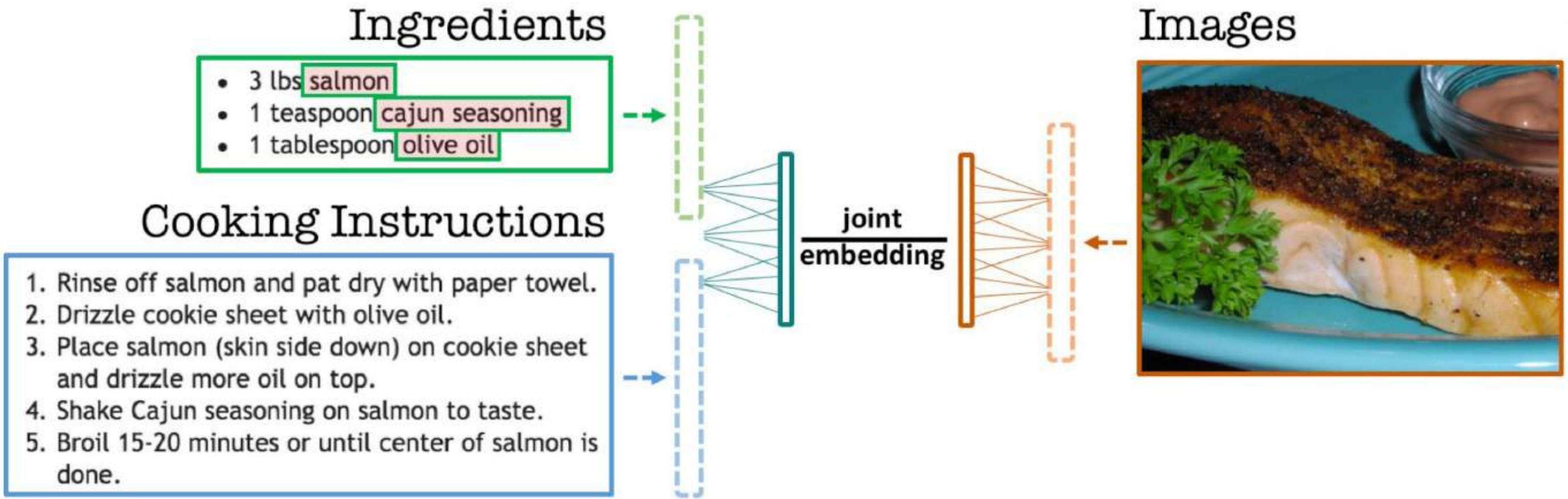
Multimedia Retrieval



Multimedia Retrieval: Shared Multimodal Representation

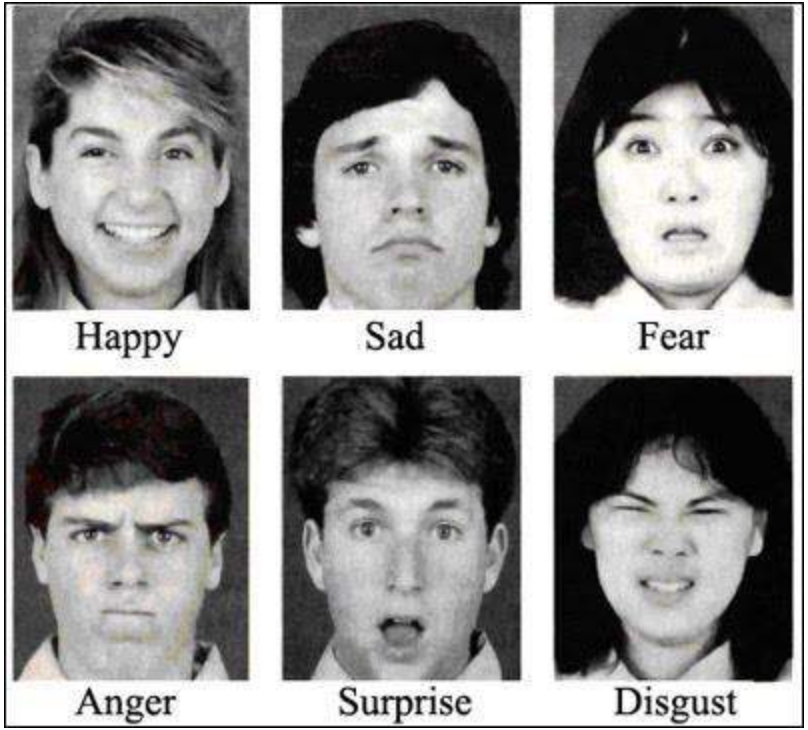


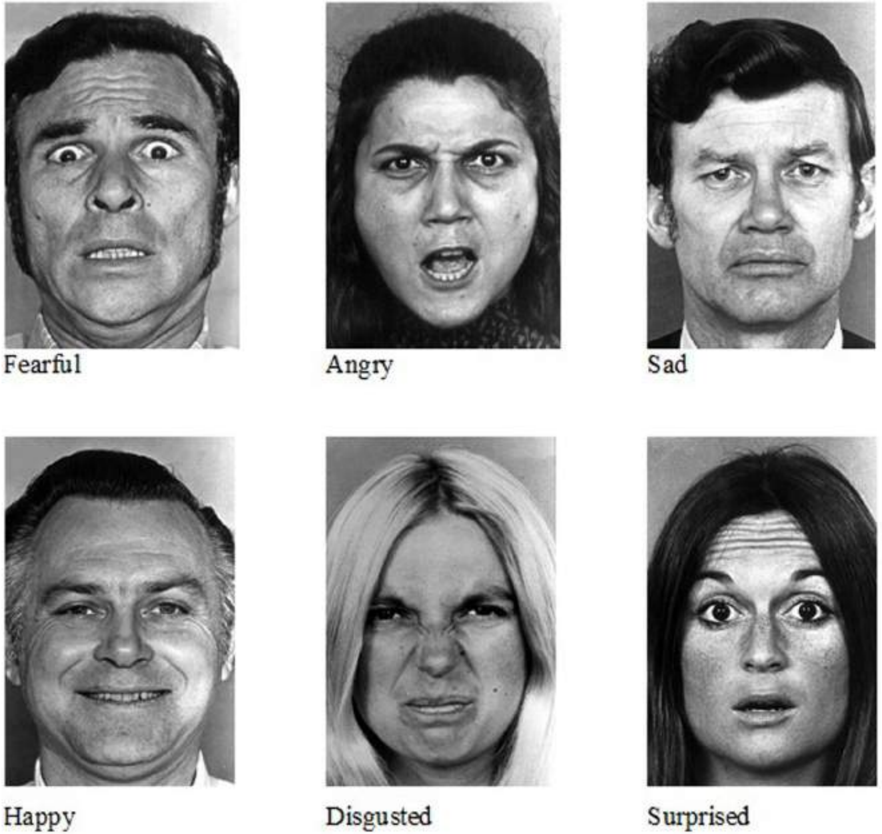
Multimedia Retrieval



4. Emotion and Affect

Affect Recognition:
Emotion, Sentiment, Persuasion, Personality





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