[course site]

#DLUPC

Day 2 Lecture 4

Language and Vision

Xavier Giro-i-Nieto
xavier.giro@upc.edu

Associate Professor
Universitat Politecnica de Catalunya
Technical University of Catalonia

Acknowledgments

Antonio Bonafonte

Santiago Pascual

TALP Research Center

Acknowledgments

Marta R. Costa-jussà

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.
#deepvision16 #cvpr16

Tradueix del anglès

RETUIT 1 AGRADA A 1

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

telecom BCN | TALP | Image Processing Group, Signal Theory and Communications Department | UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

+ info: **TelecomBCN.DeepLearning.Barcelona**

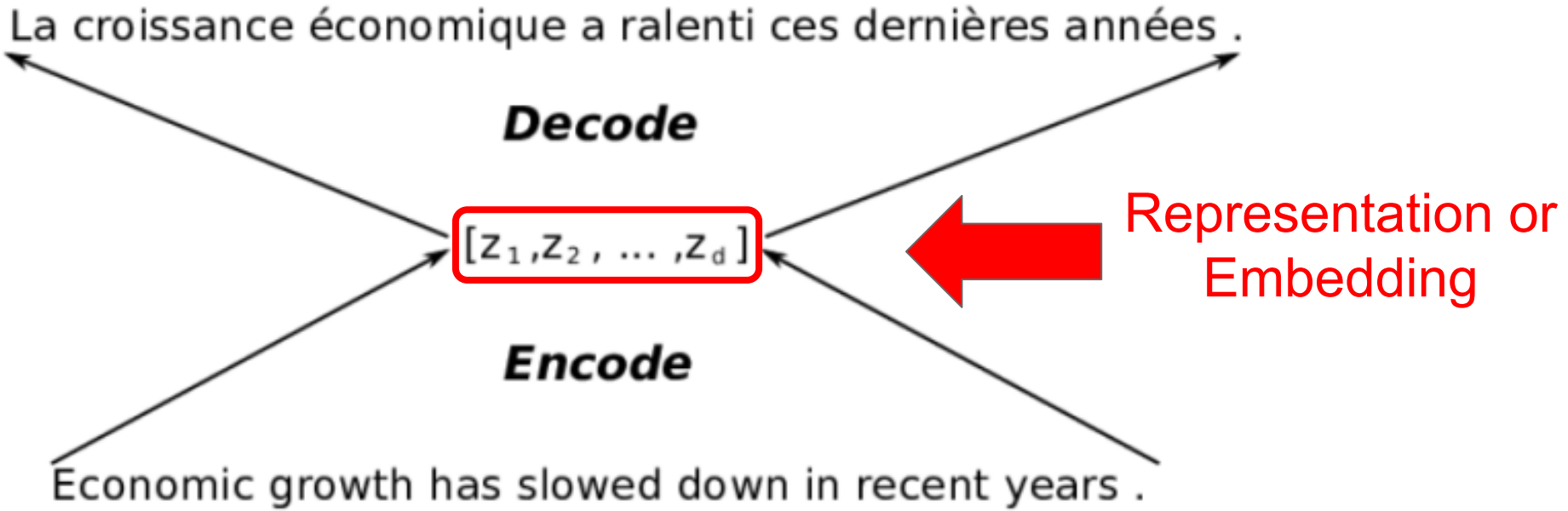
[course site]

Text (English)

Neural Machine Translation

Text (French)

Neural Machine Translation



Kyunghyun Cho, "Introduction to Neural Machine Translation with GPUs" (2015)

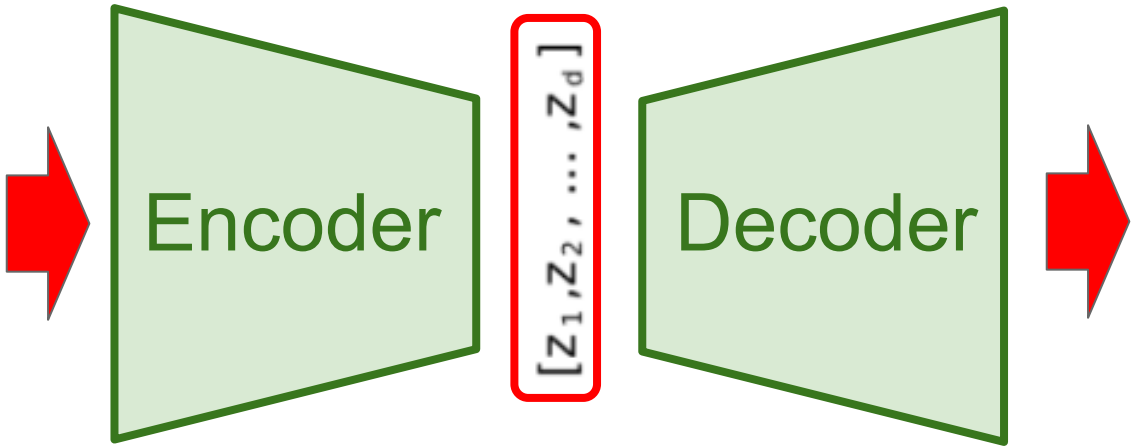
Economic growth has slowed down in recent years .

Representation or Embedding

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

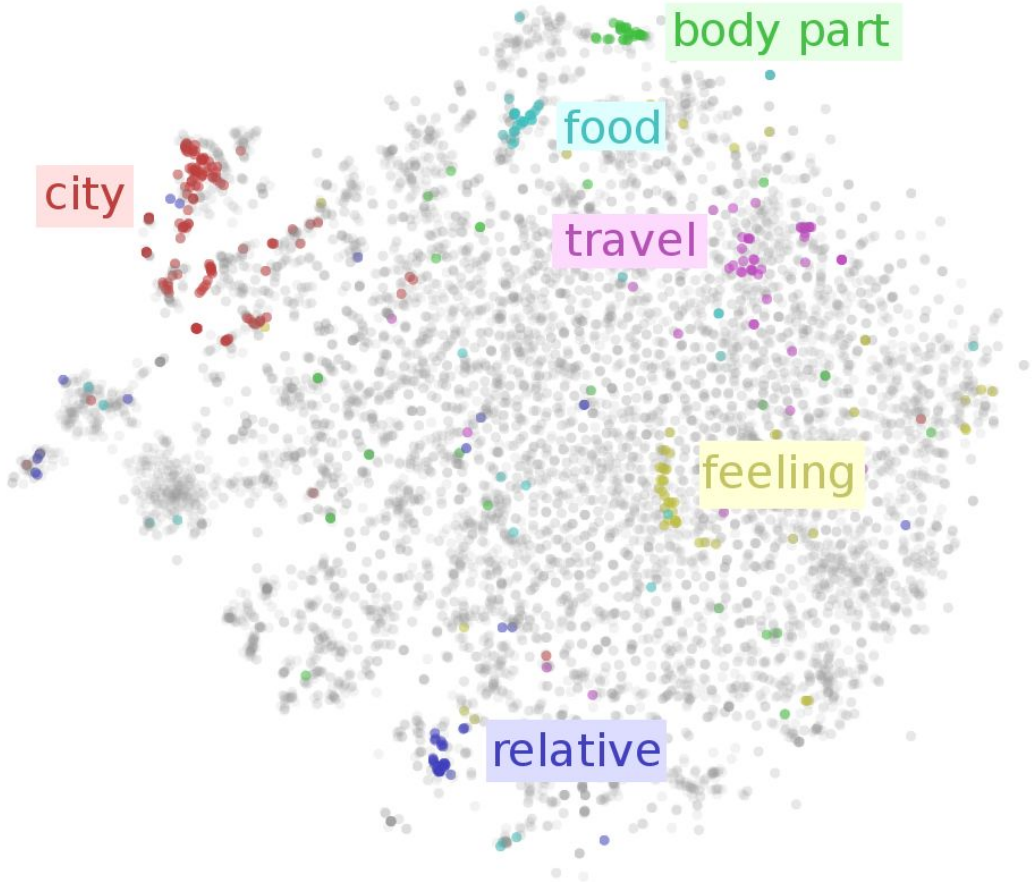


Figure:
Christopher Olah
Visualizing Representations

Outline

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

Representation or Embedding

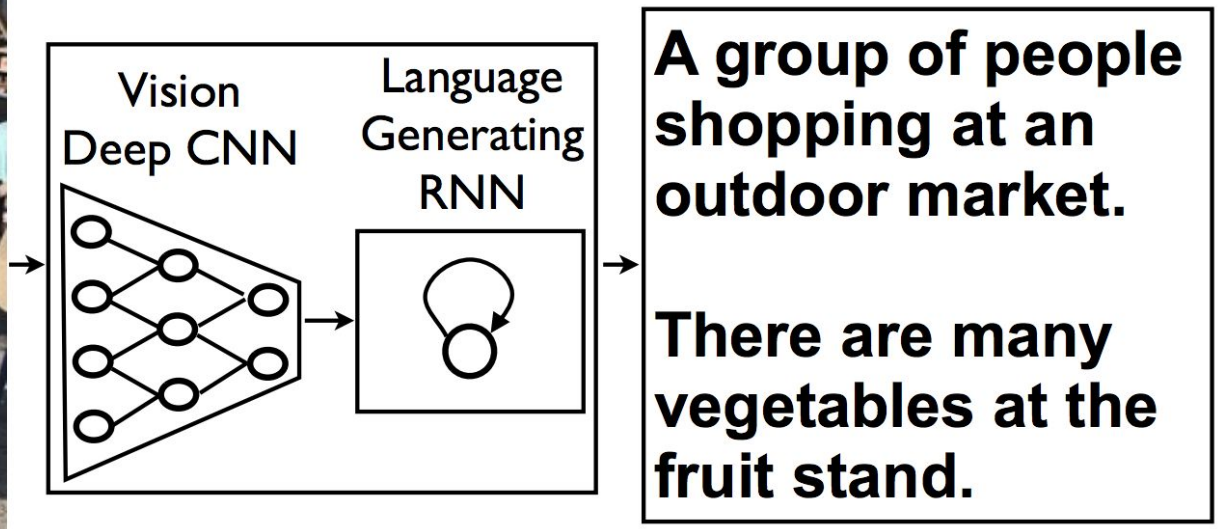
Encoder

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

Decoder

La croissance économique a ralenti ces dernières années .

Captioning: Show & Tell



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

Captioning: DeepImageSent

man in black shirt is playing guitar.

construction worker in orange safety vest is working on road.

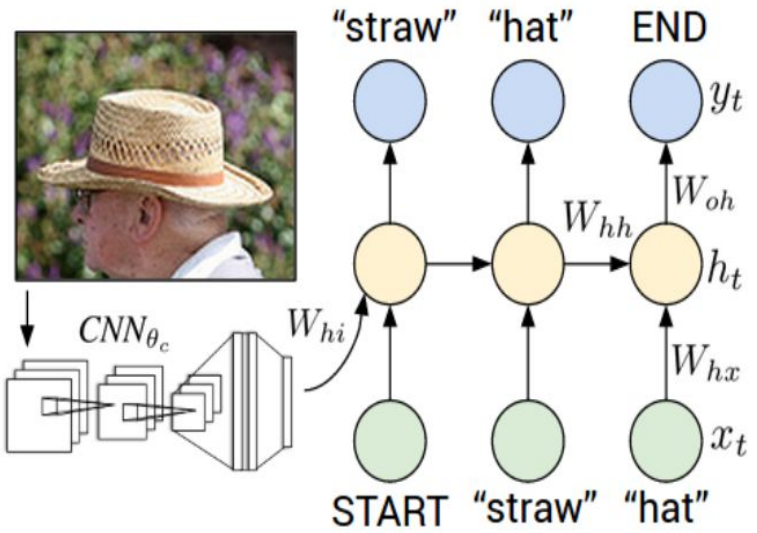
two young girls are playing with lego toy.

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

Captioning: DeepImageSent

only takes into account image features in the first hidden state

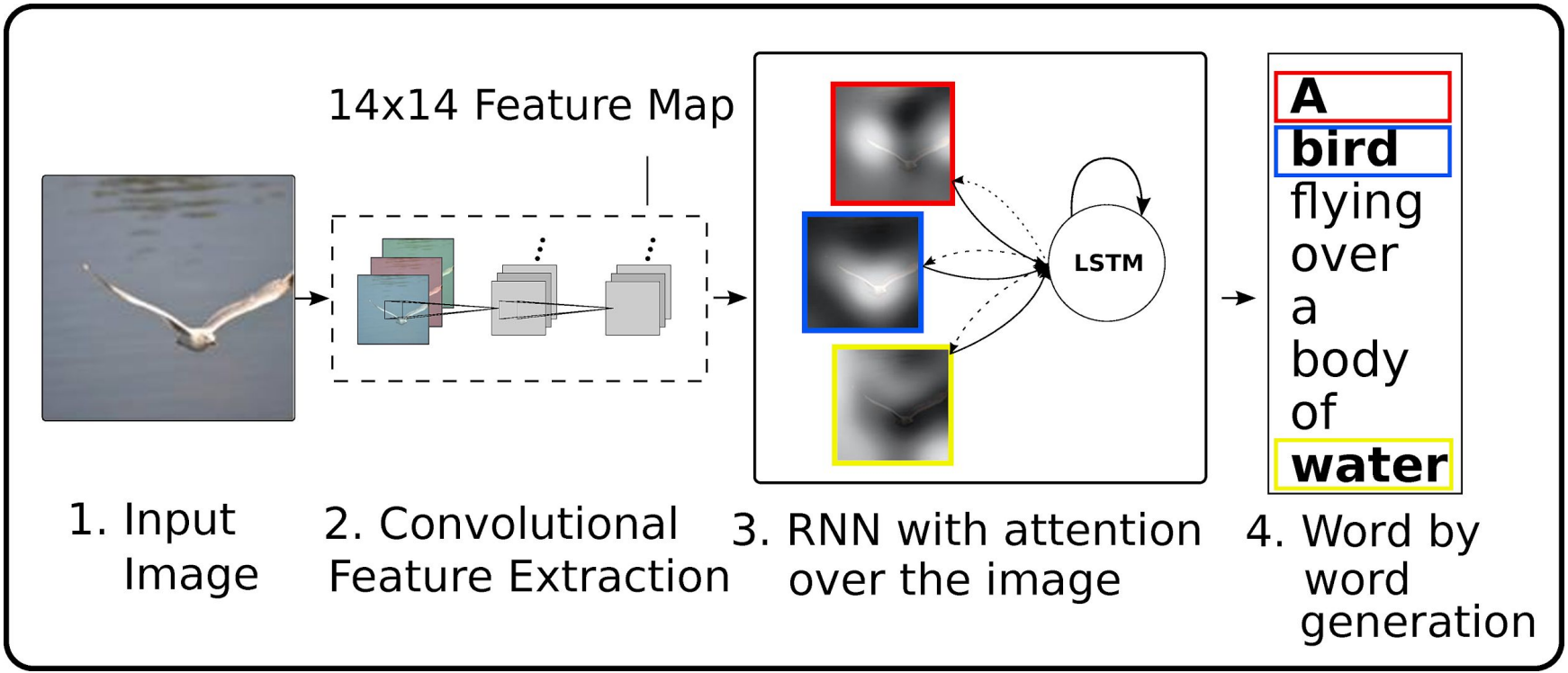
$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1) \odot b_v)$$
$$y_t = softmax(W_{oh}h_t + b_o).$$



Multimodal Recurrent Neural Network

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

Captioning: Show, Attend & Tell

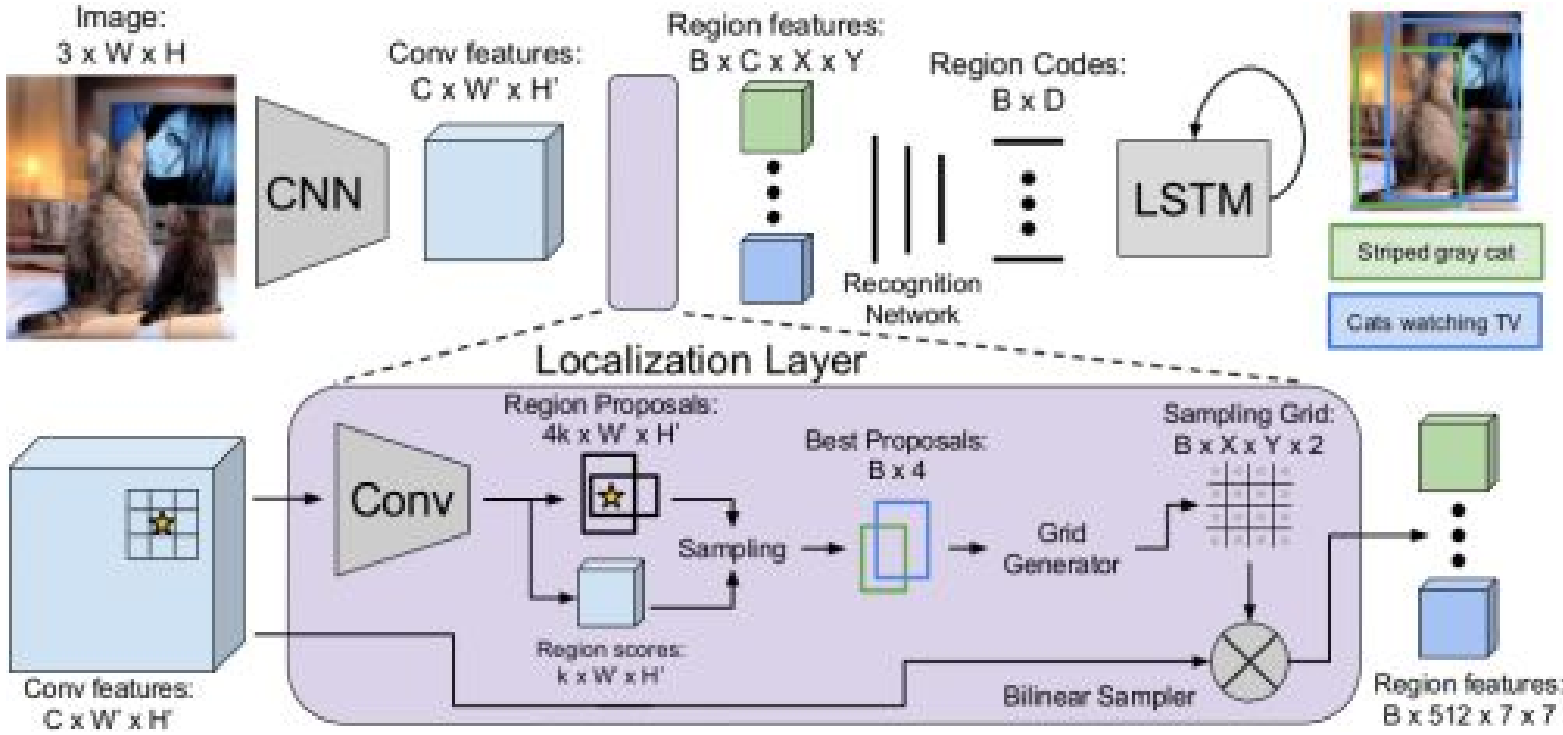


Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." ICML 2015

Captioning: Show, Attend & Tell

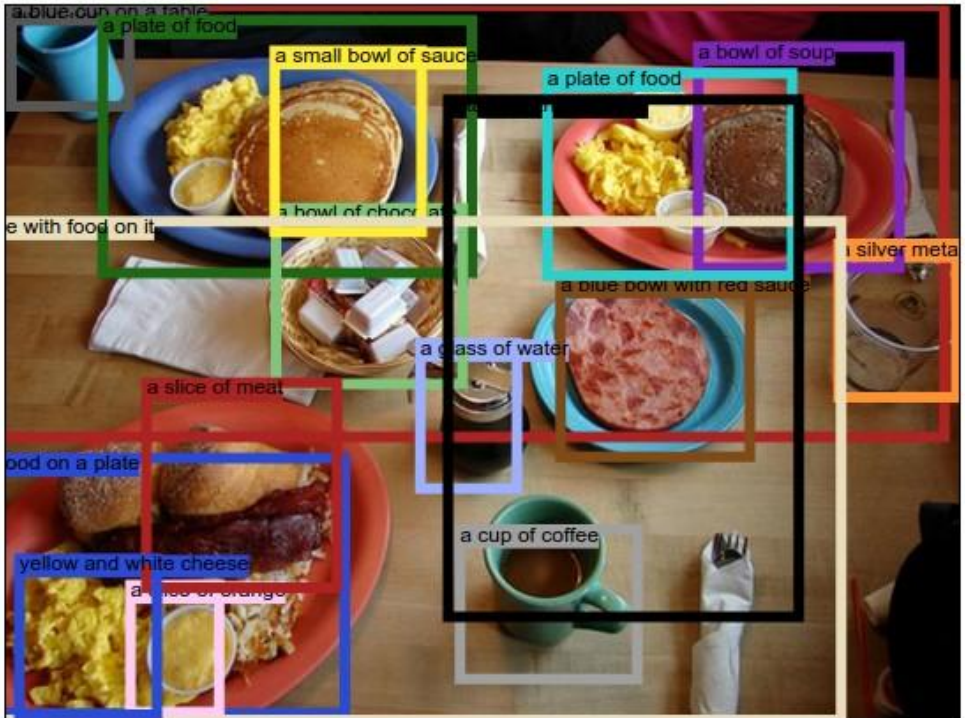
Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." ICML 2015

Captioning (+ Detection): DenseCap



Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning." *CVPR 2016*

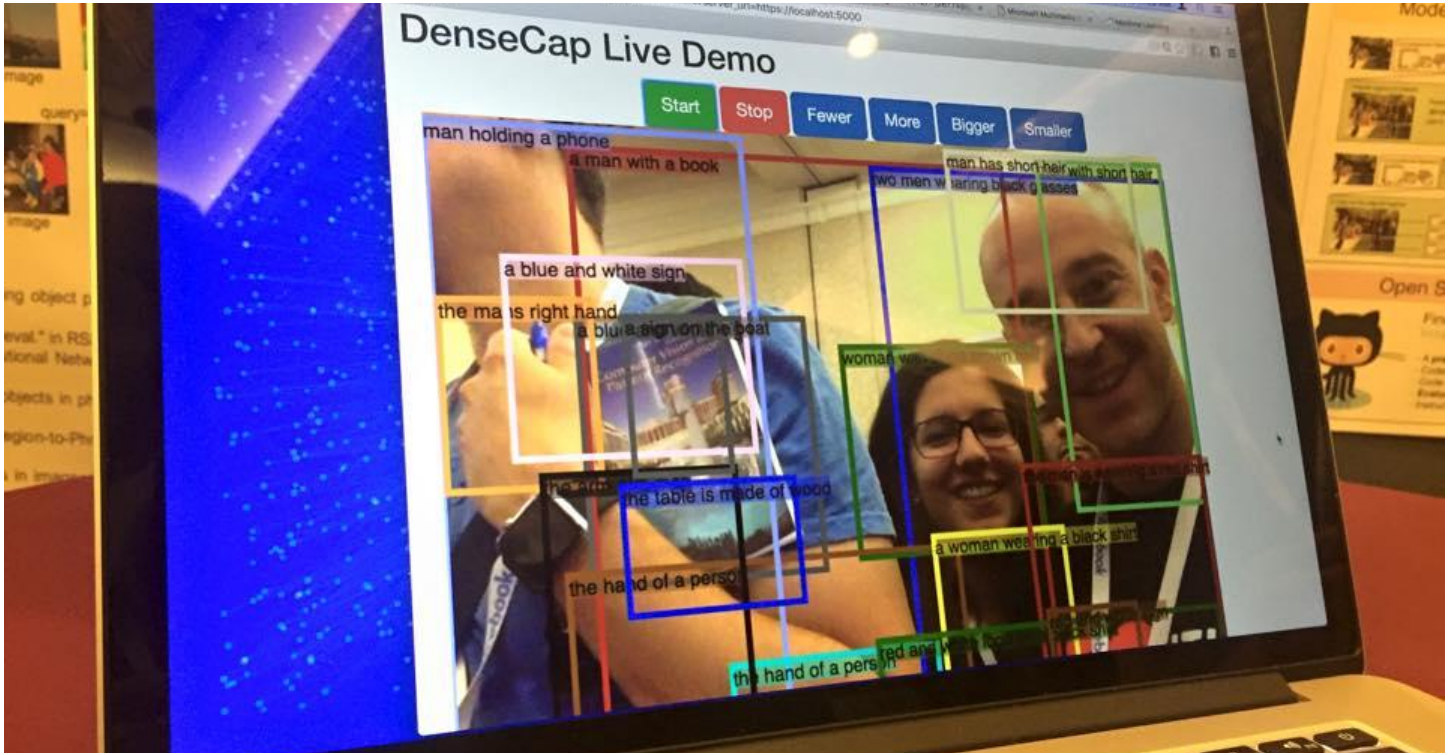
Captioning (+ Detection): DenseCap



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



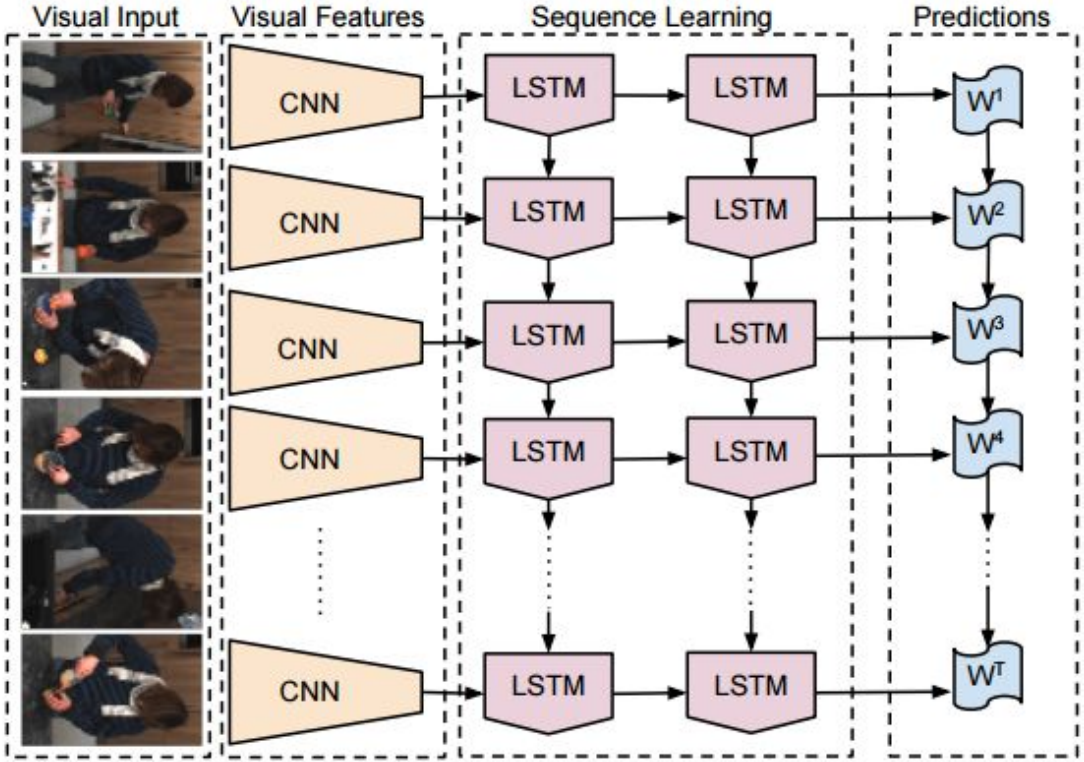
XAVI: “man has short hair”, “man with short hair”

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

BOTH: “two men wearing black glasses”

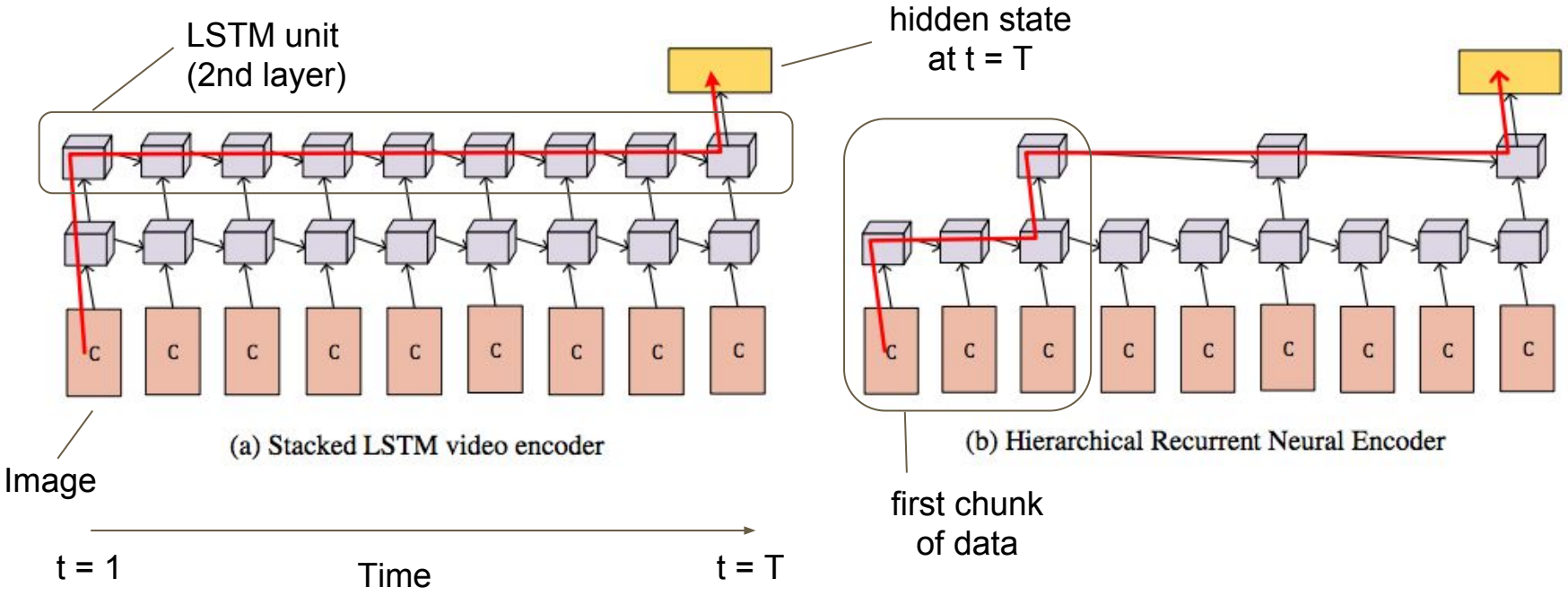
Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning." *CVPR 2016*

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,Yueting 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

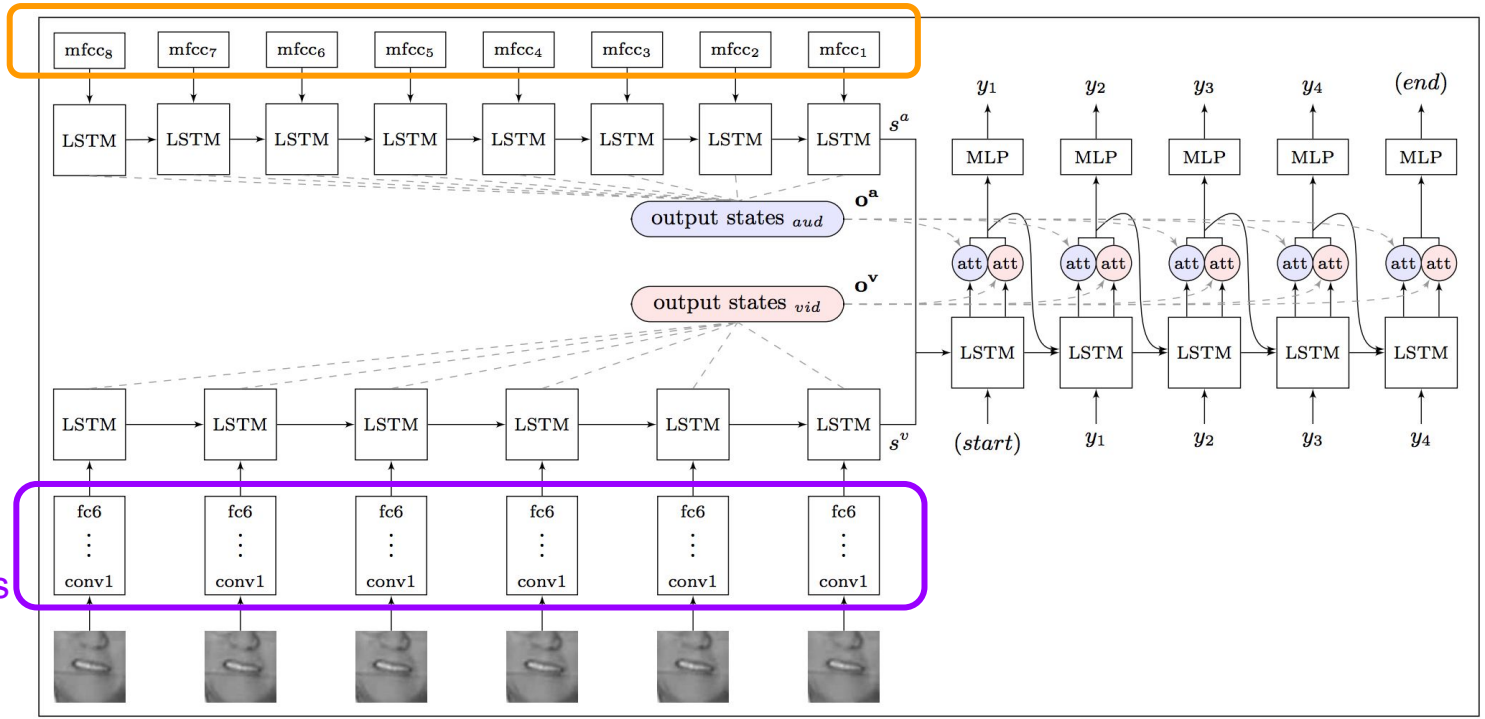


Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character y_i, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

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

Lipreading: Watch, Listen, Attend & Spell

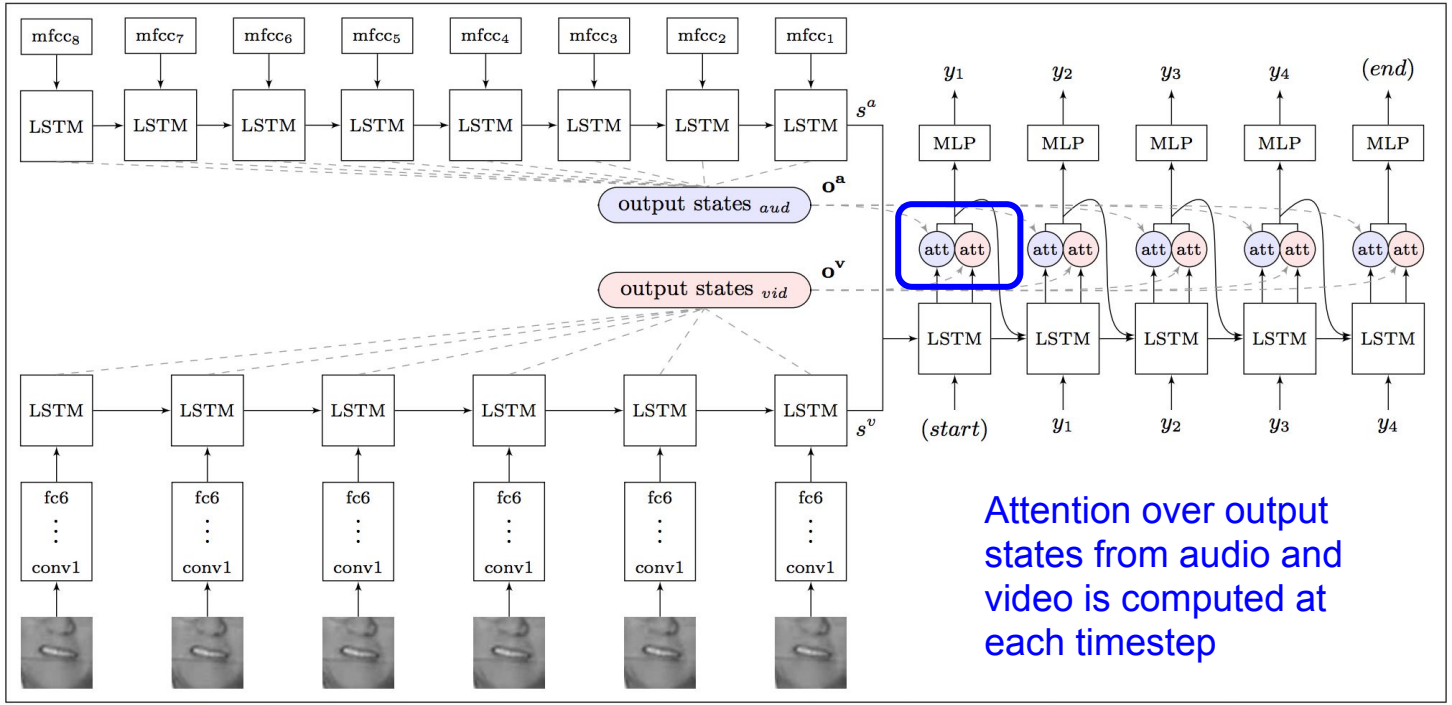


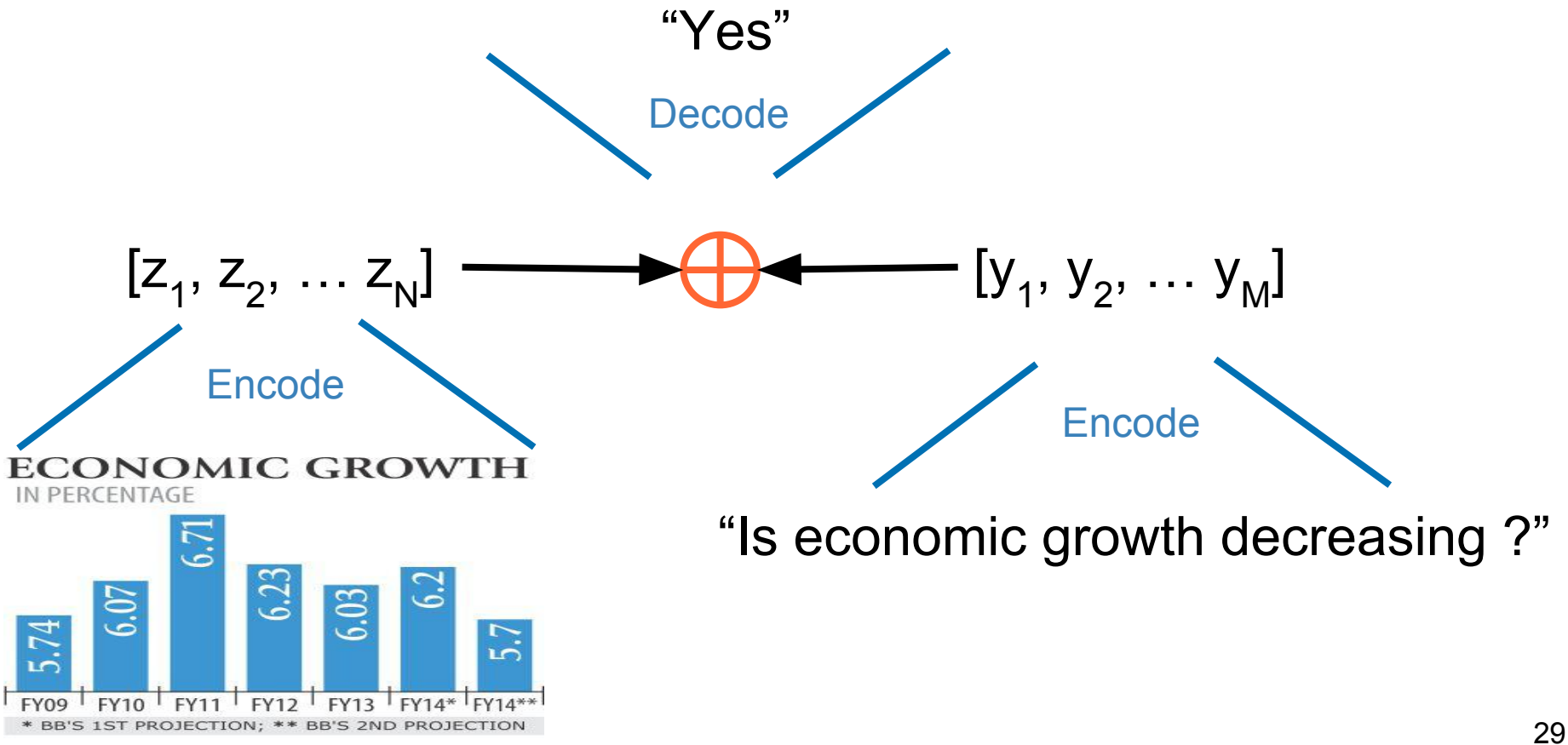
Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character y_i, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

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)

What is the mustache made of?

AI System

bananas

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. "VQA: Visual question answering." CVPR 2015.

Visual Question Answering (VQA)

Extract visual features

Question

What object is flying?

Embedding

Merge

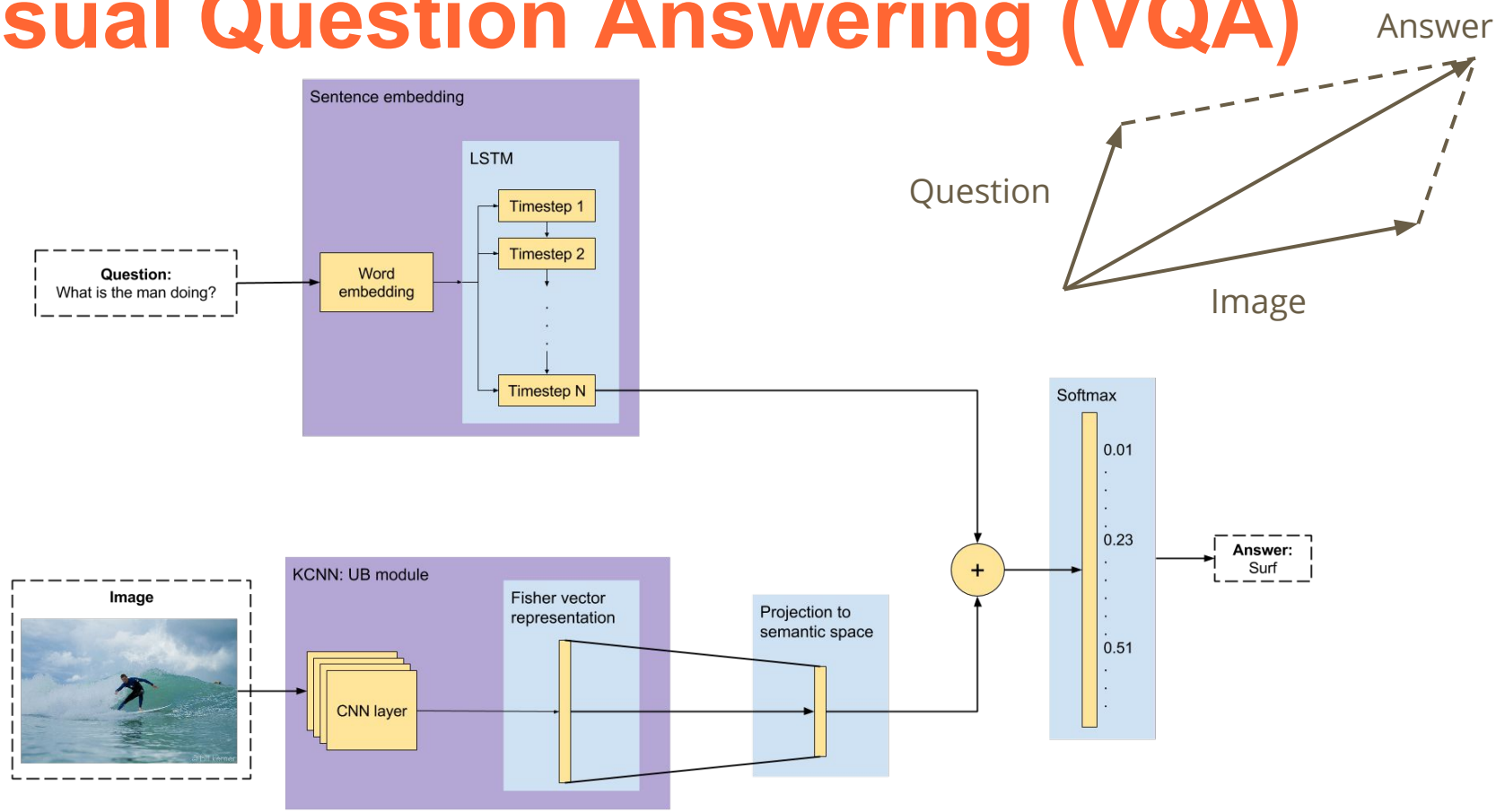
Predict answer

Answer

Kite

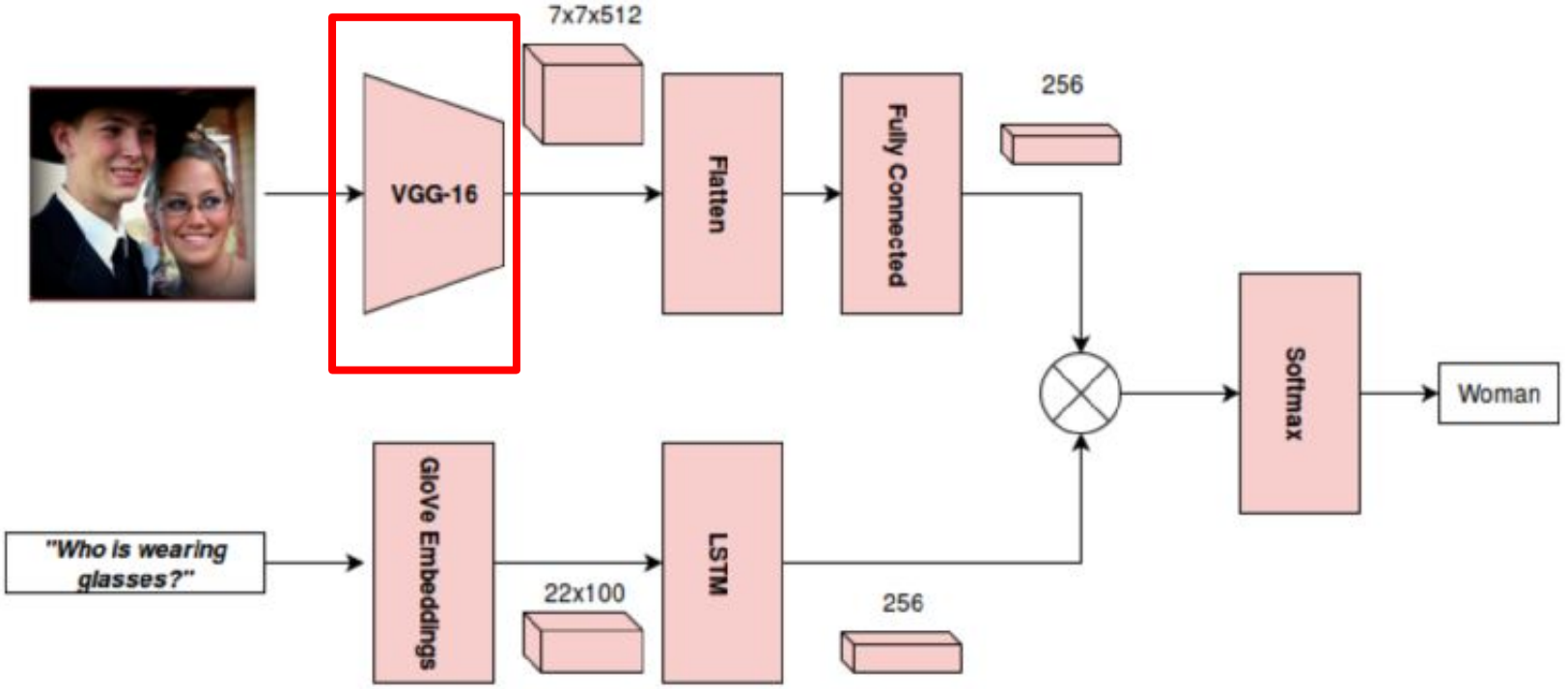
Slide credit: Issey Masuda

Visual Question Answering (VQA)



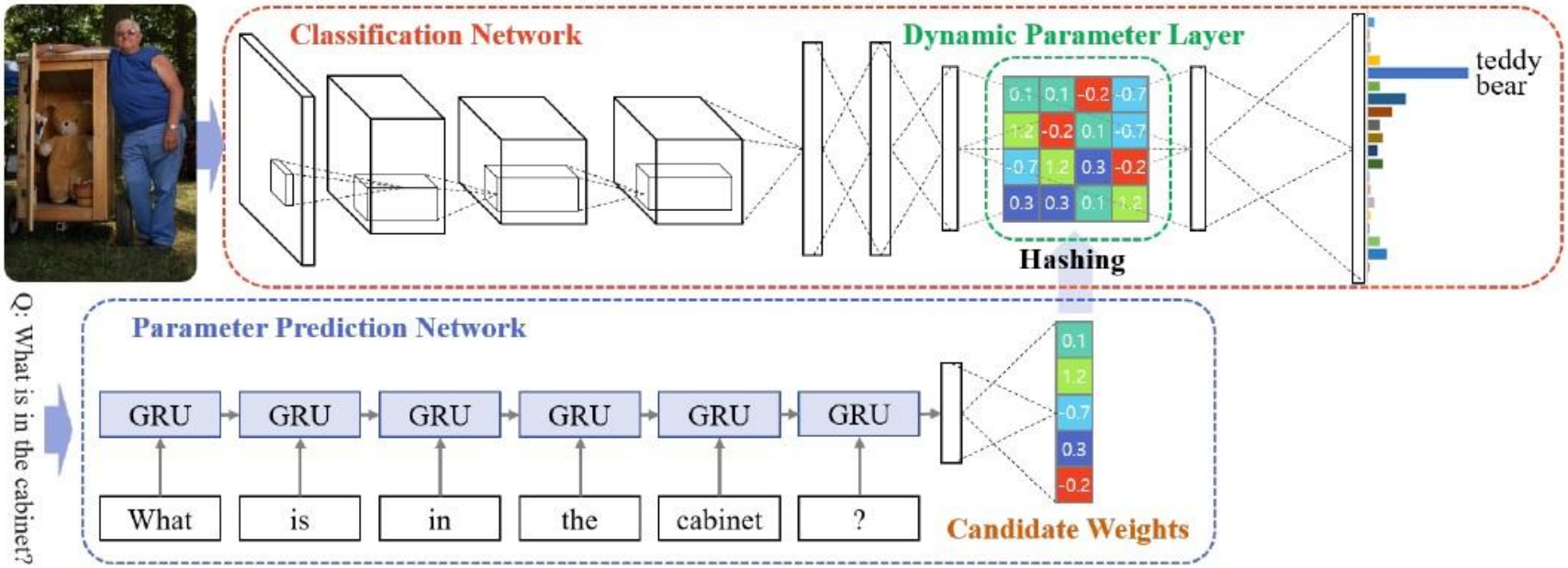
Masuda, Issey, Santiago Pascual de la Puente, and Xavier Giro-i-Nieto. "Open-Ended Visual Question-Answering." ETSETB UPC TelecomBCN (2016).

Visual Question Answering (VQA)



Francisco Roldán, Issey Masuda, Santiago Pascual de la Puente, and Xavier Giro-i-Nieto. "Visual Question-Answering 2.0." ETSETB UPC TelecomBCN (2017).

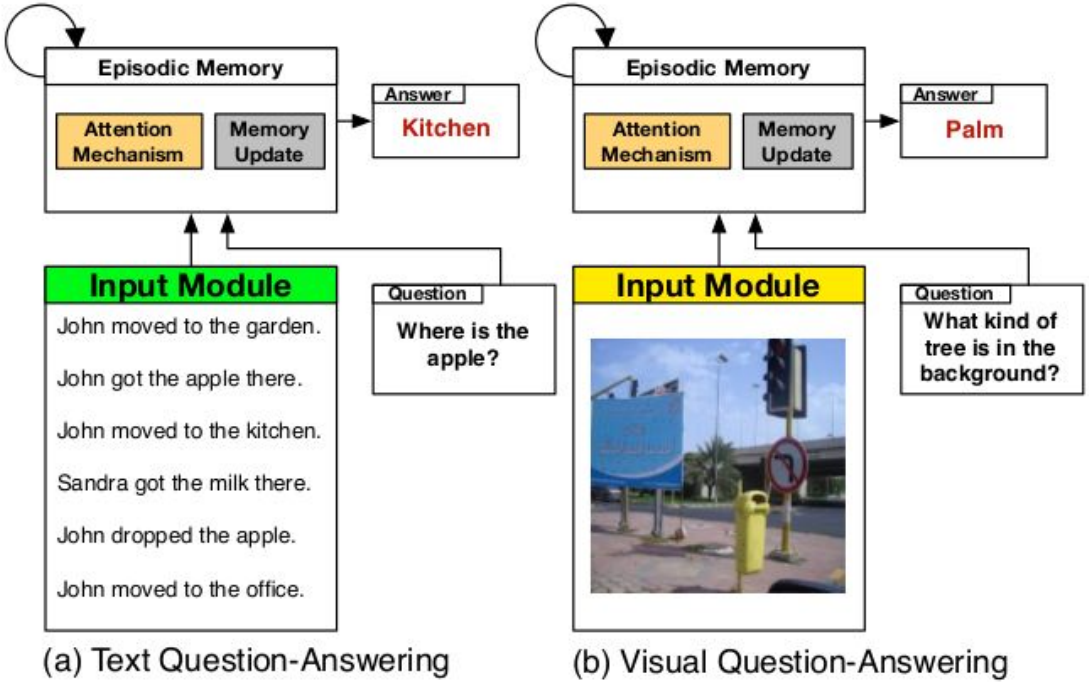
Visual Question Answering (VQA)



Dynamic Parameter Prediction Network (DPPnet)

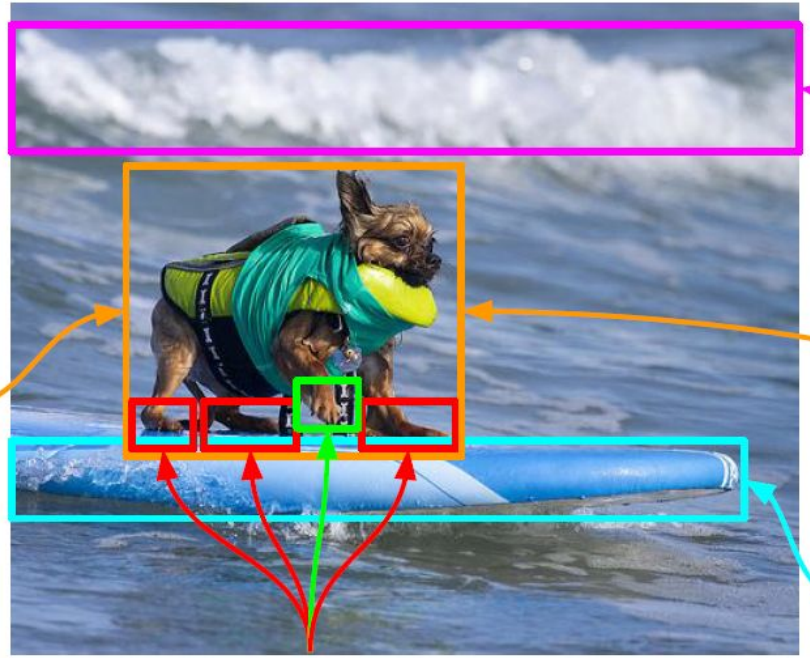
Noh, H., Seo, P. H., & Han, B. Image question answering using convolutional neural network with dynamic parameter prediction. CVPR 2016

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.

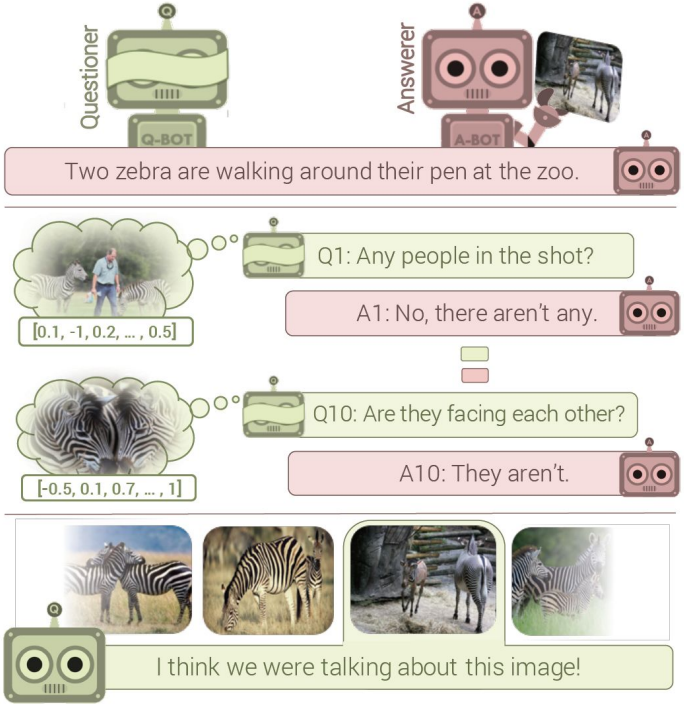
What is the dog doing?
A) Surfing. ✔
B) Sleeping.
C) Running.
D) Eating.

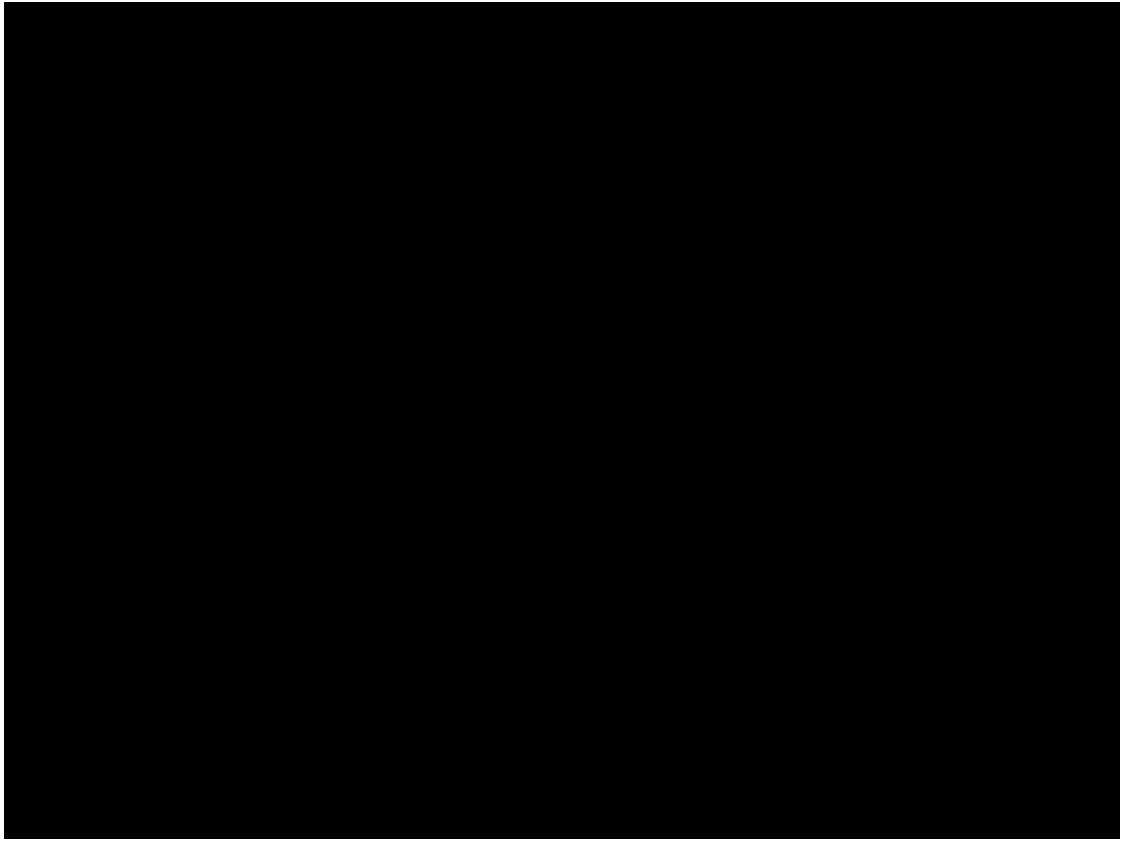
Why is there foam?
A) Because of a wave. ✔
B) Because of a boat.
C) Because of a fire.
D) Because of a leak.

What is the dog standing on?
A) On a surfboard. ✔
B) On a table.
C) On a garage.
D) On a ball.

(Slides and Screencast by Issey Masuda): Zhu, Yuke, Oliver Groth, Michael Bernstein, and Li Fei-Fei."Visual7W: Grounded Question Answering in Images." CVPR 2016.

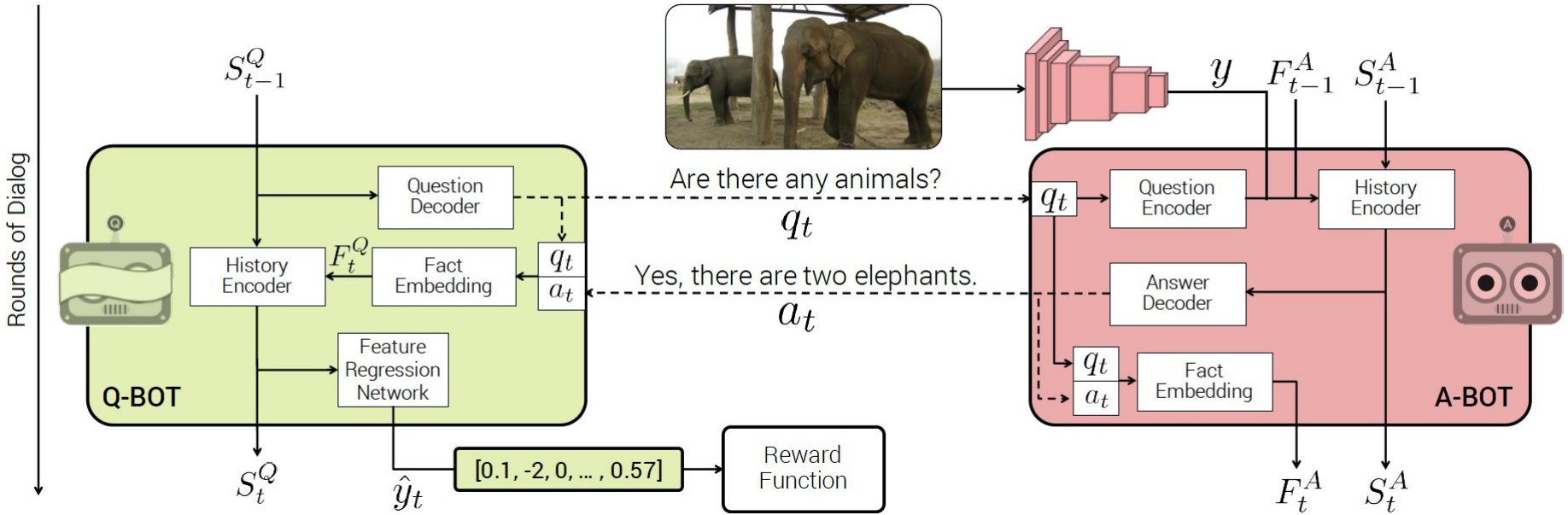
Visual Dialog (Image Guessing Game)





Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. "Visual Dialog." CVPR 2017

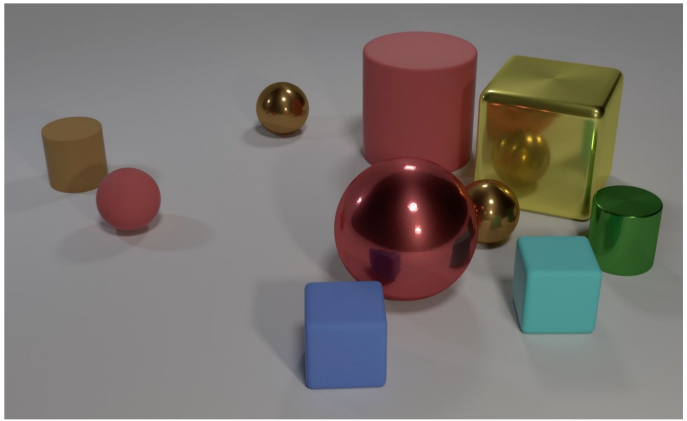
Visual Dialog (Image Guessing Game)



Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. "Visual Dialog." CVPR 2017

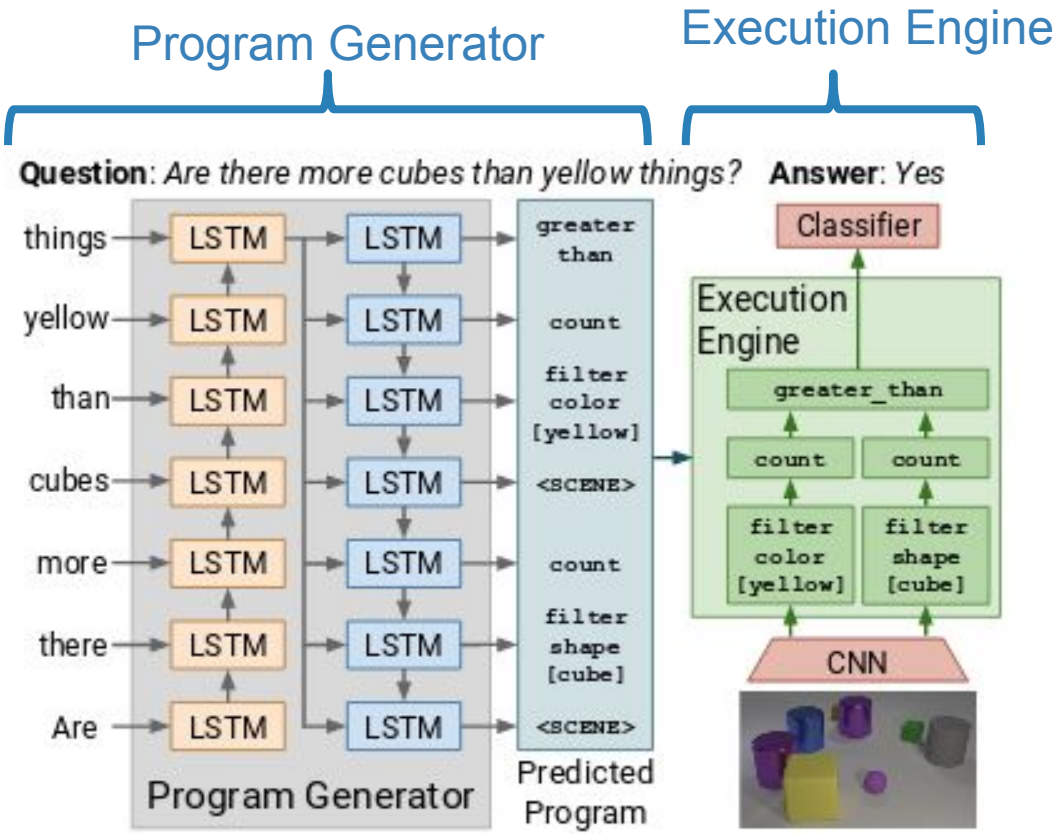
Visual Reasoning

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



(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 .

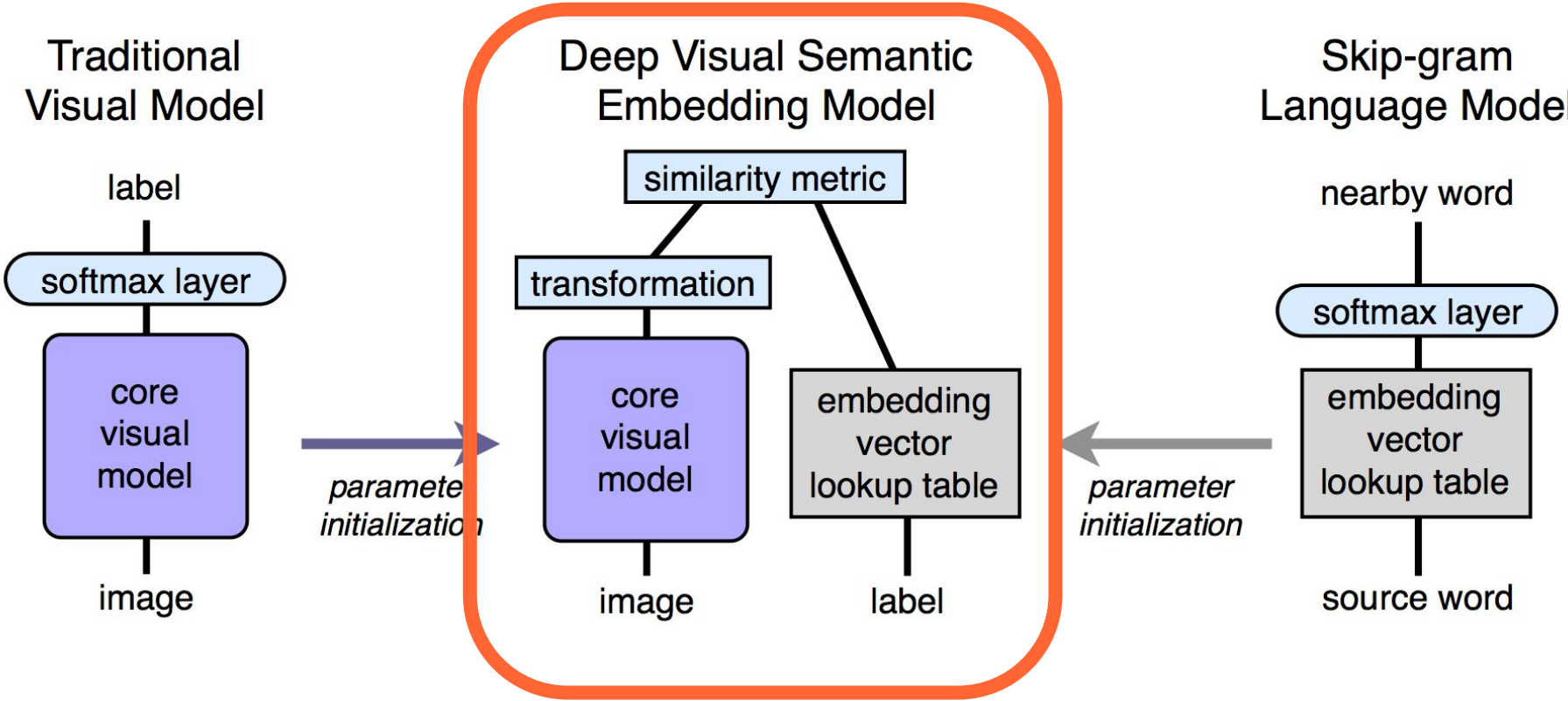
Encoder

Encoder

Representation or Embedding

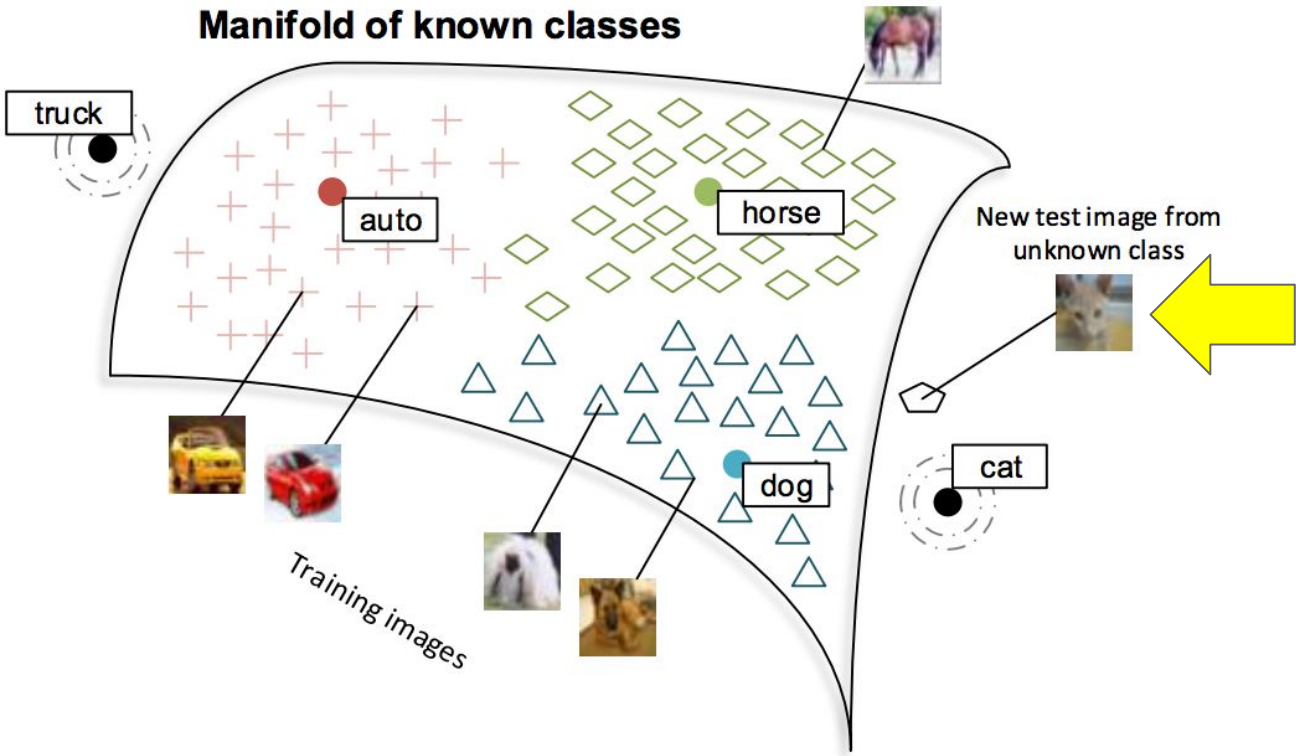
$[z_1, z_2, \ldots, z_d]$

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



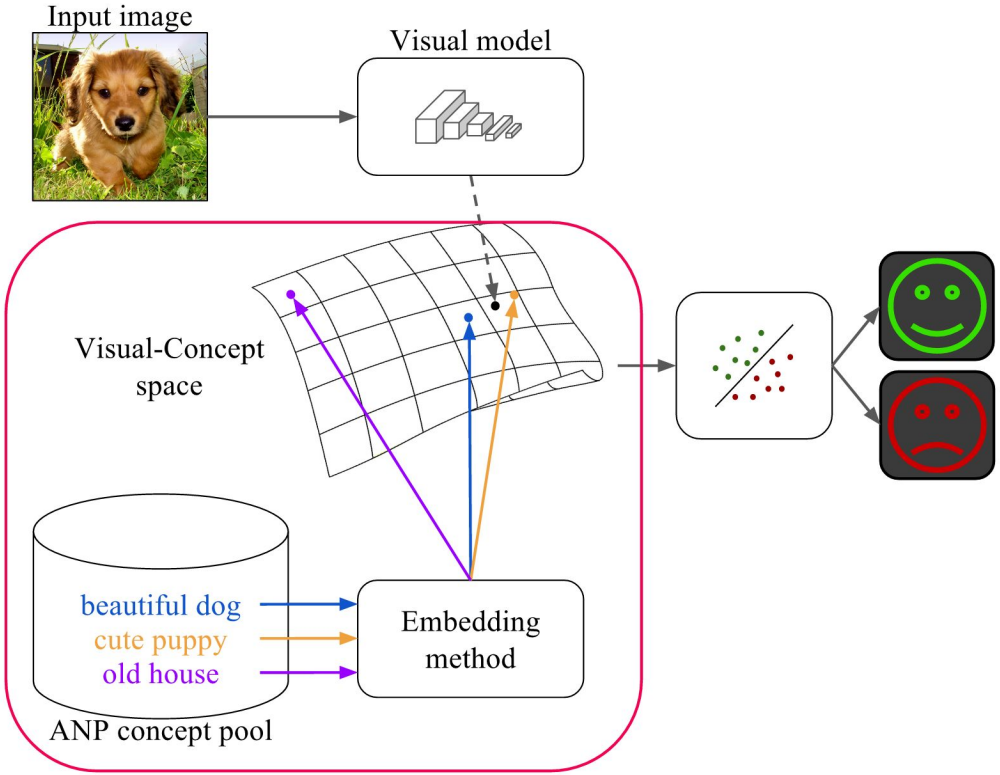
Zero-shot learning:
a class not present in the training set of images can be predicted

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

Socher, R., Ganjoo, M., Manning, C. D., & Ng, A., Zero-shot learning through cross-modal transfer. NIPS 2013 [slides] [code]

Joint Neural Embeddings

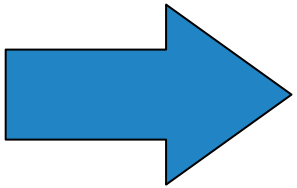
Foggy Day



Alejandro Woodward, Víctor Campos, Dèlia Fernàndez, Jordi Torres, Xavier Giró-i-Nieto, Brendan Jou and Shih-Fu Chang (submitted)

Joint Neural Embeddings

Image and text retrieval with joint embeddings.



Ingredients

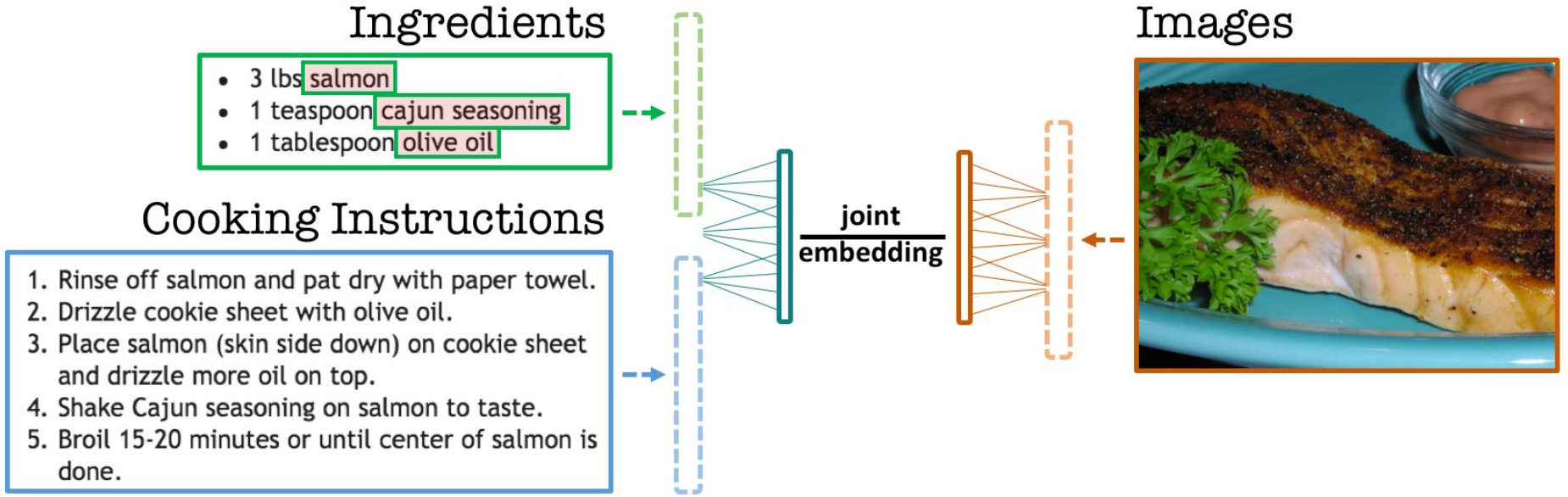
- 3 lbs salmon
- 1 teaspoon cajun seasoning
- 1 tablespoon olive oil

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.

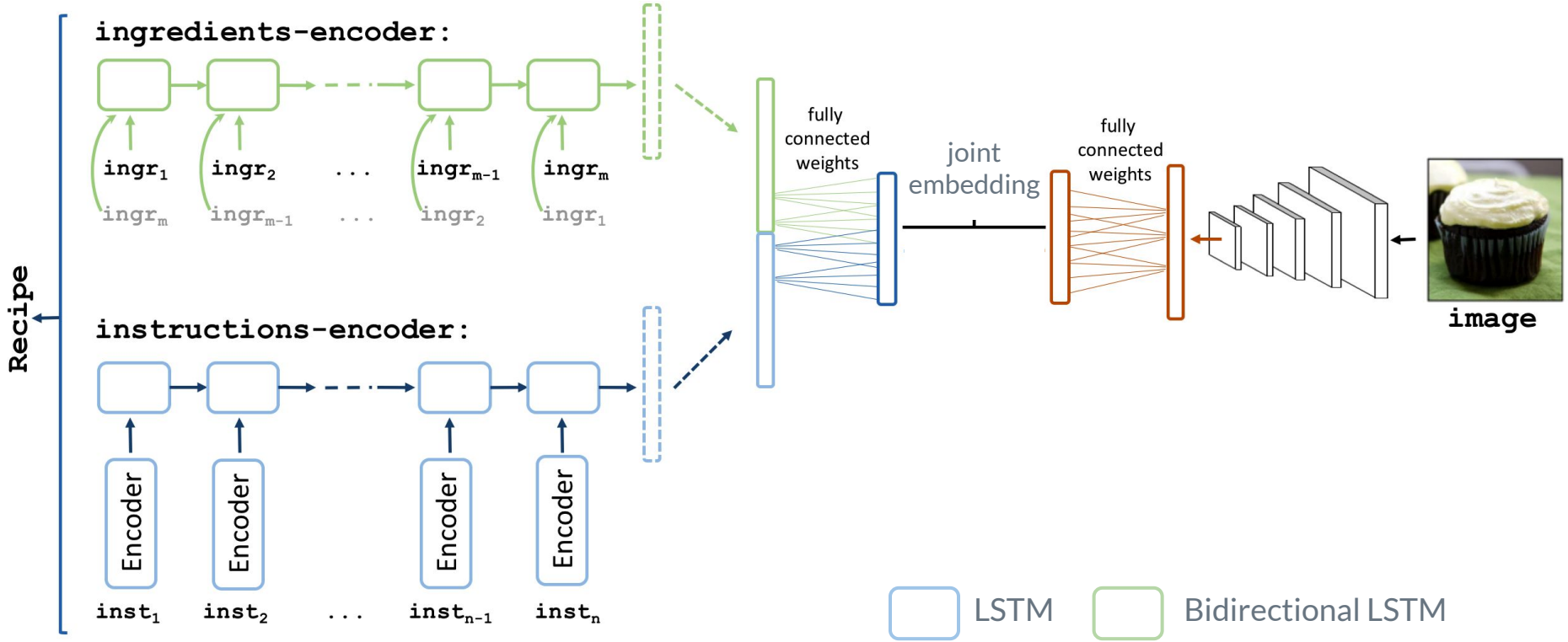
Amaia Salvador, Nicholas Haynes, Yusuf Aytar, Javier Marín, Ferda Ofli, Ingmar Weber, Antonio Torralba, "Learning Cross-modal Embeddings for Cooking Recipes and Food Images". CVPR 2017

Joint Neural Embeddings



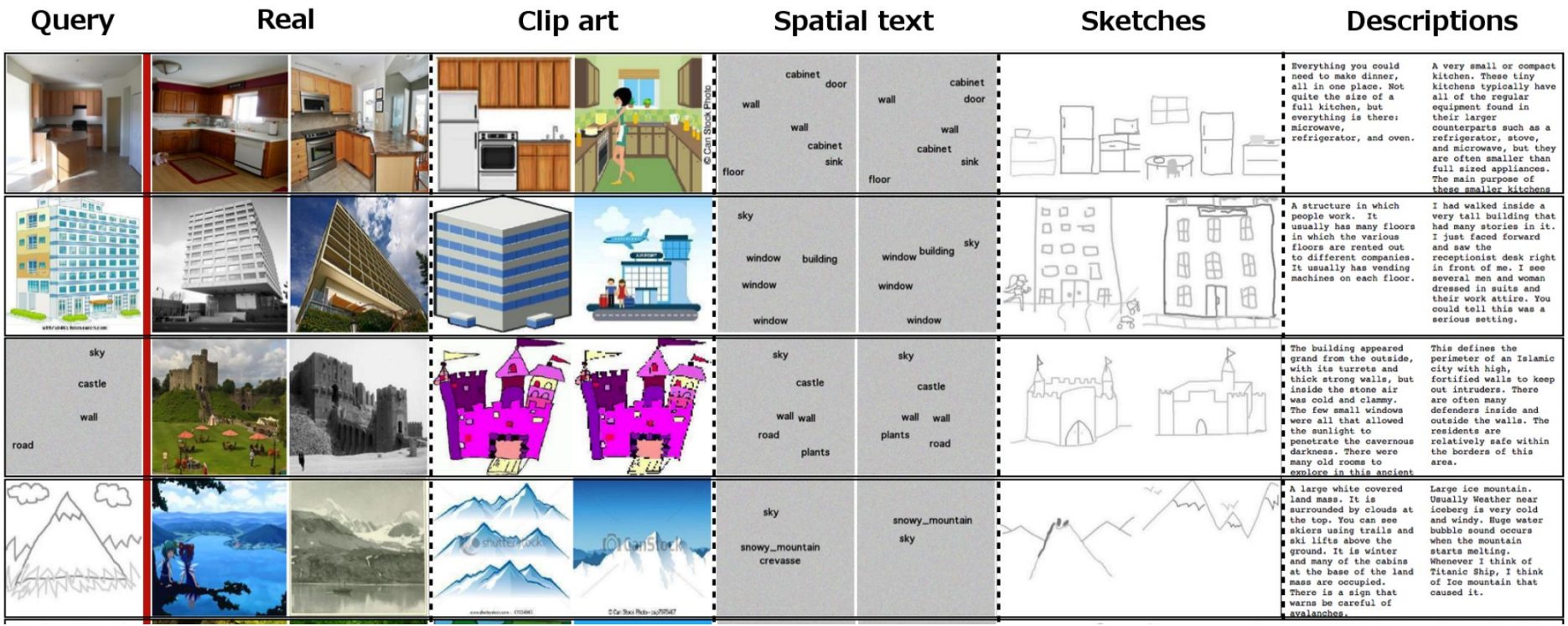
Amaia Salvador, Nicholas Haynes, Yusuf Aytar, Javier Marín, Ferda Ofli, Ingmar Weber, Antonio Torralba, "Learning Cross-modal Embeddings for Cooking Recipes and Food Images". CVPR 2017

Joint Neural Embeddings



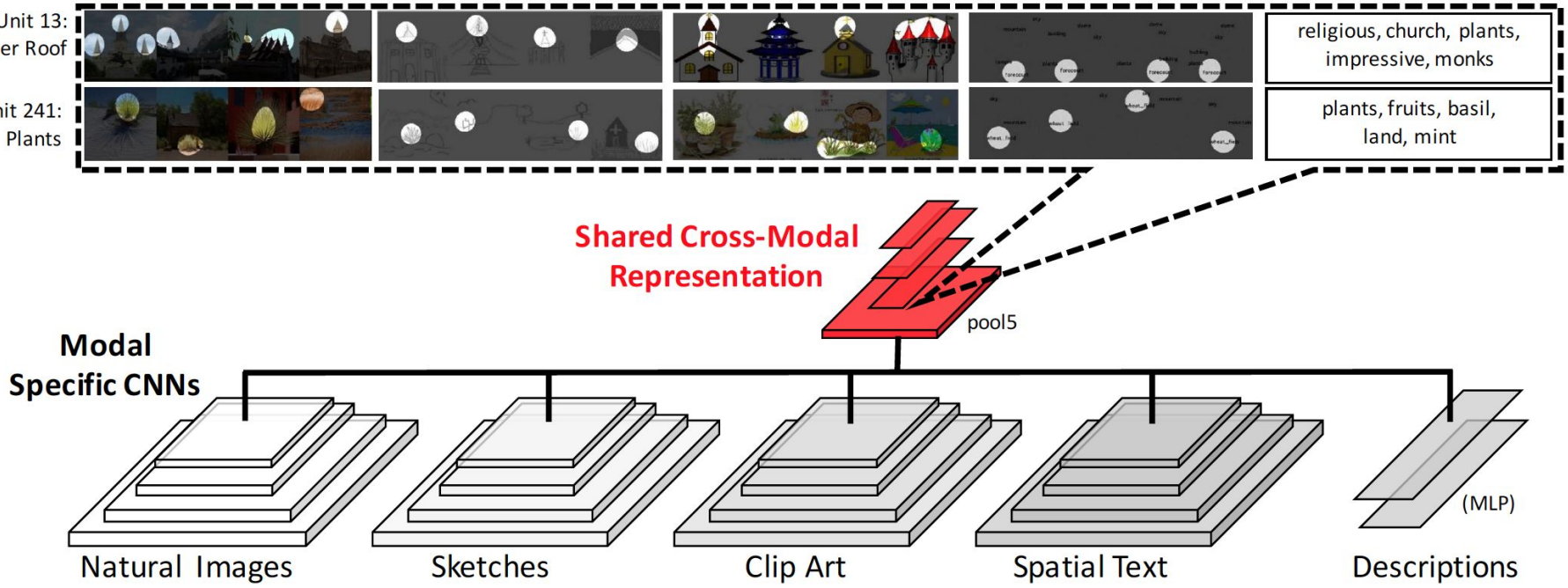
Amaia Salvador, Nicholas Haynes, Yusuf Aytar, Javier Marín, Ferda Ofli, Ingmar Weber, Antonio Torralba, "Learning Cross-modal Embeddings for Cooking Recipes and Food Images". CVPR 2017

Image to image and text



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

Image to image and text

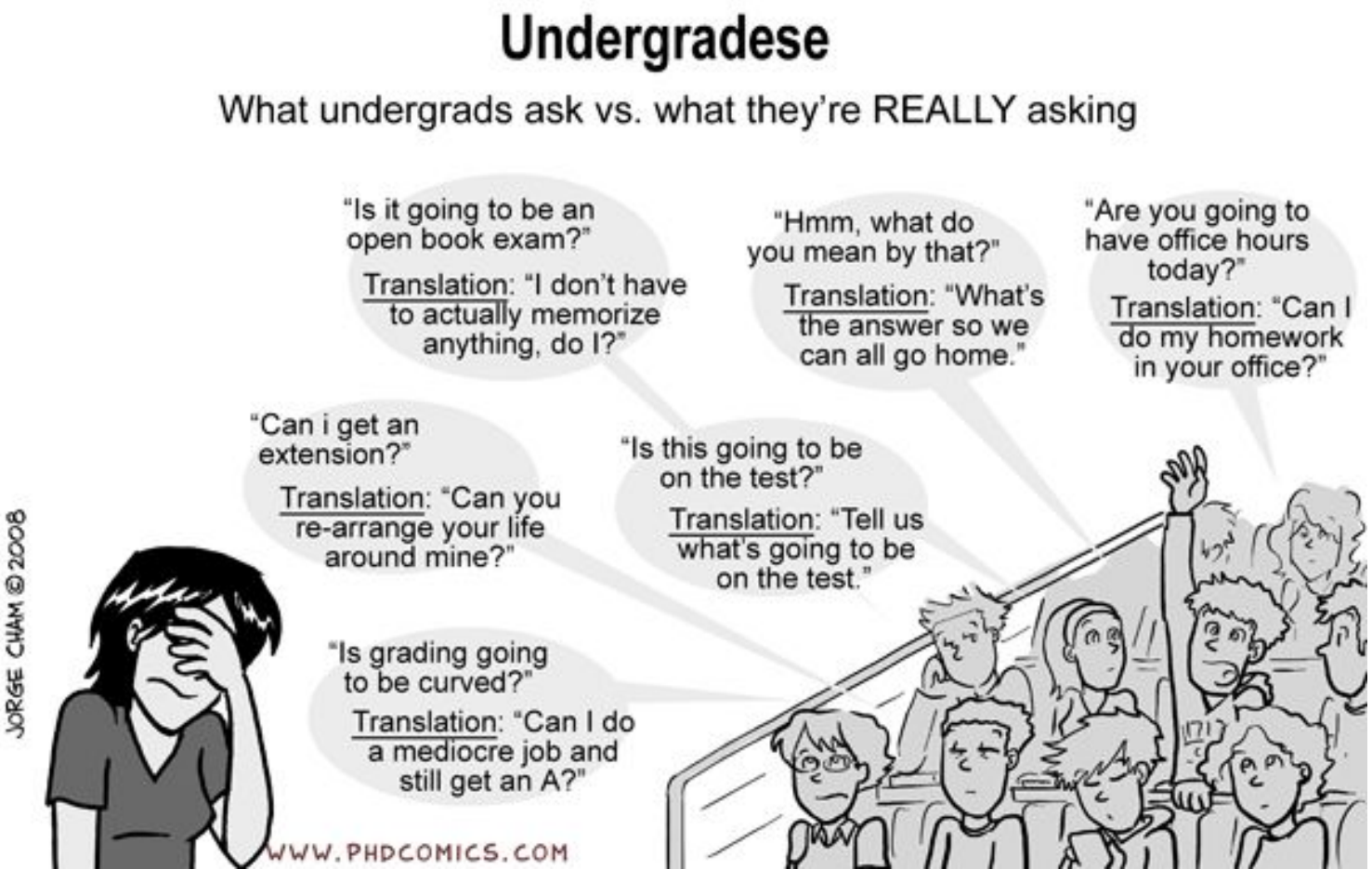


Aytar, Yusuf, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. "Cross-Modal Scene Networks." CVPR 2016.

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?"
JORGE CHAM © 2008
WWW.PHDCOMICS.COM