

# DEEP LEARNING FOR SPEECH AND LANGUAGE

Winter School at UPC TelecomBCN Barcelona. 24-30 January 2018.



## Instructors



Marta R.  
Costa-jussà



José A. R.  
Fonollosa



Santiago  
Pascual



Javier  
Hernando



Antonio  
Bonafonte



Xavier  
Giro-i-Nieto

## Organized by



## Supported by



+ info: <https://telecombcn-dl.github.io/2018-dls/>

[\[course site\]](#)



#DLUPC

## Day 2 Lecture 4

# Language and Vision



Xavier Giro-i-Nieto  
[xavier.giro@upc.edu](mailto:xavier.giro@upc.edu)



Associate Professor  
Universitat Politècnica de Catalunya  
Technical University of Catalonia



# Acknowledgments



Antonio  
Bonafonte



Santiago  
Pascual



The slide is titled "Winter Seminar UPC TelecomBCN: 24 - 25 January 2017". It shows a cityscape background with the Sagrada Família. The "Instructors" section lists several names with small profile pictures. The "Organizers" section includes logos for TALP, UPC, and other institutions. A link "+ info: TelecomBCN.DeepLearning.Barcelona" and a "[course site]" button are at the bottom. To the right, there is a smaller image of Antonio Bonafonte, the TALP logo, the UPC logo, and the TelecomBCN logo.

Day 2 Lecture 4  
**Word Embeddings**  
Word2Vec

Antonio Bonafonte

A smaller portrait of Antonio Bonafonte.

# Acknowledgments



Marta R. Costa-jussà



Day 3 Lecture 4  
**Neural Machine Translation**

Marta R. Costa-jussà

The slide is titled "Day 3 Lecture 4 Neural Machine Translation". It features a cityscape background. At the top left, there's a section for "Instructors" showing small profile pictures. In the center, there's a section for "Organizers" with logos for TALP and other organizations. At the bottom, there's a link "+ info: TelecomBCN.DeepLearning.Barcelona" and a "Course site" button. The slide is part of a larger presentation, with a video thumbnail and the Universitat Politècnica de Catalunya logo visible below it.



Day 4 Lecture 2  
**Advanced Neural Machine Translation**

Marta R. Costa-jussà

This slide is titled "Day 4 Lecture 2 Advanced Neural Machine Translation". It has a similar layout to the previous slide, with a cityscape background, sections for instructors and organizers, and links for more information. It also includes a video thumbnail and the Universitat Politècnica de Catalunya logo.

# Outline

1. Motivation
2. Image Captioning
3. Visual Question Answering / Reasoning
4. Joint Embeddings

# Outline

1. Motivation
2. Image and Video Captioning
3. Visual Question Answering / Reasoning
4. Joint Embeddings



Xavier Giró-i-Nieto  
@DocXavi



Take home message by @karpathy : read papers from machine translation community.  
**#deeplearning16 #cvpr16**

Tradueix del anglès



# DEEP LEARNING FOR SPEECH & LANGUAGE

Winter Seminar UPC TelecomBCN, 24 - 31 January 2017



## Instructors



Antonio  
Bonafonte



J. Adrián Rodríguez  
Fonollosa



Marta R.  
Costa-jussà



Javier  
Hernando



Santiago  
Pascual



Elisa  
Sayrol



Xavier  
Giró

## Organizers



Image Processing Group  
Signal Theory and Communications Department



UNIVERSITAT POLITÈCNICA  
DE CATALUNYA  
BARCELONATECH



+ info: [TelecomBCN.DeepLearning.Barcelona](http://TelecomBCN.DeepLearning.Barcelona)

[\[course site\]](#)

con-  
eius-  
re et  
n ve-  
umco  
con-  
hen-  
re eu  
cca-  
culpa  
labo-  
  
idip-  
icidi-  
a. Ut  
exer-  
ex ea  
dolor  
e cil-  
cep-  
dent,  
nollit  
n sit

sit amet, consectetur adipiscing elit,  
sed do eiusmod tempor incididunt

# Text (English)

repi-  
son-  
nifi-  
cou-  
was  
blar  
whi-  
wait  
the  
sint  
in c  
est  
sect  
tem  
na z  
nos  
aliq  
autc  
lupt  
null  
con  
tem  
na a

# Neural Machine Translation



co  
ei  
re  
n  
un  
co  
h  
re  
cc  
:u  
lab

id  
ici  
a.  
ex  
ex  
id  
e  
cc  
de  
to  
n

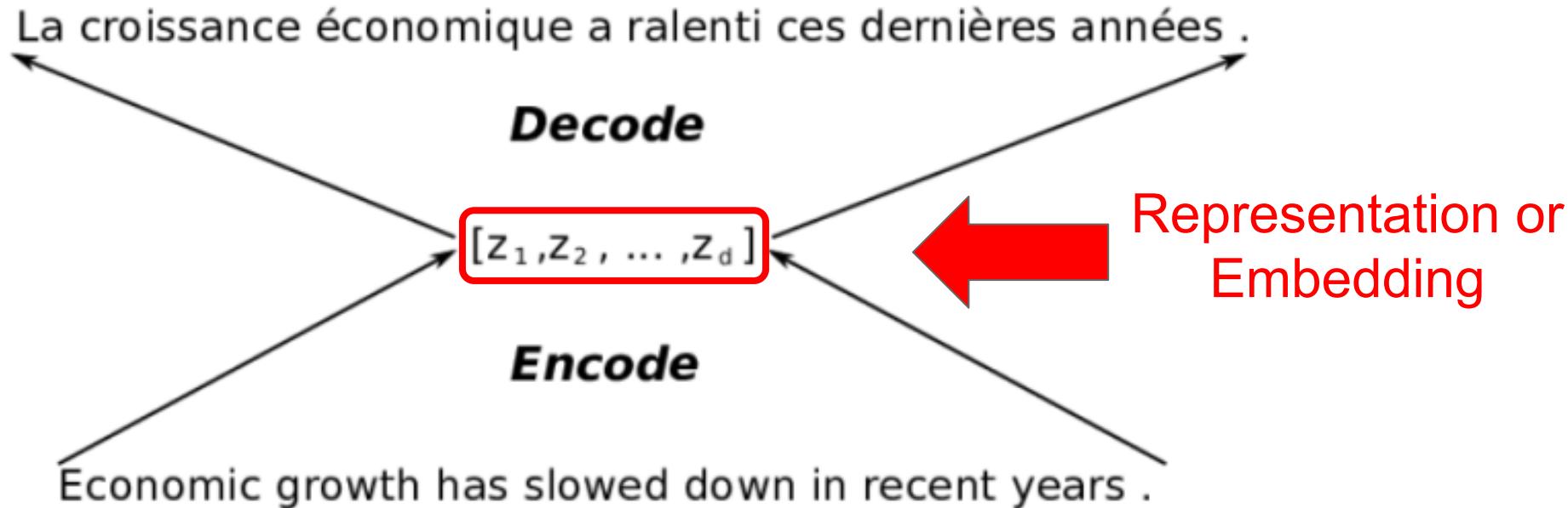
## Section 2

sit amet, consectetur adipiscing elit,  
sed do eiusmod tempor incididunt

do aliquip ex ea commodo conse

# Text (French)

# Neural Machine Translation

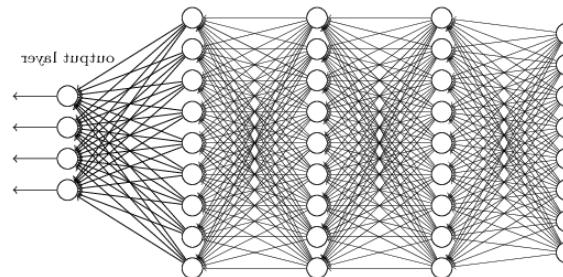
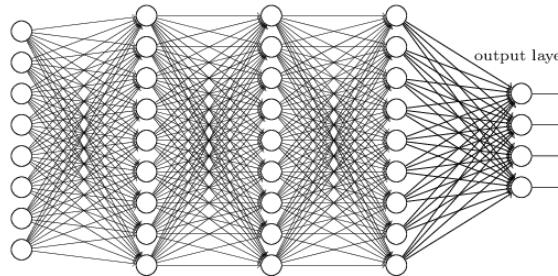


Economic growth has slowed down in recent years .



## Representation or Embedding

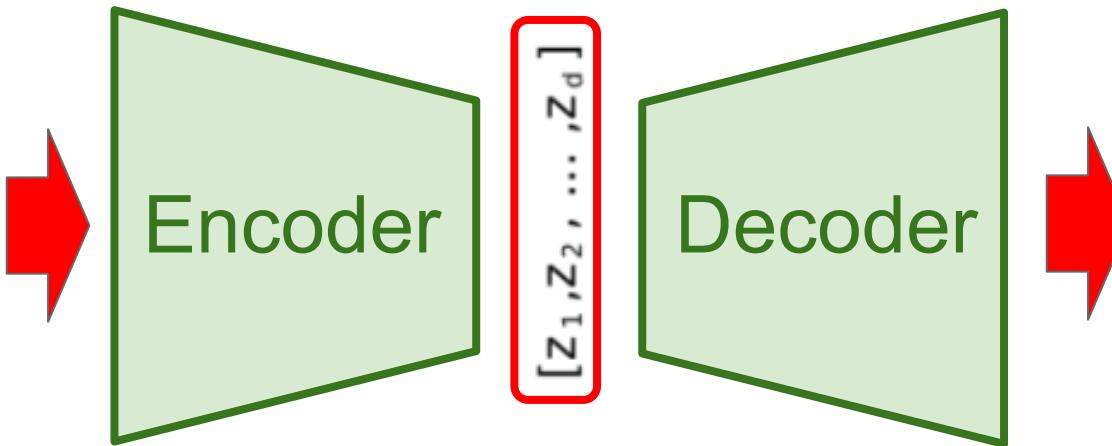
$[z_1, z_2, \dots, z_d]$



La croissance économique a ralenti ces dernières années .

Economic growth has slowed down in recent years .

## Representation or Embedding



La croissance économique a ralenti ces dernières années .

# Word Embeddings

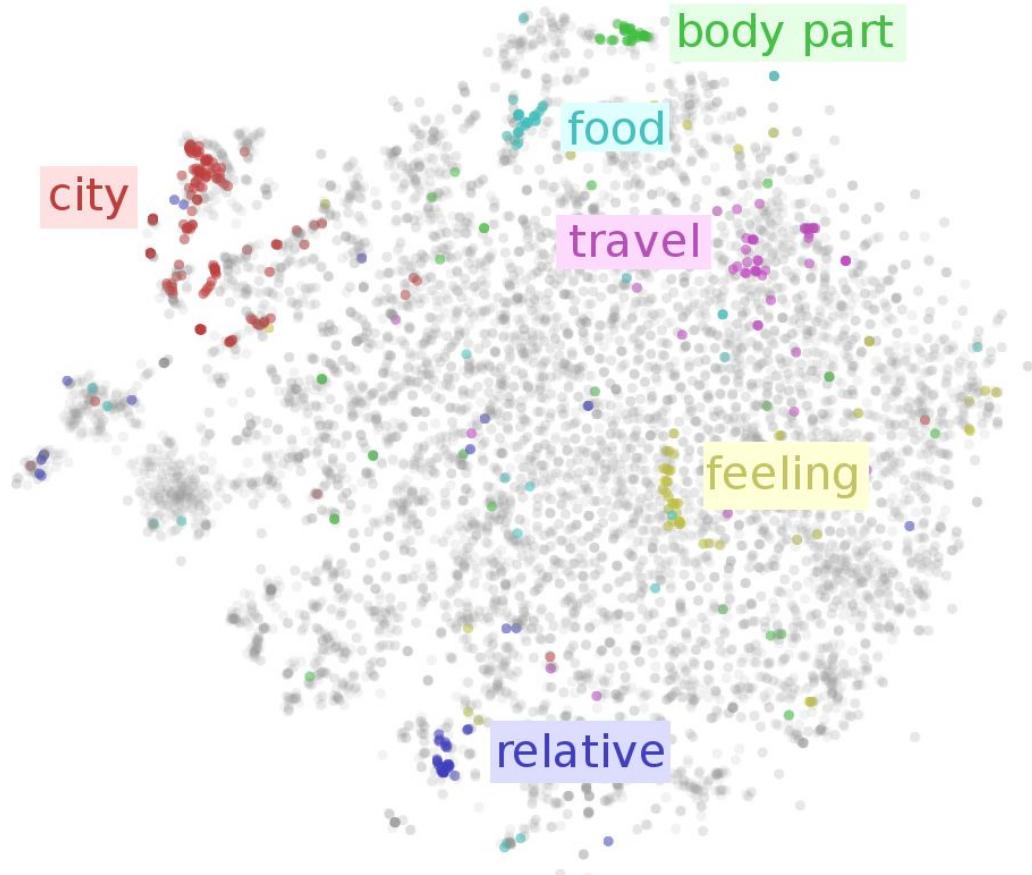


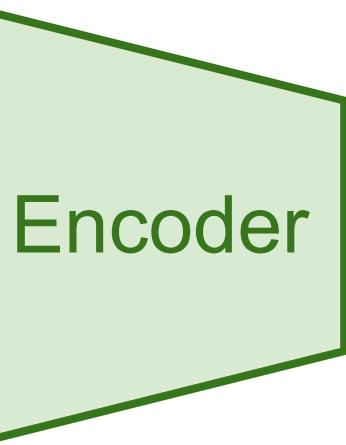
Figure:  
[Christopher Olah](#)  
[Visualizing Representations](#)

# Outline

1. Motivation
2. **Image and Video Captioning**
3. Visual Question Answering / Reasoning
4. Joint Embeddings

La croissance économique a ralenti ces dernières années.

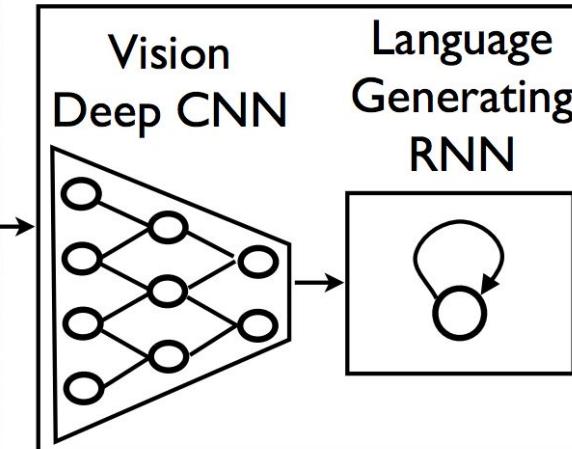
## Representation or Embedding



$[z_1, z_2, \dots, z_d]$



# Captioning: Show & Tell



**A group of people shopping at an outdoor market.**

**There are many vegetables at the fruit stand.**

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "[Show and tell: A neural image caption generator.](#)" CVPR 2015. [\[video\]](#)

# Captioning: DeeplImageSent



man in black shirt is playing guitar.



construction worker in orange safety vest is working on road.



two young girls are playing with lego toy.

(Slides by Marc Bolaños): Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." CVPR 2015

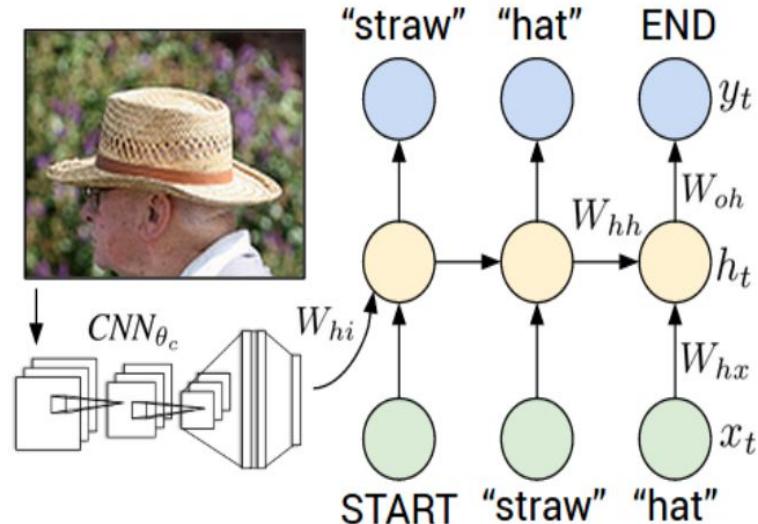
# Captioning: DeeplImageSent

only takes into account  
image features in the first  
hidden state

$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$

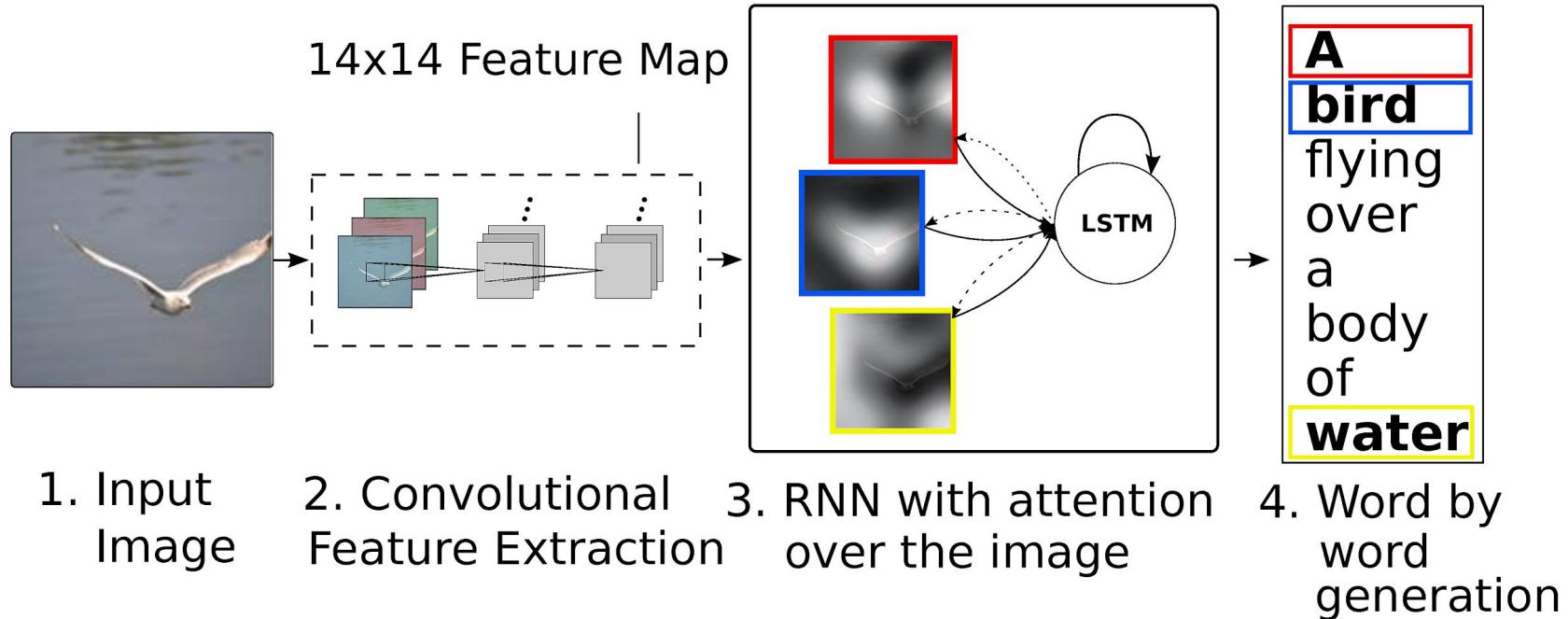
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1) \odot b_v)$$

$$y_t = \text{softmax}(W_{oh}h_t + b_o).$$



Multimodal Recurrent  
Neural Network

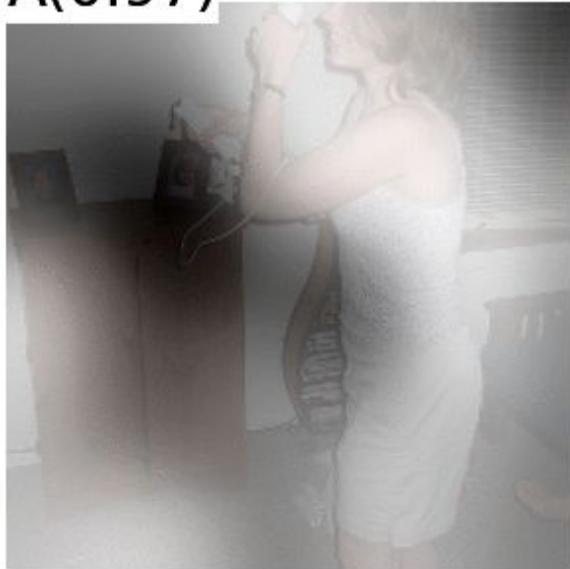
# Captioning: Show, Attend & Tell



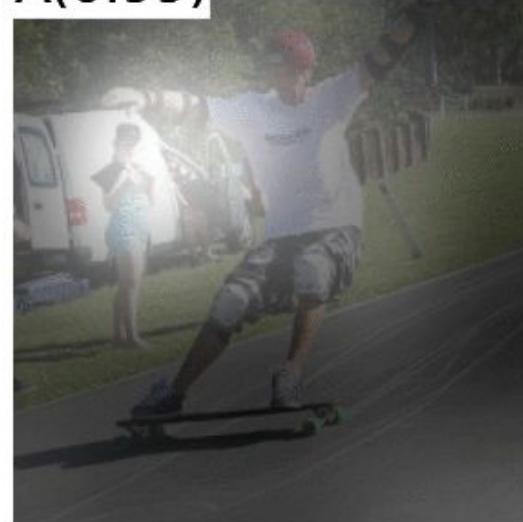
Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. "[Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.](#)" ICML 2015

# Captioning: Show, Attend & Tell

A(0.97)

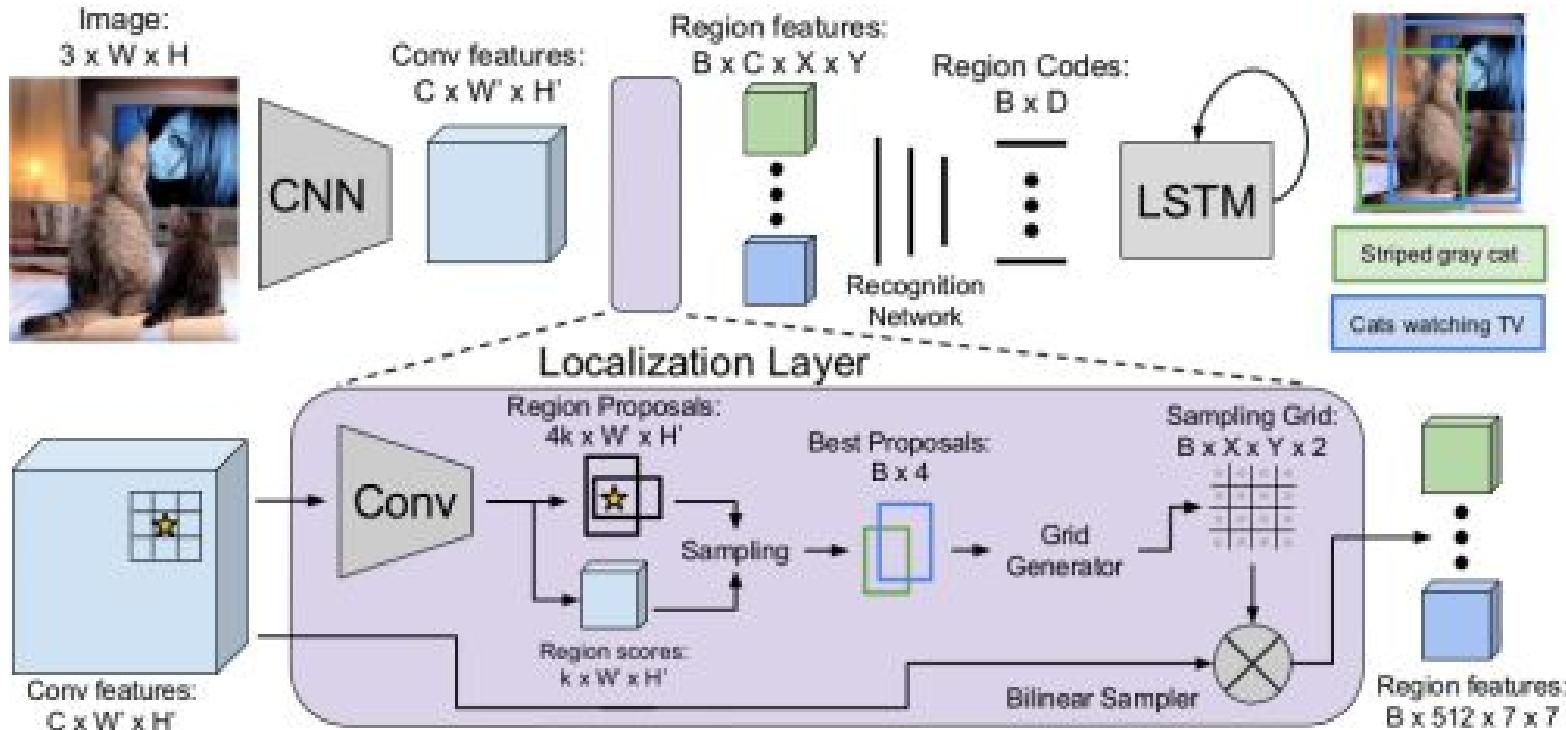


A(0.99)



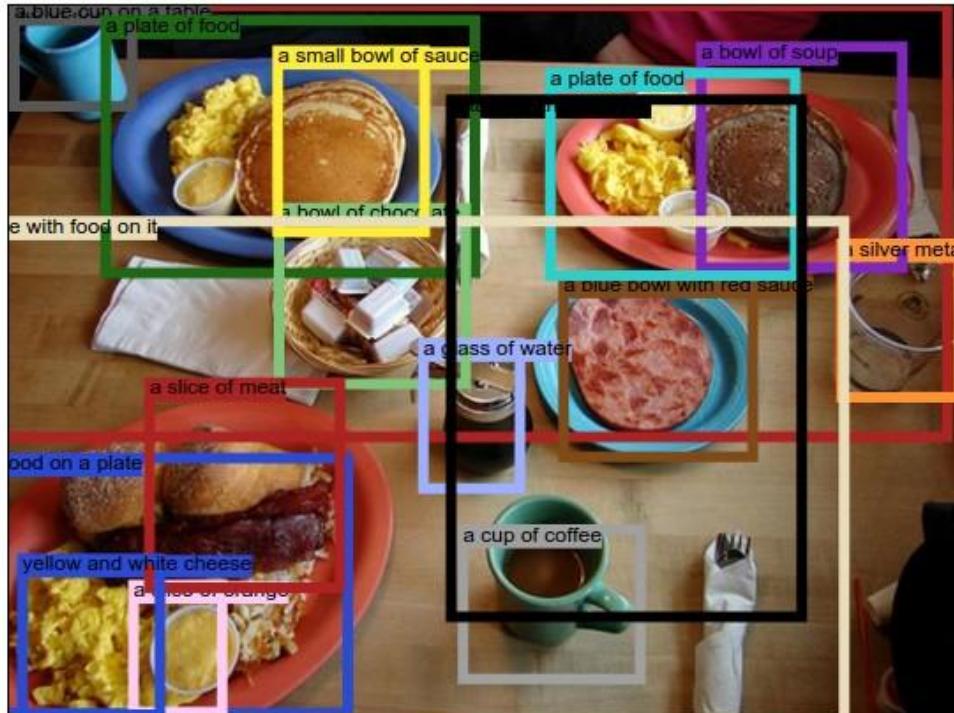
Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. "[Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.](#)" ICML 2015

# Captioning (+ Detection): DenseCap



Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. ["Densecap: Fully convolutional localization networks for dense captioning."](#) CVPR 2016

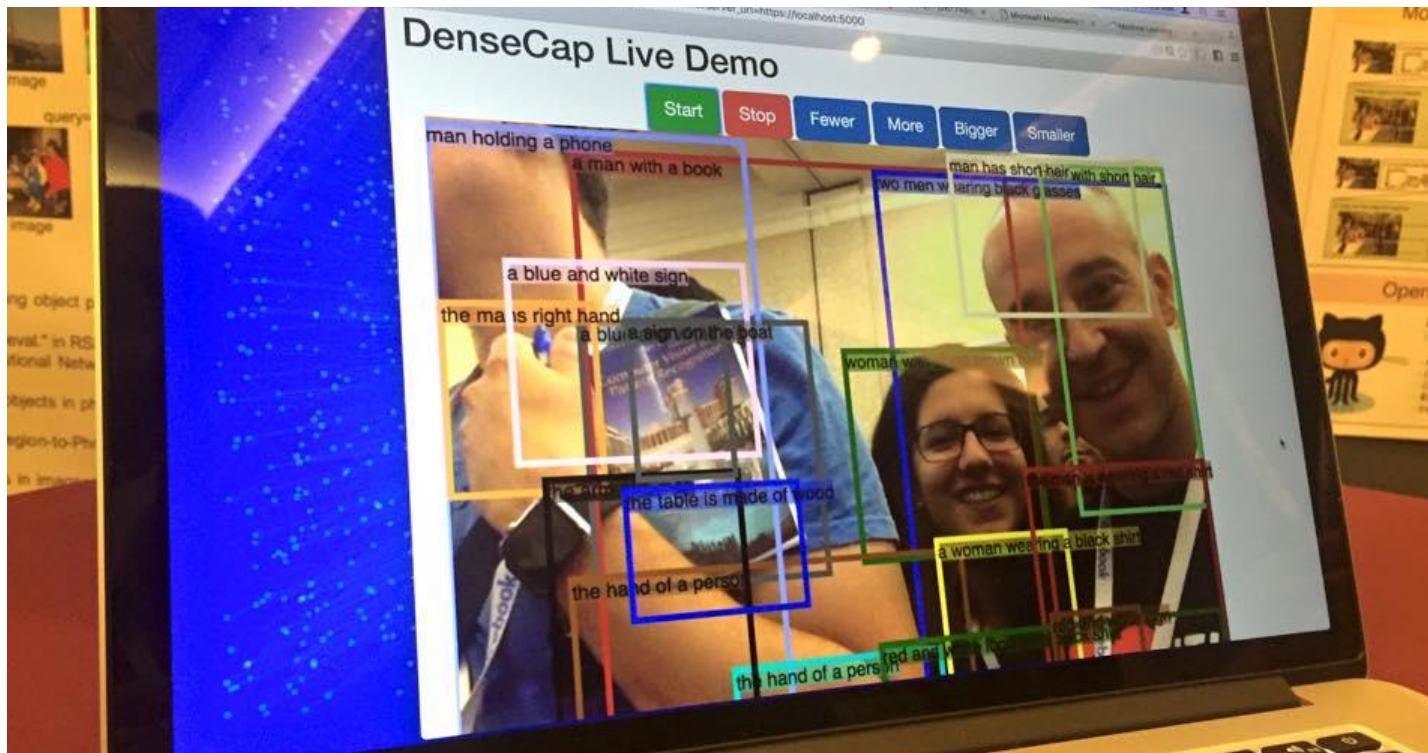
# Captioning (+ Detection): DenseCap



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, Justin, Andrej Karpathy, and Li Fei-Fei. ["Densecap: Fully convolutional localization networks for dense captioning."](#) CVPR 2016

# Captioning (+ Detection): DenseCap



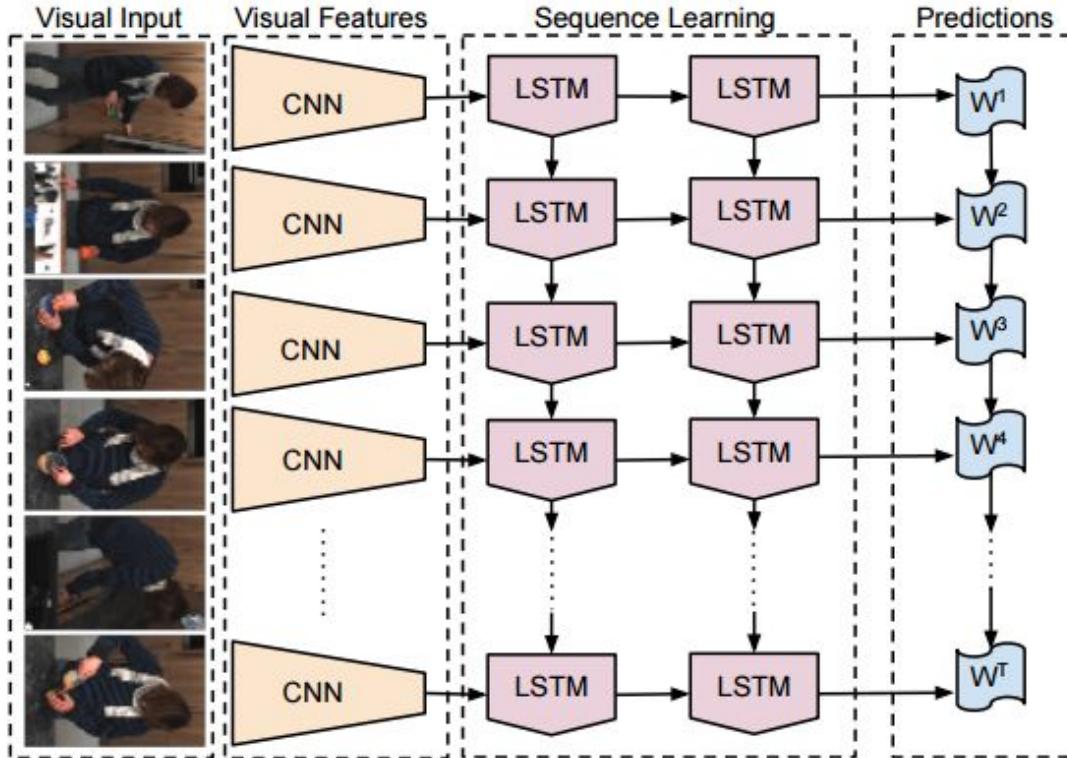
XAVI: “man has short hair”, “man with short hair”

AMAIA: “a woman wearing a black shirt”, “

BOTH: “two men wearing black glasses”

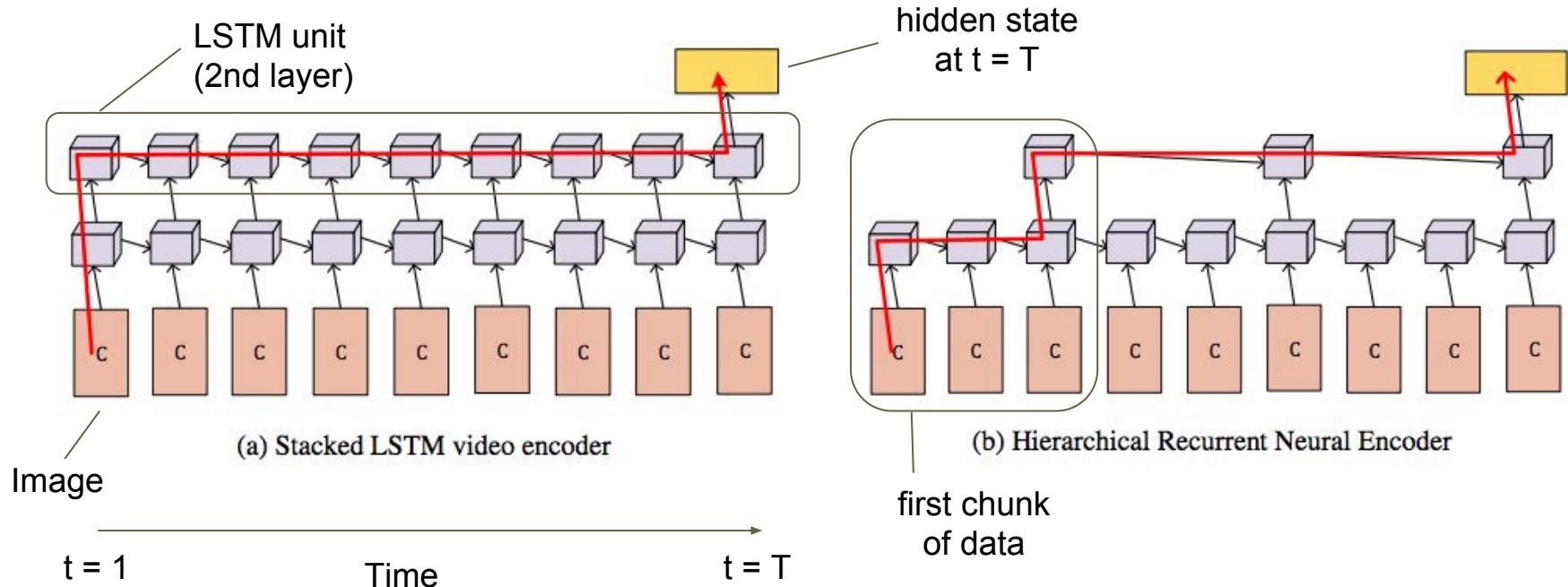
Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. [“Densecap: Fully convolutional localization networks for dense captioning.”](#) CVPR 2016

# Captioning: Video



Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrel. [Long-term Recurrent Convolutional Networks for Visual Recognition and Description](#), CVPR 2015. [code](#)

# Captioning: Video



(Slides by Marc Bolaños) Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, Yueling Zhuang [Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning](#), CVPR 2016.



Chung, Joon Son, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. "Lip reading sentences in the wild." CVPR 2017

# Lipreading: Watch, Listen, Attend & Spell

Audio  
features

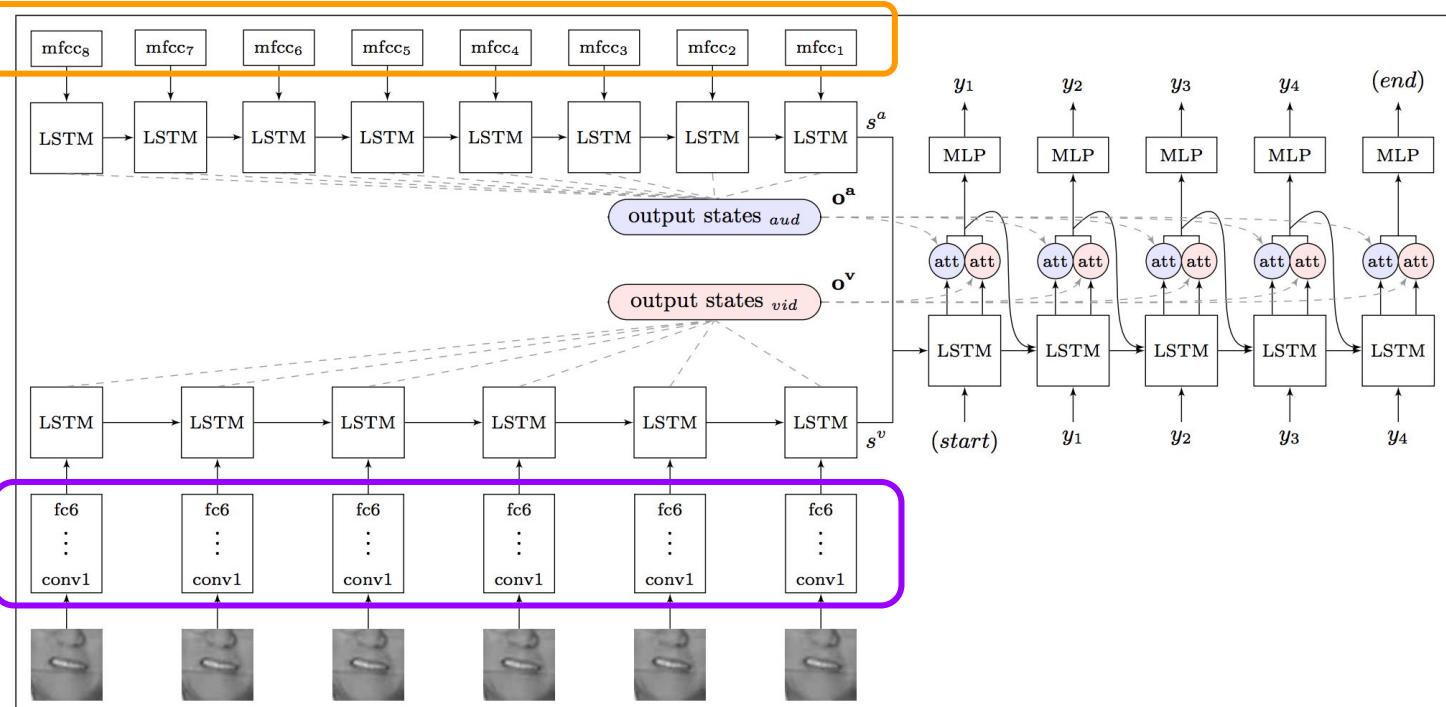


Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character  $y_i$ , as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

Chung, Joon Son, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. ["Lip reading sentences in the wild."](#) CVPR 2017

# Lipreading: Watch, Listen, Attend & Spell

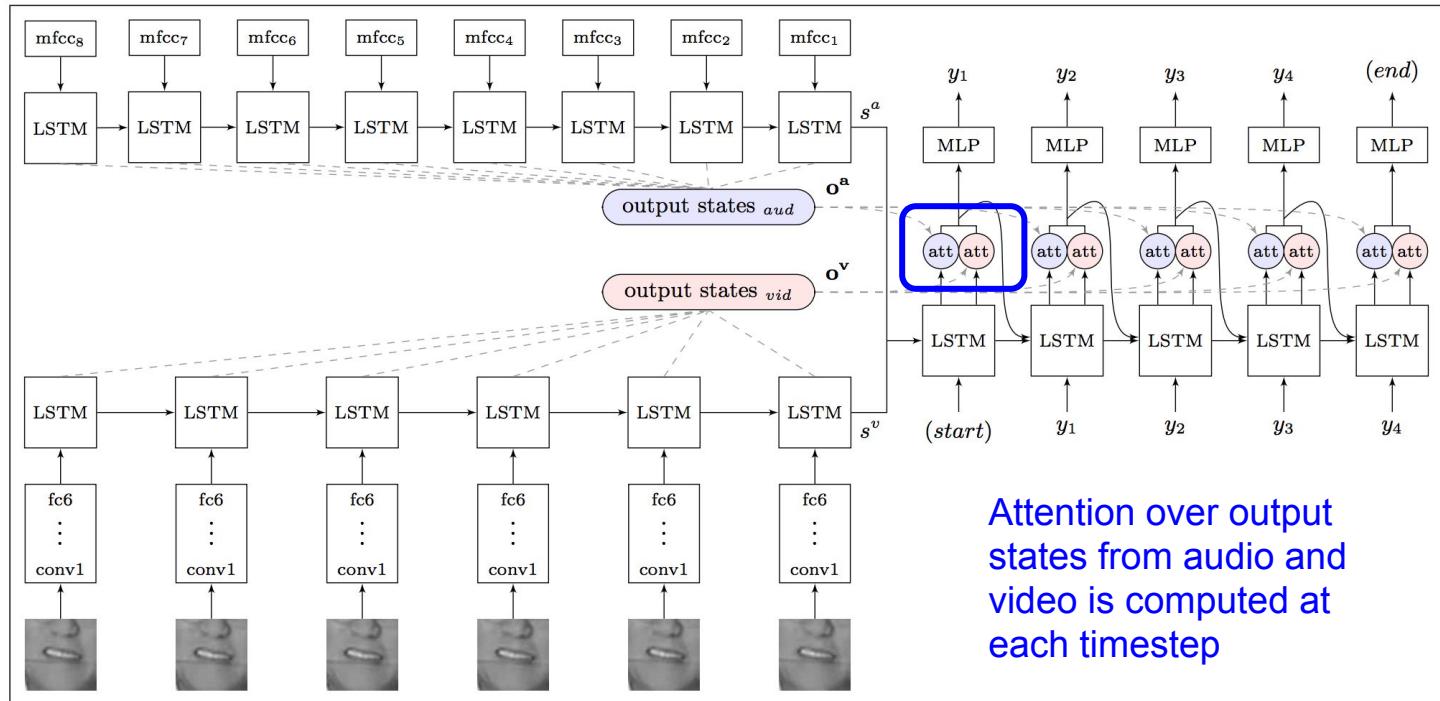


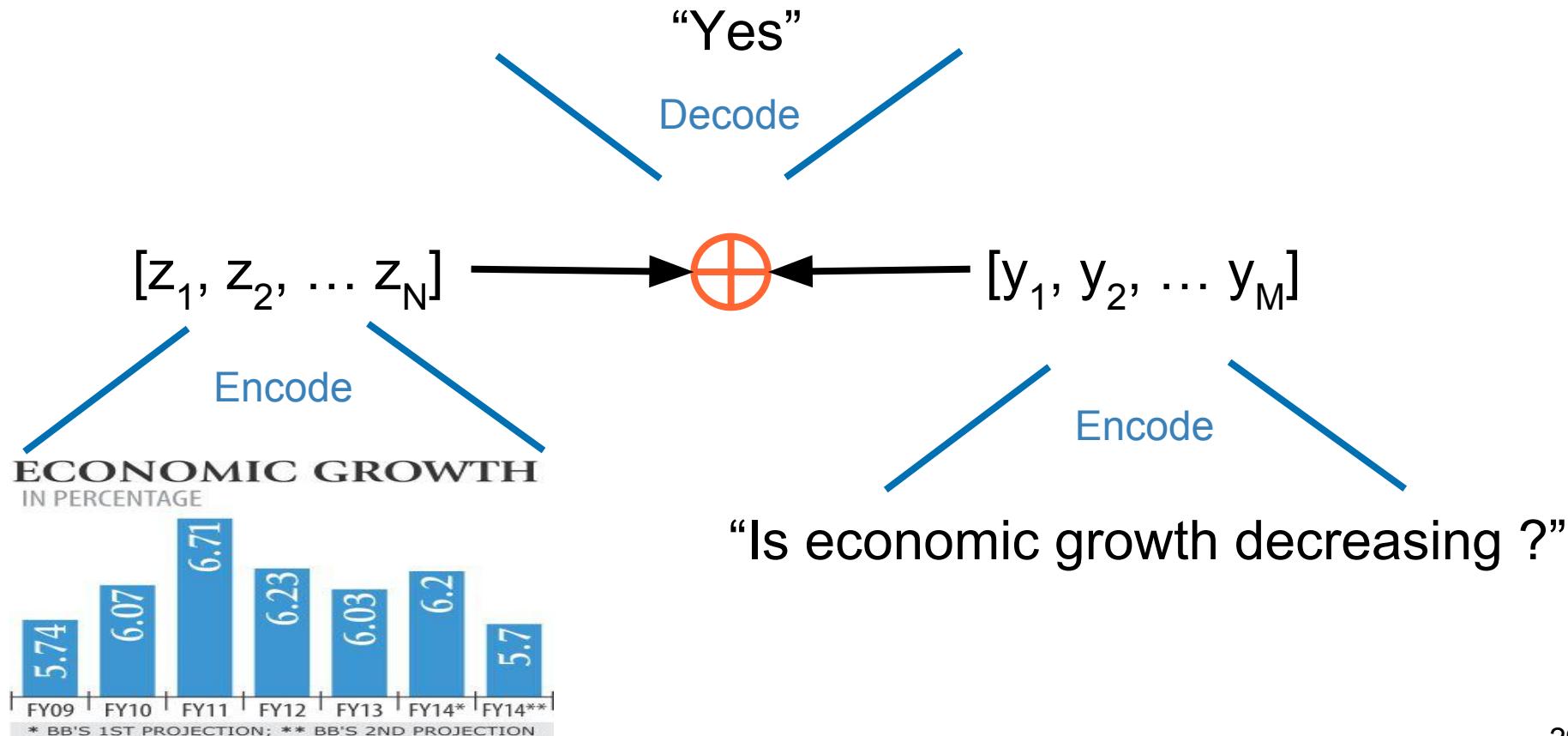
Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character  $y_i$ , as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

Chung, Joon Son, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. ["Lip reading sentences in the wild."](#) CVPR 2017

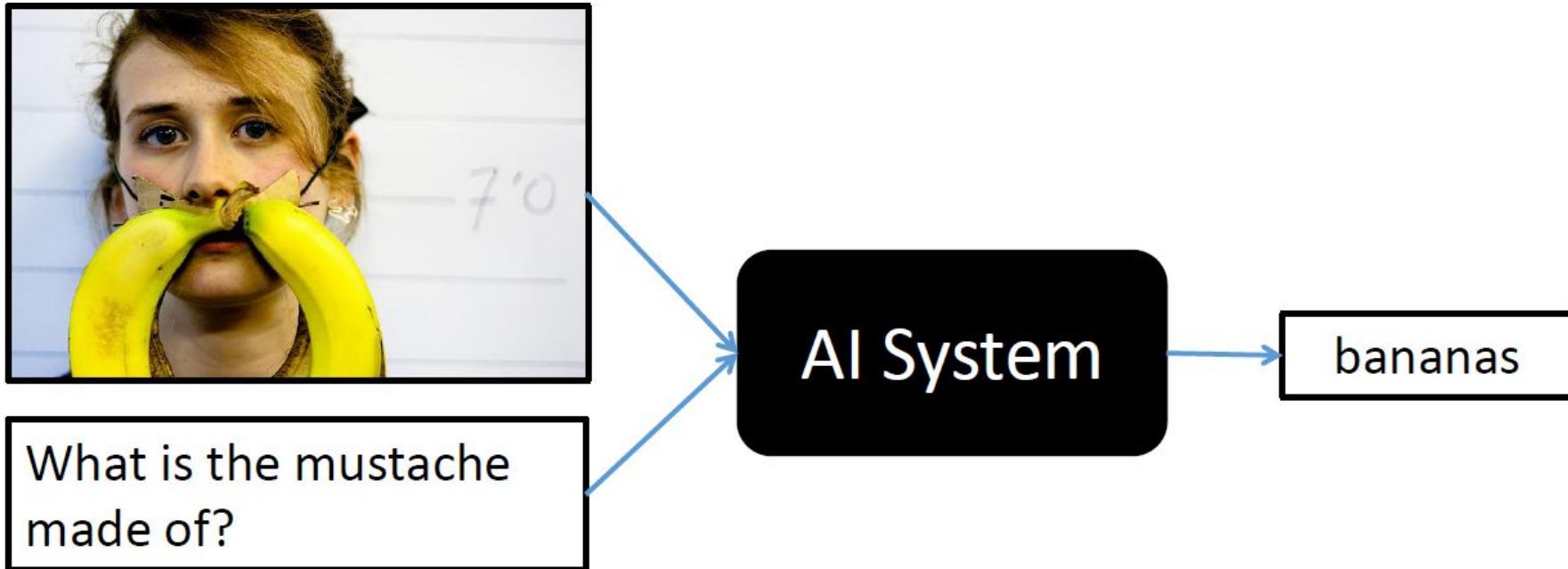
# Outline

1. Neural Machine Transaltion (no vision here !)
2. Image and Video Captioning
3. **Visual Question Answering / Reasoning**
4. Joint Embeddings

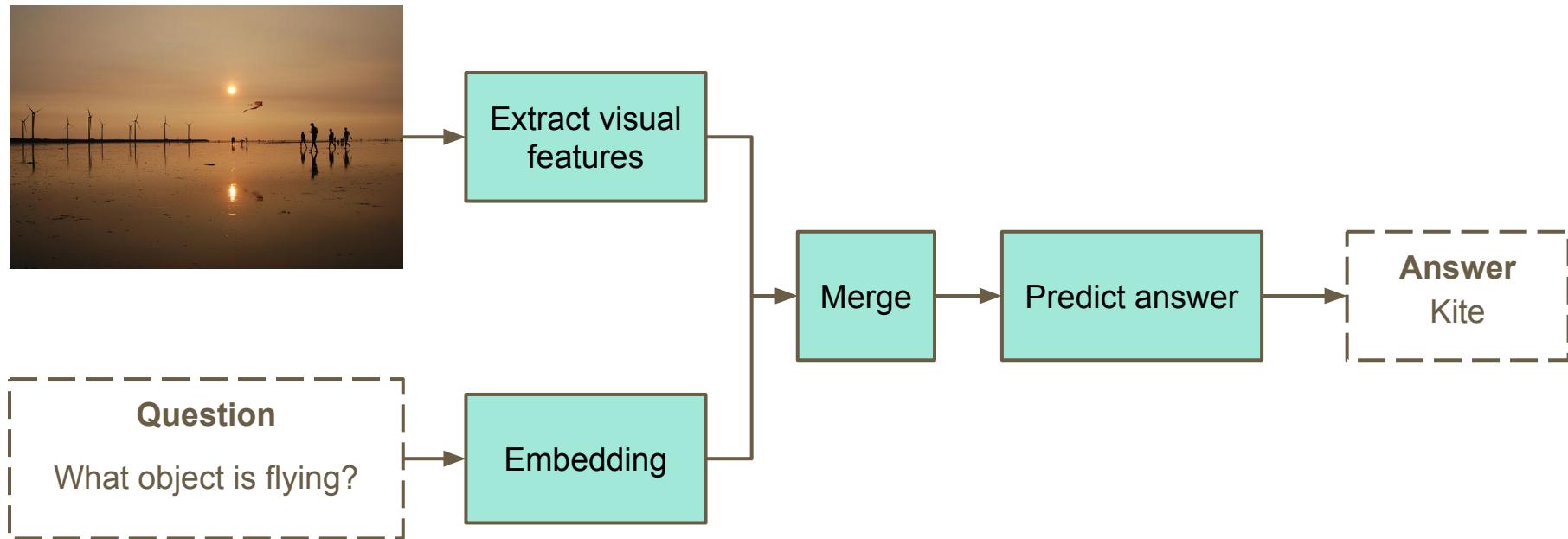
# Visual Question Answering (VQA)



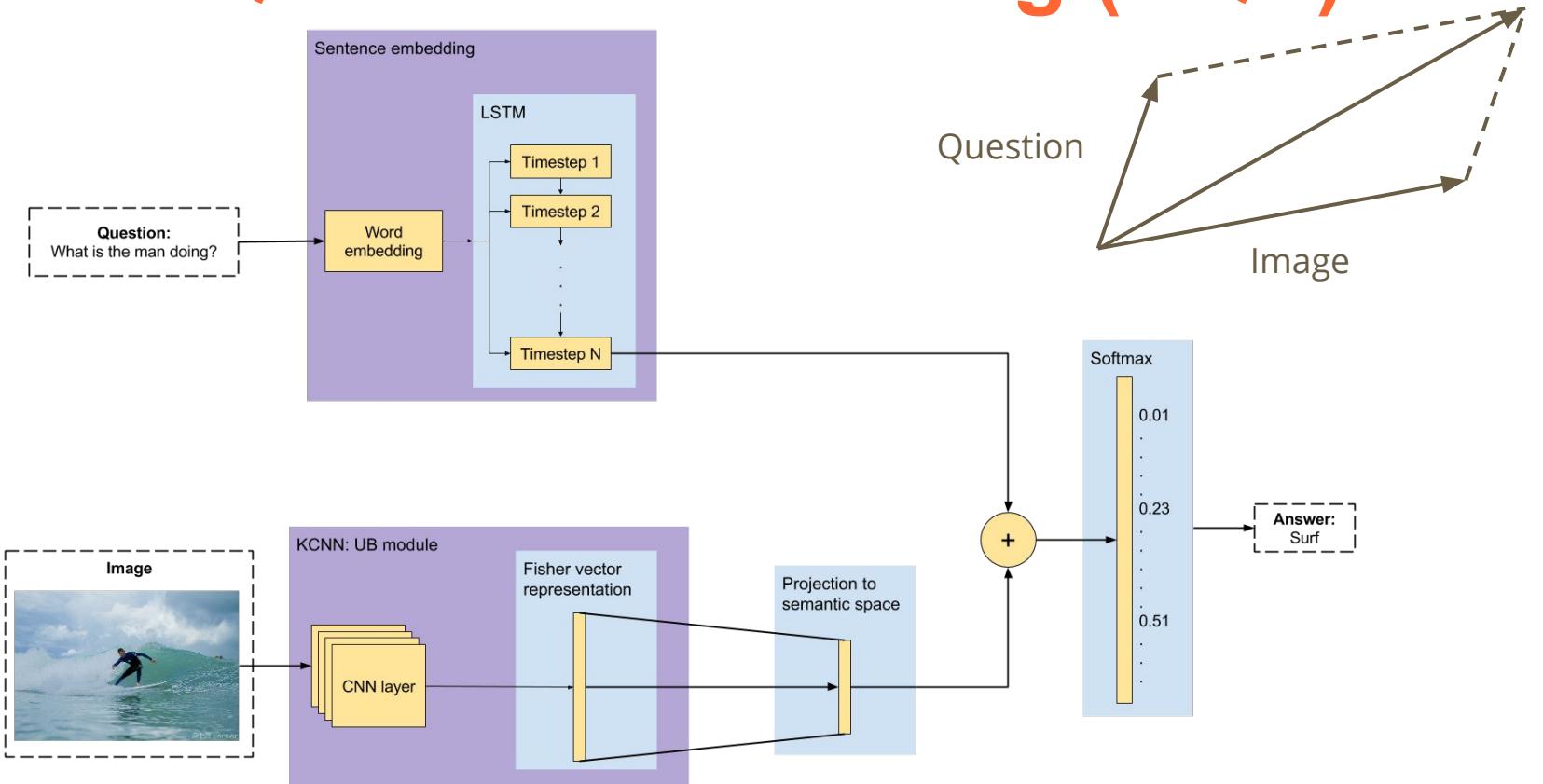
# Visual Question Answering (VQA)



# Visual Question Answering (VQA)

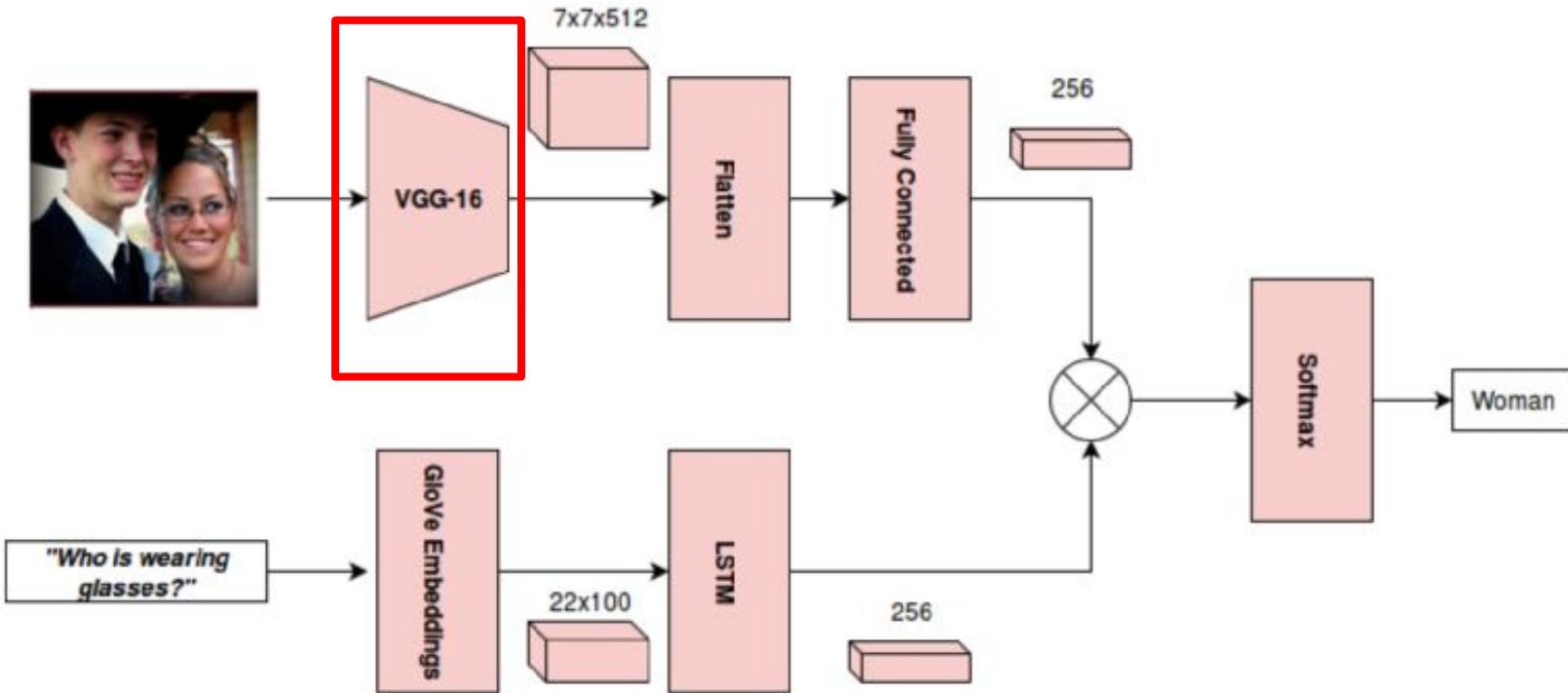


# Visual Question Answering (VQA)

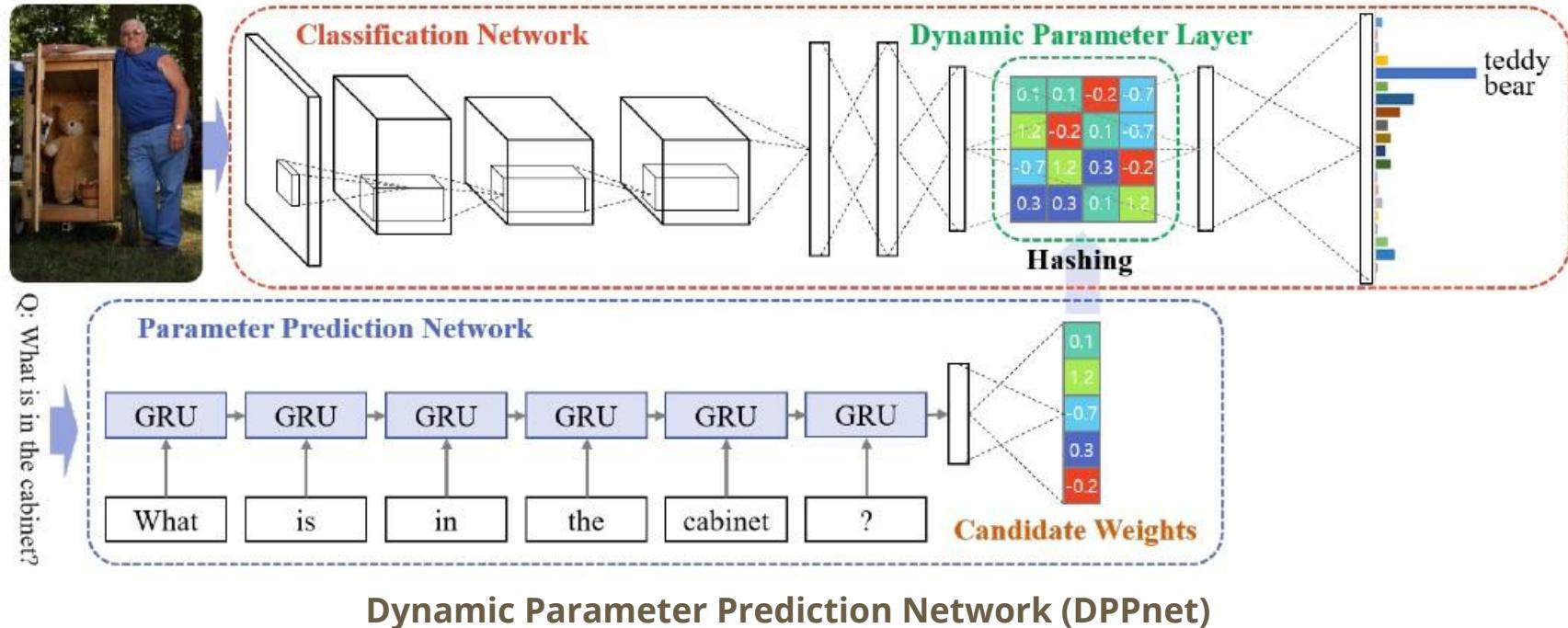


Masuda, Issey, Santiago Pascual de la Puente, and Xavier Giro-i-Nieto. ["Open-Ended Visual Question-Answering."](#) ETSETB UPC TelecomBCN (2016).

# Visual Question Answering (VQA)

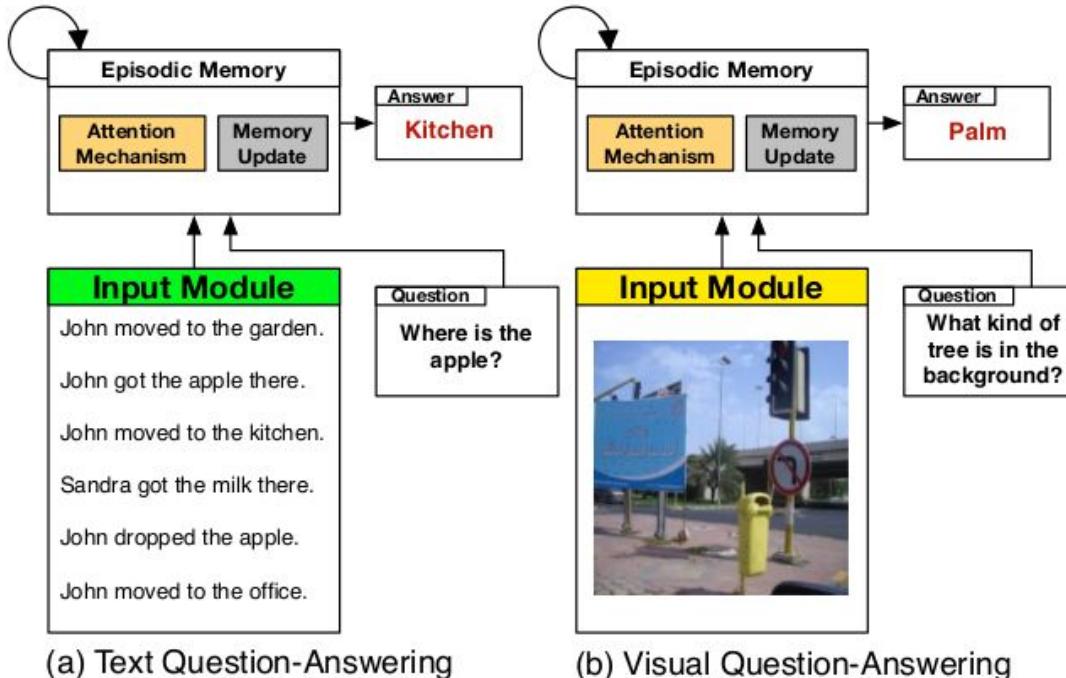


# Visual Question Answering (VQA)



Noh, H., Seo, P. H., & Han, B. [Image question answering using convolutional neural network with dynamic parameter prediction](#). CVPR 2016

# Visual Question Answering: Dynamic



(Slides and Slidecast by Santi Pascual): Xiong, Caiming, Stephen Merity, and Richard Socher. "Dynamic Memory Networks for Visual and Textual Question Answering." ICML 2016

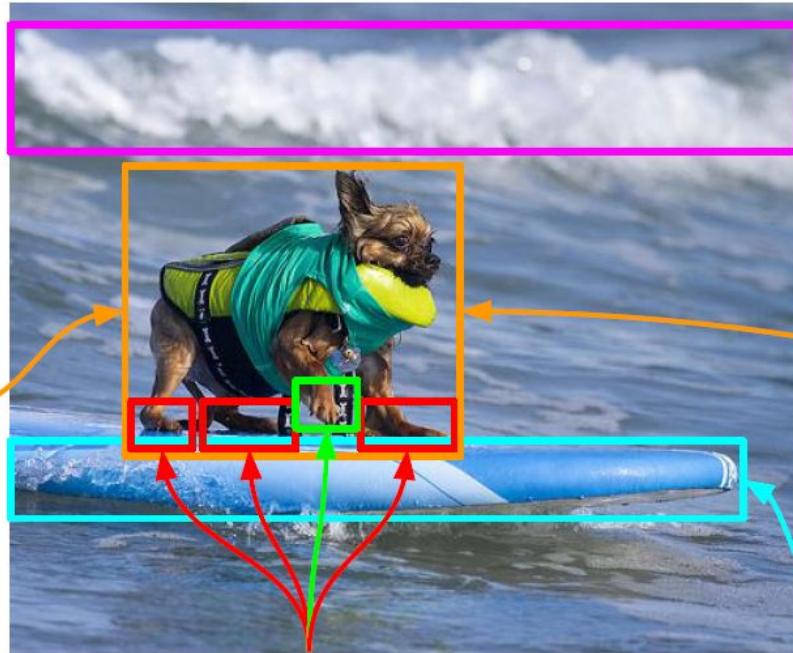
# Visual Question Answering: Grounded

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.



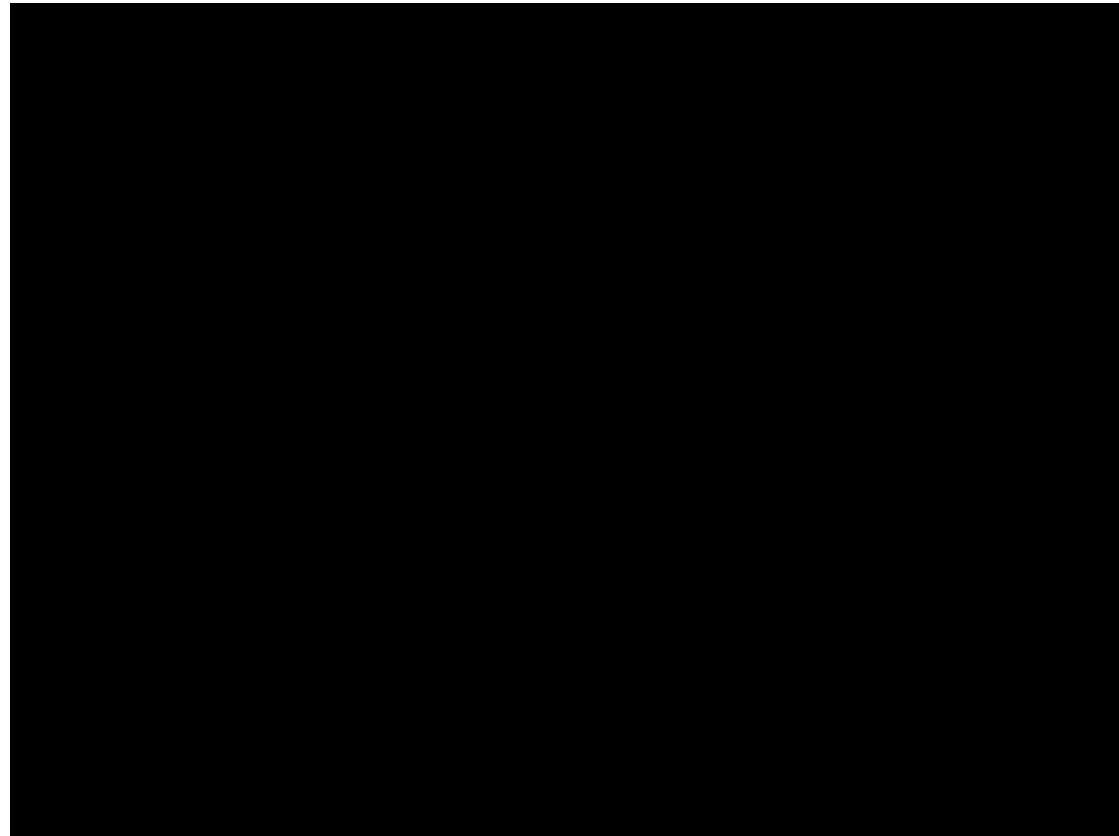
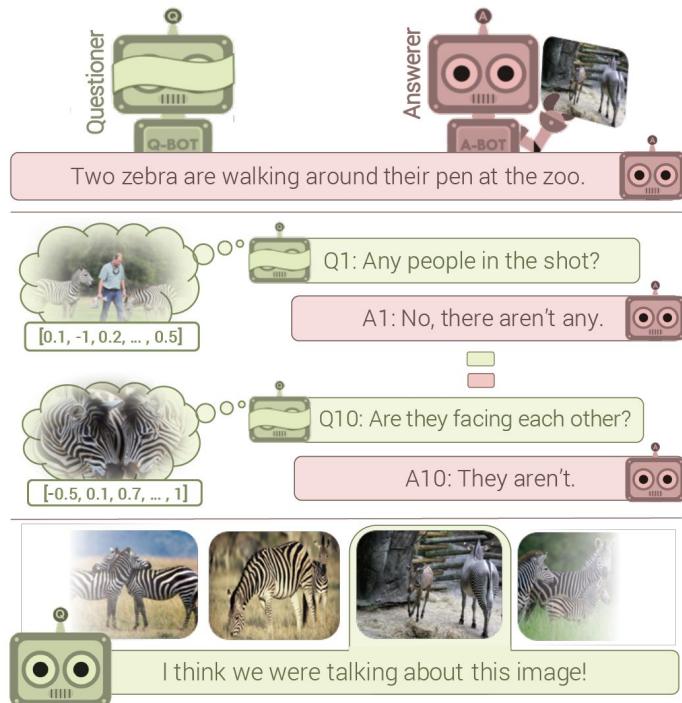
Why is there foam?

- 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.

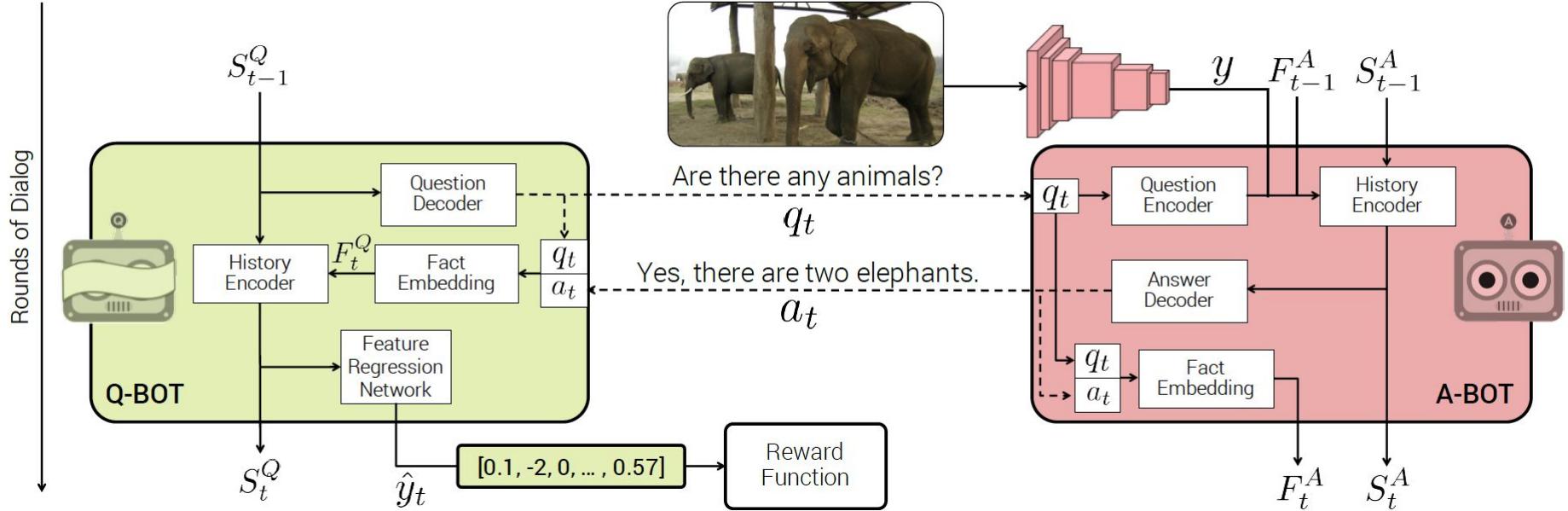
# Visual Dialog (Image Guessing Game)



Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra.

["Visual Dialog."](#) CVPR 2017

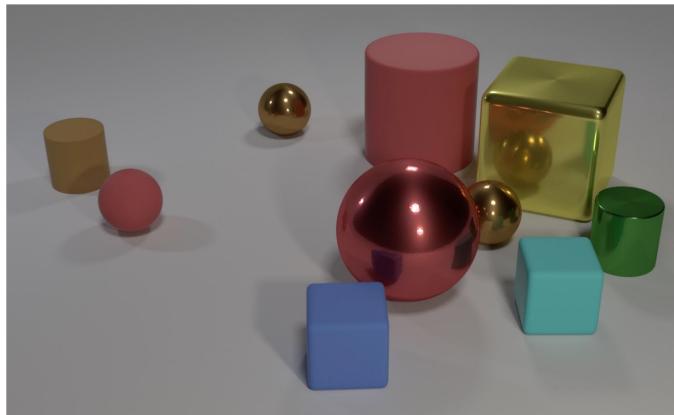
# Visual Dialog (Image Guessing Game)



# Visual Reasoning

**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?



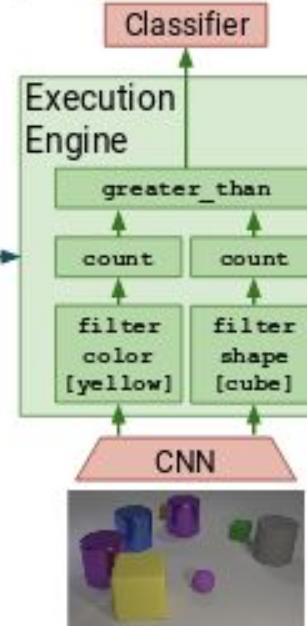
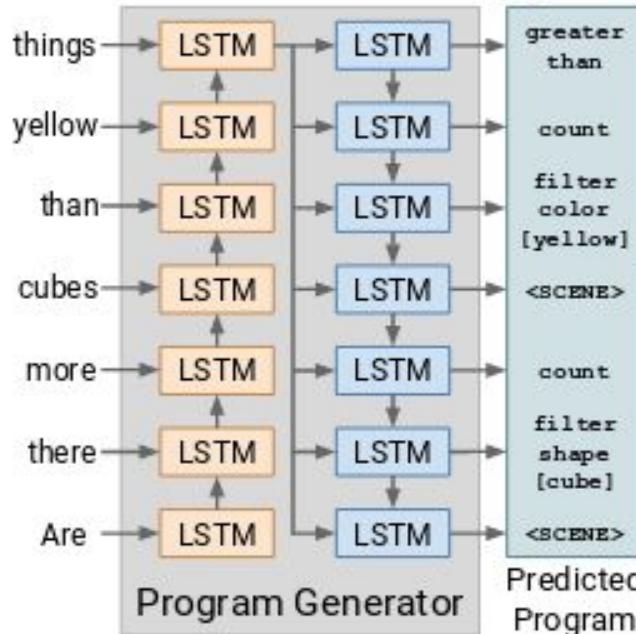
Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. "[CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning.](#)" CVPR 2017

# Visual Reasoning

Program Generator

Execution Engine

Question: Are there more cubes than yellow things? Answer: Yes



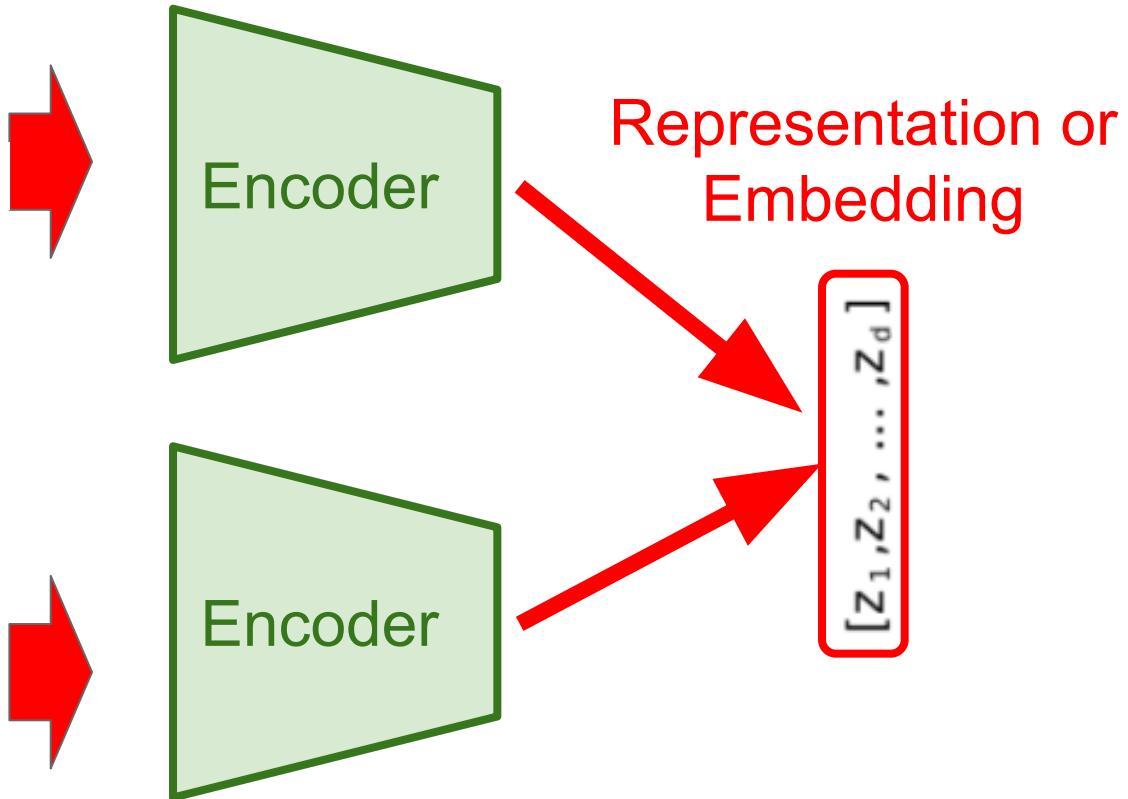
(Slides by Fran Roldan) Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Fei-Fei Li, Larry Zitnick, Ross Girshick , ["Inferring and Executing Programs for Visual Reasoning"](#). ICCV 2017

# Outline

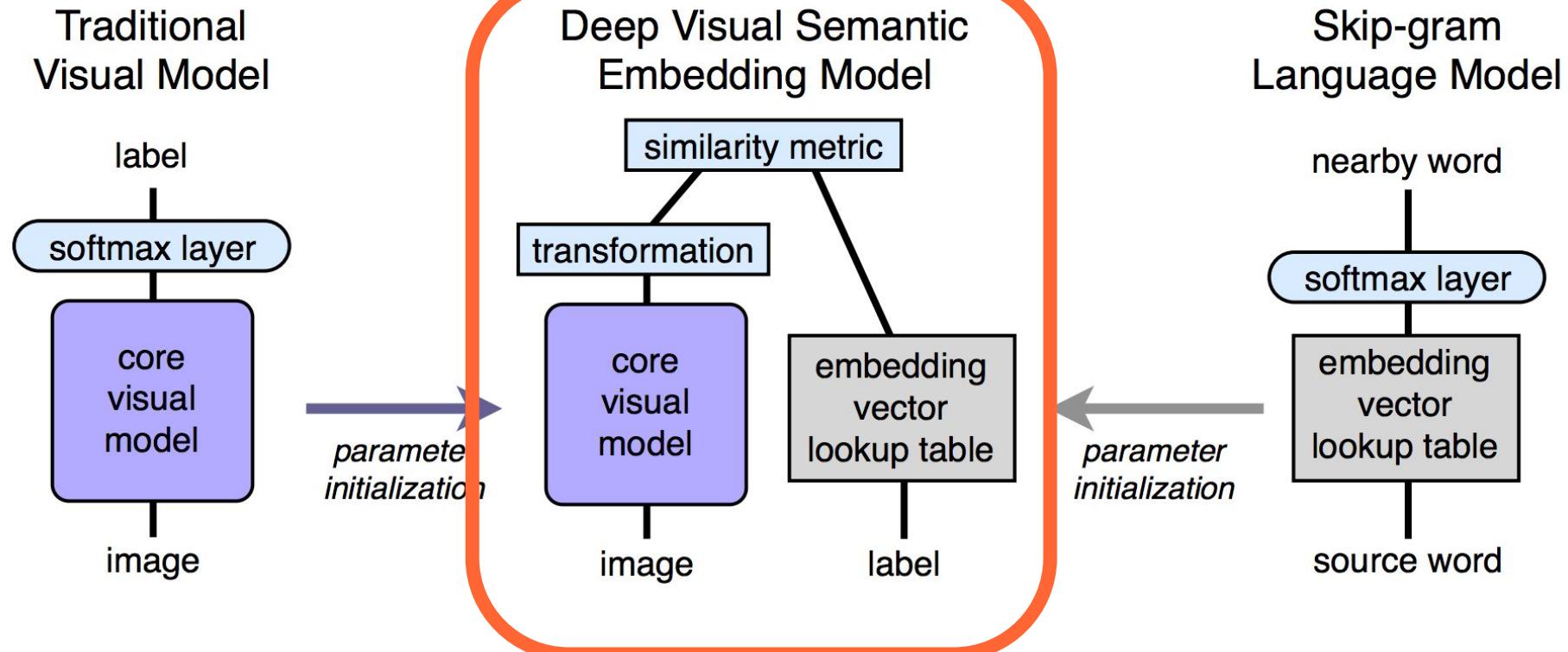
1. Neural Machine Transaltion (no vision here !)
2. Image and Video Captioning
3. Visual Question Answering / Reasoning
4. Joint Embeddings

# Joint Neural Embeddings

Economic growth has slowed down in recent years .



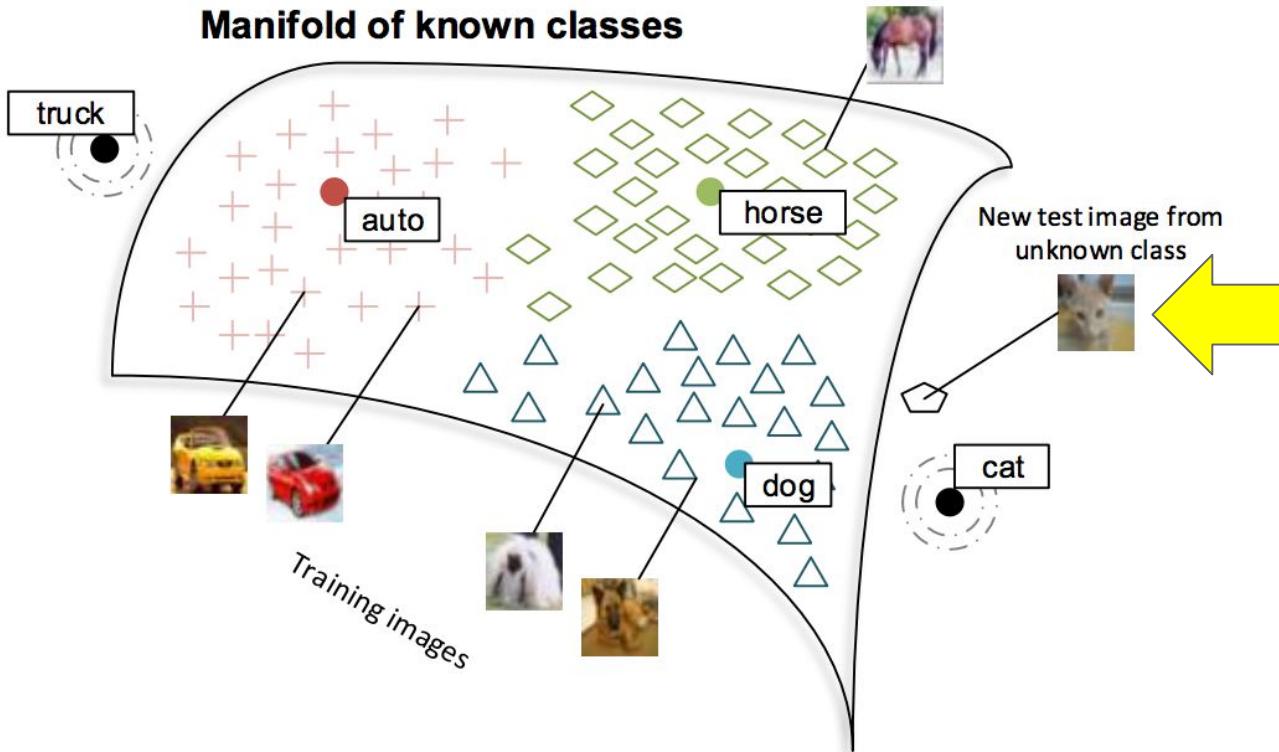
# Joint Neural Embeddings



Frome, Andrea, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, and Tomas Mikolov. "[Devise: A deep visual-semantic embedding model.](#)" NIPS 2013

# Joint Neural Embeddings

Manifold of known classes



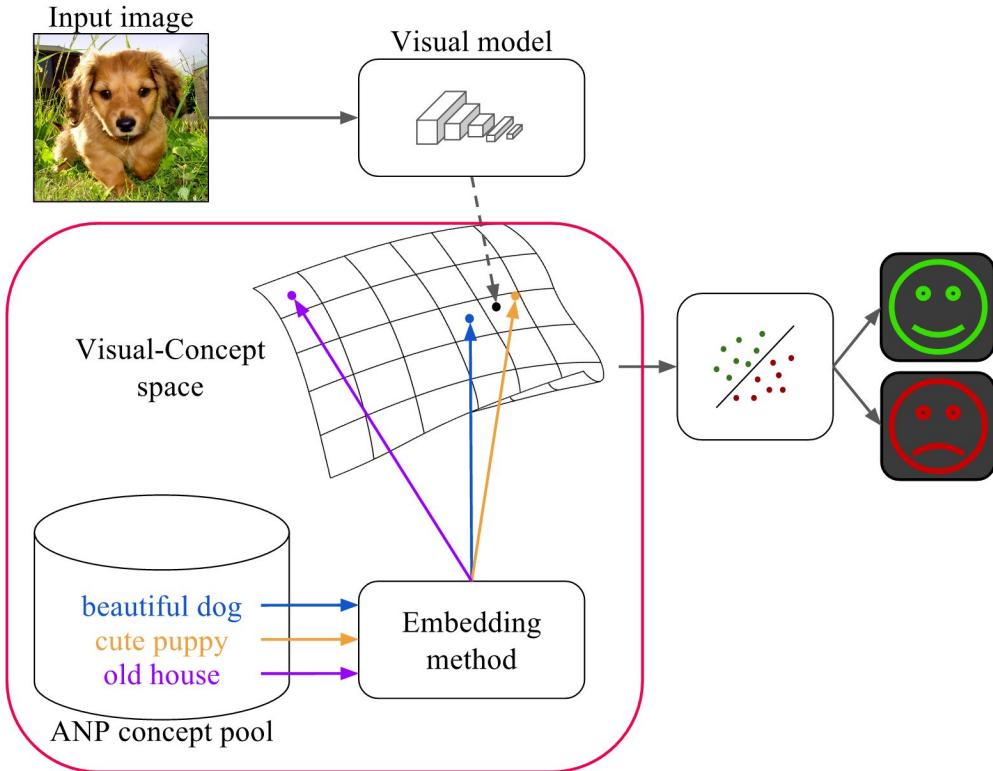
**Zero-shot learning:**  
a class not present in the  
training set of images  
can be predicted

(eg. no images from  
“cat” in the training set)

# Joint Neural Embeddings



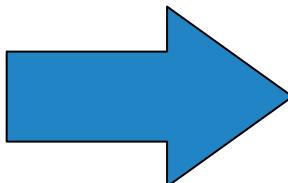
Foggy Day





# Joint Neural Embeddings

Image and text retrieval with joint embeddings.



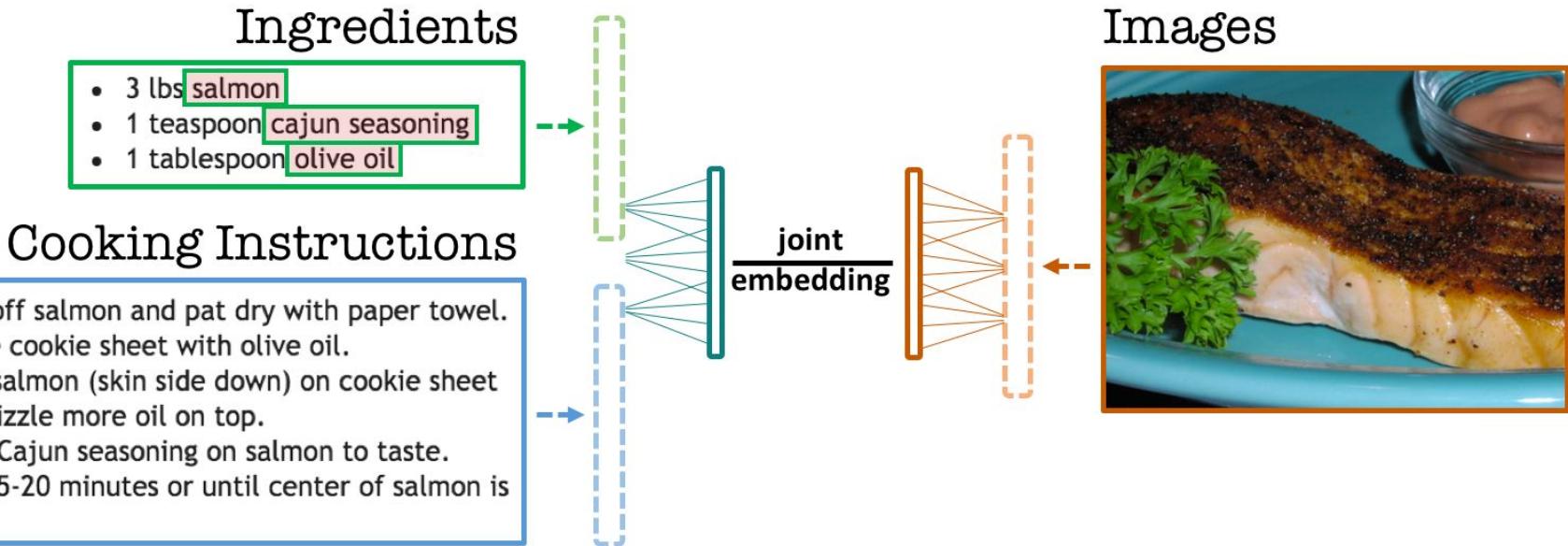
## Ingredients

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

## Cooking Instructions

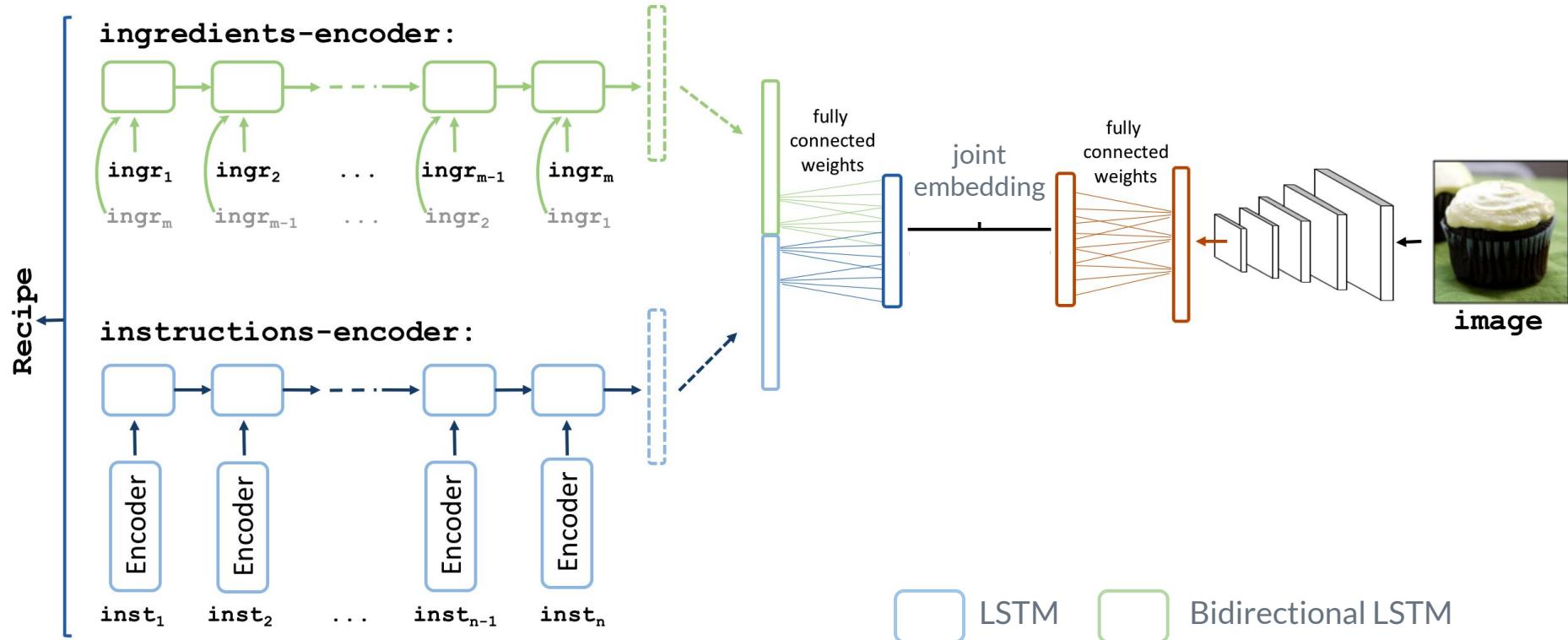
1. Rinse off salmon and pat dry with paper towel.
2. Drizzle cookie sheet with olive oil.
3. Place salmon (skin side down) on cookie sheet and drizzle more oil on top.
4. Shake Cajun seasoning on salmon to taste.
5. Broil 15-20 minutes or until center of salmon is done.

# Joint Neural Embeddings



Amaia Salvador, Nicholas Haynes, Yusuf Aytar, Javier Marín, Ferda Ofli, Ingmar Weber, Antonio Torralba, "[Learning Cross-modal Embeddings for Cooking Recipes and Food Images](#)". CVPR 2017

# Joint Neural Embeddings



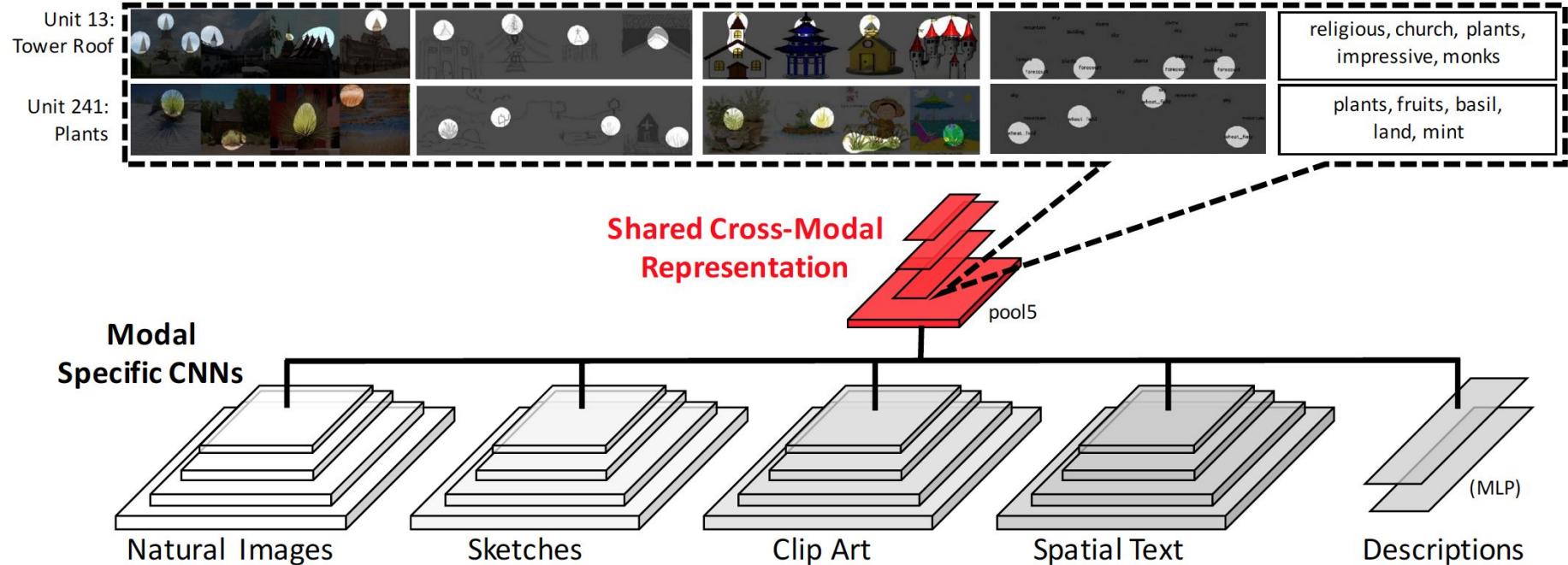
Amaia Salvador, Nicholas Haynes, Yusuf Aytar, Javier Marín, Ferda Ofli, Ingmar Weber, Antonio Torralba, “[Learning Cross-modal Embeddings for Cooking Recipes and Food Images](#)”. CVPR 2017

# Image to image and text

Query	Real	Clip art	Spatial text	Sketches	Descriptions	
			cabinet door wall wall cabinet sink floor	cabinet door wall wall cabinet sink floor		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.
			sky window building window window	window building sky window		A very small or compact kitchen. These little 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
sky castle wall road			sky castle wall wall road plants	sky castle wall wall plants road		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 women dressed in suits and their work attire. You could tell this was a serious setting.
			sky snowy...mountain crevasse	snowy...mountain sky		The building appeared grand from the outside, with its turrets and thick stone 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
						This defines the perimeter of a Islamic city with high, fort like 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.

Aytar, Yusuf, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. ["Cross-Modal Scene Networks."](#) CVPR 2016.

# Image to image and text



Aytar, Yusuf, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. ["Cross-Modal Scene Networks."](#) CVPR 2016.

# Outline

1. Motivation
2. Image Captioning
3. Visual Question Answering / Reasoning
4. Joint Embeddings

# Questions ?

## Undergradese

What undergrads ask vs. what they're REALLY asking

"Is it going to be an open book exam?"

Translation: "I don't have to actually memorize anything, do I?"

"Hmm, what do you mean by that?"

Translation: "What's the answer so we can all go home."

"Are you going to have office hours today?"

Translation: "Can I do my homework in your office?"

"Can i get an extension?"

Translation: "Can you re-arrange your life around mine?"

"Is this going to be on the test?"

Translation: "Tell us what's going to be on the test."

"Is grading going to be curved?"

Translation: "Can I do a mediocre job and still get an A?"

