DEEP LEARNING FOR COMPUTER VISION

Summer Seminar UPC TelecomBCN, 4 - 8 July 2016

Instructors

Xavier Giró-i-Nieto | Elisa Sayrol | Amaia Salvador | Jordi Torres | Eva Mohedano | Kevin McGuinness

Organizers

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

telecom BCN

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

DCU Dublin City University Ollscoil Chathair Bhaile Átha Cliath

Insight Centre for Data Analytics

NVIDIA GPU CENTER OF EXCELLENCE

Co-funded by the Erasmus+ Programme of the European Union

+ info: TelecomBCN.DeepLearning.Barcelona

Day 4 Lecture 3

Language and Vision

Xavier Giró-i-Nieto

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Department of Signal Theory and Communications

Image Processing Group

Acknowledgments

Santi Pascual

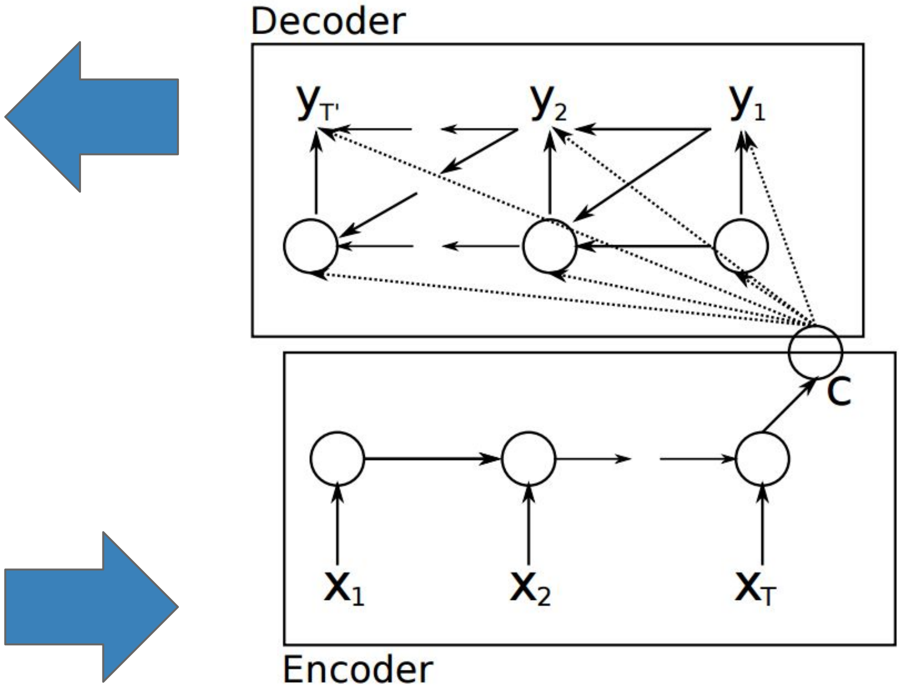
UPC TALP

TALP Research Center

In lecture D2L6 RNNs...

Language OUT

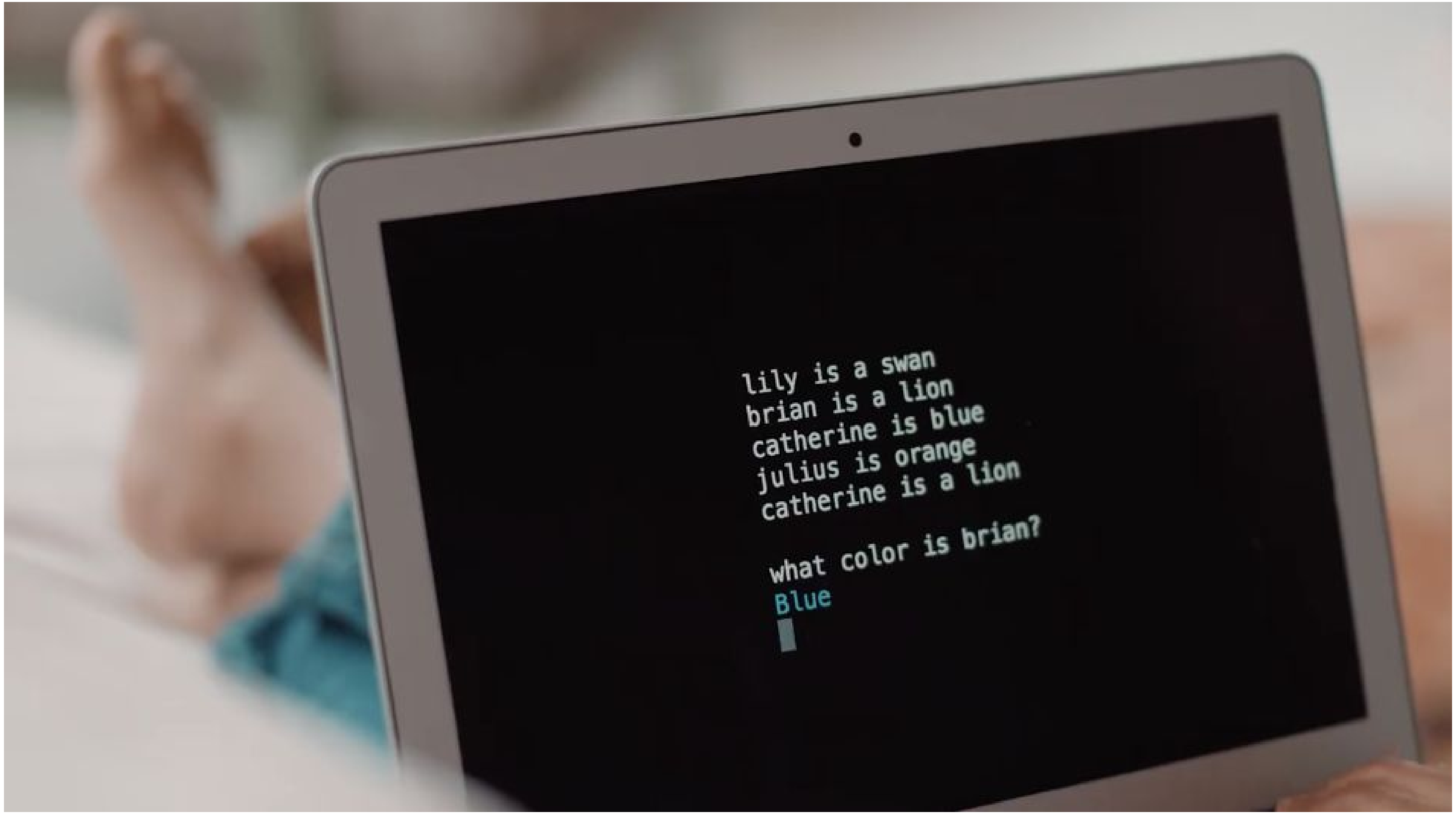
Language IN



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

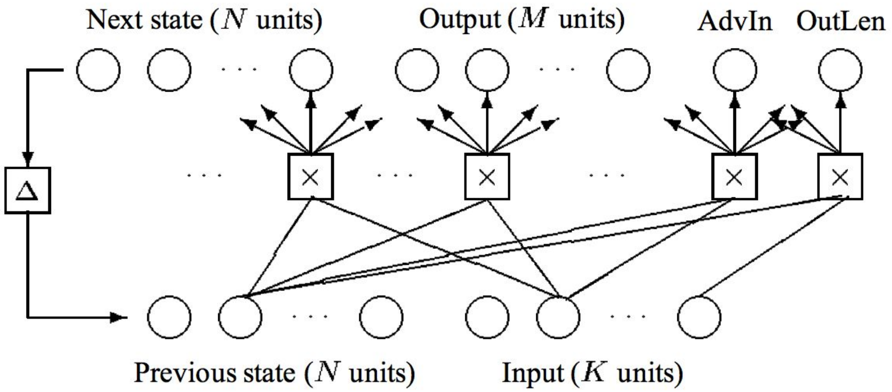
Facebook AI Research



Much earlier than lecture D2L6 RNNs...

Asynchronous translations with recurrent neural nets

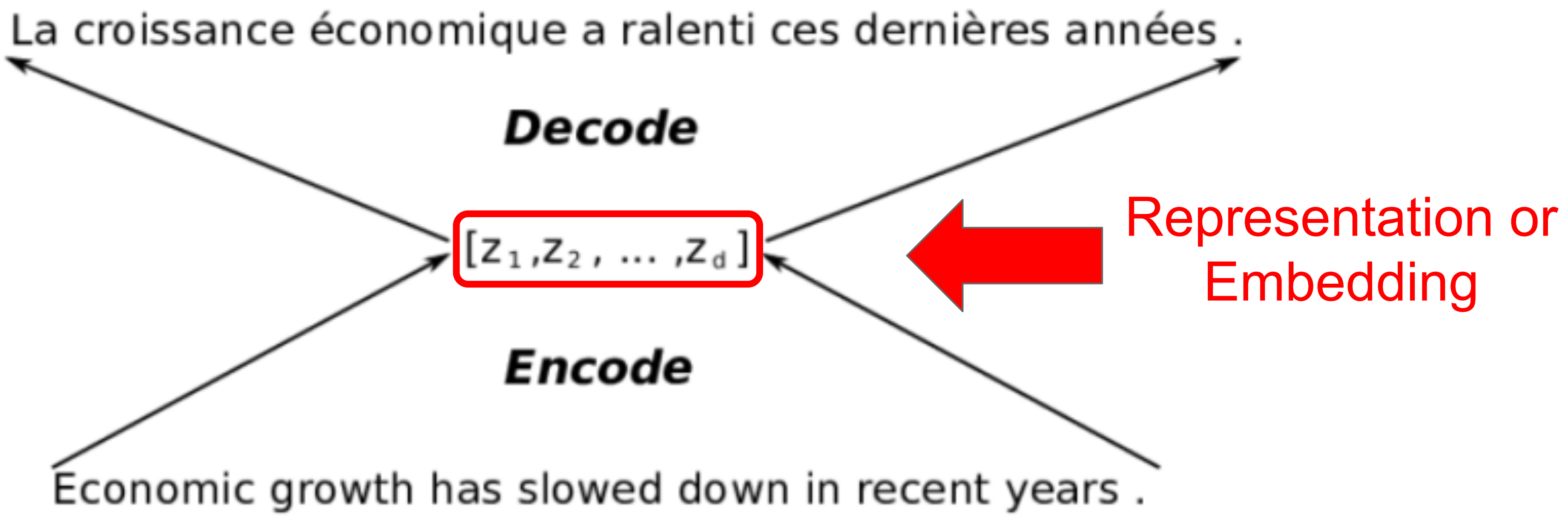
Ramón P. Ñeco, Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics,
Universitat d'Alacant, E-03071 Alacant, Spain.
E-mail: {neco,mlf}@dlsi.ua.es



Neco, R.P. and Forcada, M.L., 1997, June. Asynchronous translations with recurrent neural nets. In Neural Networks, 1997., International Conference on (Vol. 4, pp. 2535-2540). IEEE.

Encoder-Decoder

For clarity, let's study a Neural Machine Translation (NMT) case:



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

Encoder: One-hot encoding

One-hot encoding: Binary representation of the words in a vocabulary, where the only combinations with a single hot (1) bit and all other cold (0) bits are allowed.

Word	Binary	One-hot encoding
zero	00	0000
one	01	0010
two	10	0100
three	11	1000

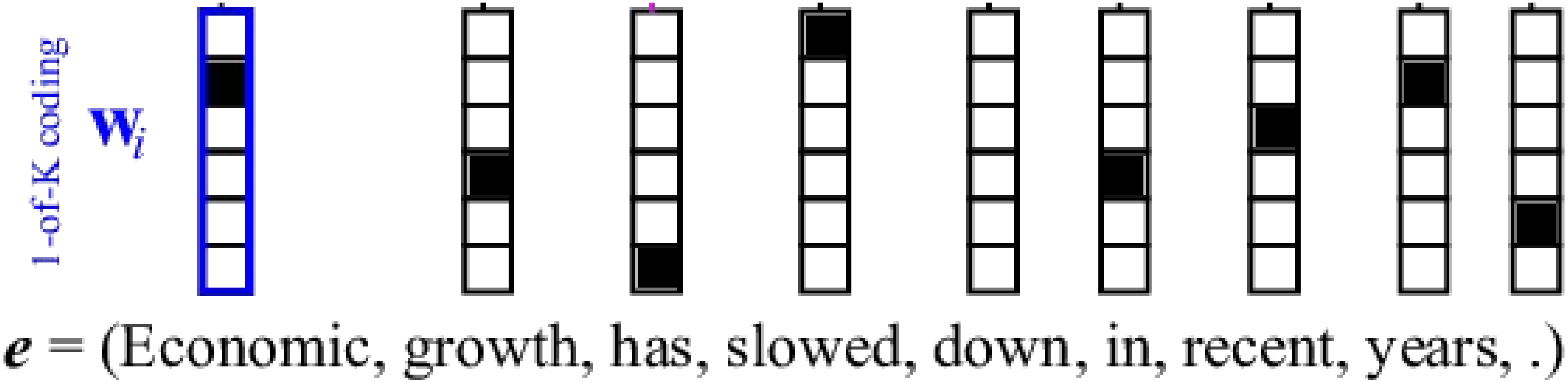
Encoder: One-hot encoding

Natural language words can also be one-hot encoded on a vector of dimensionality equal to the size of the dictionary (K).

Word	One-hot encoding
economic	000010...
growth	001000...
has	100000...
slowed	000001...

Encoder: One-hot encoding

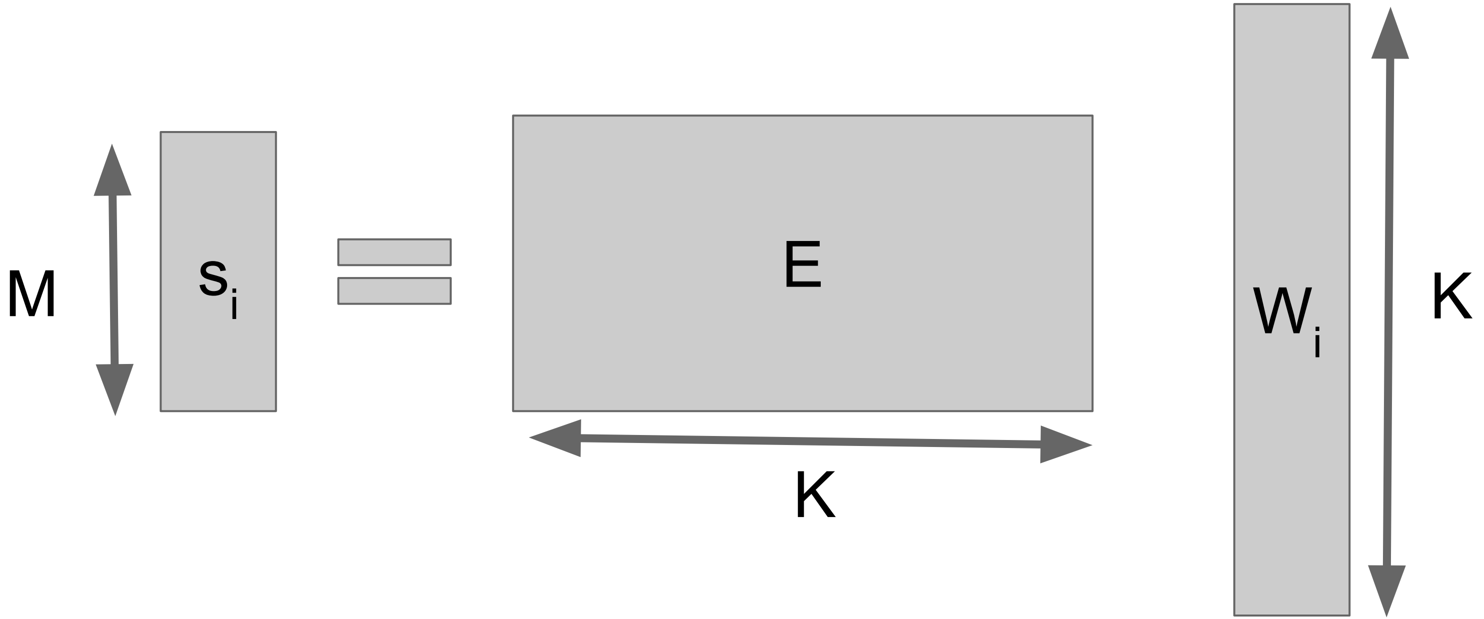
One-hot is a very simple representation: every word is equidistant from every other word.



Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”](#) (2015)

Encoder: Projection to continious space

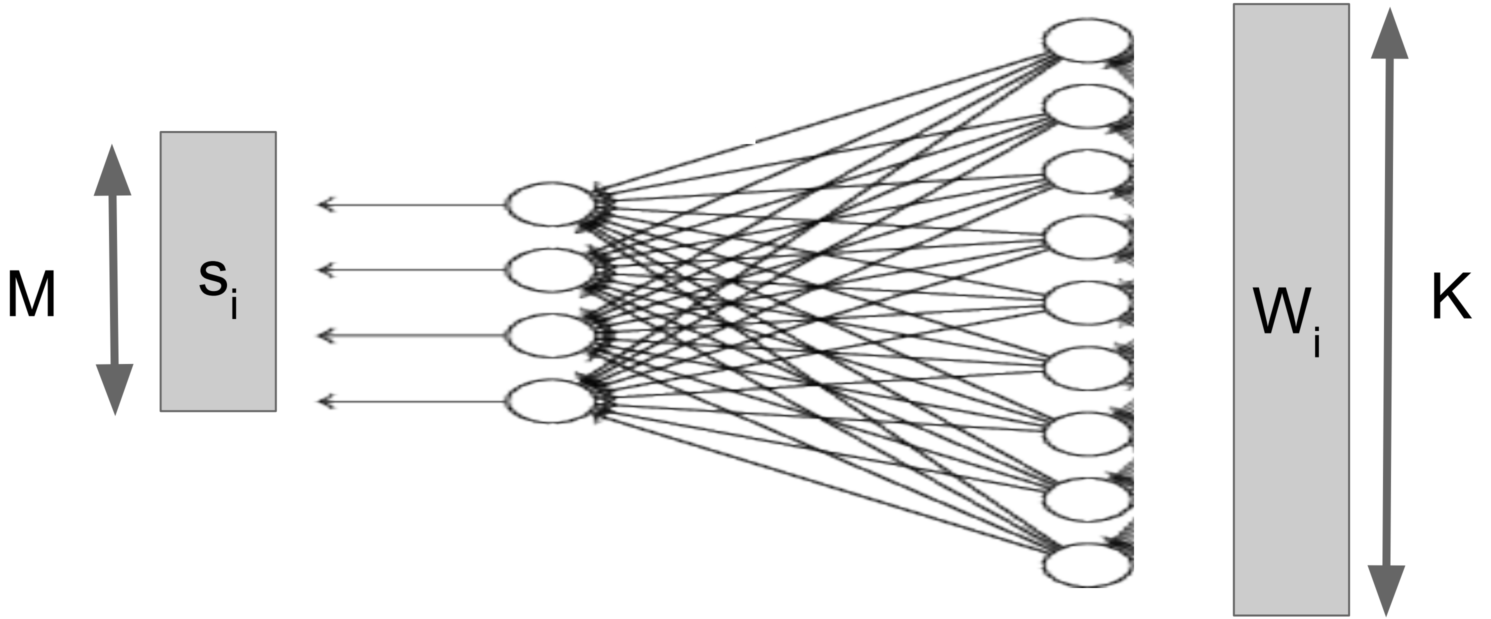
The one-hot is linearly projected to a space of lower dimension (typically 100-500) with matrix E for learned weights.



Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”]() (2015)

Encoder: Projection to continious space

Projection matrix E corresponds to a fully connected layer, so its parameters will be learned with a training process.

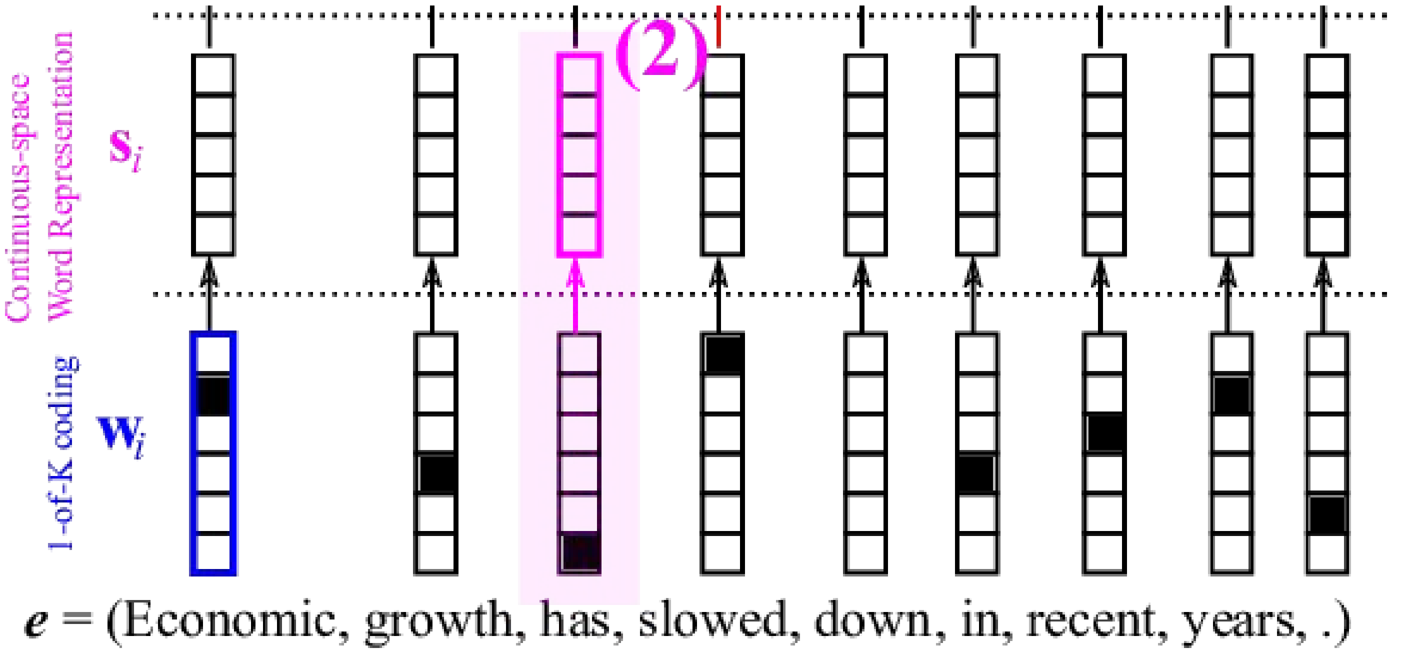


Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”]() (2015)

Encoder: Projection to continious space

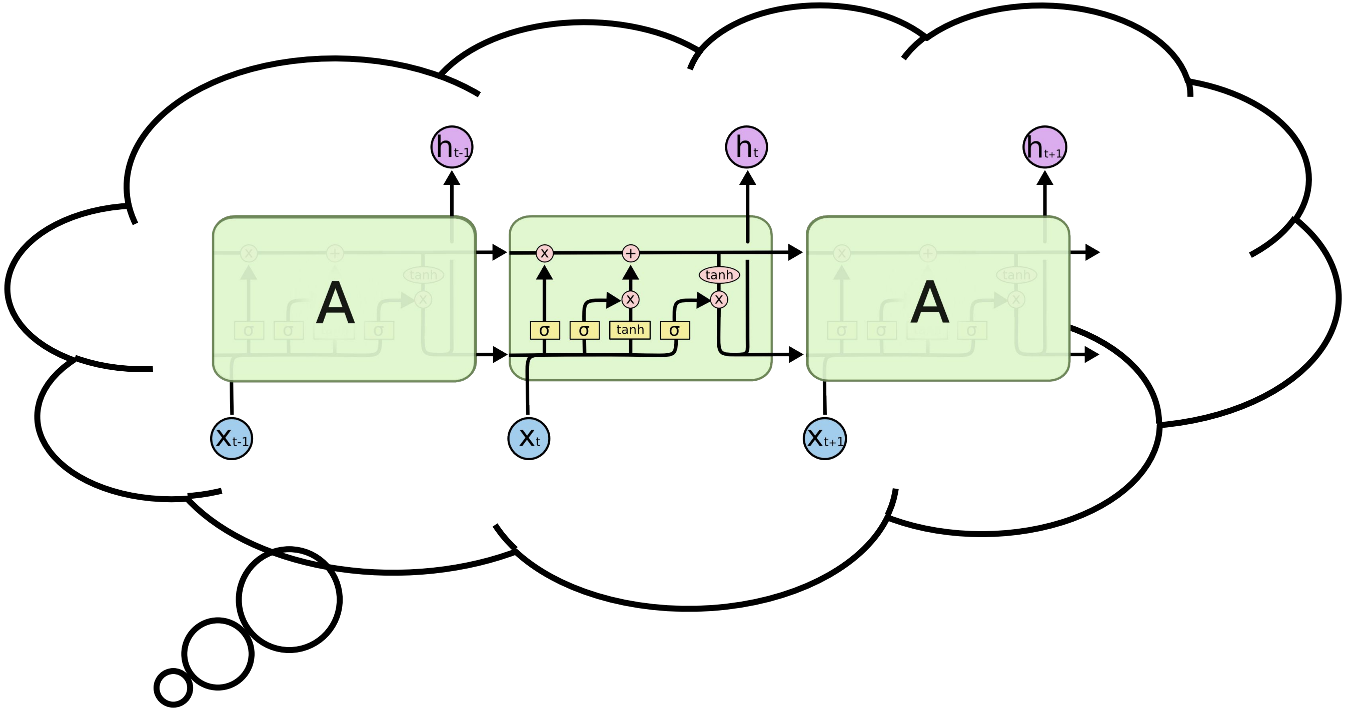
Sequence of continious-space word representations

Sequence of words



Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”]() (2015)

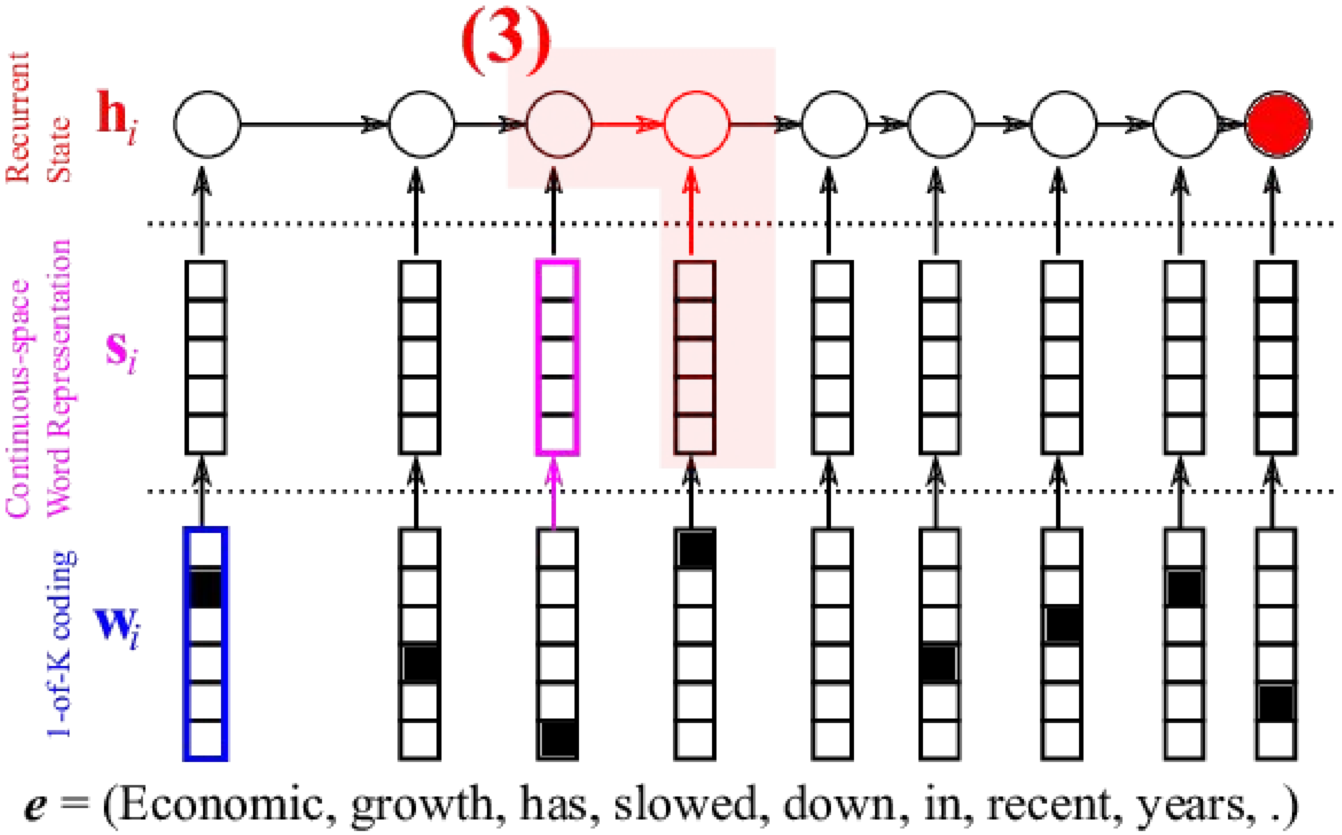
Encoder: Recurrence



Sequence

Figure: Cristopher Olah, “Understanding LSTM Networks” (2015)

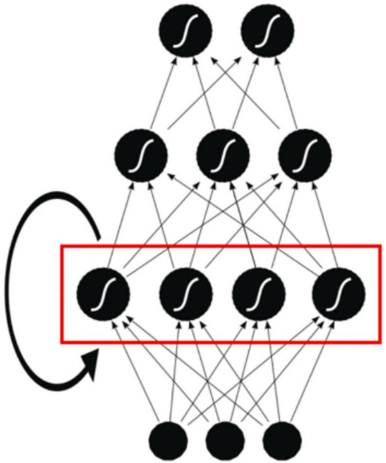
Encoder: Recurrence



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

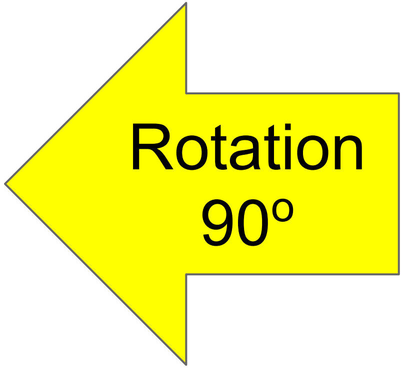
Encoder: Recurrence

Front View

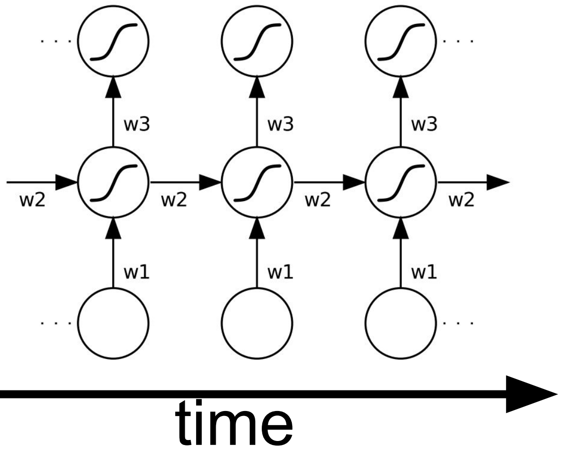


time

⊗



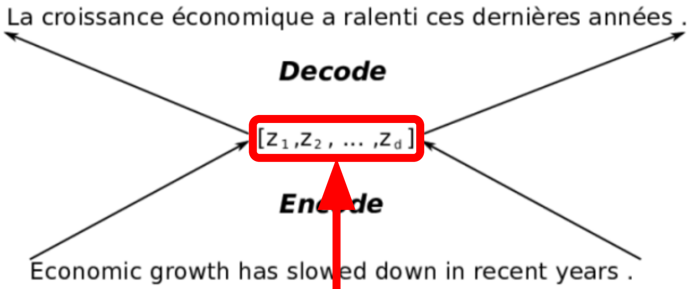
Side View

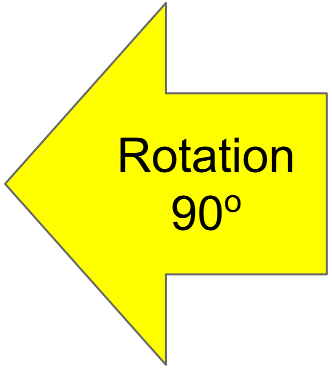


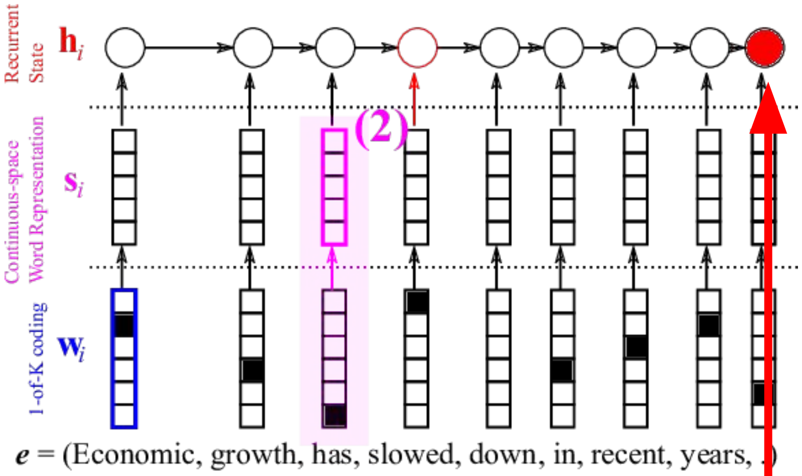
Encoder: Recurrence

Front View

Side View



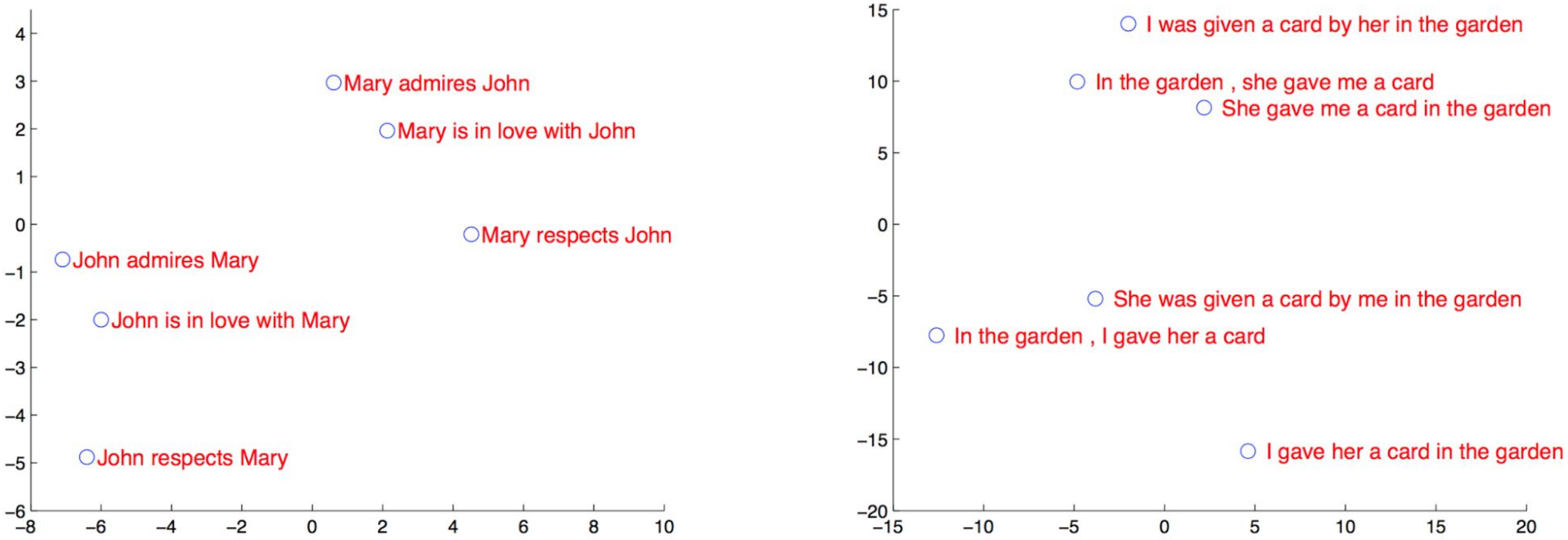




Representation or embedding of the sentence

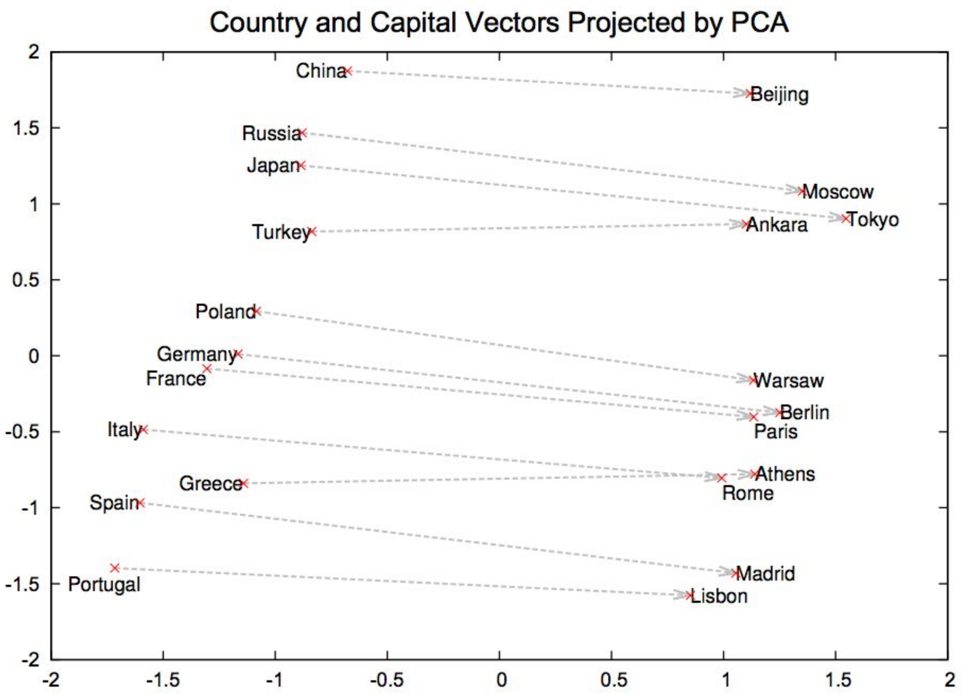
Sentence Embedding

Clusters by meaning appear on 2-dimensional PCA of LSTM hidden states



Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." NIPS 2014

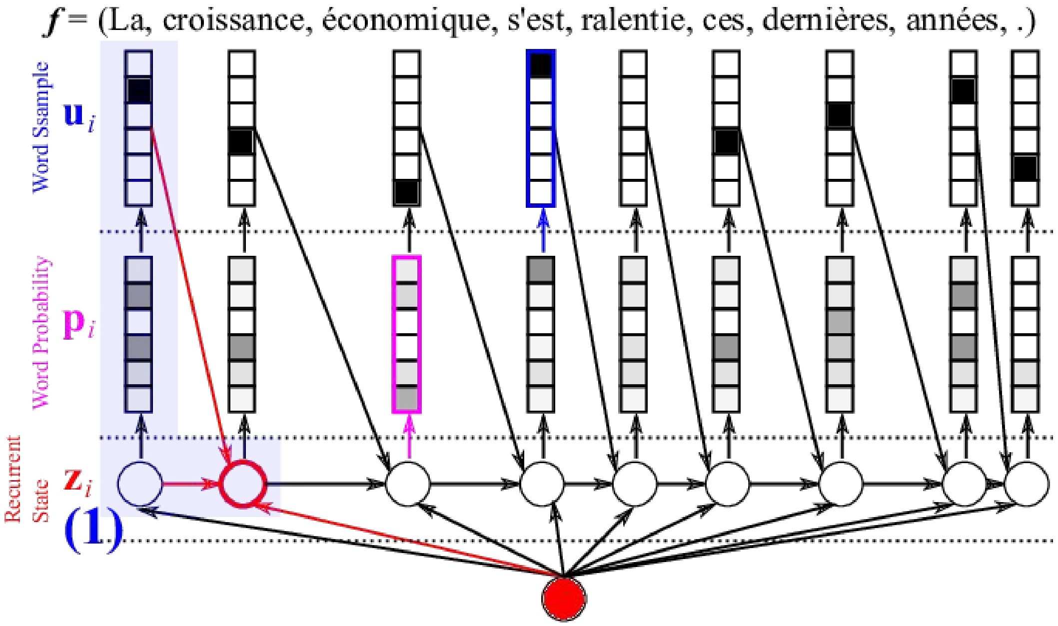
(Word Embeddings)



Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. ["Distributed representations of words and phrases and their compositionality."]() In *Advances in neural information processing systems*, pp. 3111-3119. 2013.

Decoder

RNN's internal state z_i depends on: sentence embedding h_t, previous word u_{i-1} and previous internal state z_{i-1}.



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

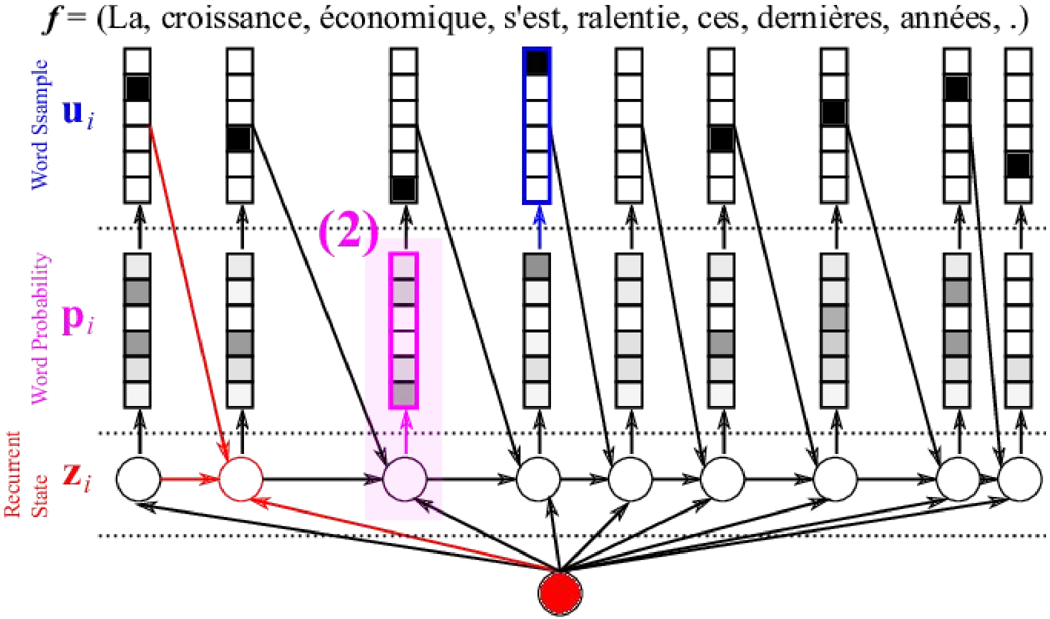
Decoder

With z_i ready, we can score each word k in the vocabulary with a dot product...

$$e(k) = w_k^\top z_i + b_k,$$

Neuron weights for word k

RNN internal state



Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”] (2015)

Decoder

...and finally normalize to word probabilities with a softmax.

Score for word k

$$e(k) = w_k^\top z_i + b_k,$$

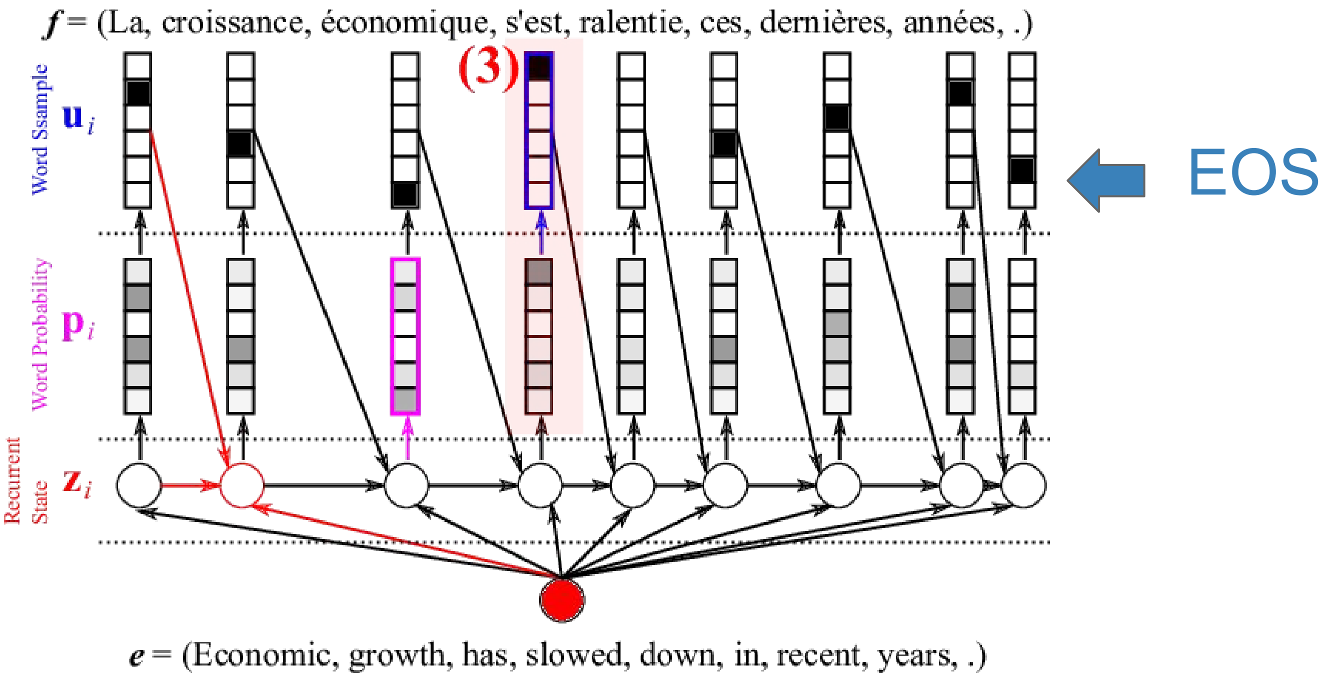
Probability that the ith word is word k

$$p(w_i = k | \underbrace{w_1, w_2, \ldots, w_{i-1}}_{\text{Previous words}}, \underbrace{h_T}_{\text{Hidden state}}) = \frac{\exp(e(k))}{\sum_j \exp(e(j))}.$$

Bridle, John S. "Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters." NIPS 1989

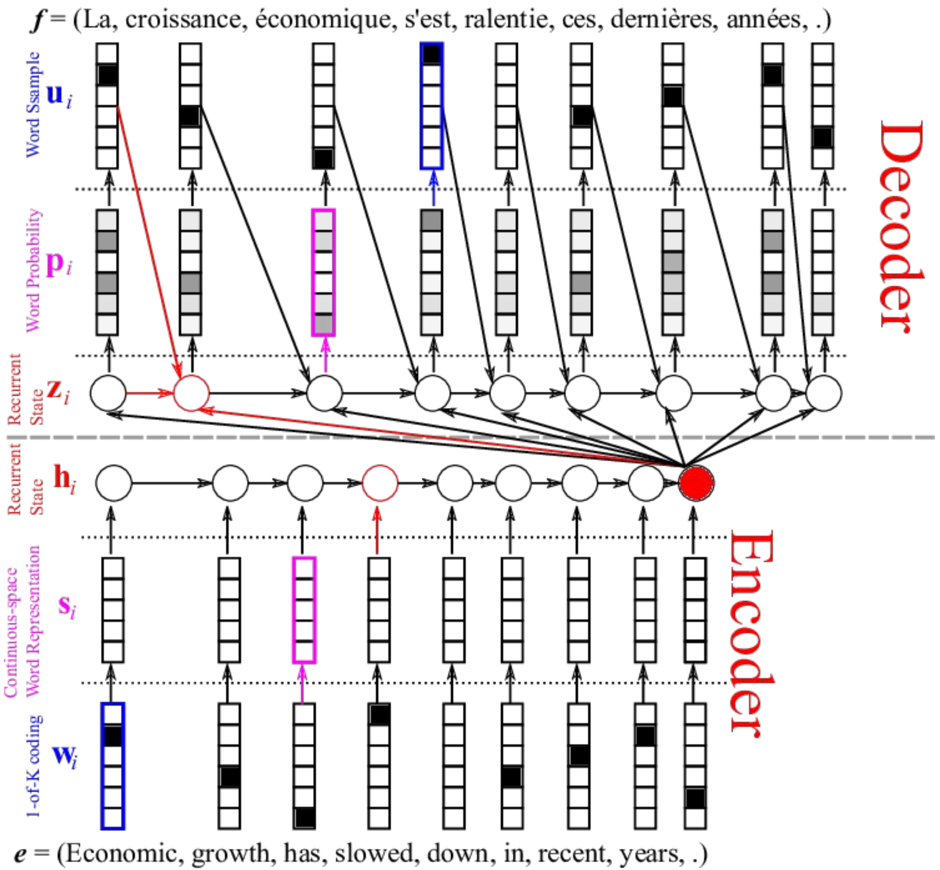
Decoder

More words for the decoded sentence are generated until a <EOS> (End Of Sentence) “word” is predicted.



Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”]() (2015)

Encoder-Decoder



Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”] (2015)

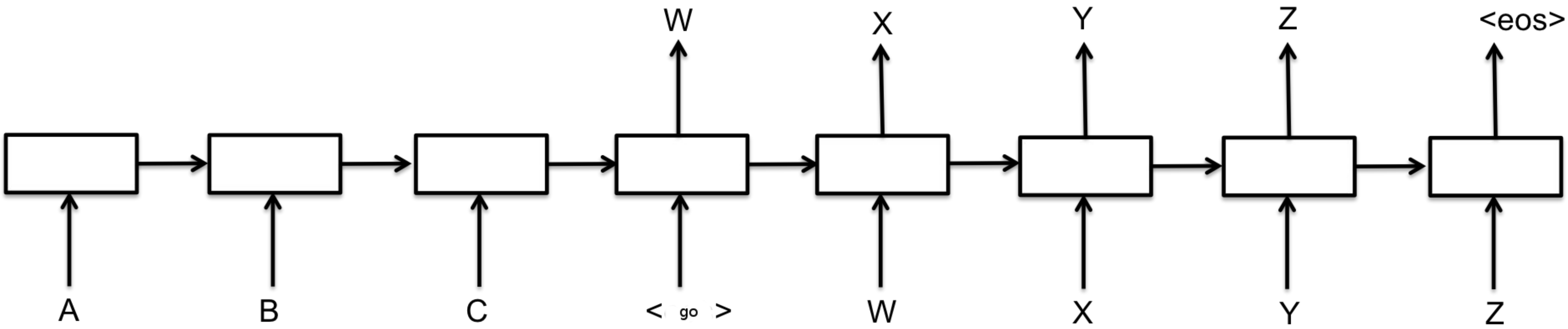
Encoder-Decoder: Training

Dataset of pairs of sentences in the two languages to translate.

Source	Translation Model
at the end of the	[a la fin de la] [ŕ la fin des années] [être supprimés à la fin de la]
for the first time	[r © pour la premirëre fois] [été donnés pour la première fois] [été commémorée pour la première fois]
in the United States and	[? aux ?tats-Unis et] [été ouvertes aux États-Unis et] [été constatées aux États-Unis et]
, as well as	[?s , qu'] [?s , ainsi que] [?re aussi bien que]
one of the most	[?t ?l' un des plus] [?l' un des plus] [être retenue comme un de ses plus]

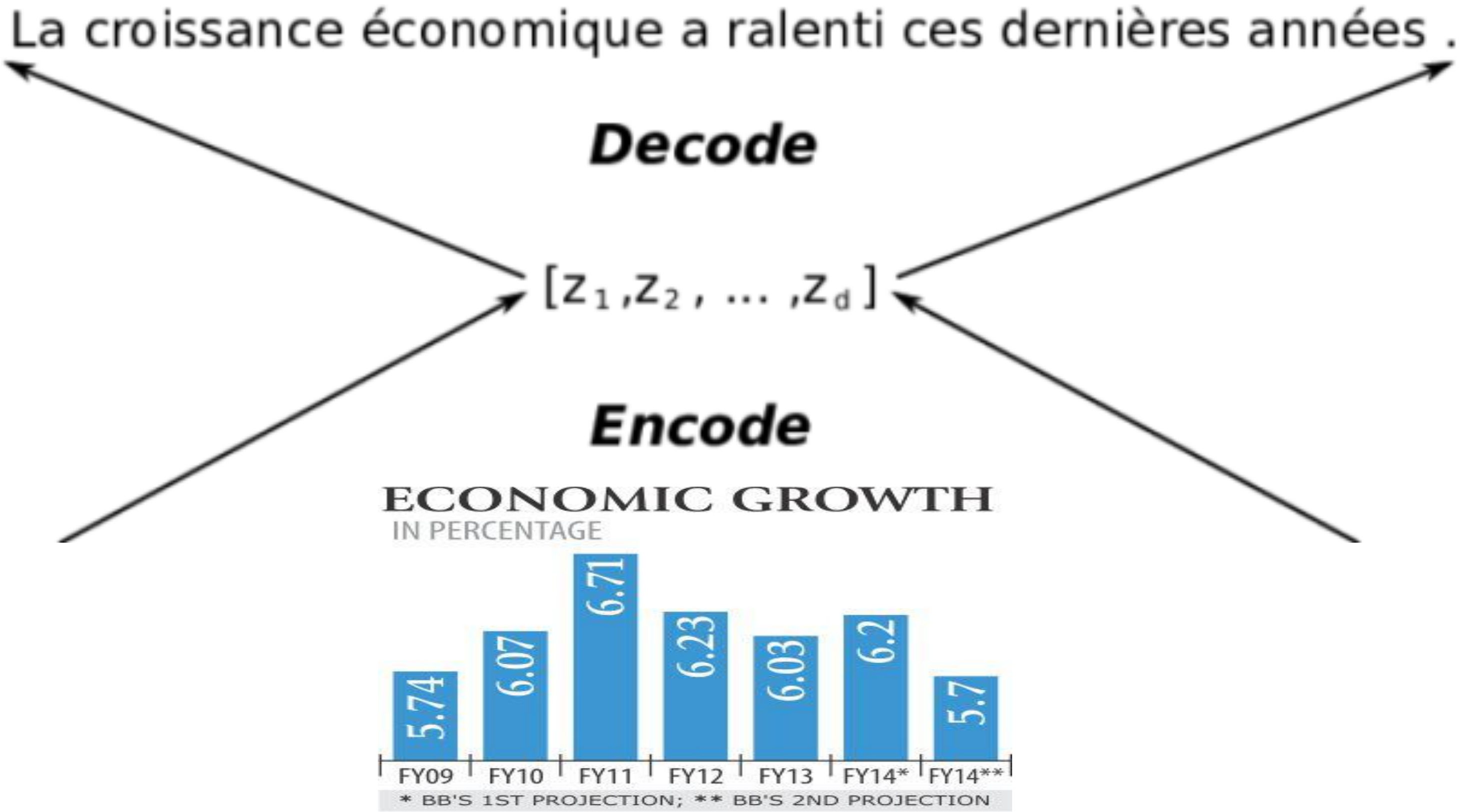
Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. ["Learning phrase representations using RNN encoder-decoder for statistical machine translation."]() AMNLP 2014.

Encoder-Decoder: Seq2Seq



Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." NIPS 2014.

Encoder-Decoder: Beyond text



Captioning: DeepImageSent

man in black shirt is playing guitar.

construction worker in orange safety vest is working on road.

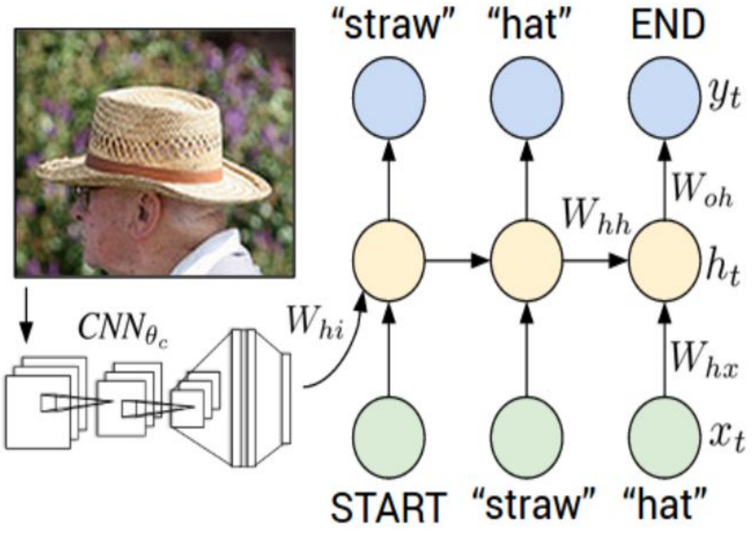
two young girls are playing with lego toy.

(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

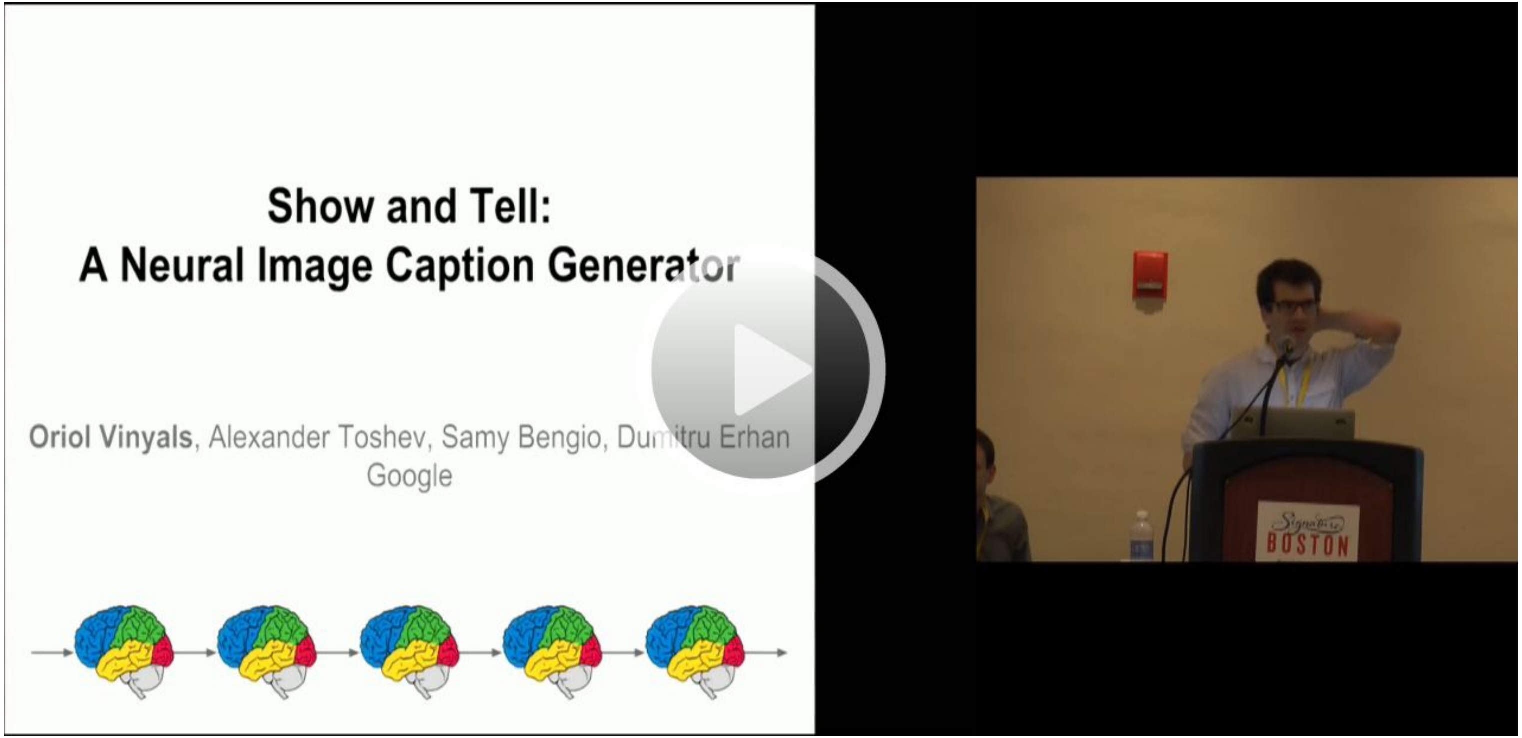
$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \boxed{\mathbb{1}(t=1) \odot b_v})$$
$$y_t = softmax(W_{oh}h_t + b_o).$$



Multimodal Recurrent Neural Network

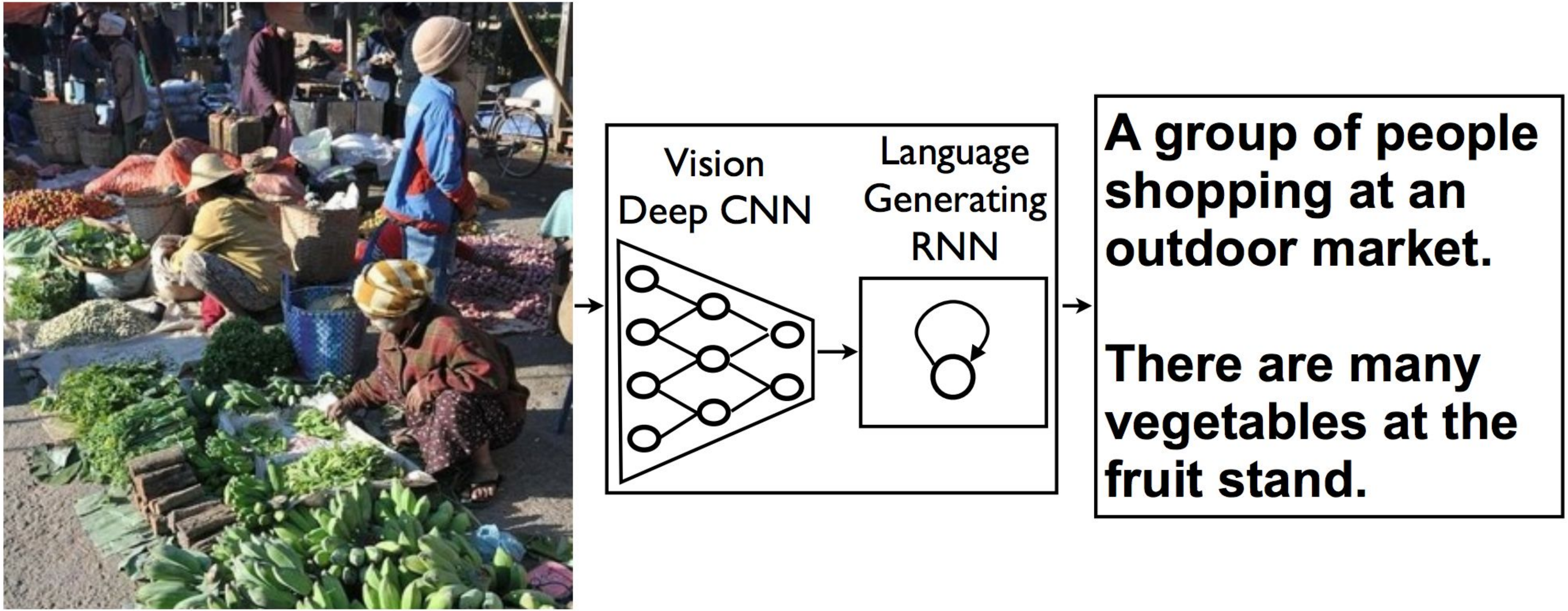
(Slides by Marc Bolaños): Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." CVPR 2015

Captioning: Show & Tell



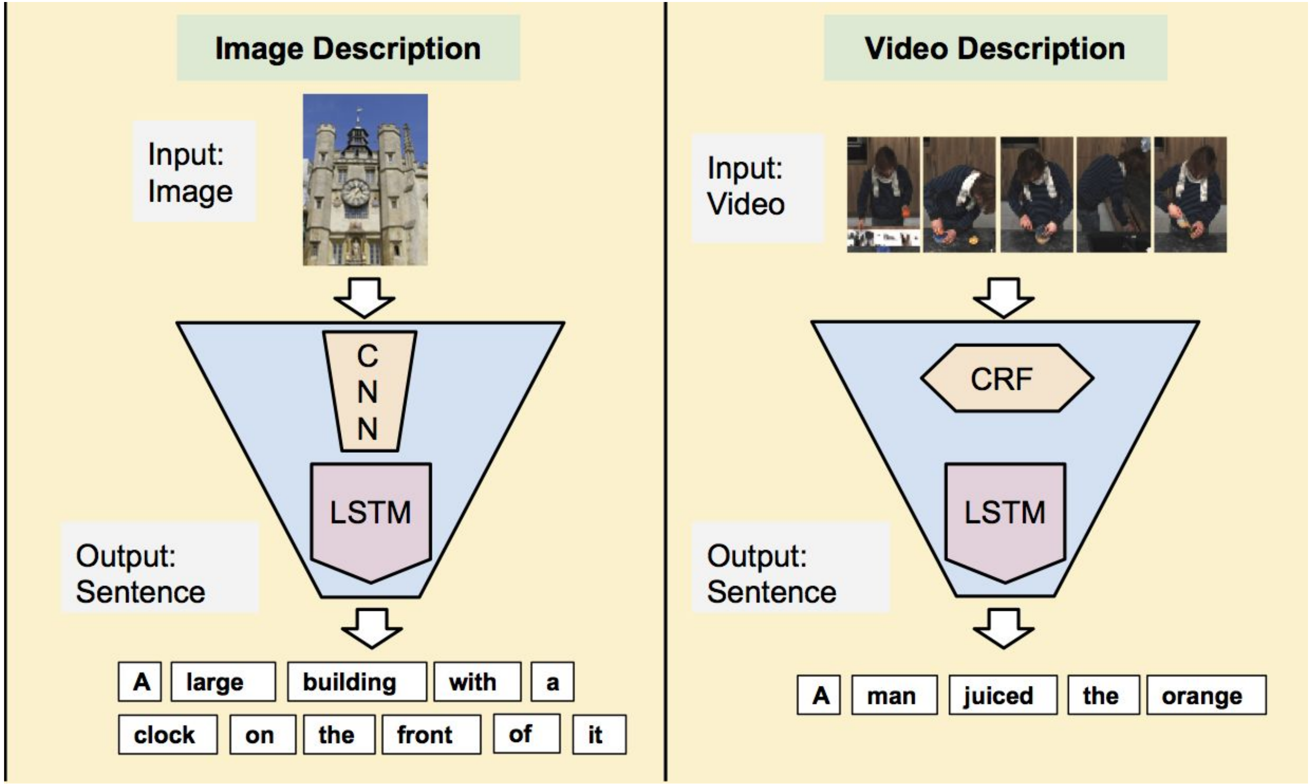
Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." CVPR 2015.

Captioning: Show & Tell



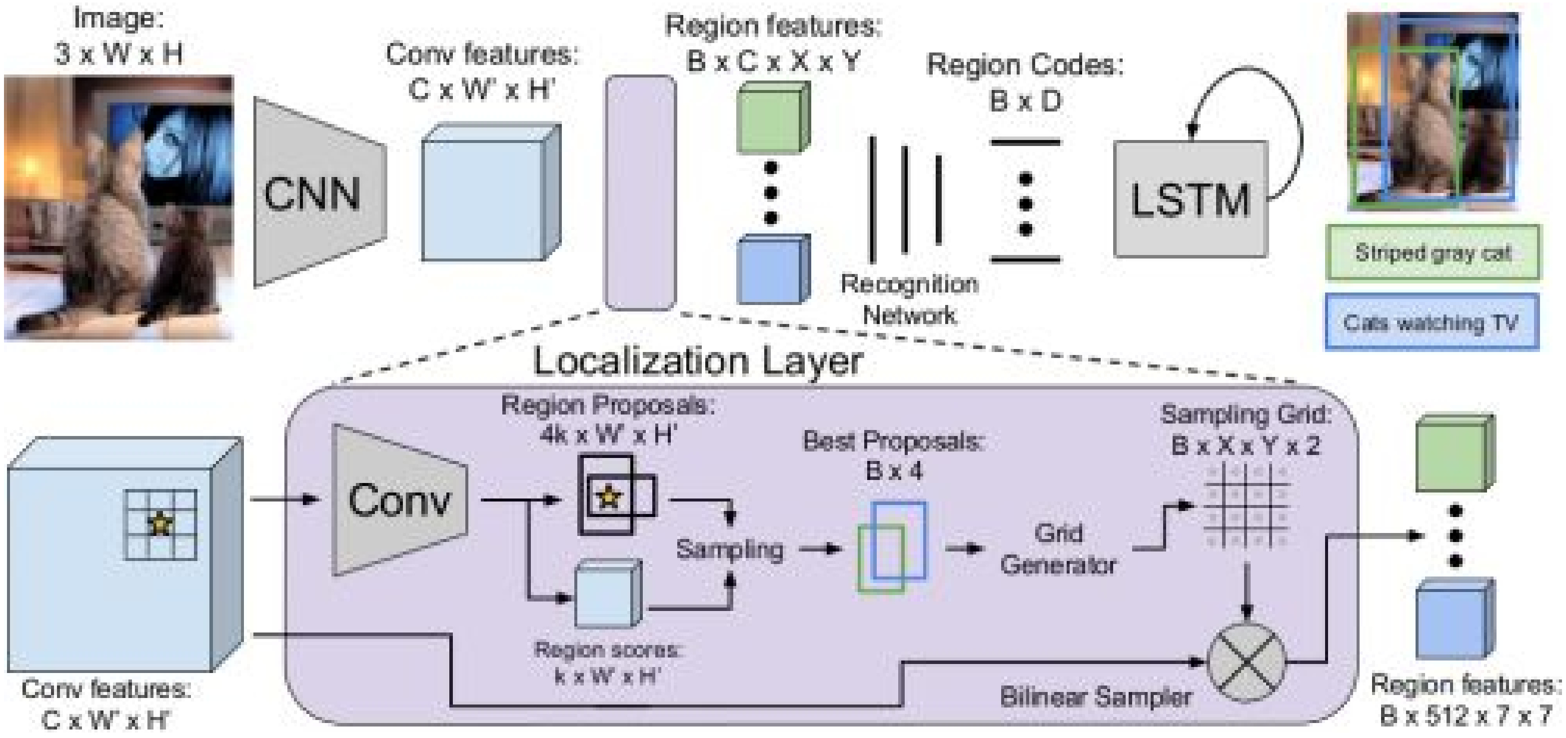
Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." CVPR 2015.

Captioning: LSTM for image & video



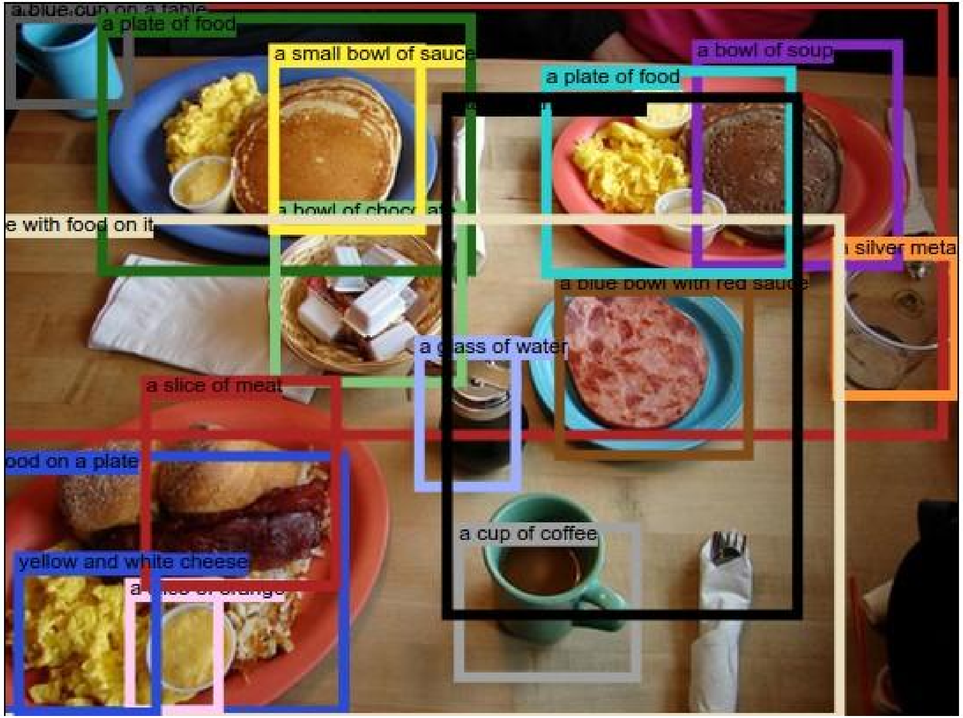
Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrel. Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015. code

Captioning (+ Detection): DenseCap



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

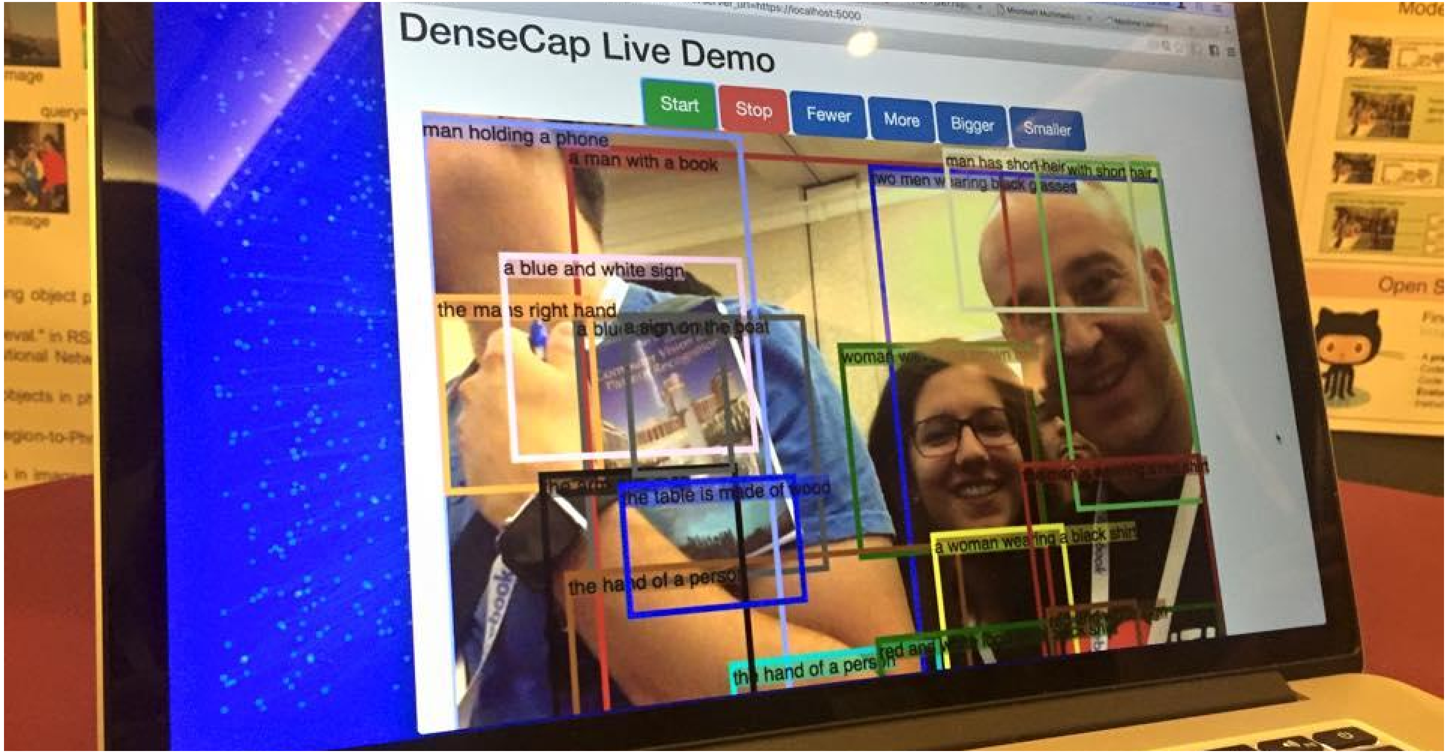
Captioning (+ Detection): DenseCap



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



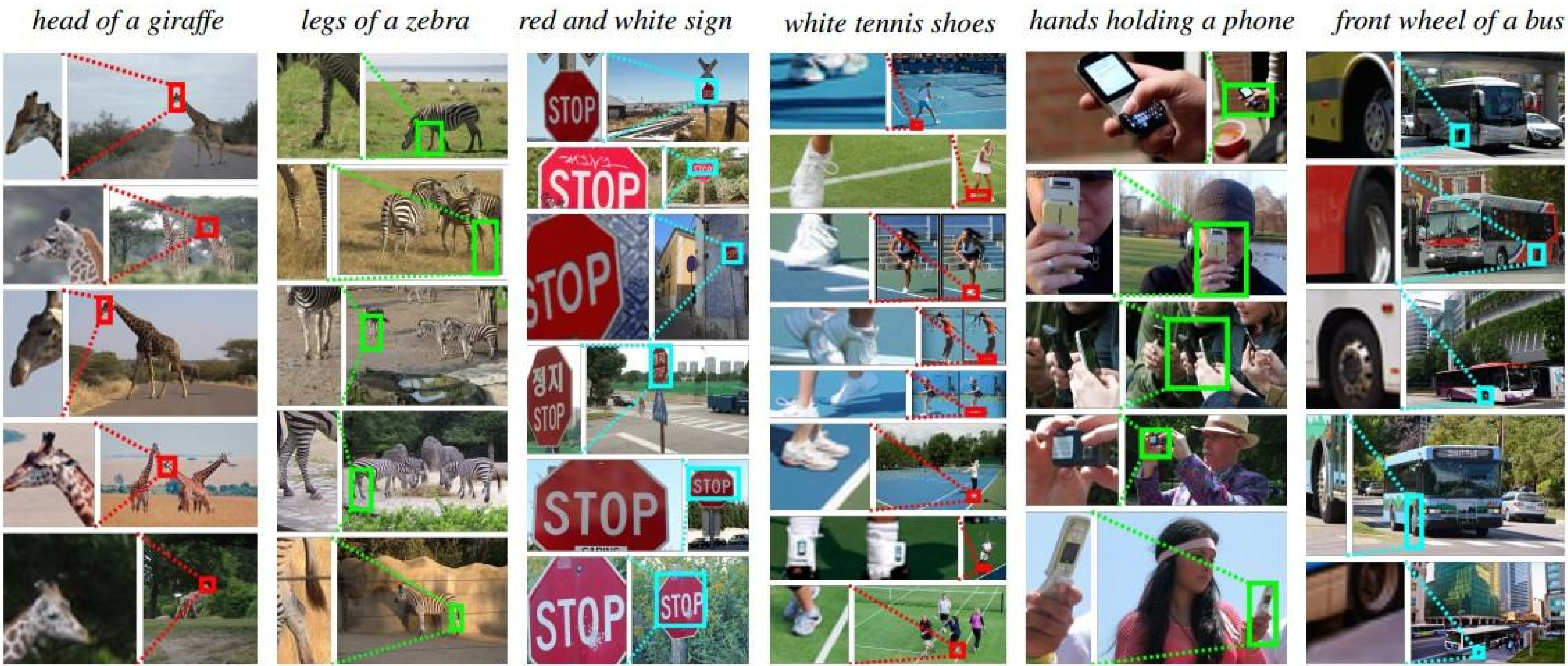
XAVI: “man has short hair”, “man with short hair”

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

BOTH: “two men wearing black glasses”

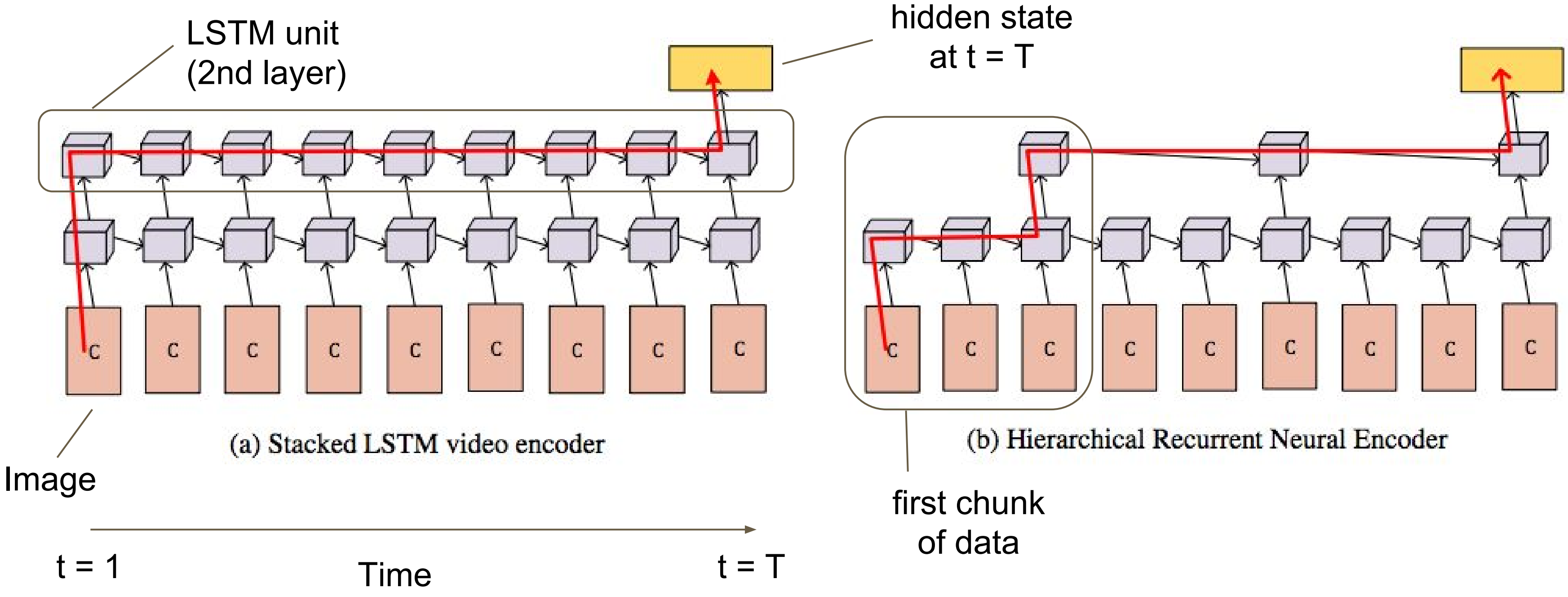
Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. ["Densecap: Fully convolutional localization networks for dense captioning."]() *CVPR 2016*

Captioning (+ Retrieval): DenseCap



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

Captioning: HRNE



(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

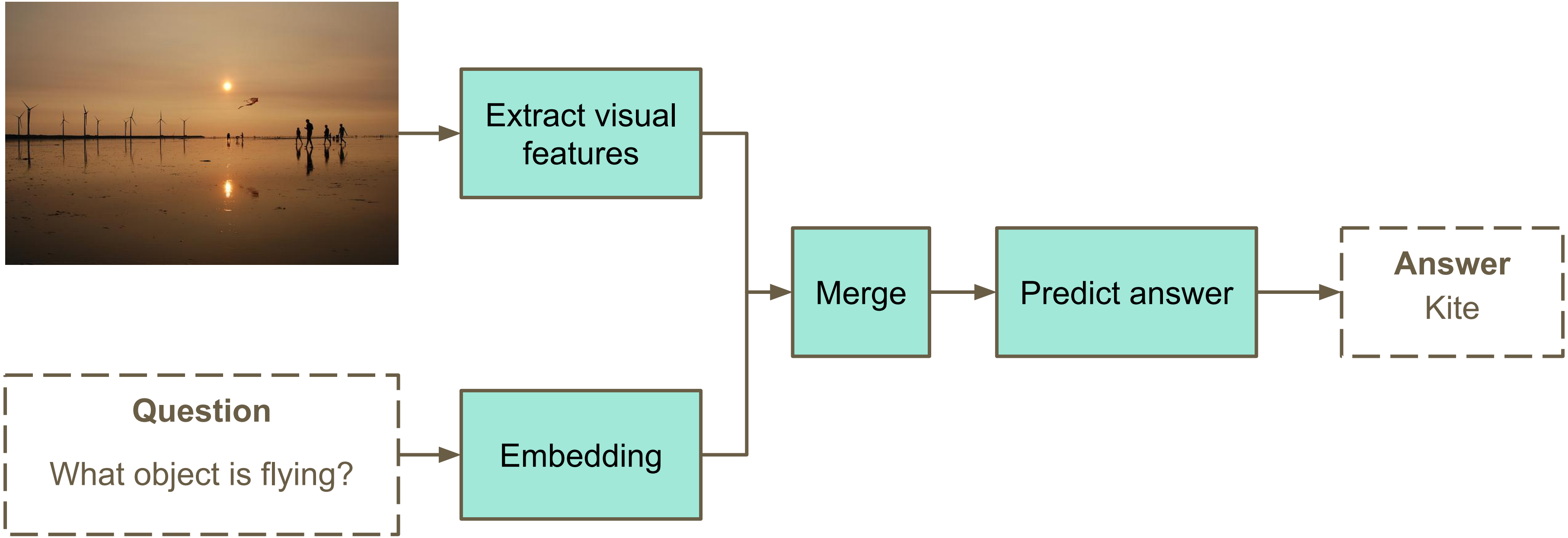
$[z_1, z_2, \ldots z_N] \rightarrow \oplus \leftarrow [y_1, y_2, \ldots y_M]$

Encode

Encode

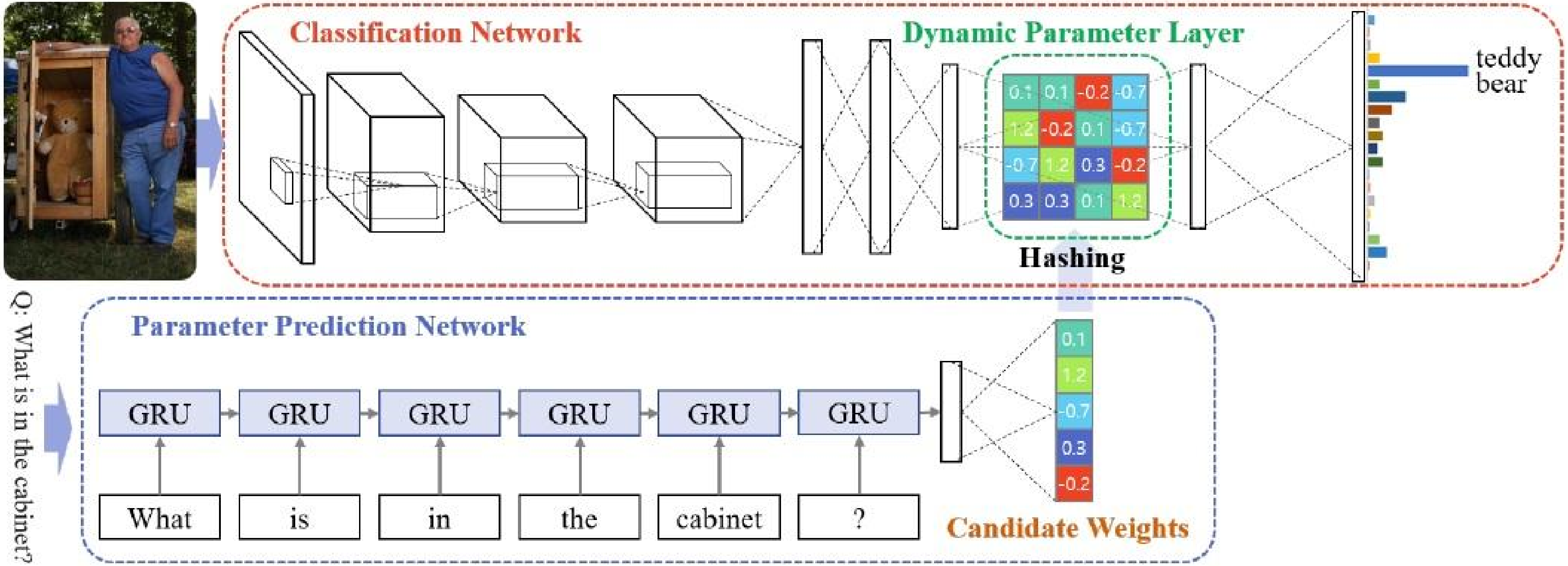
"Is economic growth decreasing ?"

Visual Question Answering



Slide credit: Issey Masuda

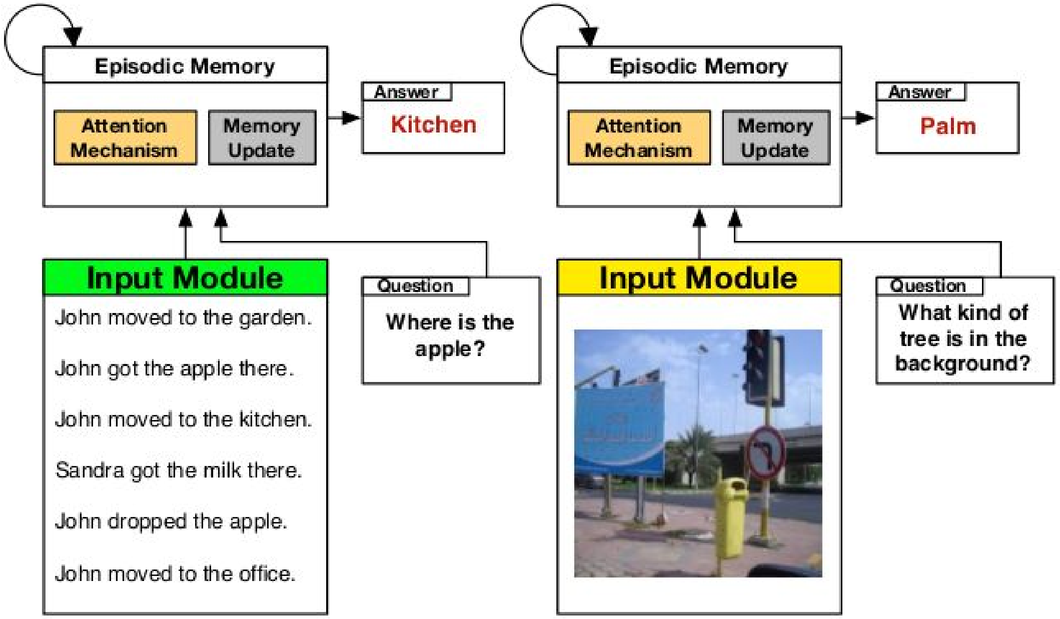
Visual Question Answering



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.

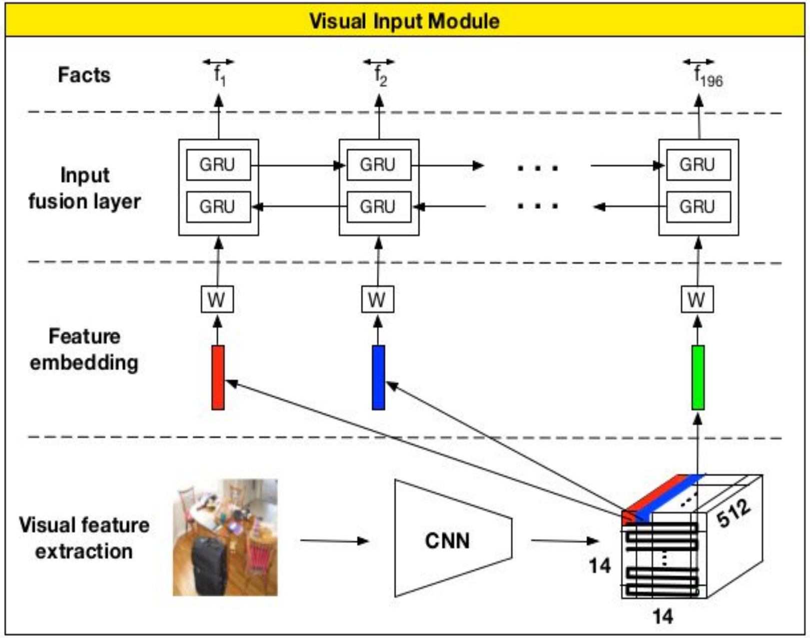
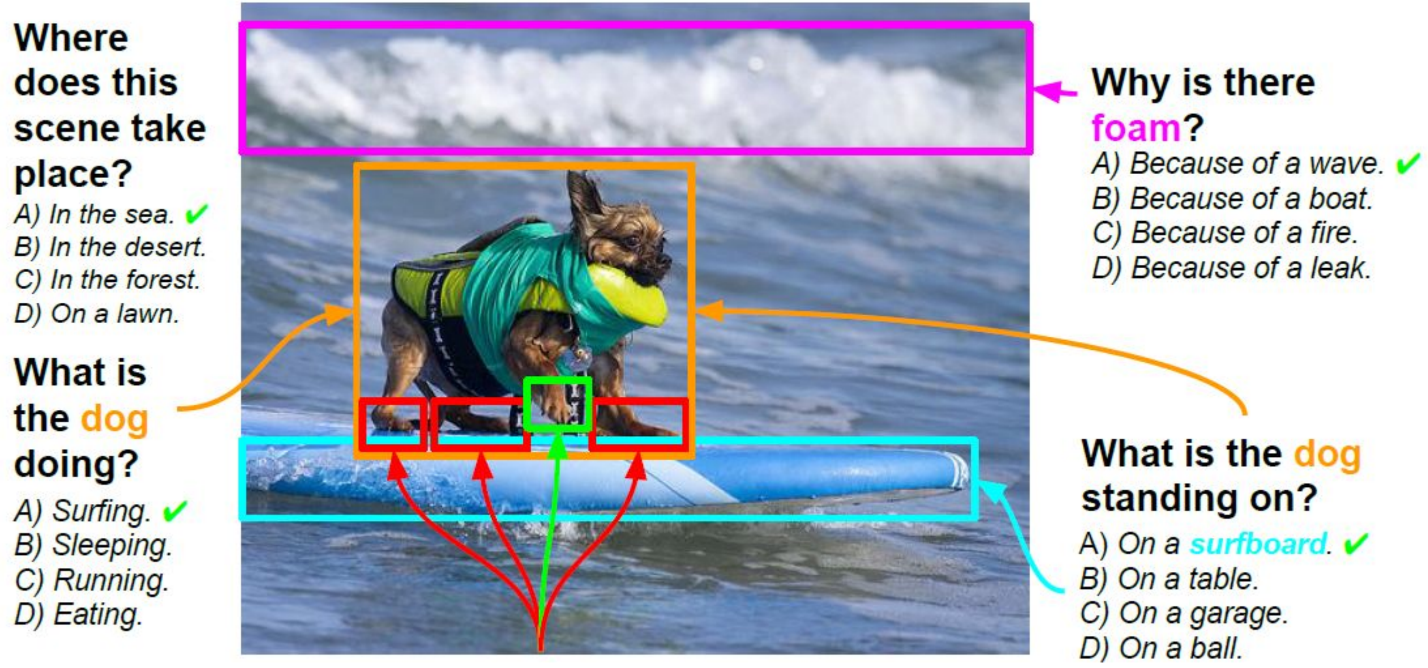


Figure 3. VQA input module to represent images for the DMN.

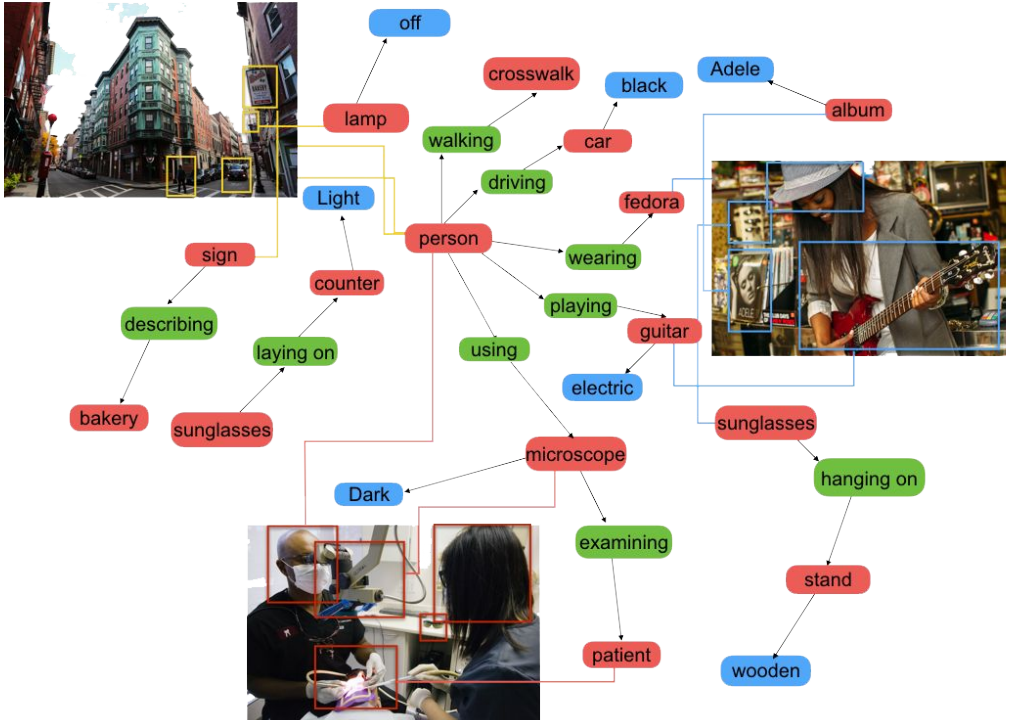
(Slides and Slidecast by Santi Pascual): Xiong, Caiming, Stephen Merity, and Richard Socher. "Dynamic Memory Networks for Visual and Textual Question Answering." ICML 2016.

Visual Question Answering: Grounded



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

Datasets: Visual Genome



Krishna, Ranjay, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." *arXiv preprint arXiv:1602.07332* (2016).

Datasets: Microsoft SIND

Example Generated Story

1

The dog was ready to go.

2

He had a great time on the hike.

3

And was very happy to be in the field.

4

His mom was so proud of him.

5

It was a beautiful day for him.

[Microsoft SIND]

Challenge: Microsoft Coco

COCO
Common Objects in Context

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Captioning

Challenge: Storytelling

Storytelling

Challenge: Movie Description

LSMDC 2016

[Movie Description, Retrieval and Fill-in-the-blank]

Challenges: Movie Question Answering

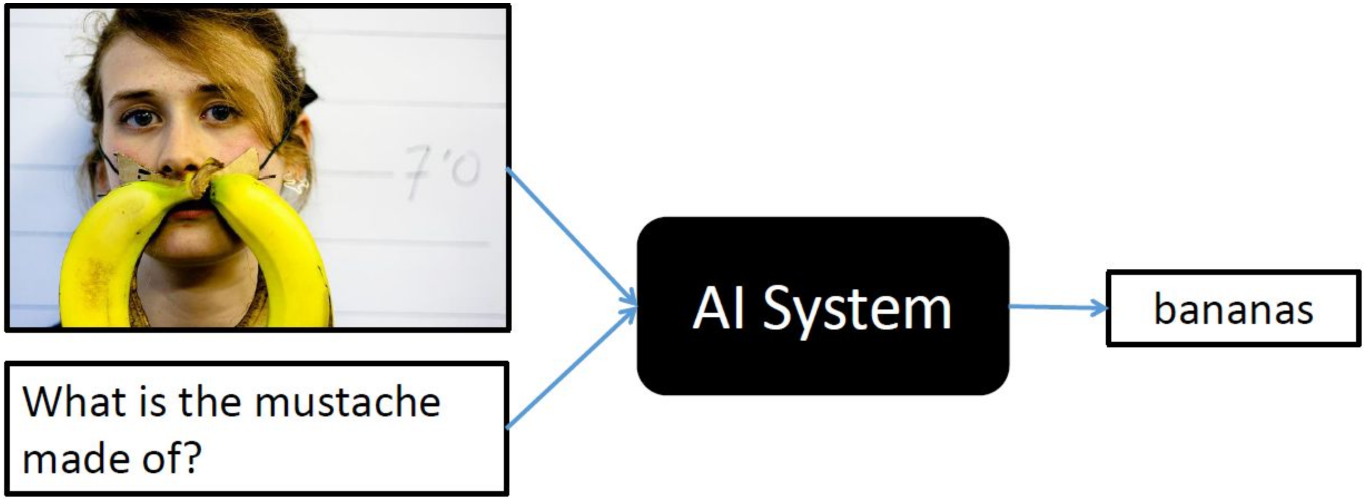
The Lord of the Rings: The Return of the King
Who sees Denethor trying to kill himself and Faramir on a bonfire?
– Gandalf! – Gandalf! – Denethor has lost his mind! – He's burning Faramir alive!
Pippin
Aragorn
Gandalf
Eowyn
Sam

Movie	E.T. the Extra-Terrestrial
Question	Do aliens leave one of their own on Earth on purpose?
Story	
Correct answer	No, they leave it accidentally
Wrong answer 1	Yes, they leave it on purpose
Wrong answer 2	No, it falls off the spaceship
Wrong answer 3	Yes, they leave it as a spy
Wrong answer 4	They don't leave any of their kind on Earth

[Movie Question Answering]

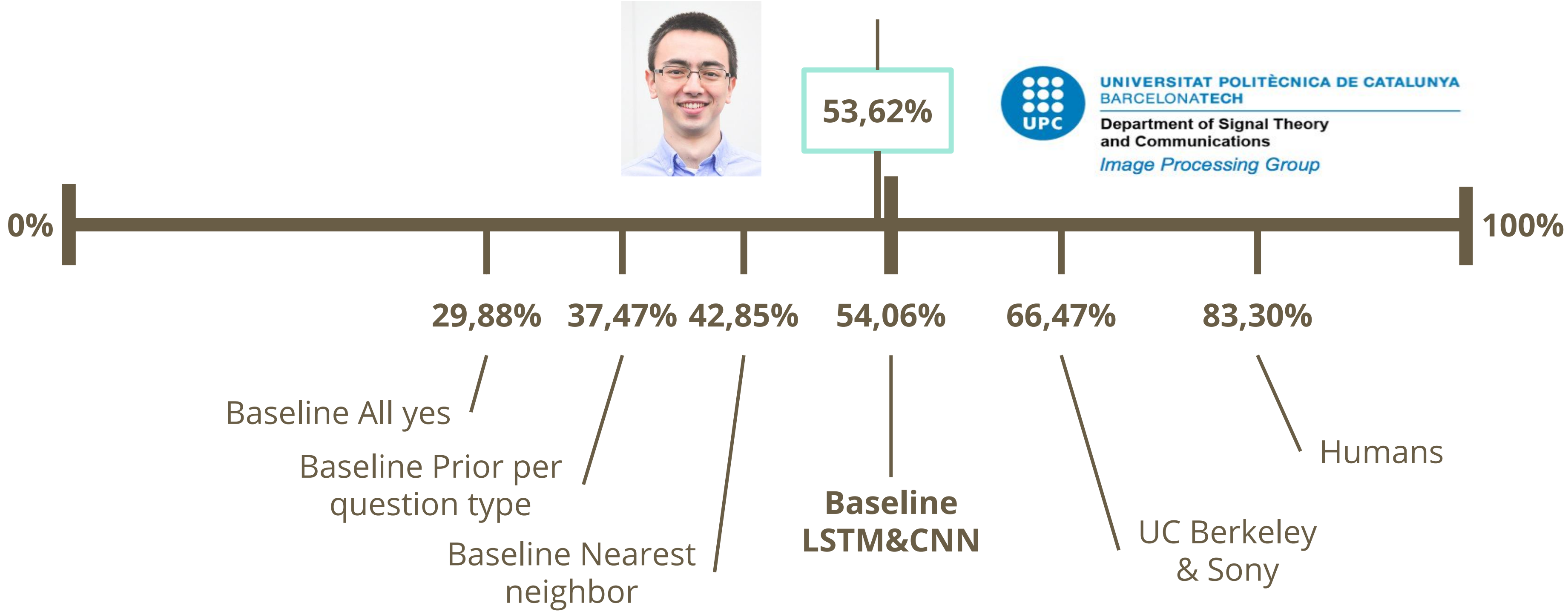
Challenges: Visual Question Answering

VQA Visual Question Answering



[Visual Question Answering](#)

Challenges: Visual Question Answering

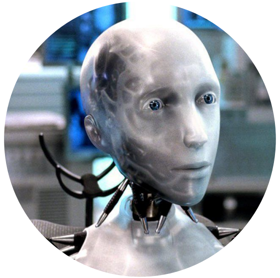


I. Masuda-Mora, "**Open-Ended Visual Question-Answering**". Submitted as BSc ETSETB thesis.
[clean code in Keras, perfect for beginners !]

Summary

- Embedding language and vision into semantic embeddings allows fusion learning.
- Very high interest among researchers. Great topic for your thesis.
- Will vision and language (and multimedia) communities be merged with (absorbed by) the machine learning one ?

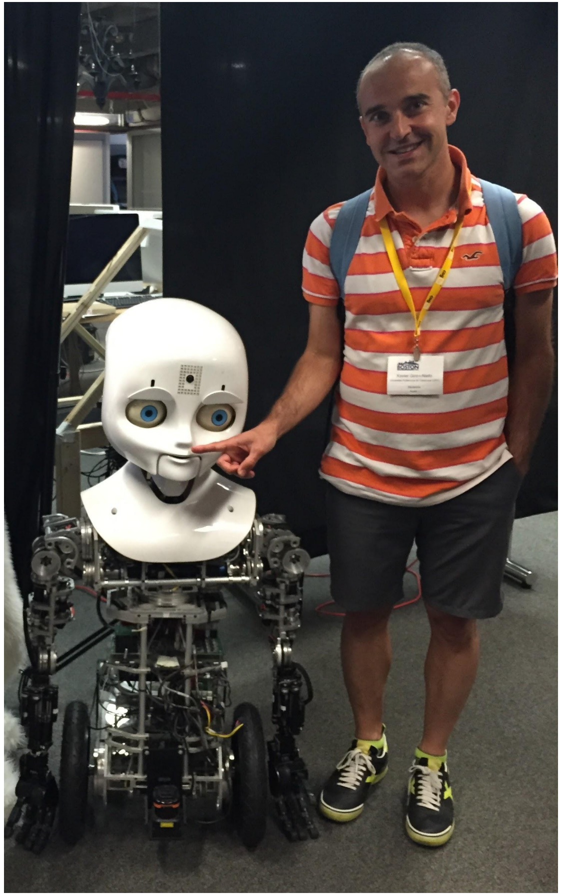
Conclusions



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

Slide credit: Issey Masuda

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