Insight
Centre for Data Analytics

DEEP
LEARNING
WORKSHOP

Dublin City University
27-28 April 2017

DCU

#InsightDL2017

Day 2 Lecture 11

Language and Vision

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

Associate Professor
Universitat Politecnica de Catalunya
Technical University of Catalonia

UPC

Acknowledgments

Santi Pascual

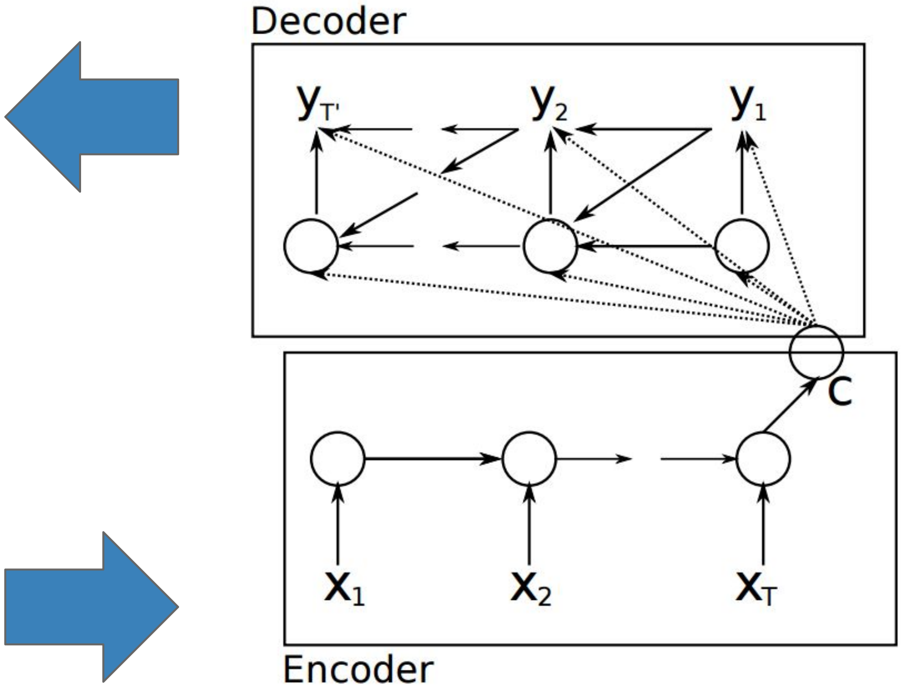
UPC TALP

TALP Research Center

Previously in the RNN lecture...

Language OUT

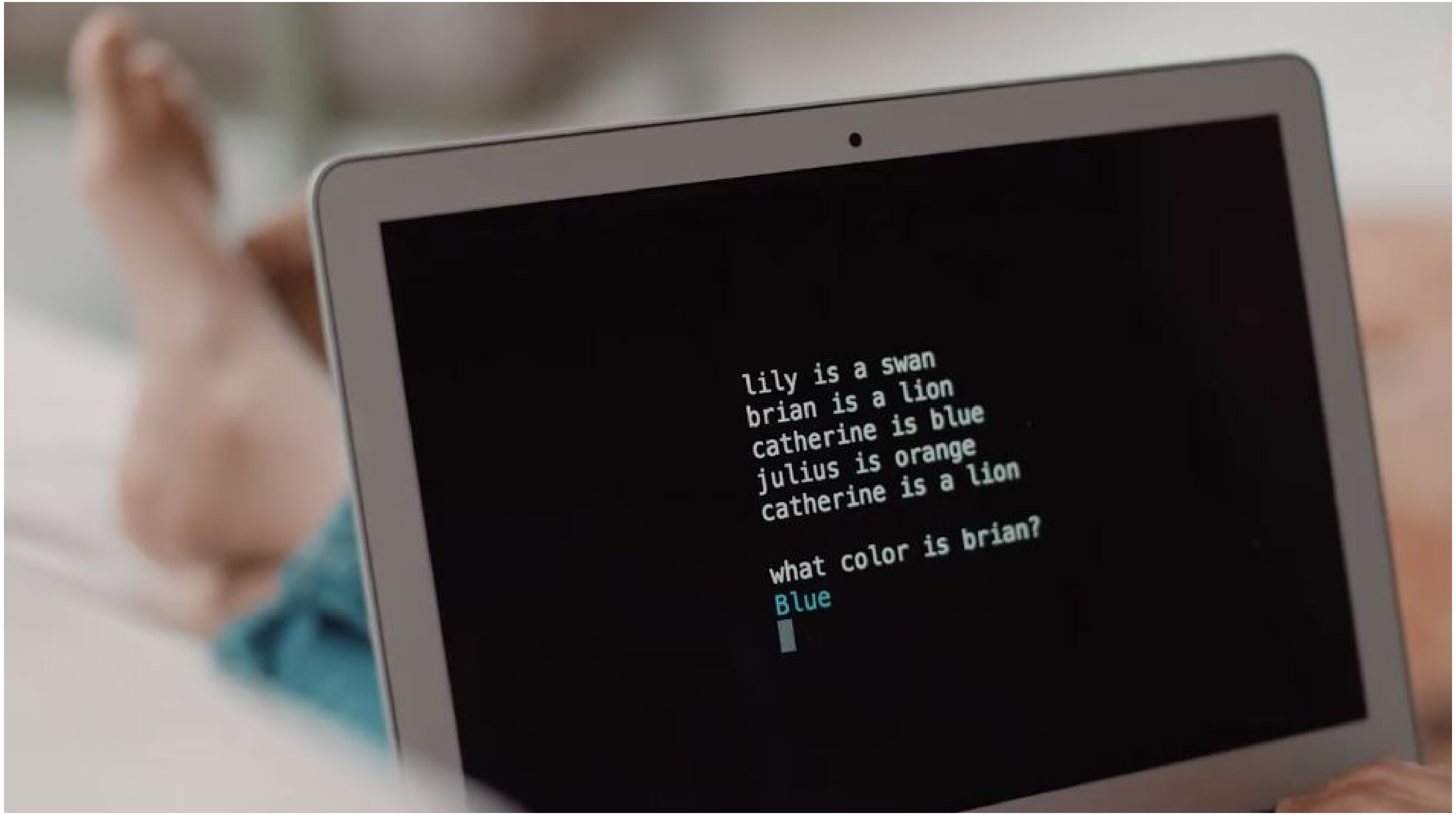
Language IN



Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

Motivation

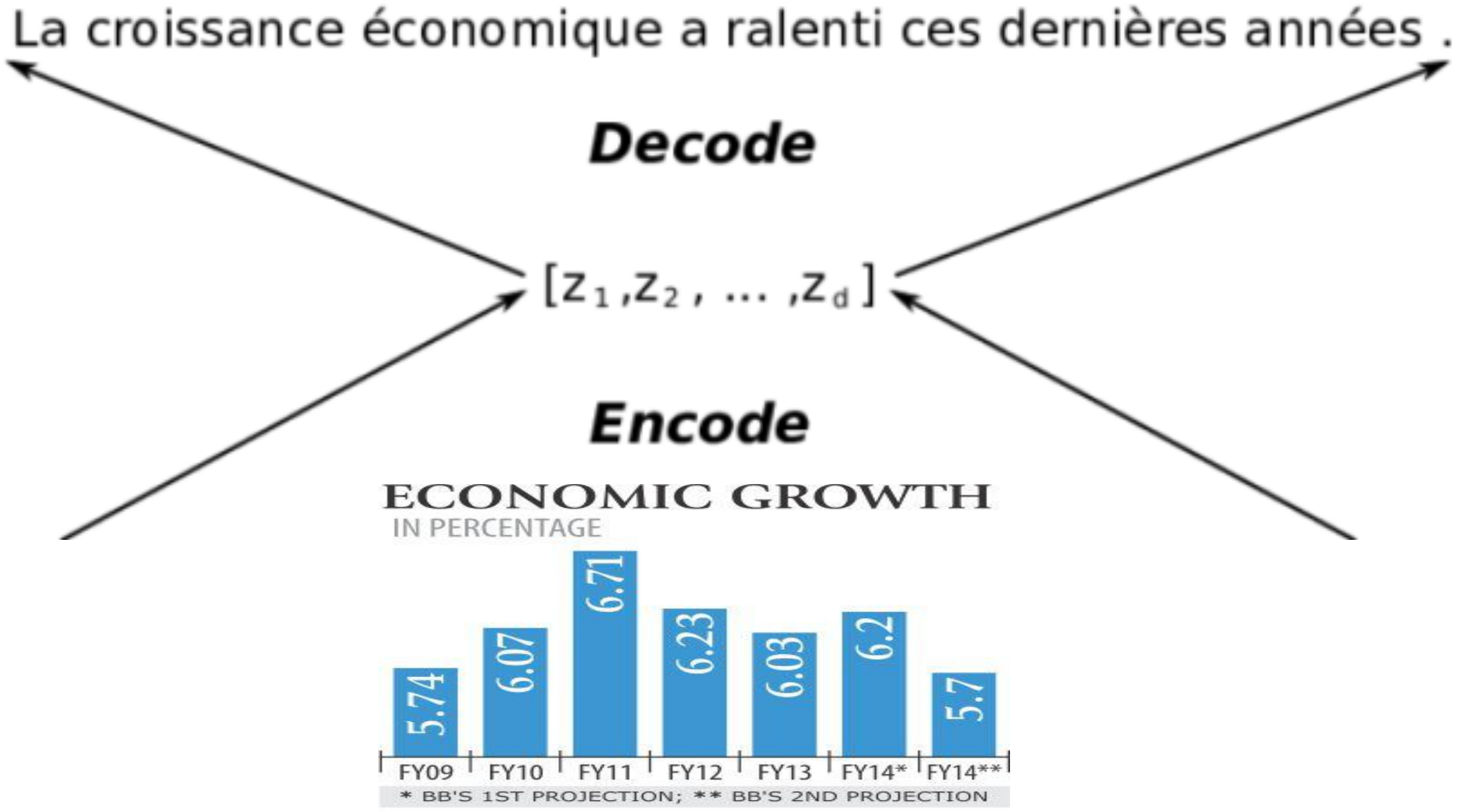
Facebook AI Research



```
lily is a swan
brian is a lion
catherine is blue
julius is orange
catherine is a lion

what color is brian?
Blue
```

Encoder-Decoder: Beyond text



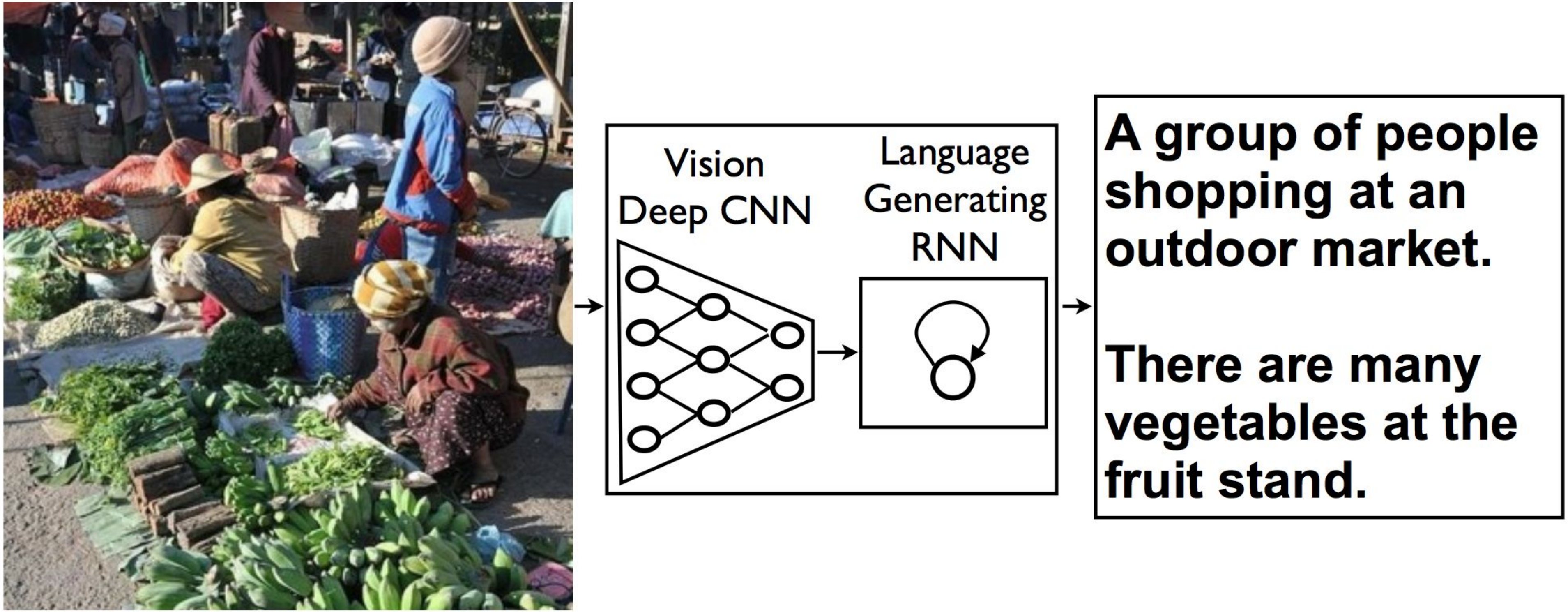
La croissance économique a ralenti ces dernières années .

Decode

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

Encode

Captioning: Show & Tell

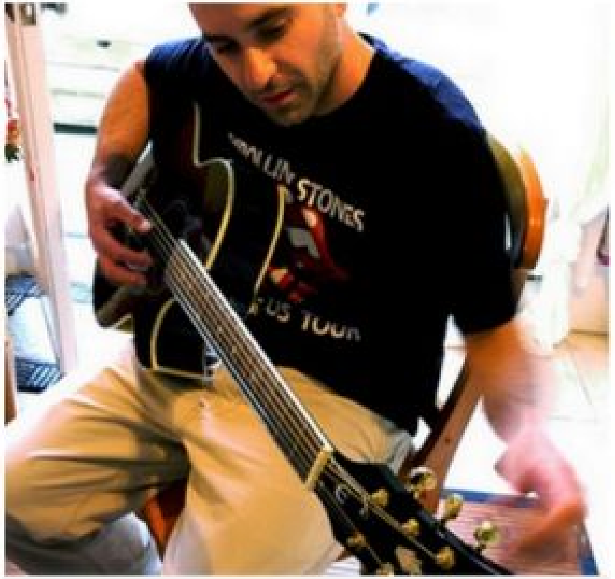


A group of people shopping at an outdoor market.

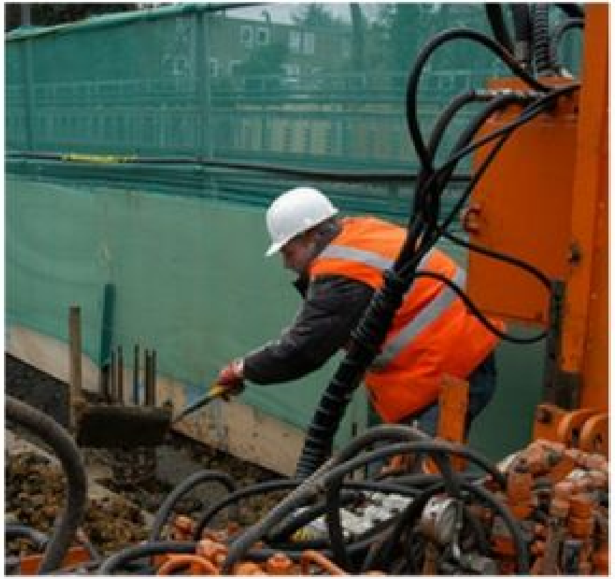
There are many vegetables at the fruit stand.

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

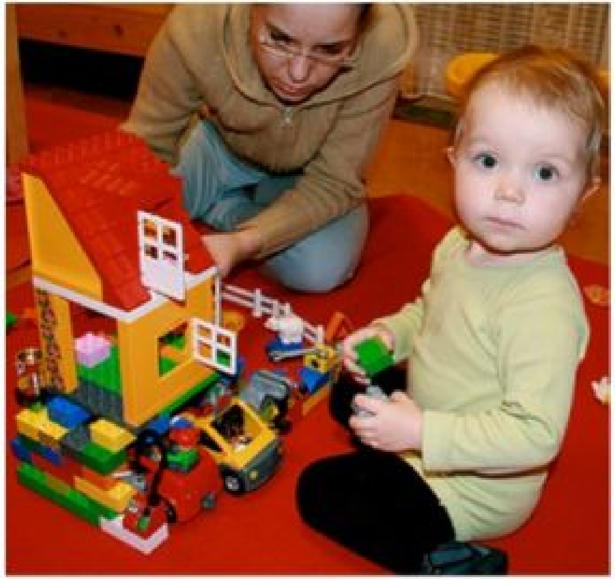
Captioning: DeepImageSent



man in black shirt is playing guitar.



construction worker in orange safety vest is working on road.



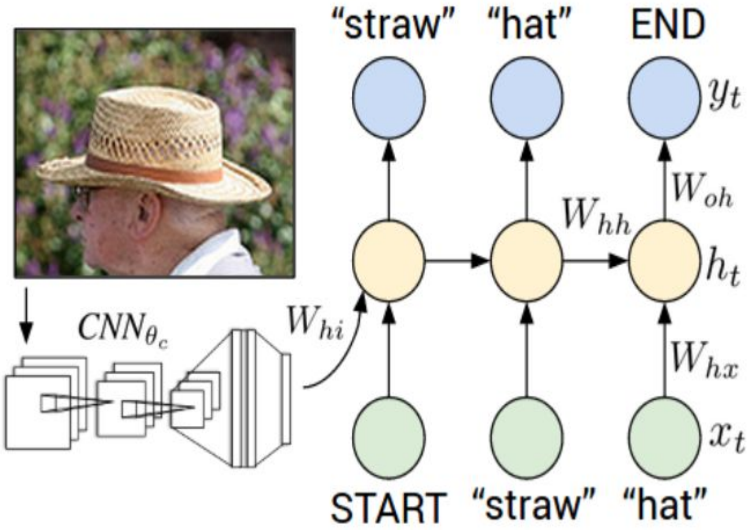
two young girls are playing with lego toy.

(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

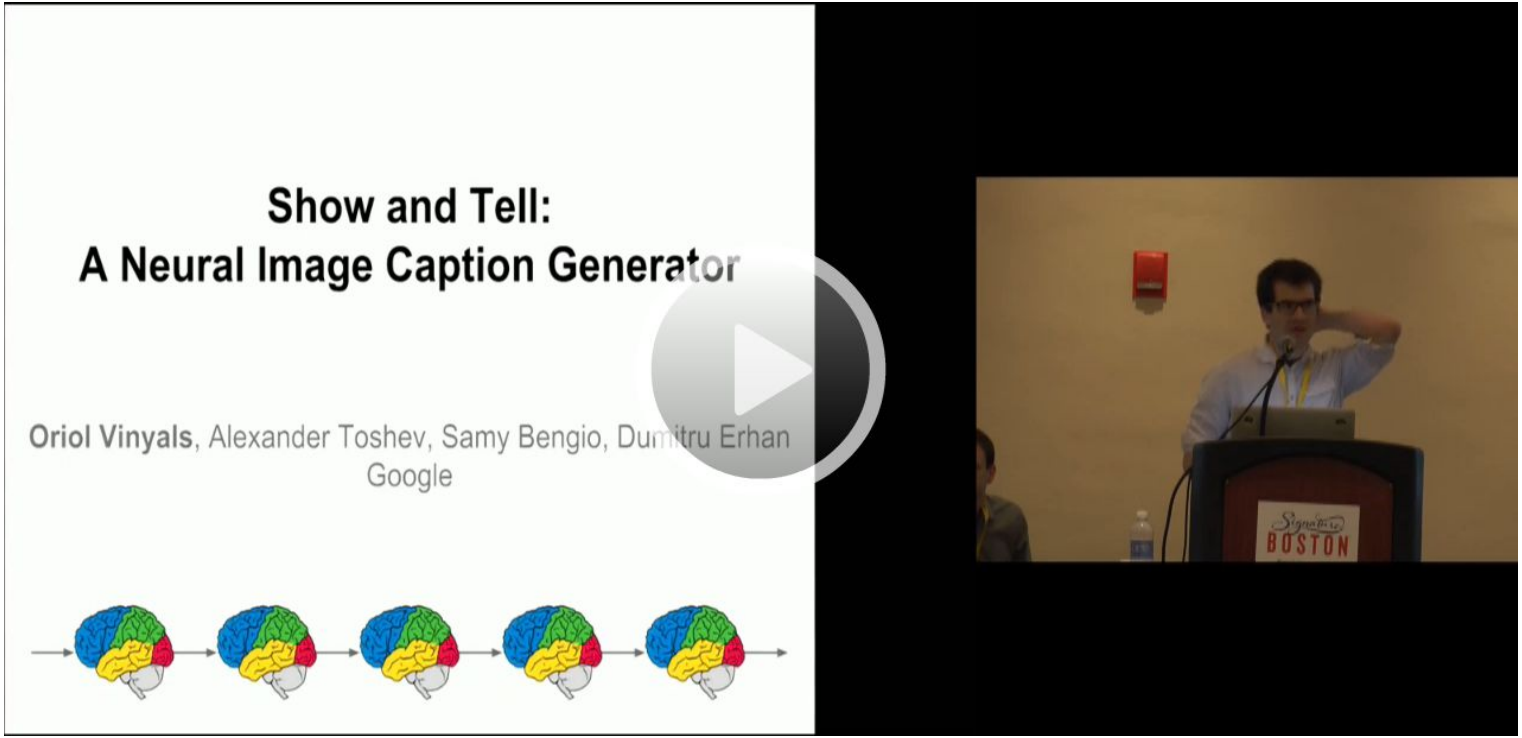
$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \boxed{\mathbb{1}(t=1) \odot b_v})$$
$$y_t = softmax(W_{oh}h_t + b_o).$$



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 & Tell

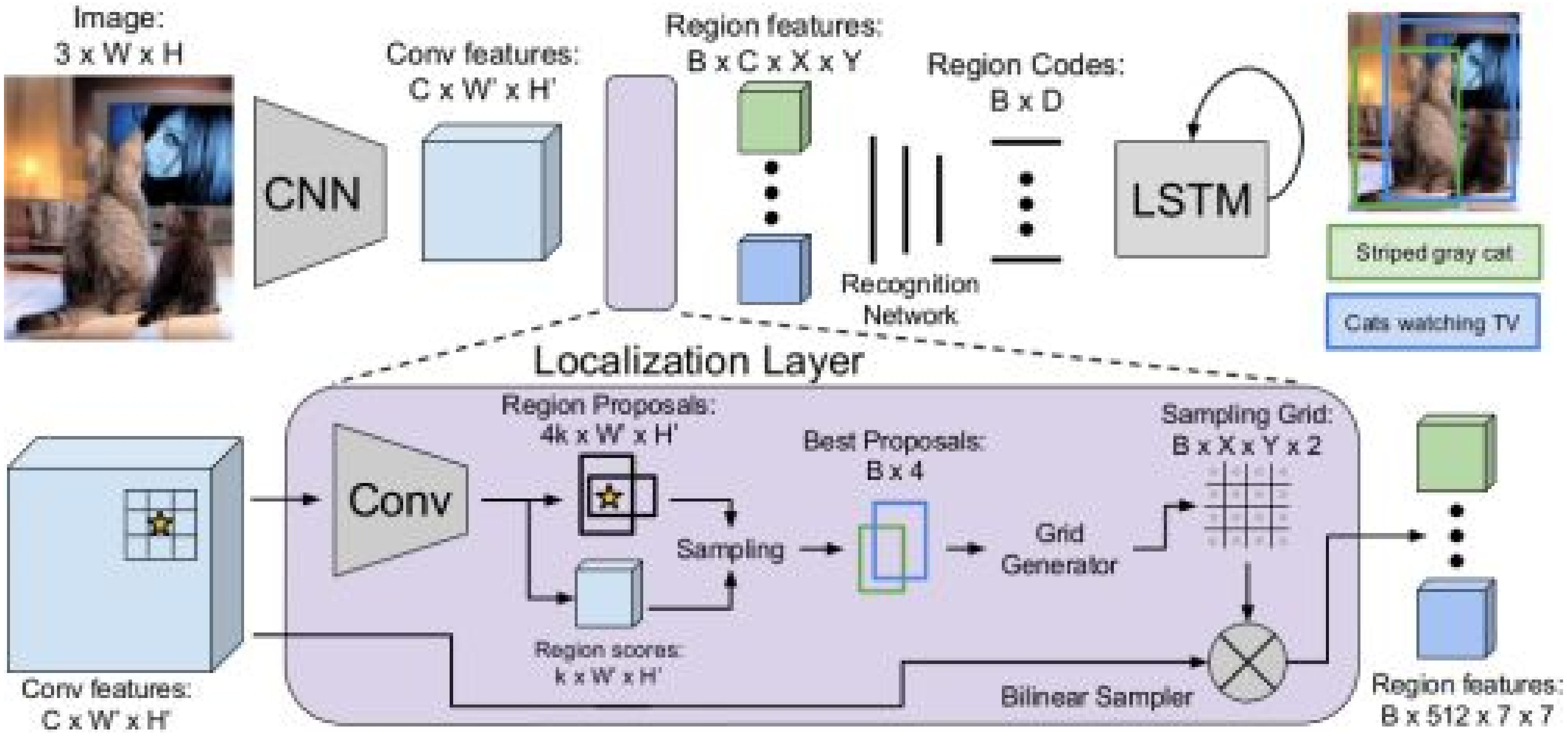


Show and Tell:
A Neural Image Caption Generator

Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan
Google

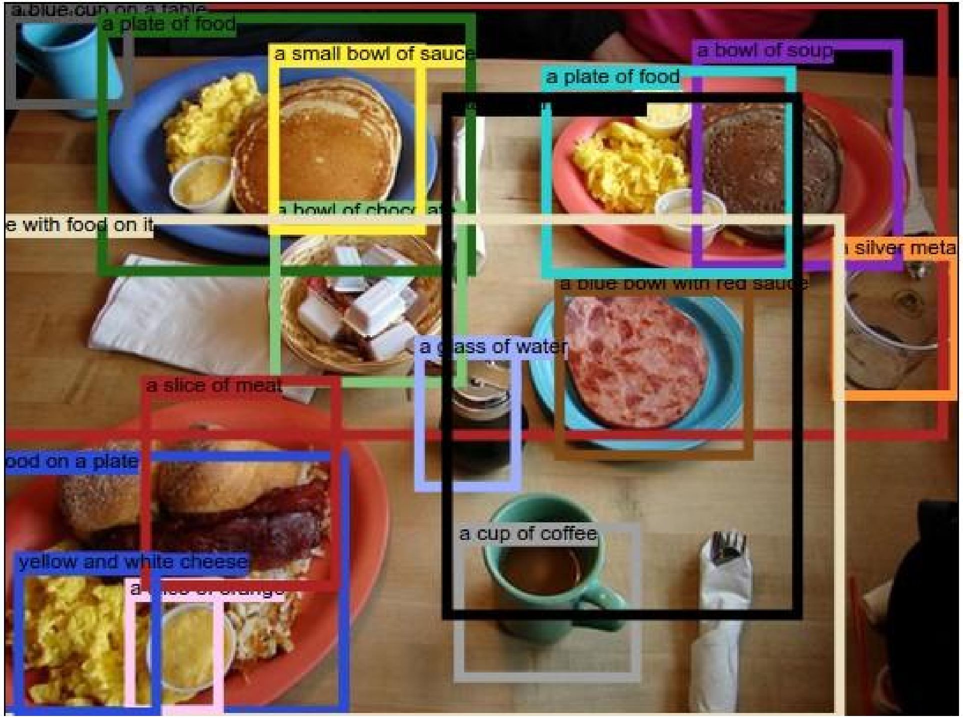
Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." CVPR 2015.

Captioning (+ Detection): DenseCap



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

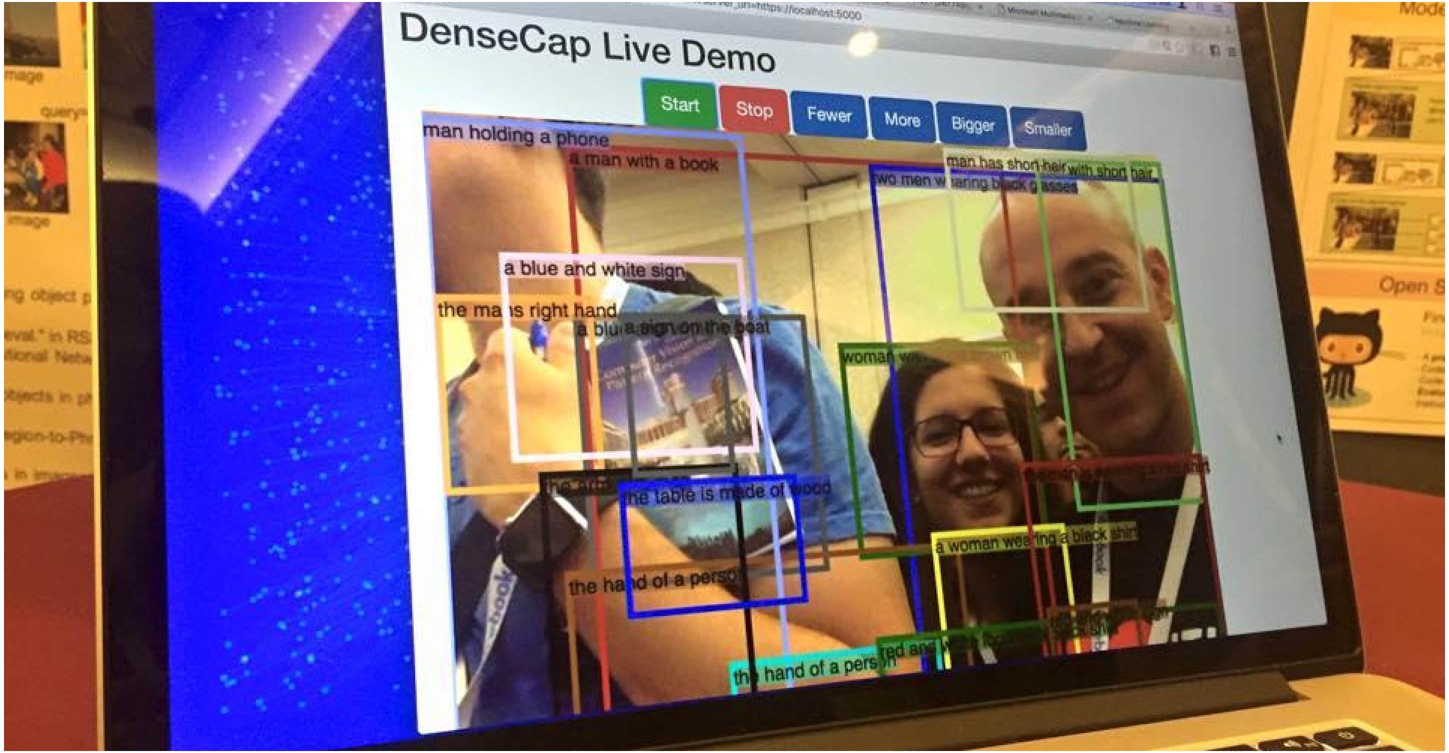
Captioning (+ Detection): DenseCap



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



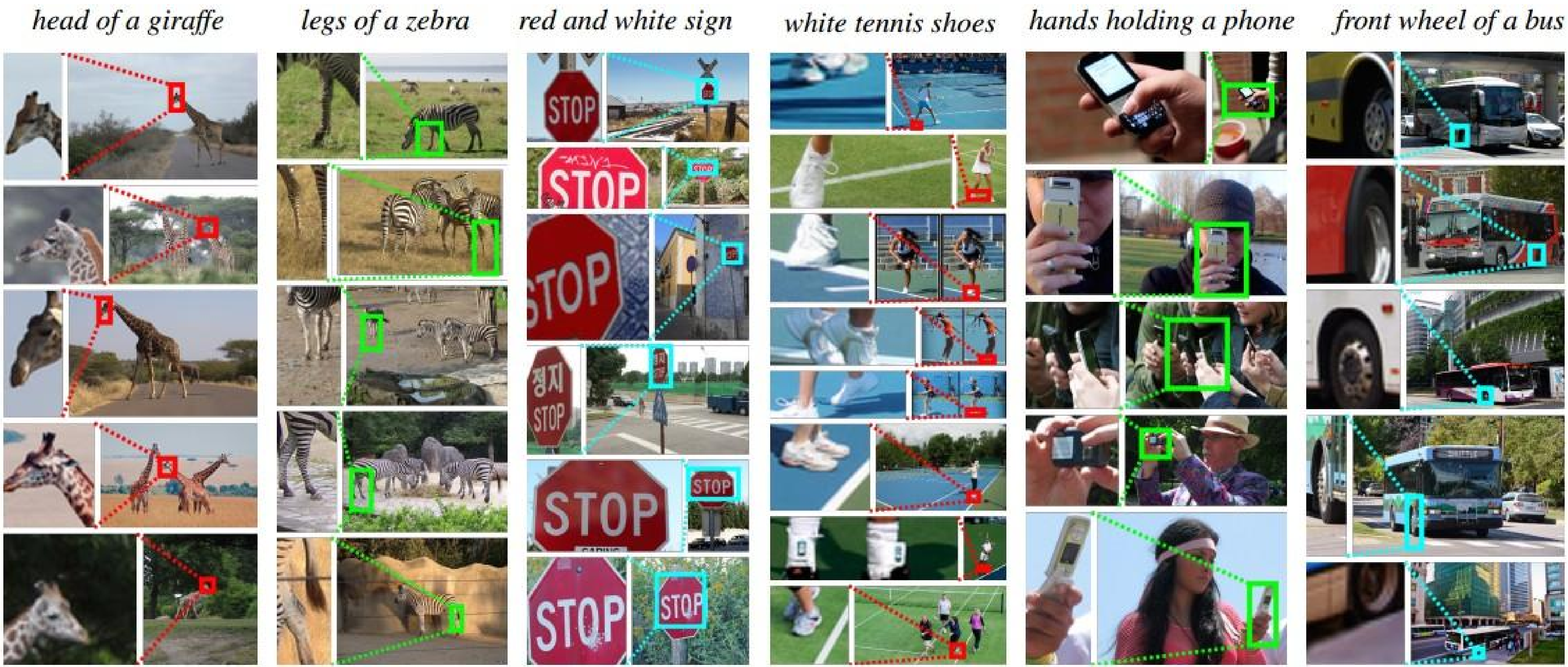
XAVI: “man has short hair”, “man with short hair”

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

BOTH: “two men wearing black glasses”

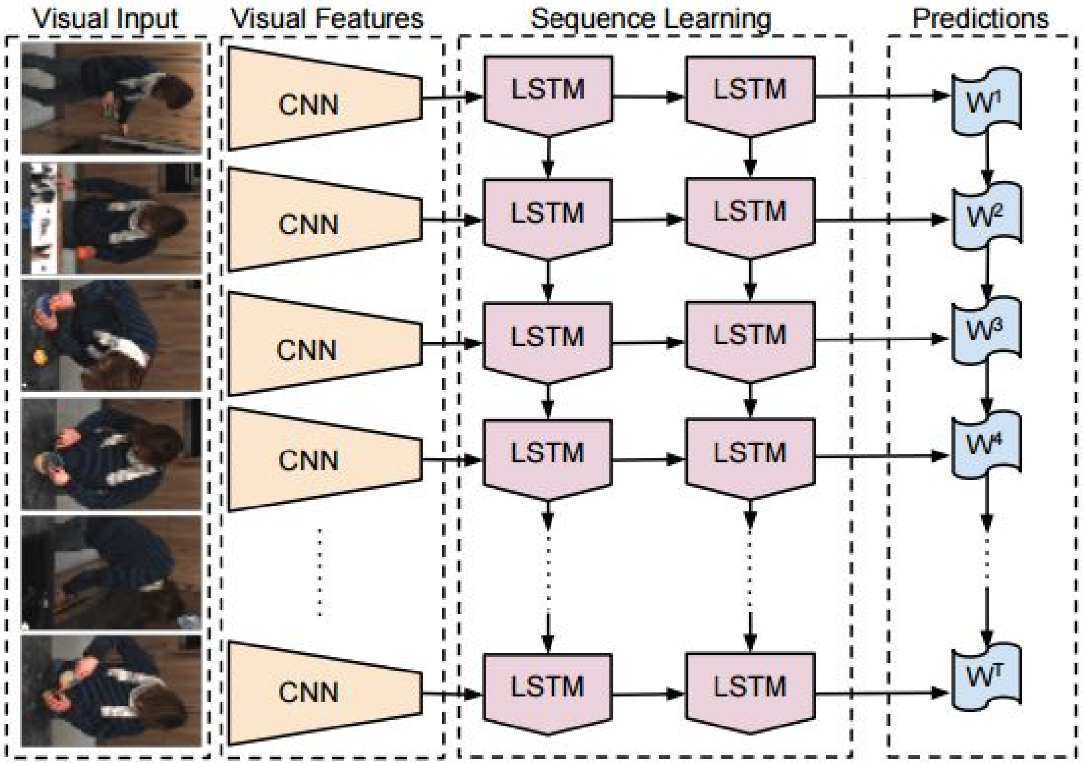
Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning." *CVPR 2016*

Captioning (+ Retrieval): DenseCap



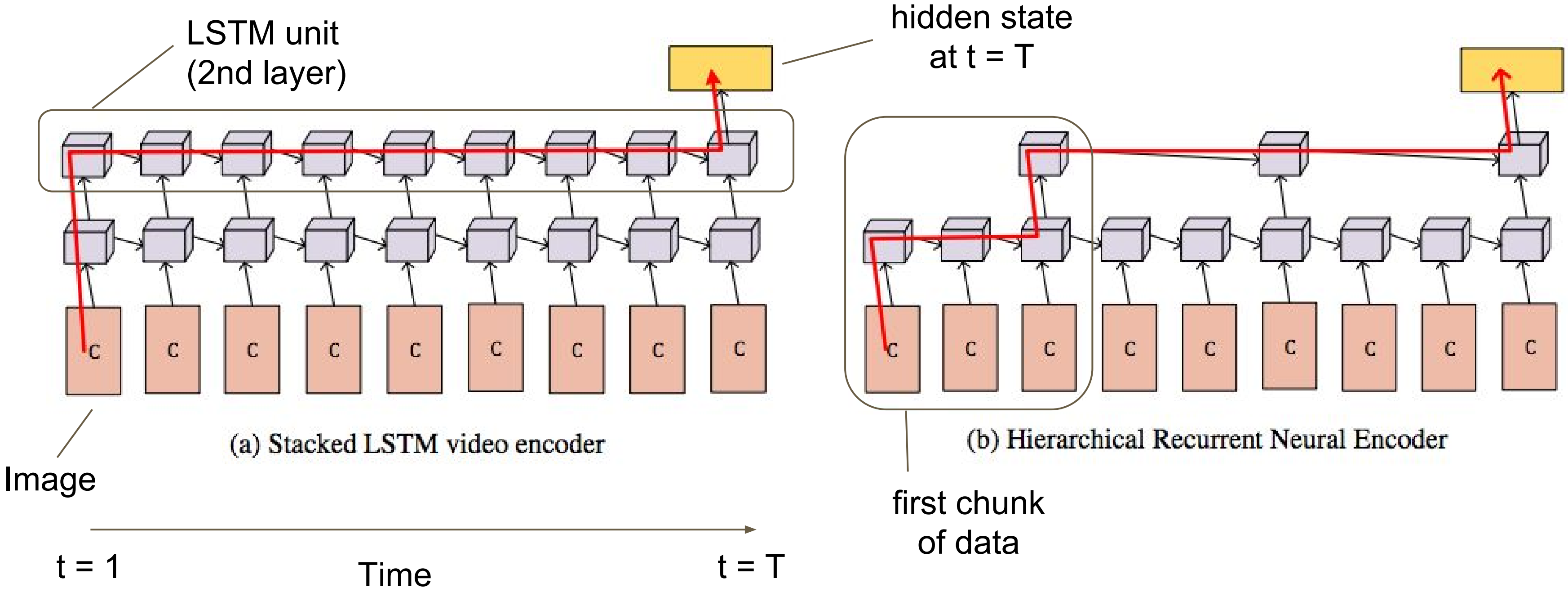
Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning." *CVPR 2016*

Captioning: Video



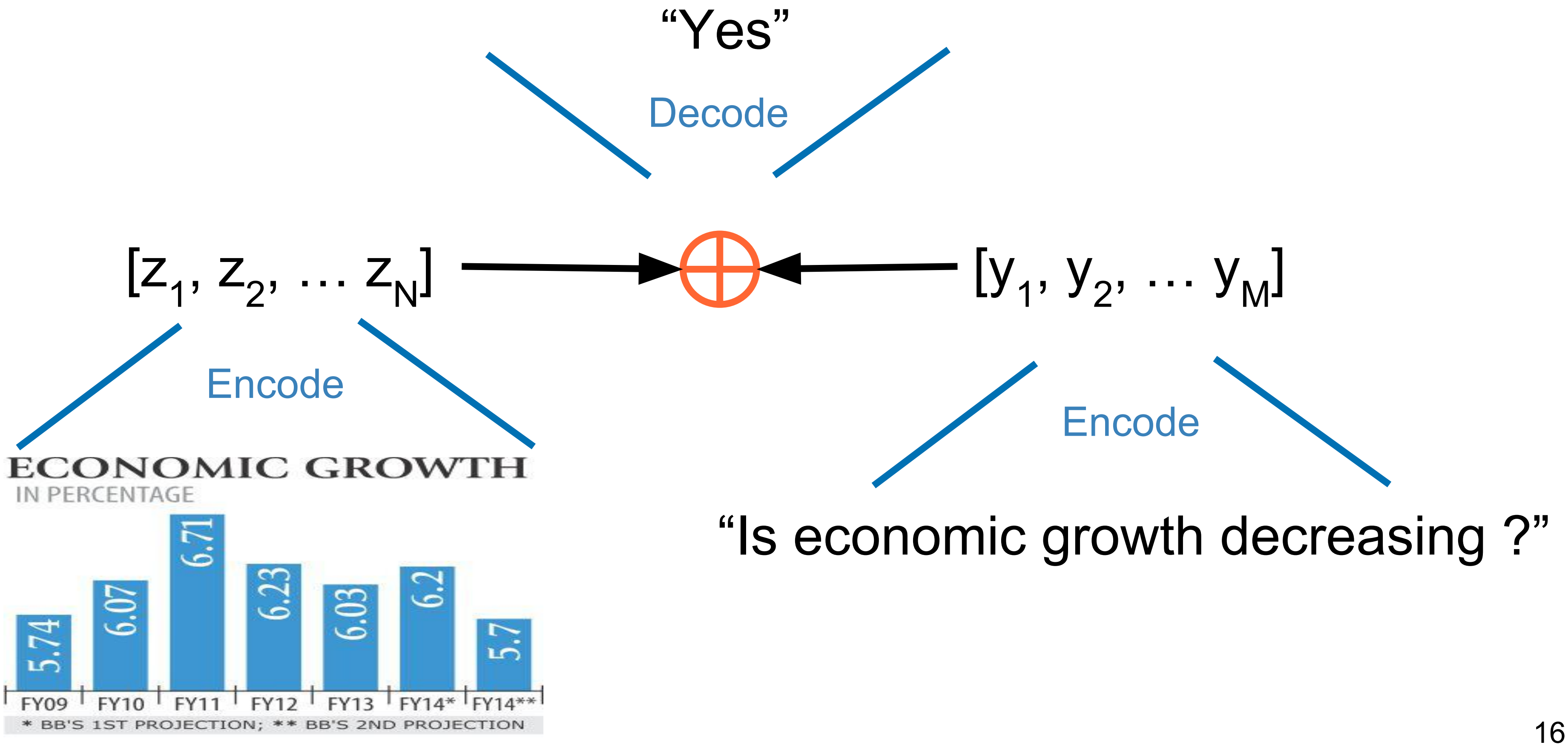
Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrel. Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015. code

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.

Visual Question Answering



“Yes”

Decode

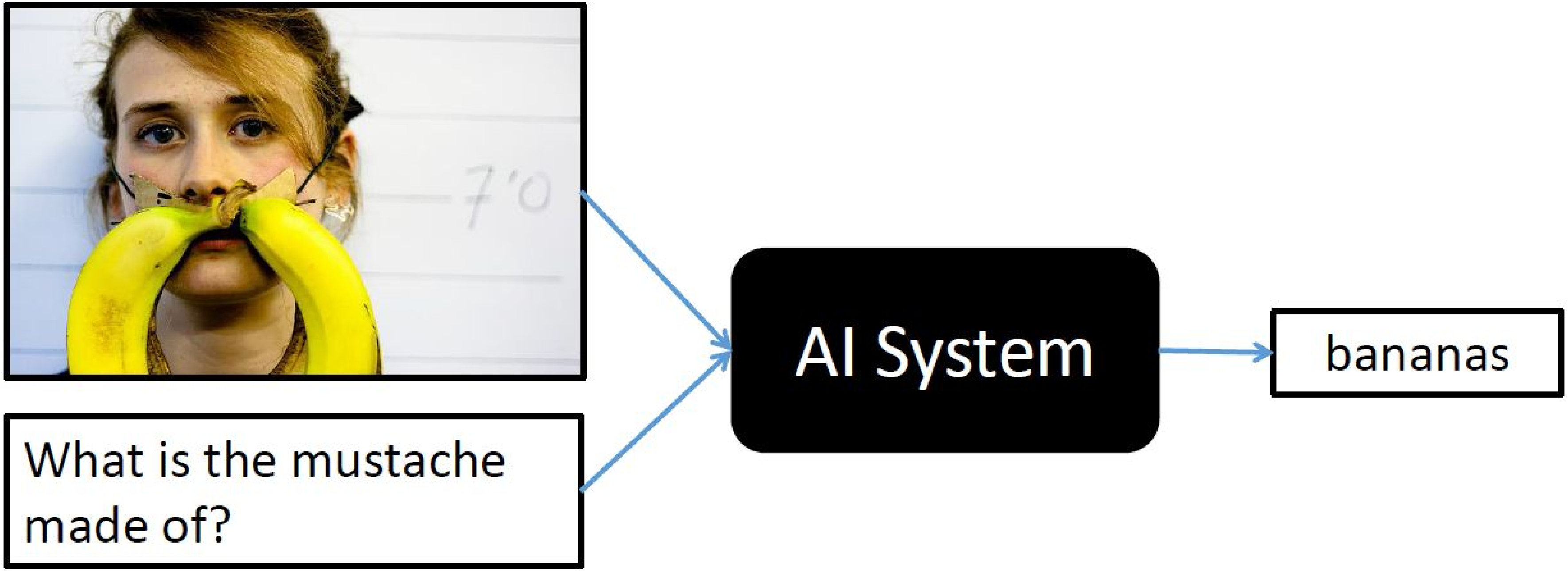
$[z_1, z_2, \ldots z_N] \rightarrow \oplus \leftarrow [y_1, y_2, \ldots y_M]$

Encode

Encode

“Is economic growth decreasing ?”

Visual Question Answering



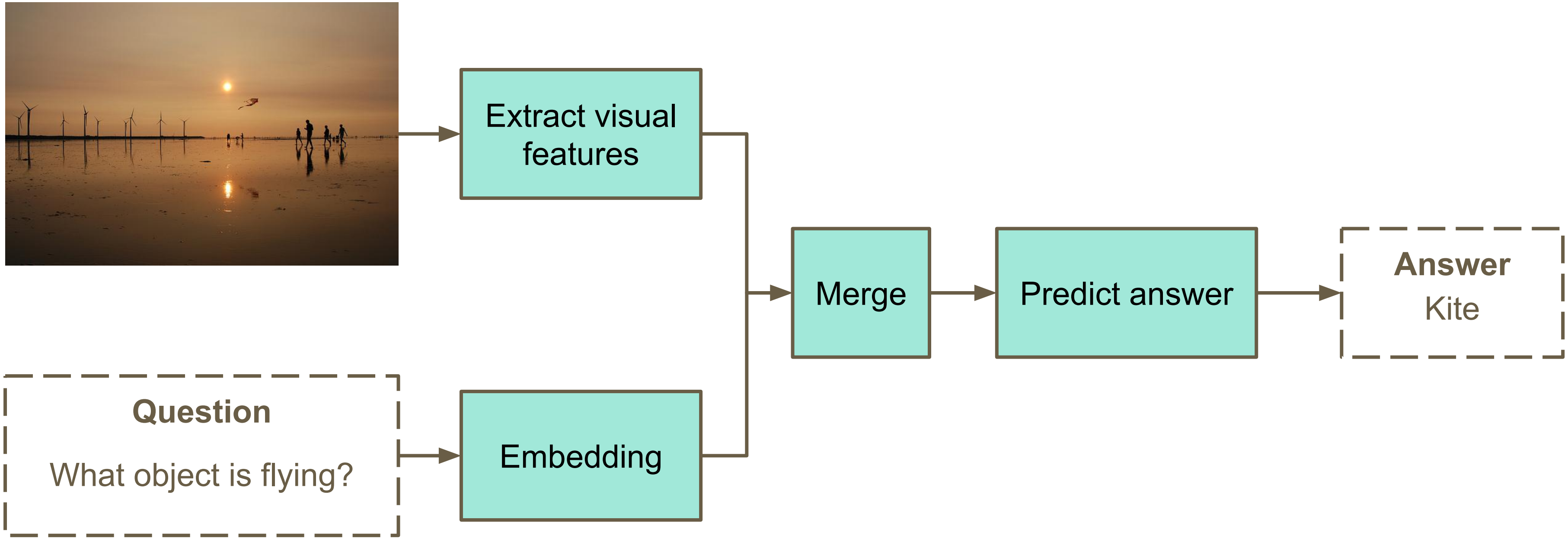
What is the mustache made of?

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



Extract visual features

Embedding

Merge

Predict answer

Answer

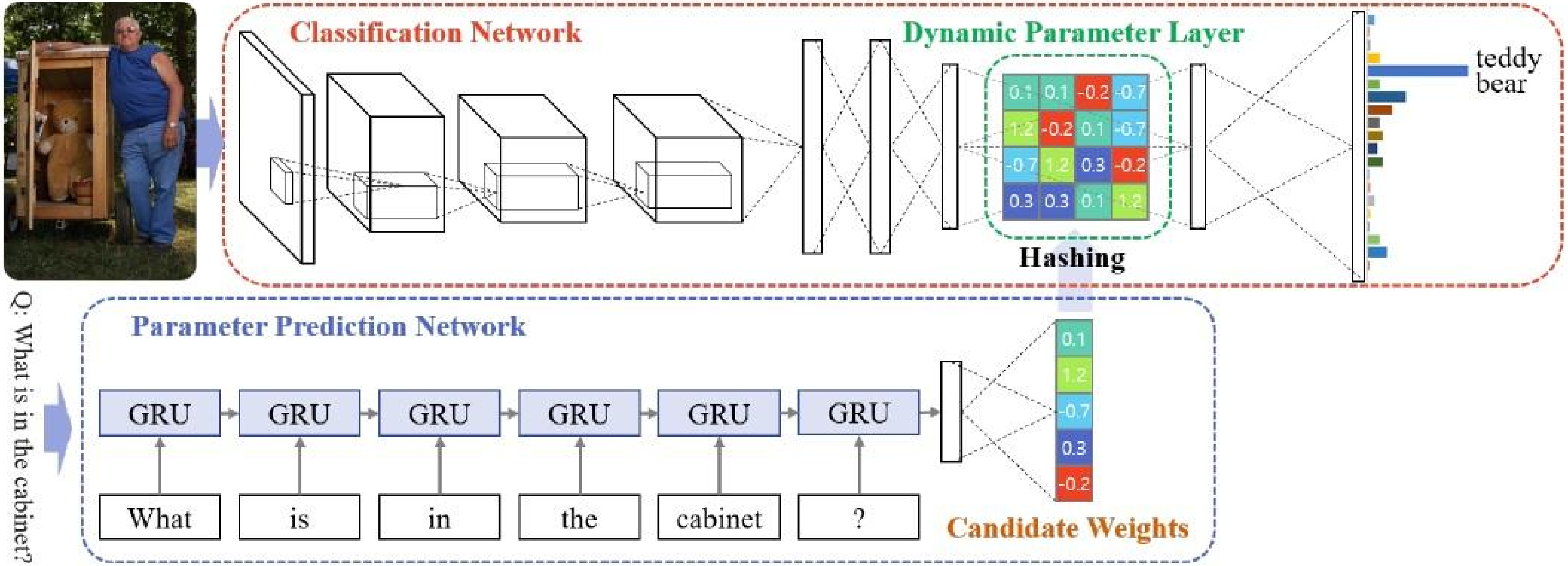
Kite

Question

What object is flying?

Slide credit: Issey Masuda

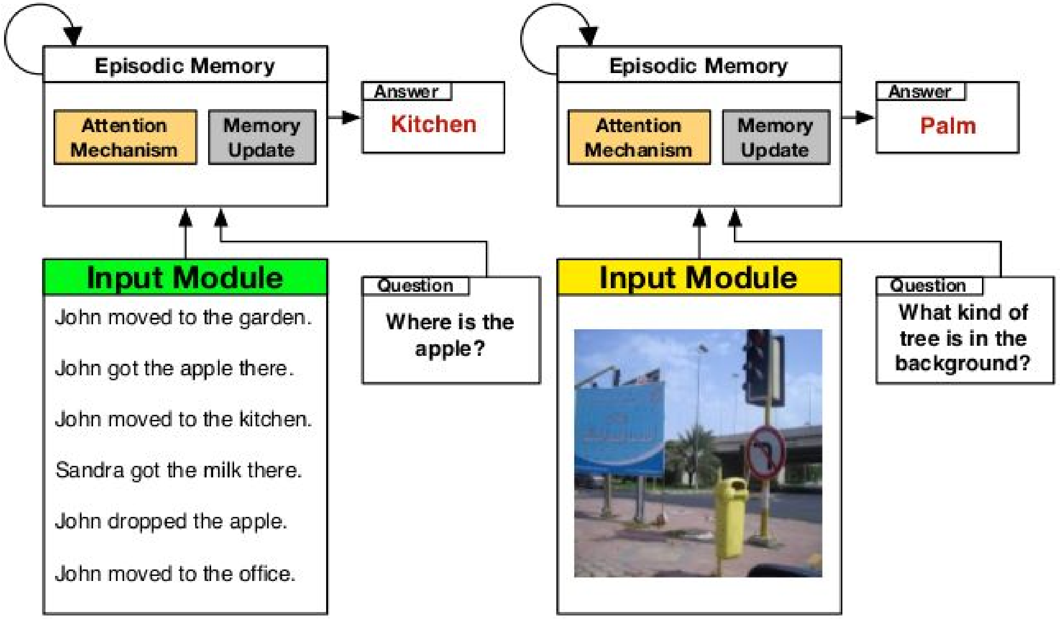
Visual Question Answering



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



(a) Text Question-Answering (b) Visual Question-Answering

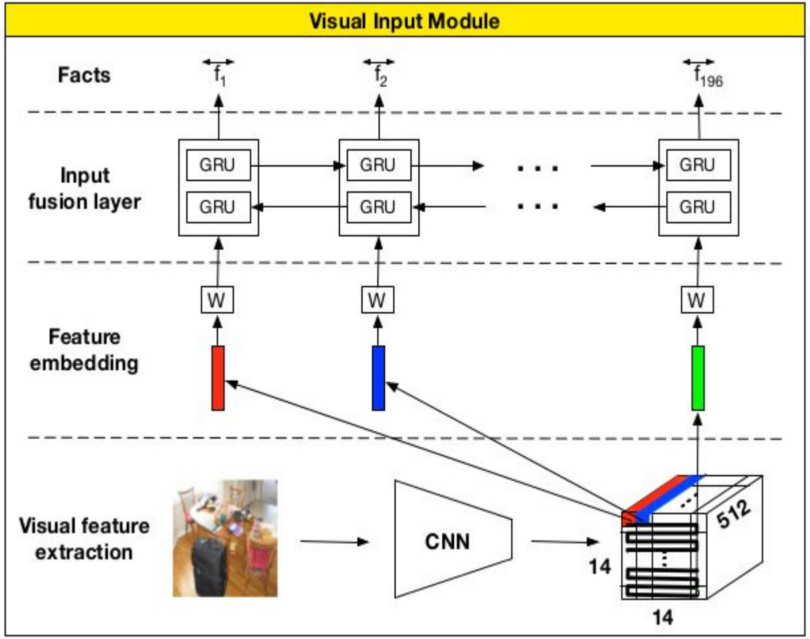
(Slides and Slidecast by Santi Pascual): Xiong, Caiming, Stephen Merity, and Richard Socher. "Dynamic Memory Networks for Visual and Textual Question Answering." arXiv preprint arXiv:1603.01417 (2016).

Visual Question Answering: Dynamic

Main idea: split image into local regions. Consider **each region equivalent to a sentence.**

Local Region Feature Extraction: CNN (VGG-19):
(1) Rescale input to 448x448.
(2) Take output from last pooling layer → D=512x14x14 → 196 512-d local region vectors.

Visual feature embedding: W matrix to project image features to *“**q**”-textual* space.

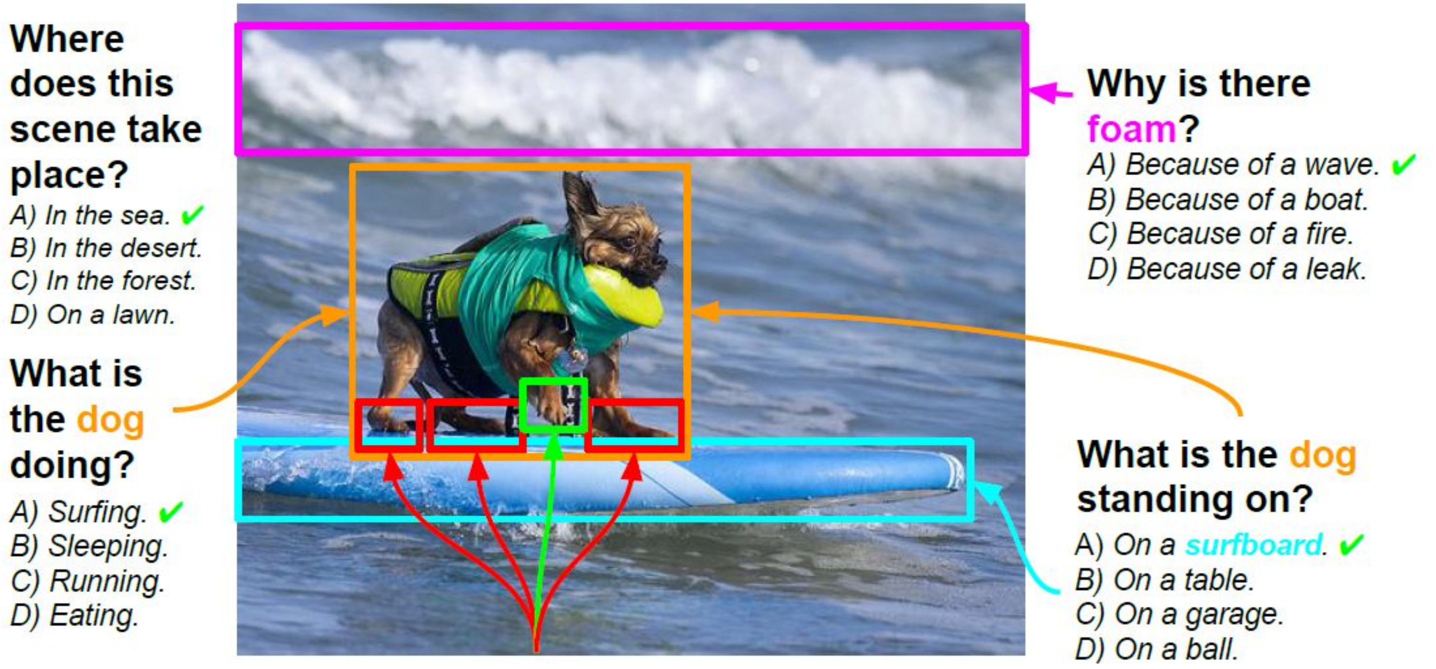


Visual Input Module
Facts
$\overleftrightarrow{f_1}$
$\overleftrightarrow{f_2}$
$\overleftrightarrow{f_{196}}$
Input fusion layer
GRU
GRU
GRU
GRU
GRU
GRU
Feature embedding
W
W
W
Visual feature extraction
CNN
512
14
14

Figure 3. VQA input module to represent images for the DMN.

(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?

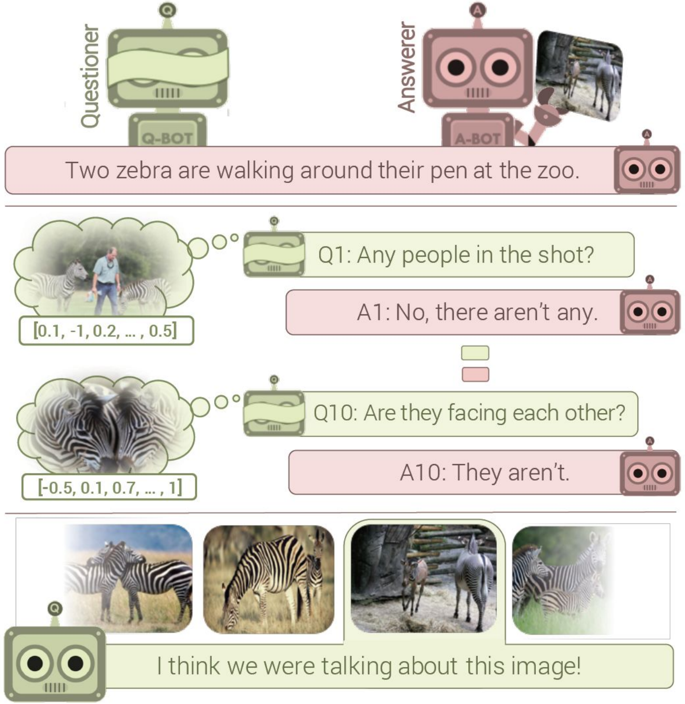
A) Because of a wave. ✓
B) Because of a boat.
C) Because of a fire.
D) Because of a leak.

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)



Two zebra are walking around their pen at the zoo.

Q1: Any people in the shot?

A1: No, there aren't any.

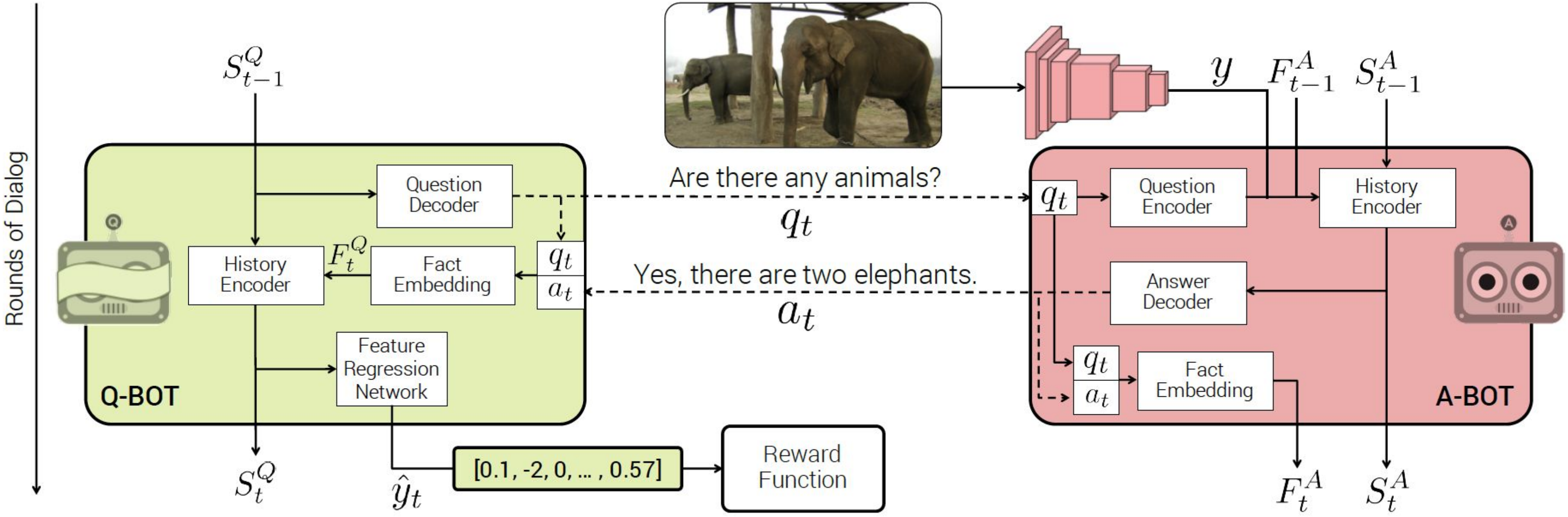
Q10: Are they facing each other?

A10: They aren't.

I think we were talking about this image!

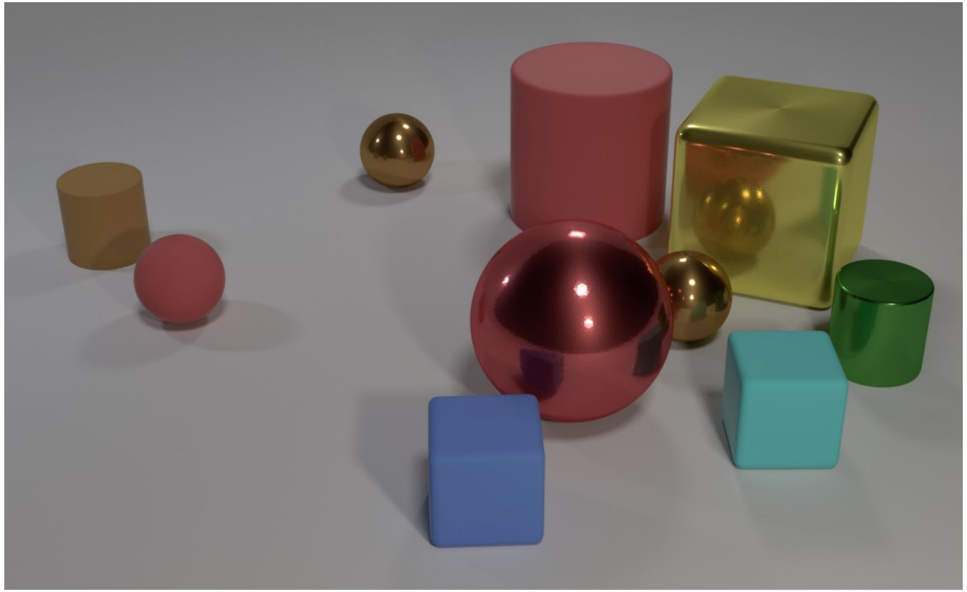
Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. "Visual Dialog." CVPR 2017

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

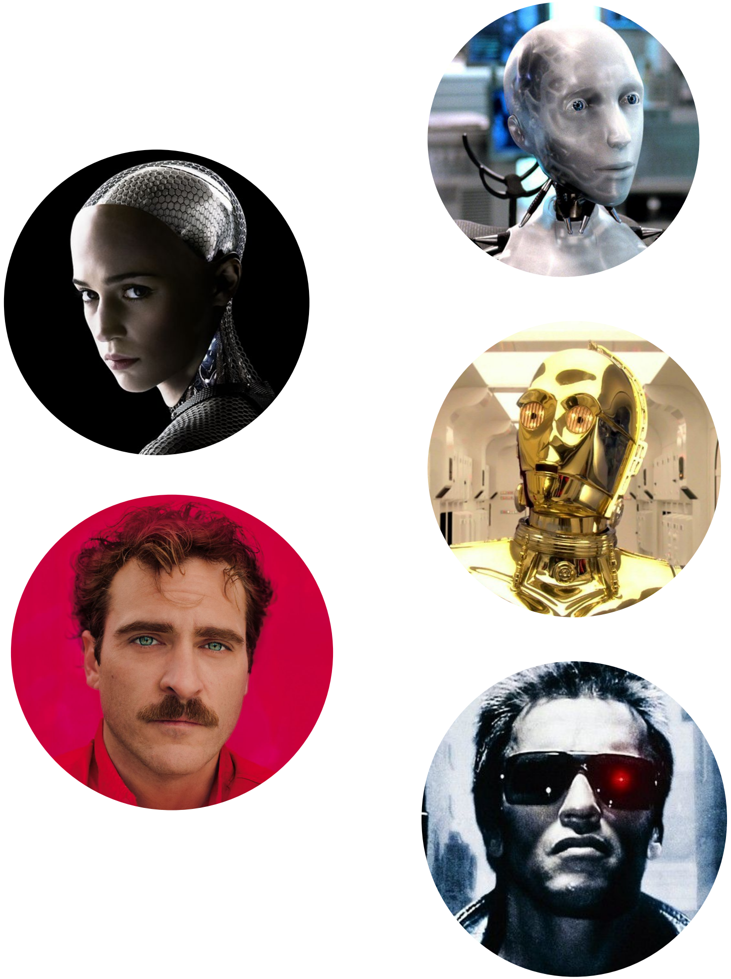
Visual Reasoning



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

Conclusions



New Turing test? How to evaluate AI's image understanding?

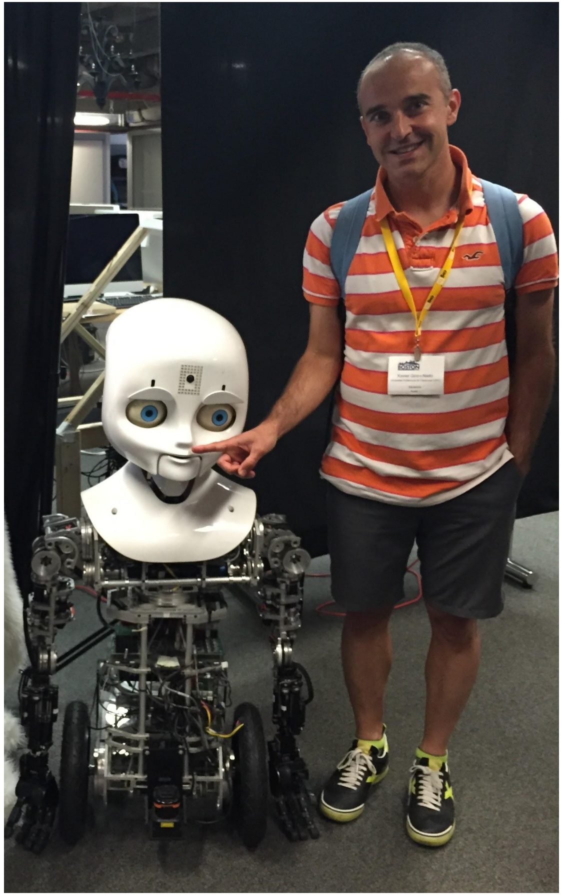
Slide credit: Issey Masuda

Learn more

Julia Hockenmeirer

Flickr8k
[Hodosh et al. 2013]
~8,000 Flickr images
>40,000 captions
5 captions/image

Thanks ! Q&A ?



Follow me at

[/ProfessorXavi]

[@DocXavi]

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Department of Signal Theory
and Communications

Image Processing Group

https://imatge.upc.edu/web/people/xavier-giro