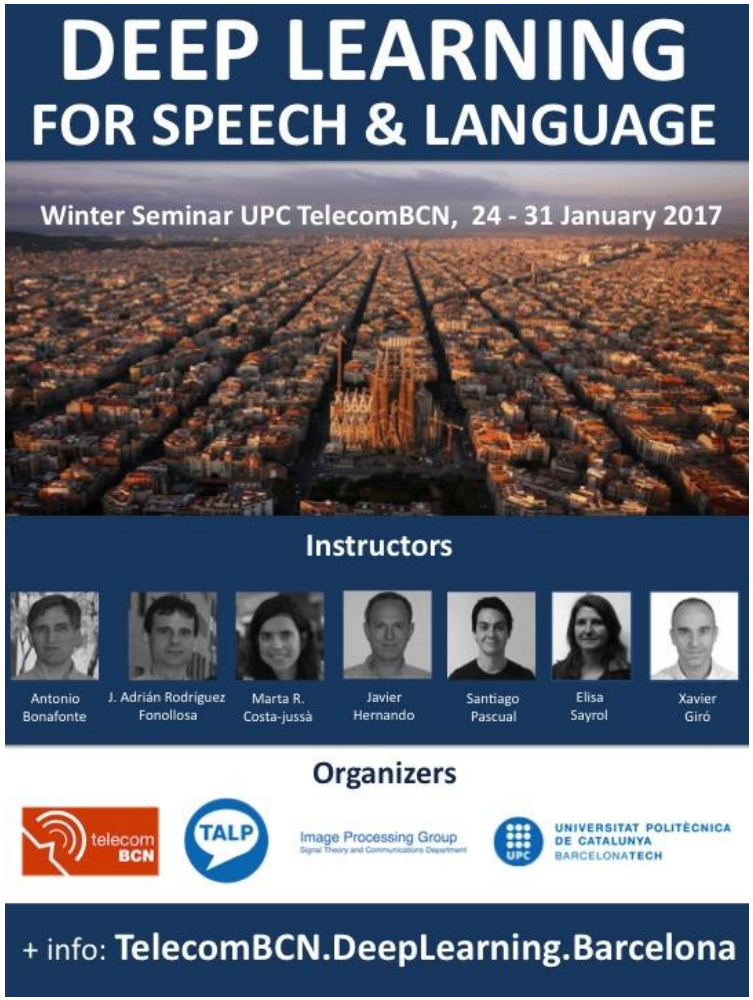


[course site]

Day 4 Lecture 4

Multimodal Deep Learning

Xavier Giró-i-Nieto

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Signal Theory and Communications
Image Processing Group

Multimedia

Text

Vision

Audio

and ratings, geolocation, time stamps...

Multimedia

Text

Vision

Audio

and ratings, geolocation, time stamps...

Xavier Giró-i-Nieto
@DocXavi

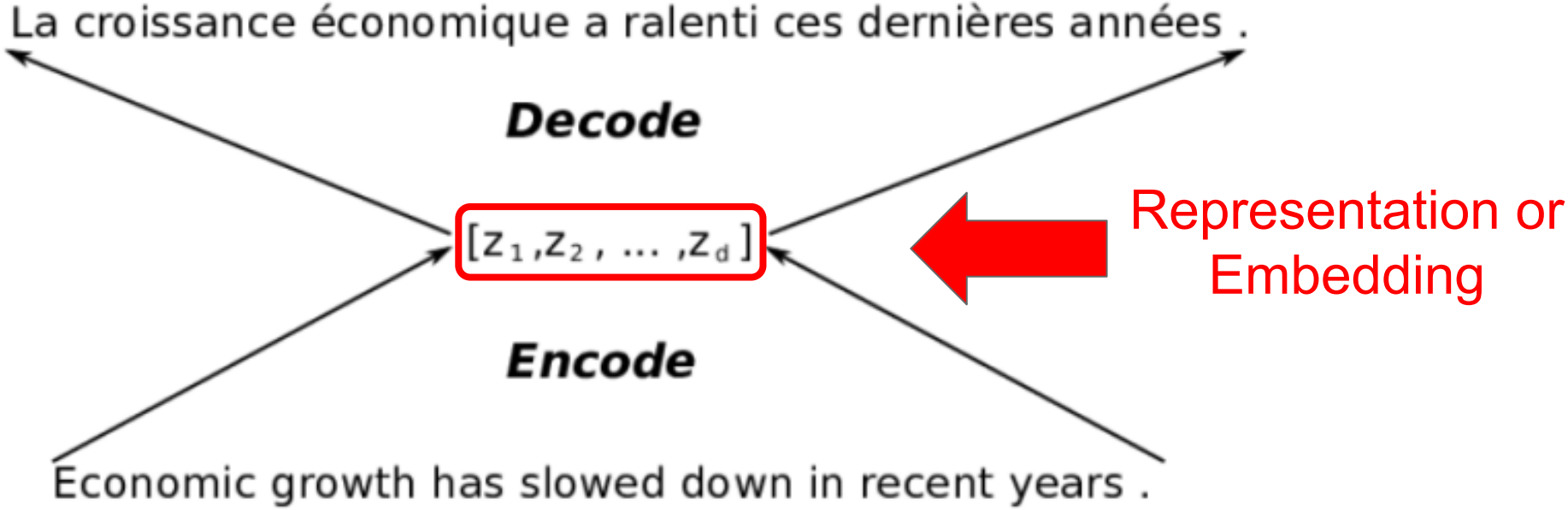
Take home message by @karpathy : read papers from machine translation community.
#deepvision16 #cvpr16

Tradueix del anglès

RETUIT 1 AGRADA A 1

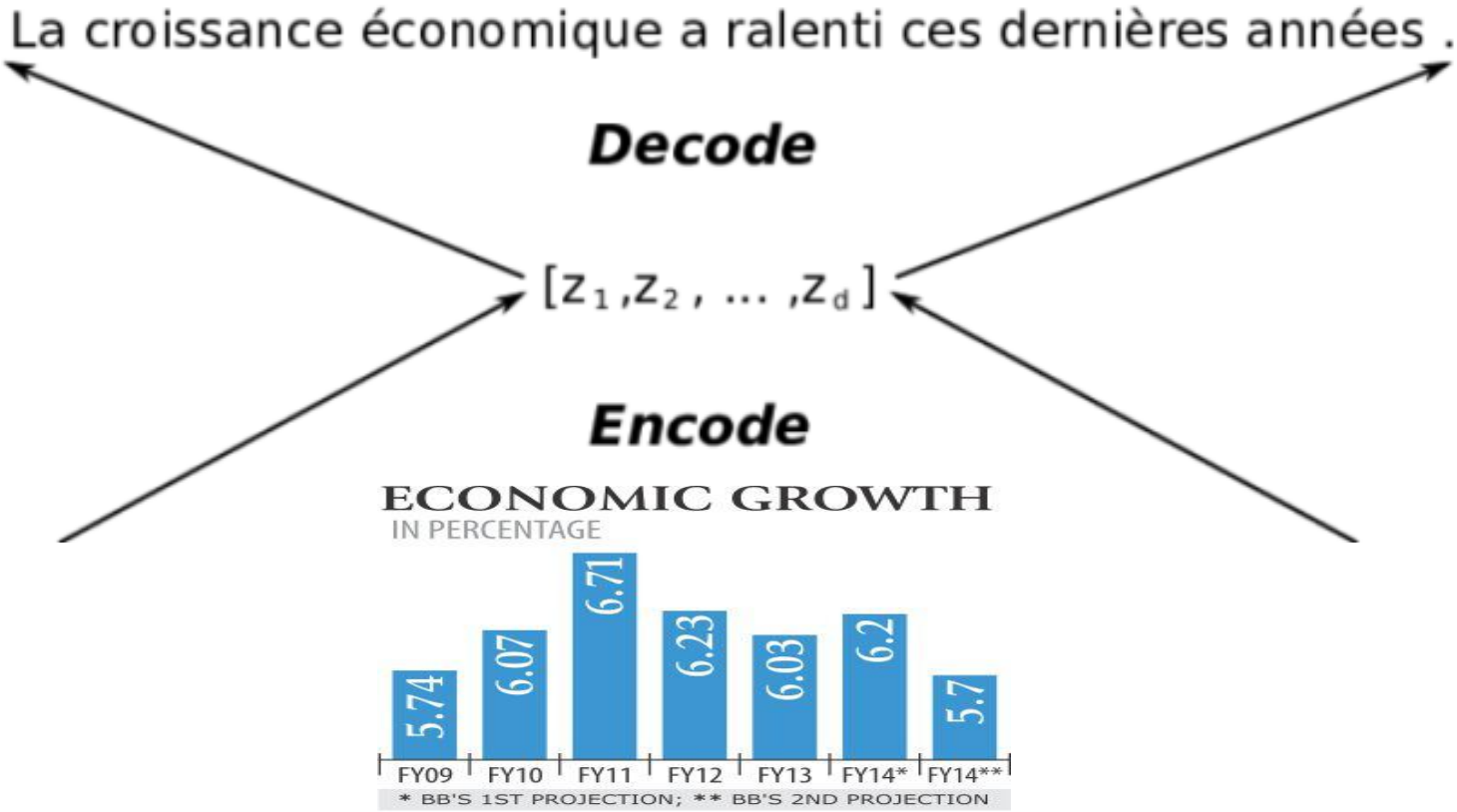
Language & Vision: Encoder-Decoder

Lectures D3L4 & D4L2 by Marta Ruiz
on Neural Machine Translation



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

Language & Vision: Encoder-Decoder



Captioning: DeepImageSent

man in black shirt is playing guitar.

construction worker in orange safety vest is working on road.

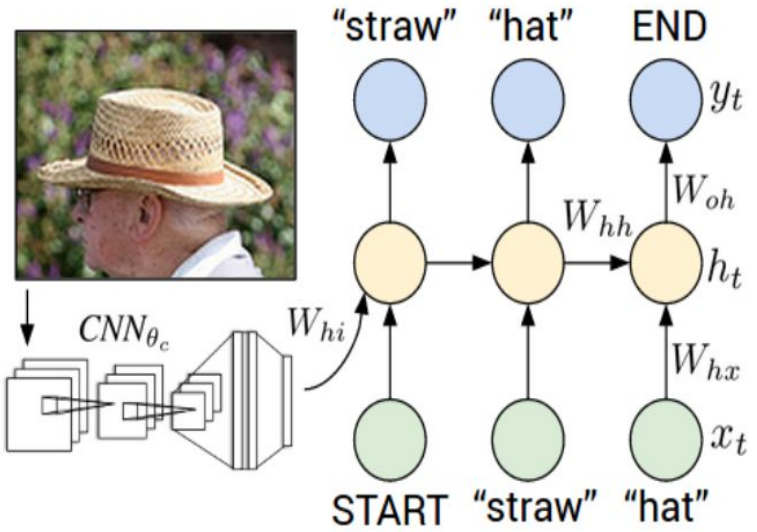
two young girls are playing with lego toy.

(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

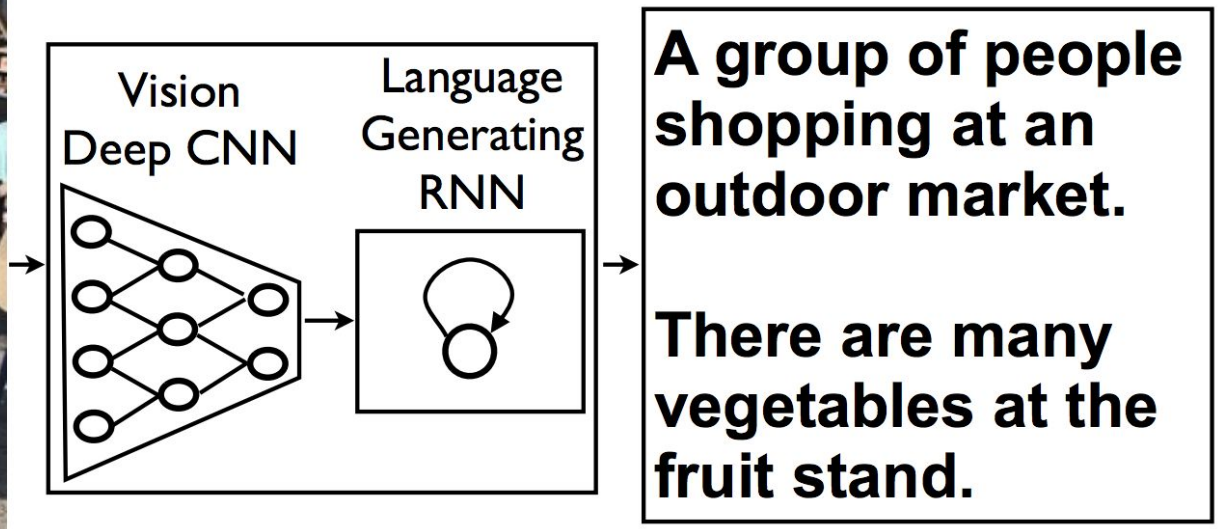
$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1) \odot b_v)$$
$$y_t = softmax(W_{oh}h_t + b_o).$$



Multimodal Recurrent Neural Network

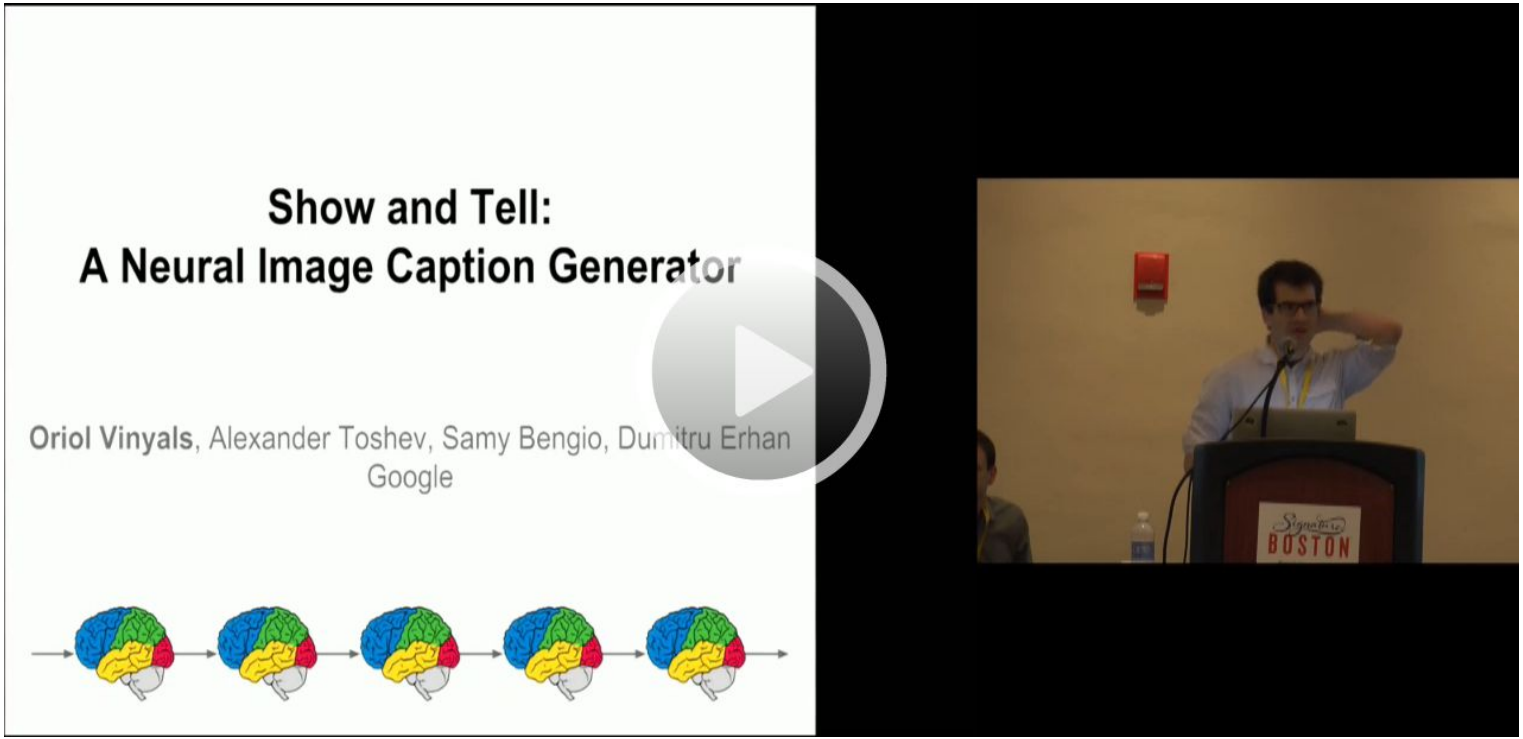
(Slides by Marc Bolaños): Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." CVPR 2015

Captioning: Show & Tell



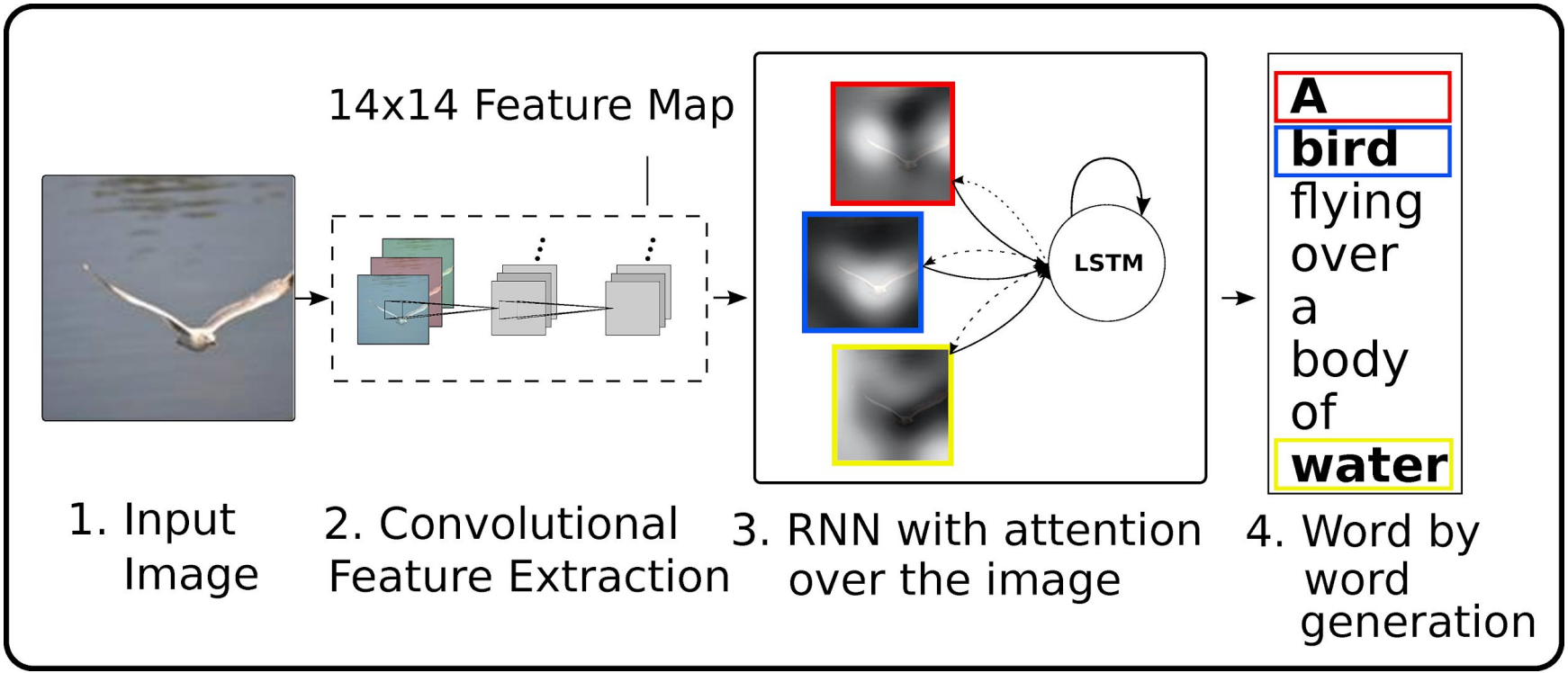
Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." CVPR 2015.

Captioning: Show & Tell



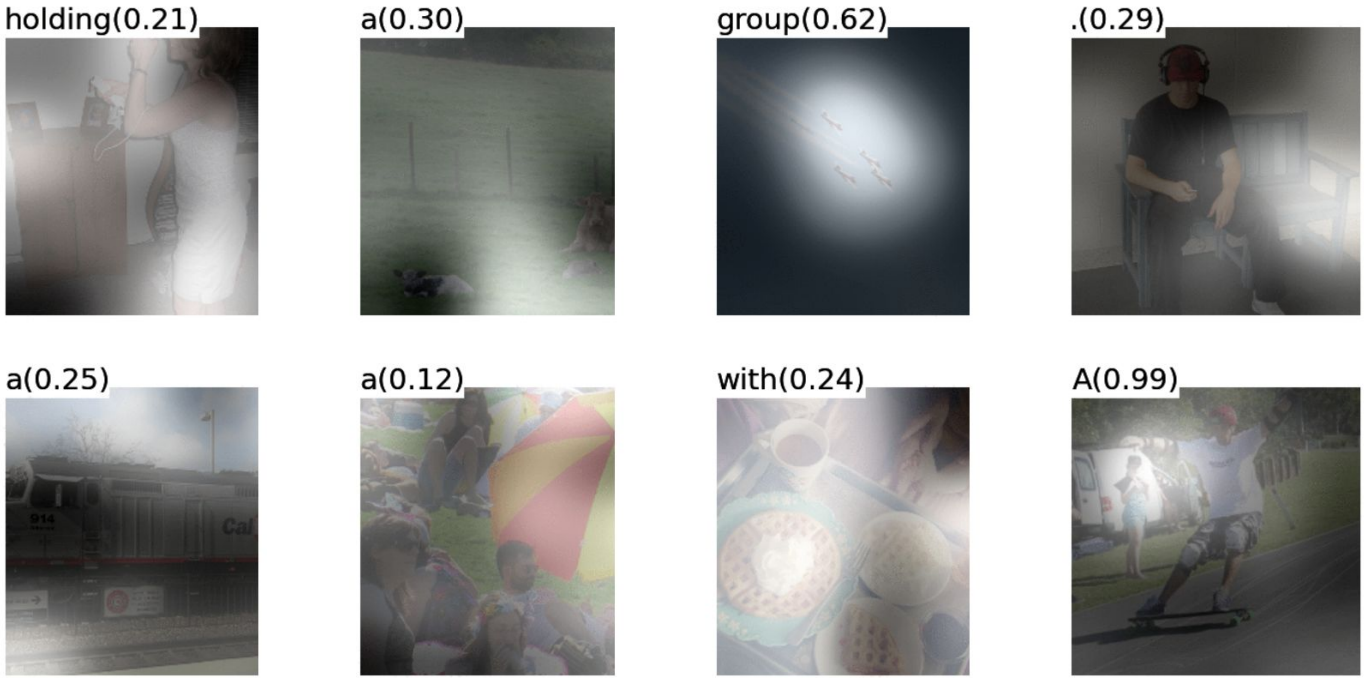
Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and tell: A neural image caption generator." CVPR 2015.

Captioning: Show, Attend & Tell



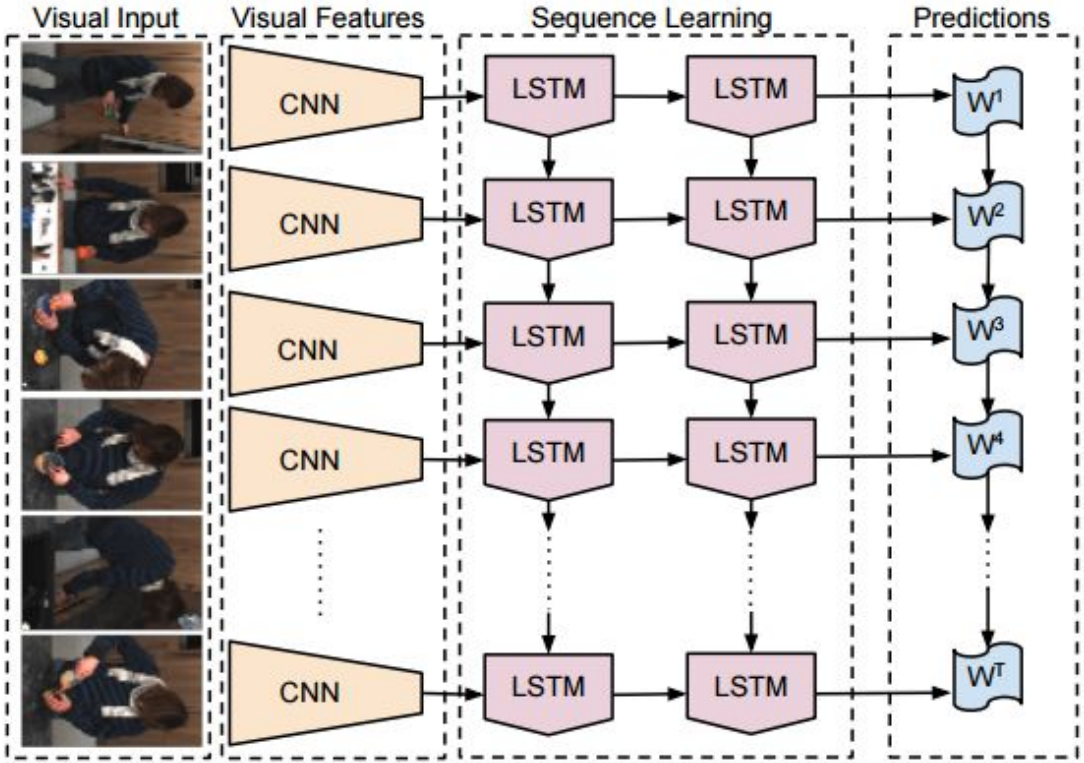
Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." ICML 2015

Captioning: Show, Attend & Tell



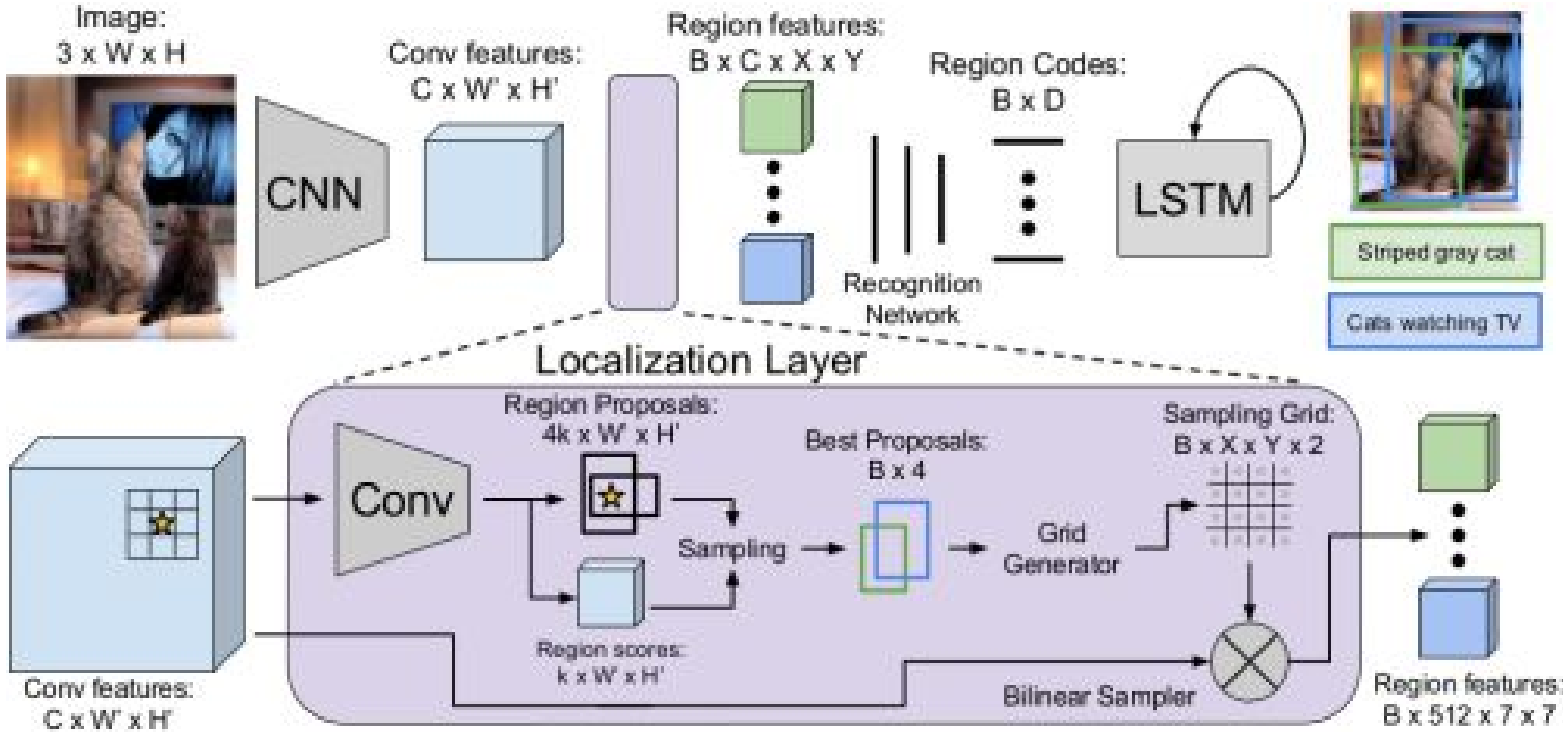
Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." ICML 2015

Captioning: LSTM with image & video



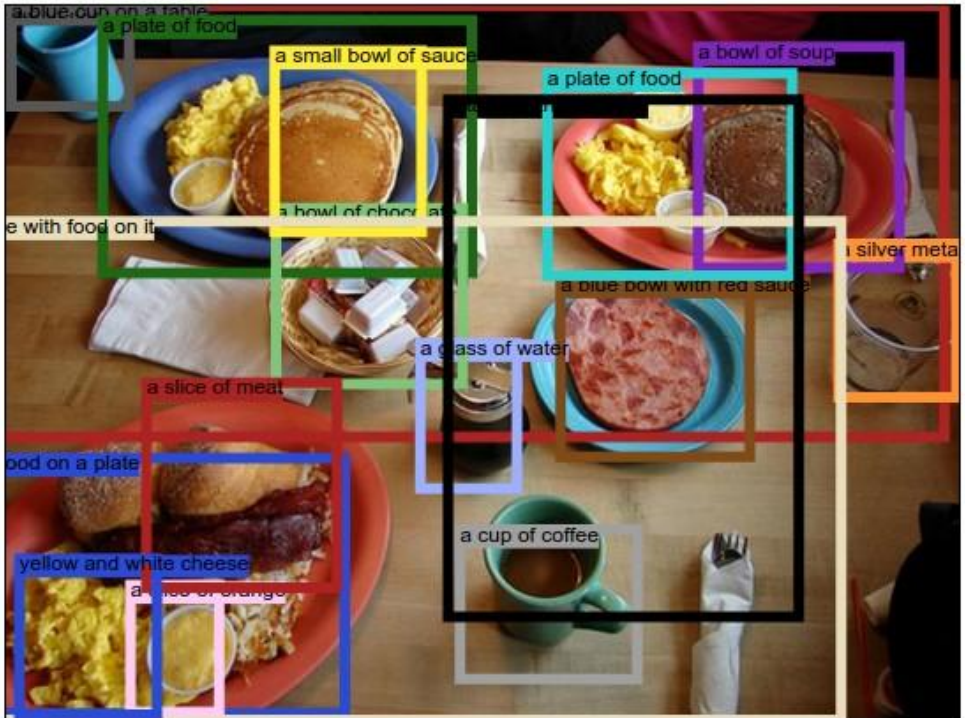
Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrel. Long-term Recurrent Convolutional Networks for Visual Recognition and Description, CVPR 2015. code

Captioning (+ Detection): DenseCap



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

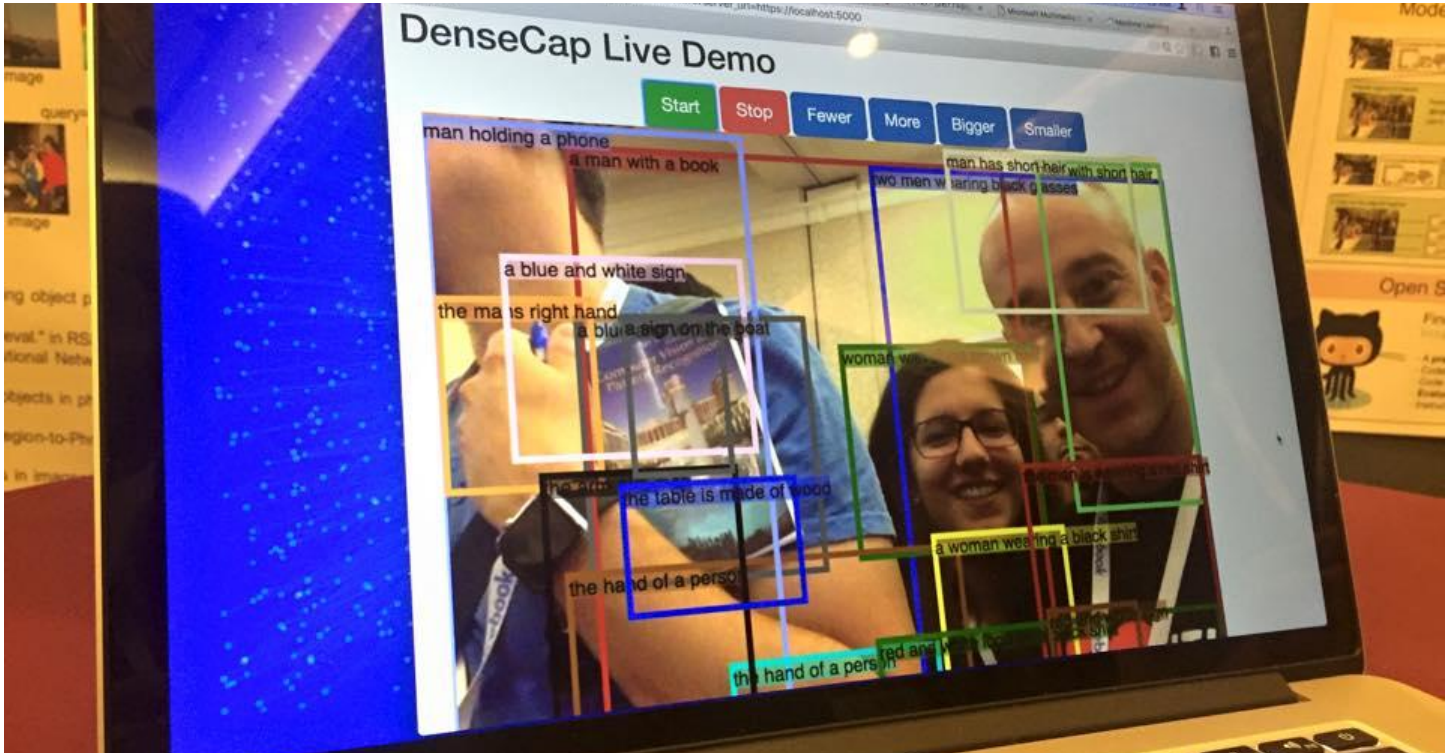
Captioning (+ Detection): DenseCap



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



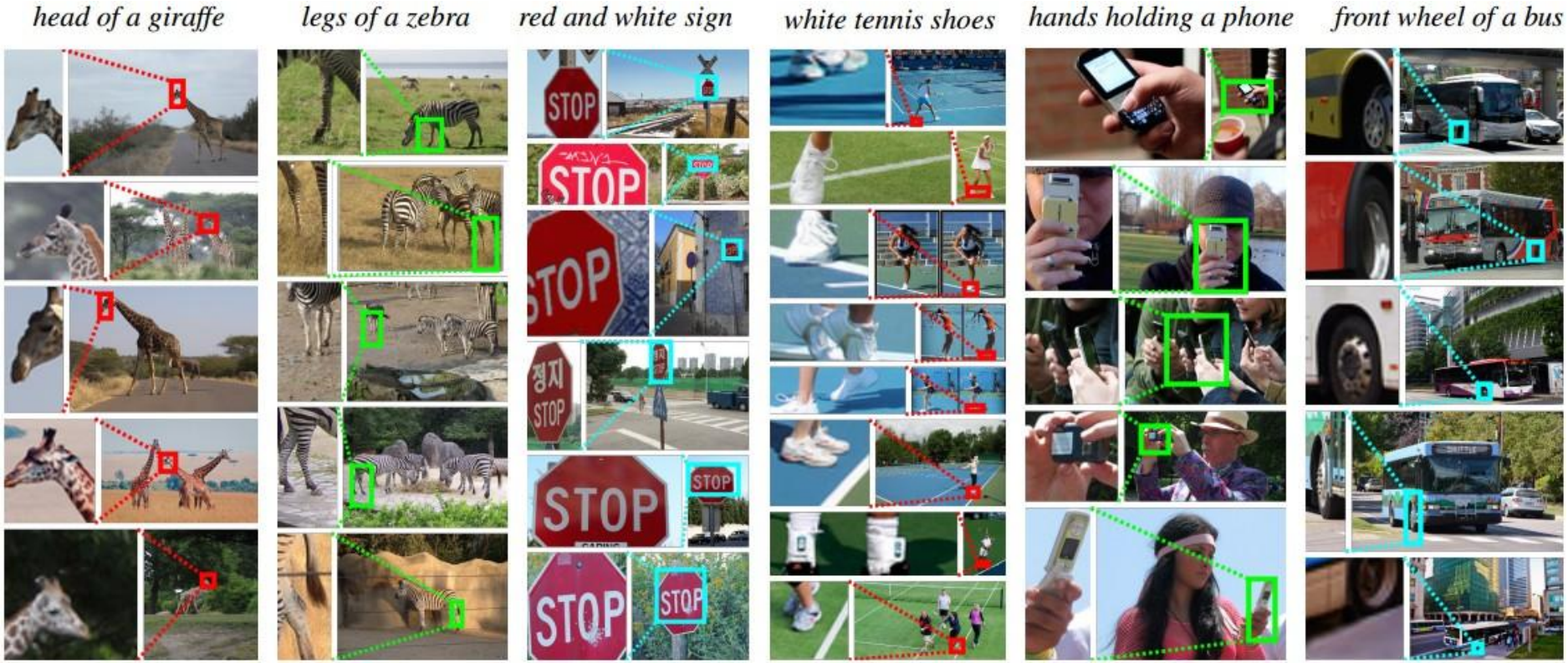
XAVI: “man has short hair”, “man with short hair”

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

BOTH: “two men wearing black glasses”

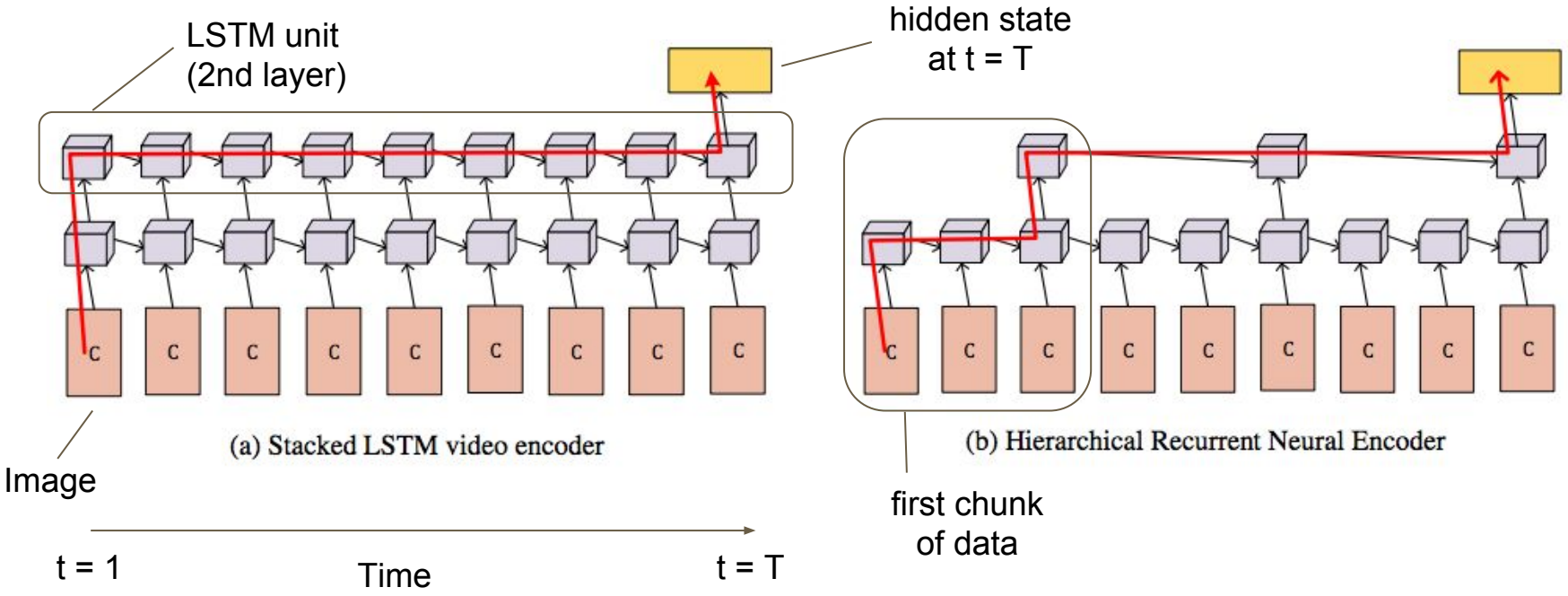
Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning." *CVPR 2016*

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

$[z_1, z_2, \ldots z_N]$ ⊕ $[y_1, y_2, \ldots y_M]$

Encode

Encode

"Is economic growth decreasing ?"

Visual Question Answering

Extract visual features

Question

What object is flying?

Embedding

Merge

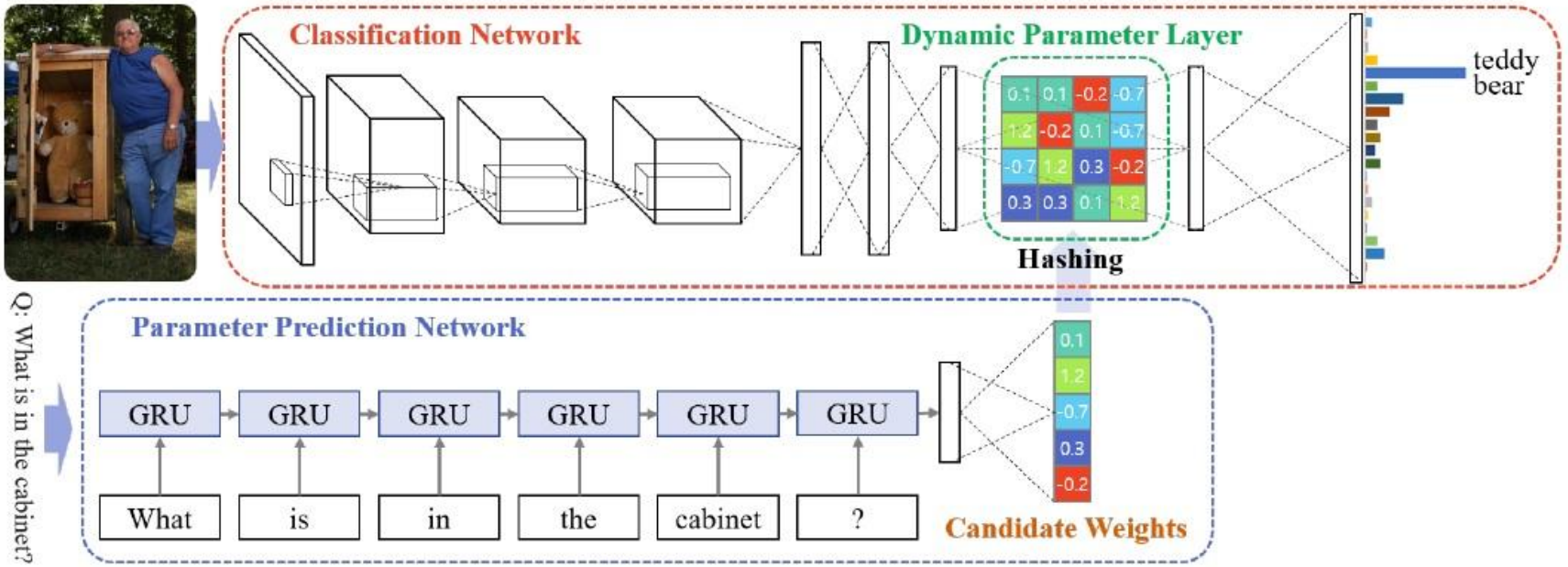
Predict answer

Answer

Kite

Slide credit: Issey Masuda

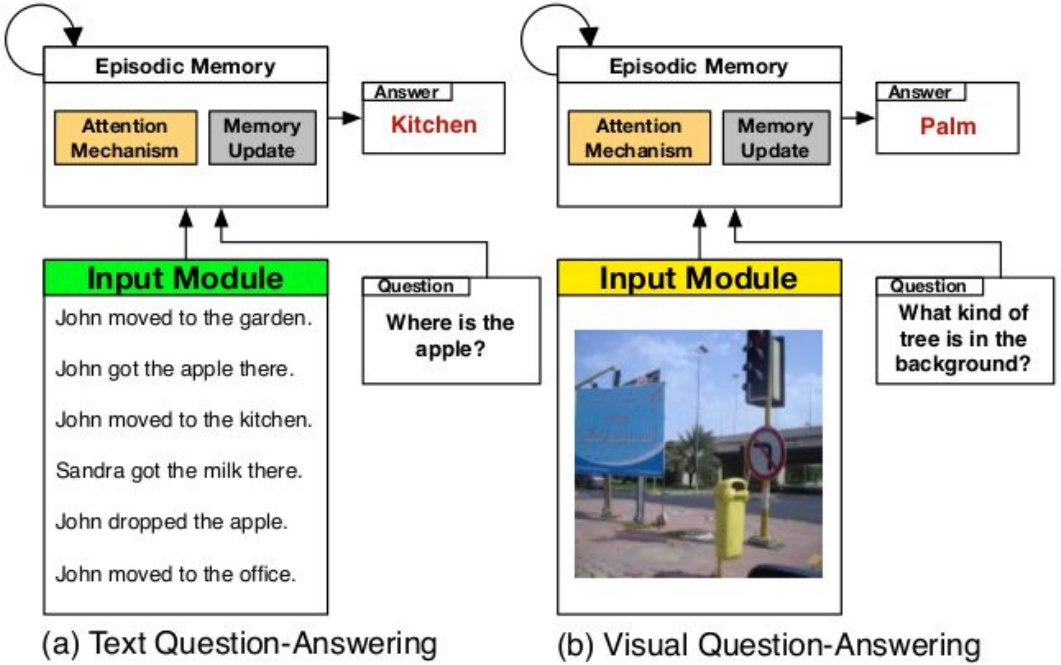
Visual Question Answering



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



(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.

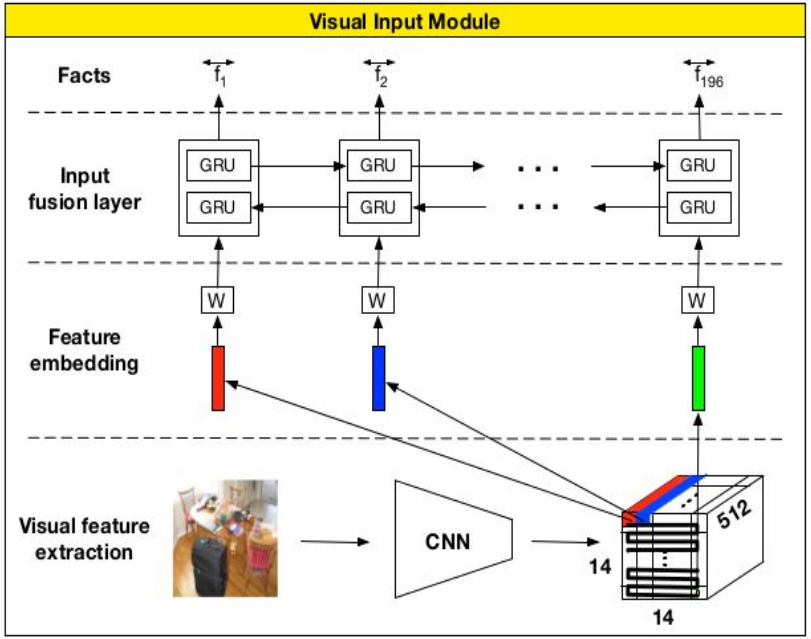
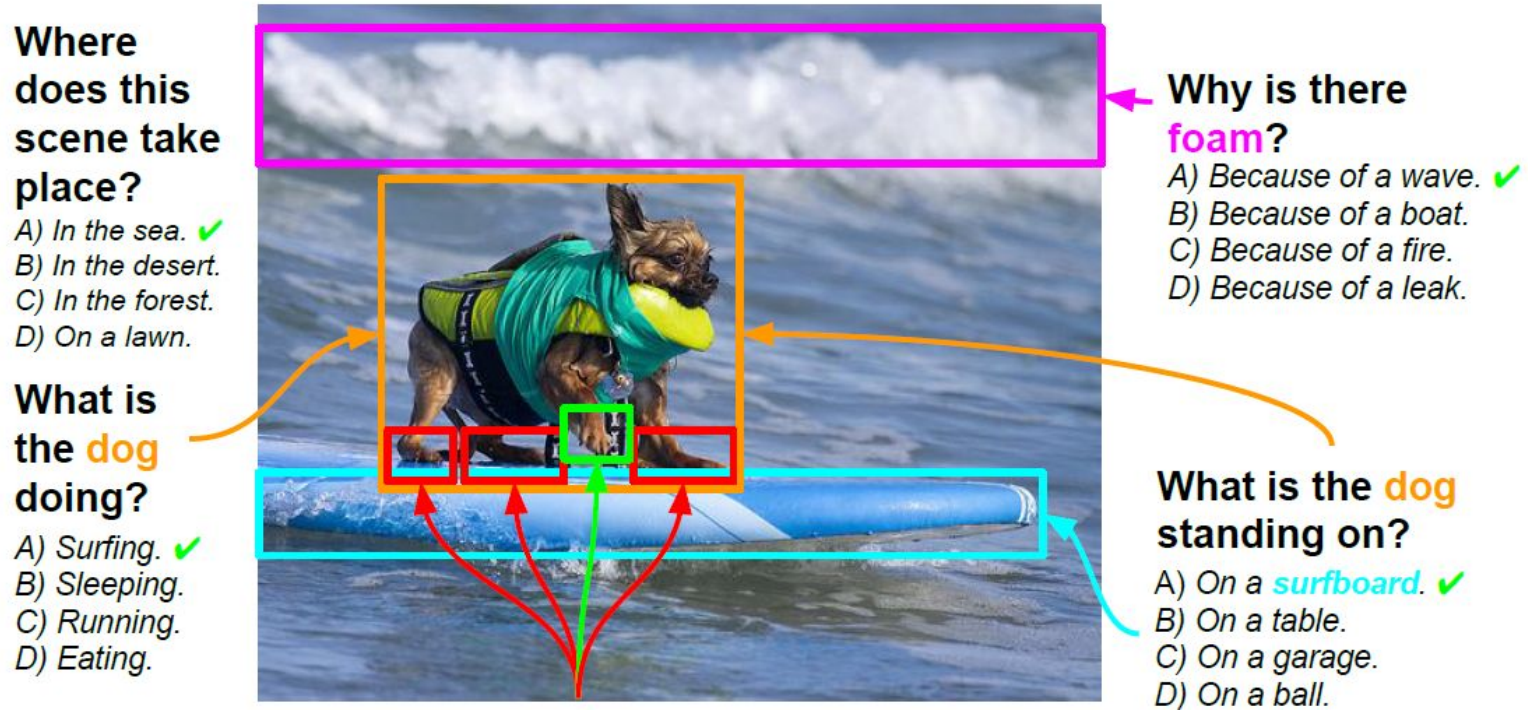


Figure 3. VQA input module to represent images for the DMN.

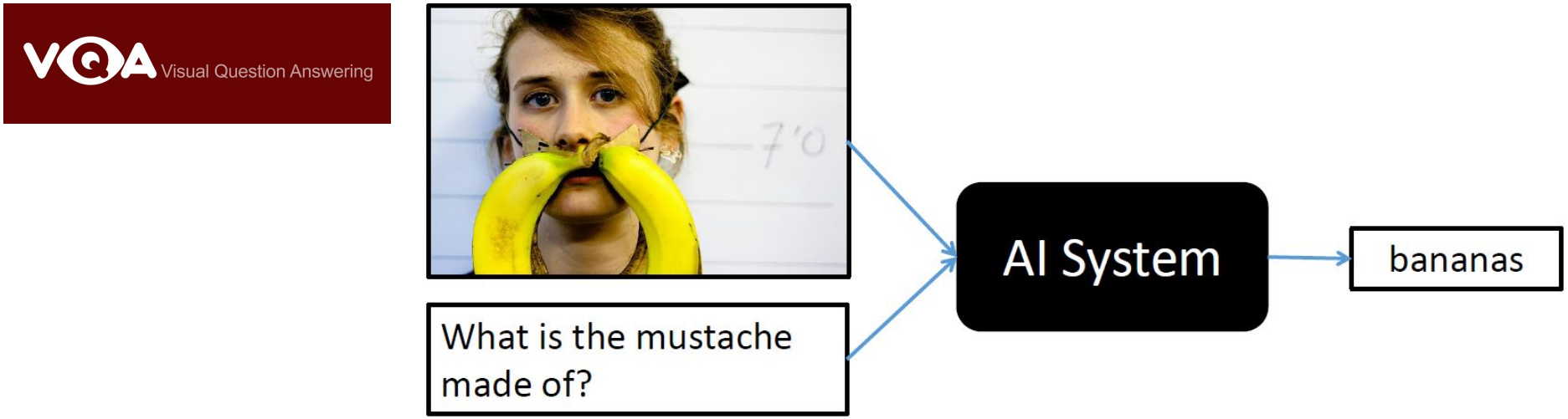
(Slides and Slidecast by Santi Pascual): Xiong, Caiming, Stephen Merity, and Richard Socher. "Dynamic Memory Networks for Visual and Textual Question Answering." ICML 2016.

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.

Challenges: Visual Question Answering



Visual Question Answering

Challenges: Visual Question Answering

53,62%

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Department of Signal Theory and Communications
Image Processing Group

0% — 100%

29,88% Baseline All yes

37,47% Baseline Prior per question type

42,85% Baseline Nearest neighbor

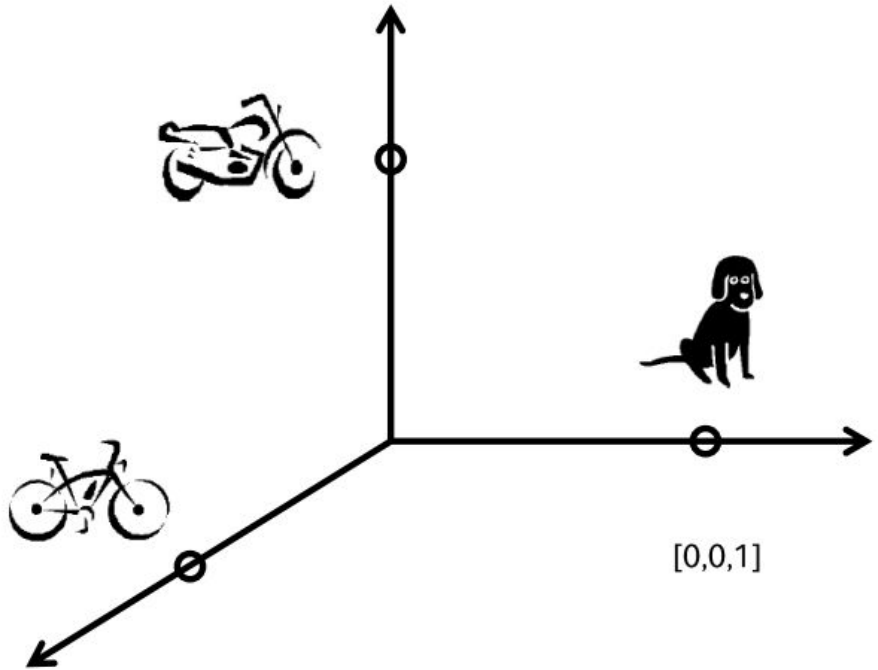
54,06% **Baseline LSTM&CNN**

66,47% UC Berkeley & Sony

83,30% Humans

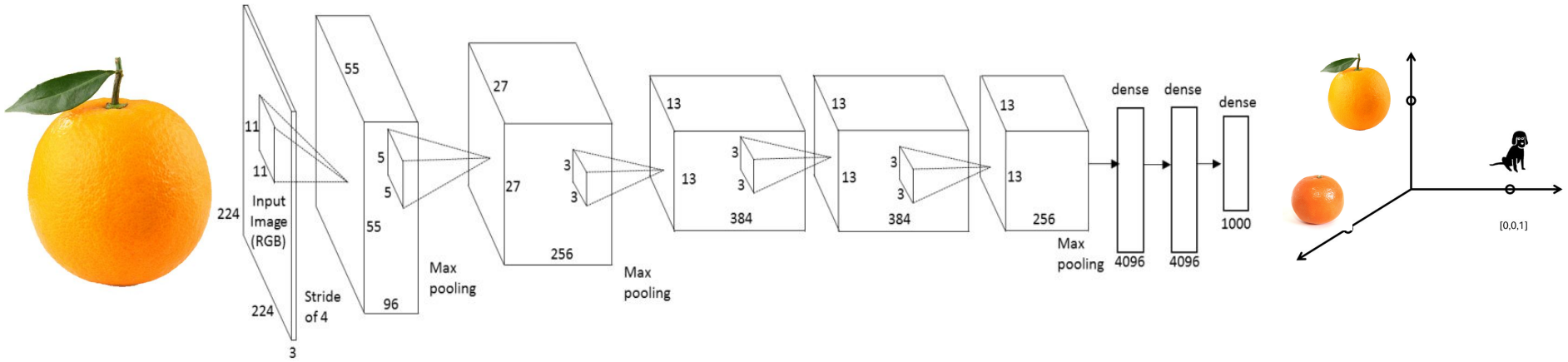
I. Masuda-Mora, "**Open-Ended Visual Question-Answering**". BSc ETSETB 2016. [Keras]

Vision and language: Embeddings



Perronin, F., CVPR Tutorial on LSVR @ CVPR'14, Output embedding for LSVR

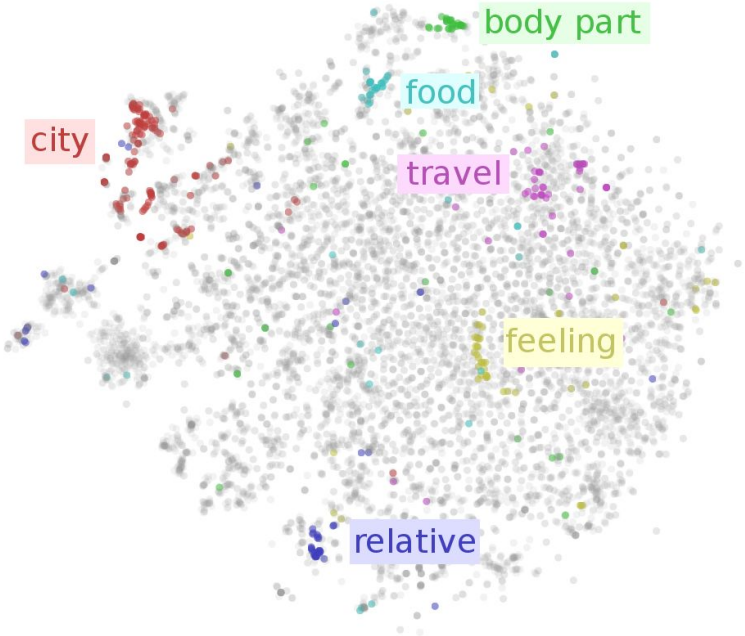
Vision and language: Embeddings



A Krizhevsky, I Sutskever, GE Hinton "Imagenet classification with deep convolutional neural networks" Part of: Advances in Neural Information Processing Systems 25 (NIPS 2012)

Vision and language: Embeddings

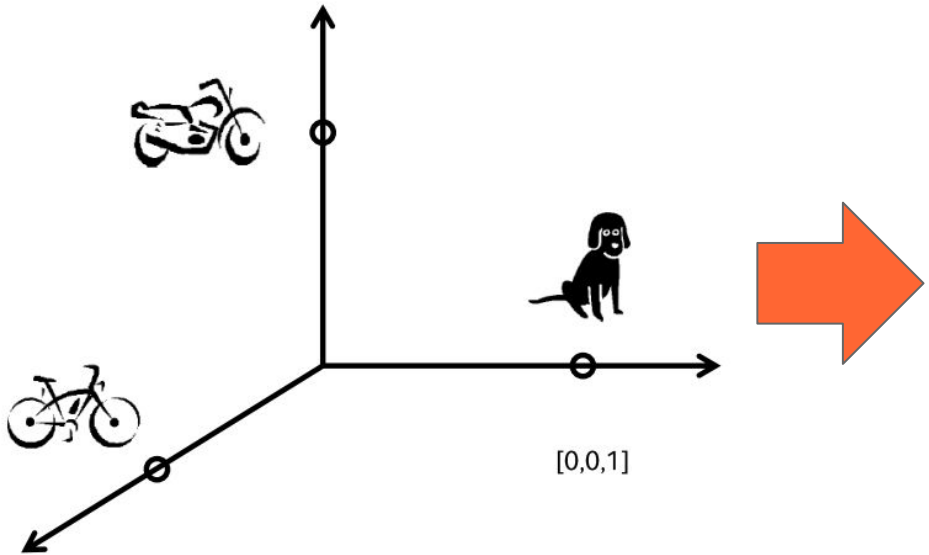
Lecture D2L4 by Toni Bonafonte
on Word Embeddings



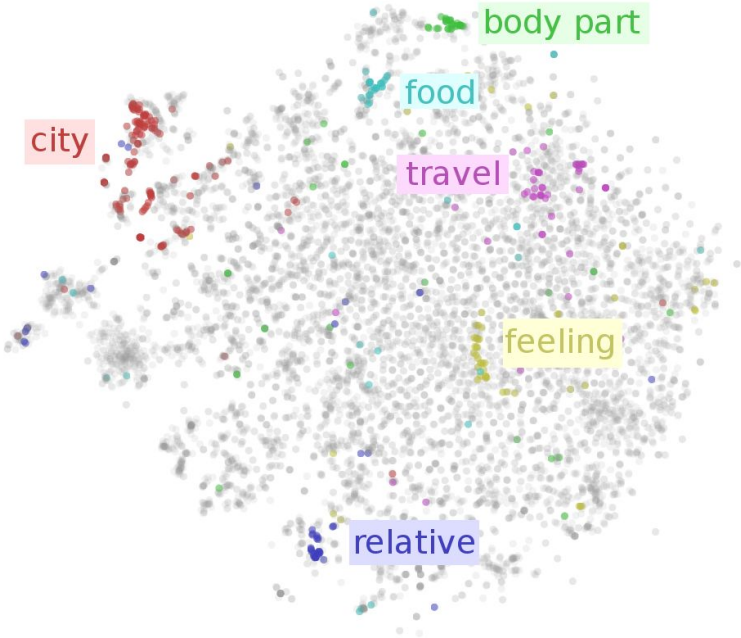
[Christopher Olah]

[Visualizing Representations]

Vision and language: Devise

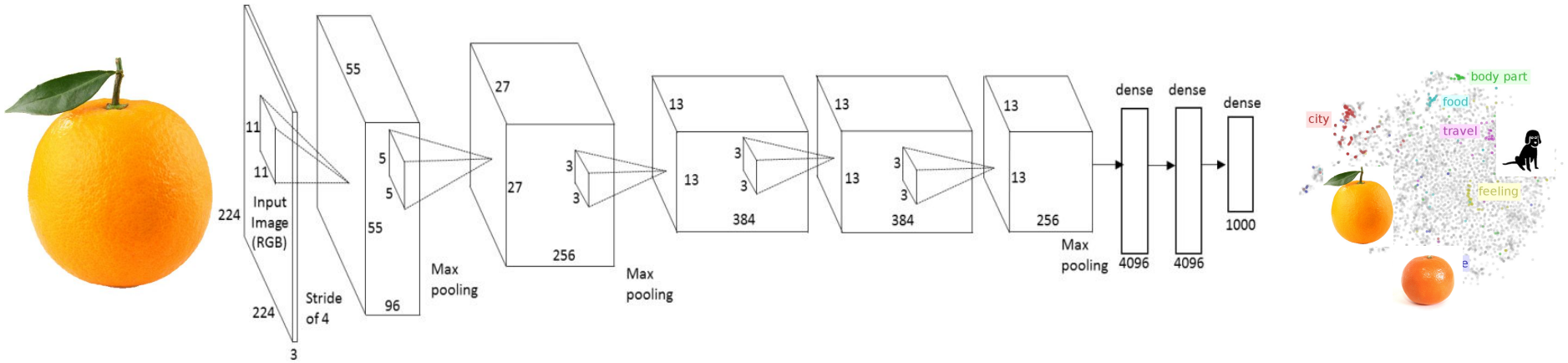


One-hot encoding



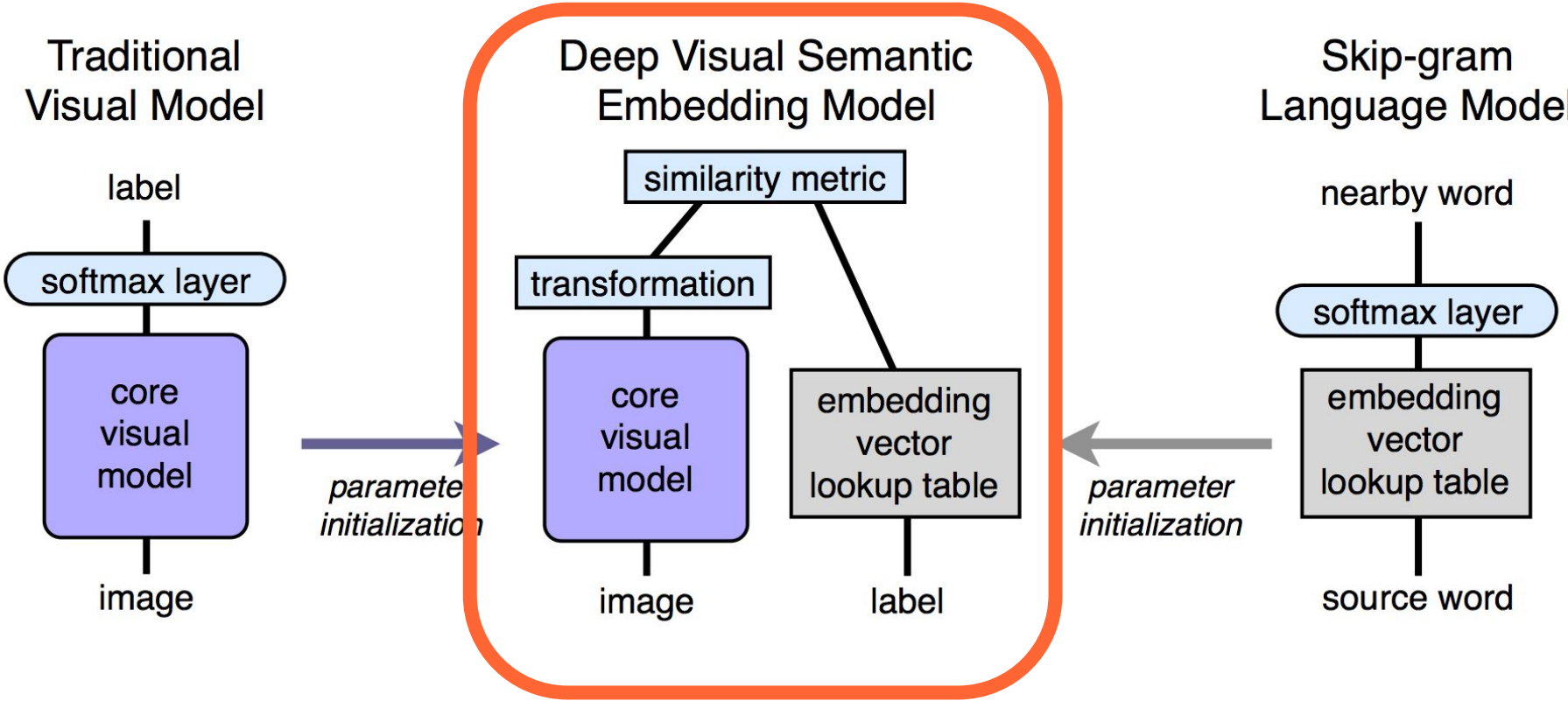
Embedding space

Vision and language: Devise



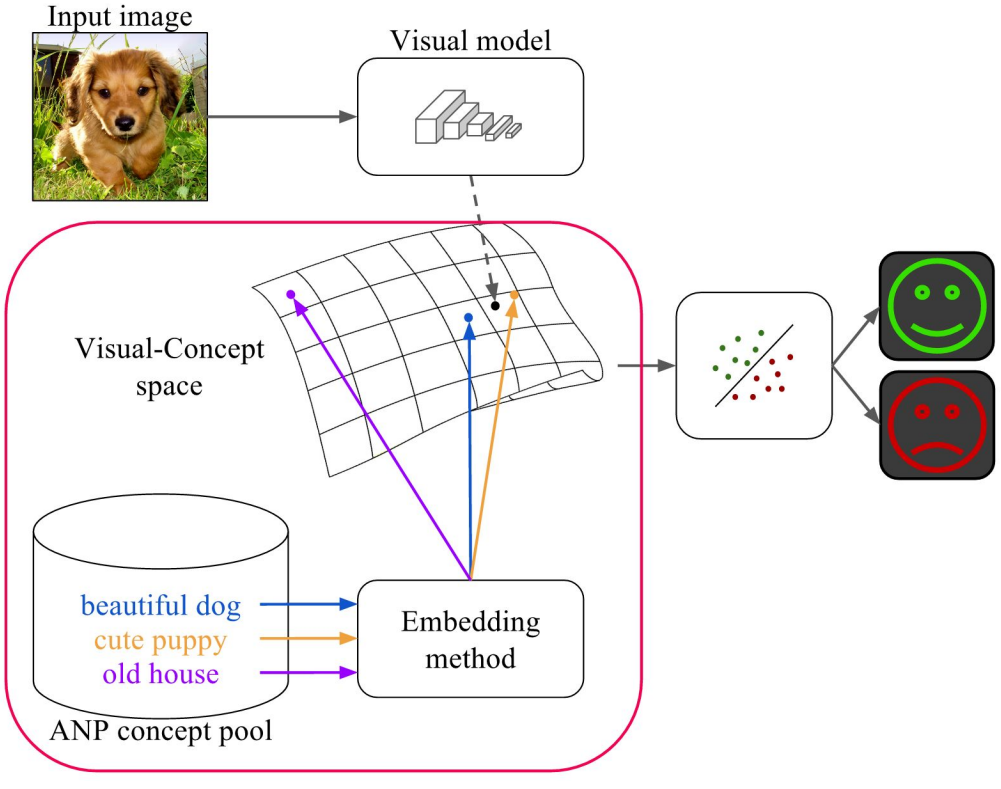
A Krizhevsky, I Sutskever, GE Hinton "Imagenet classification with deep convolutional neural networks" Part of: Advances in Neural Information Processing Systems 25 (NIPS 2012)

Vision and language: Devise



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

Vision and language: Embedding



Víctor Campos, Dèlia Fernàndez, Jordi Torres, Xavier Giró-i-Nieto, Brendan Jou and Shih-Fu Chang (work under progress)

Learn more

Julia Hockenmeirer (UIUC), Vision to Language (@ Microsoft Research)

Multimedia

Text

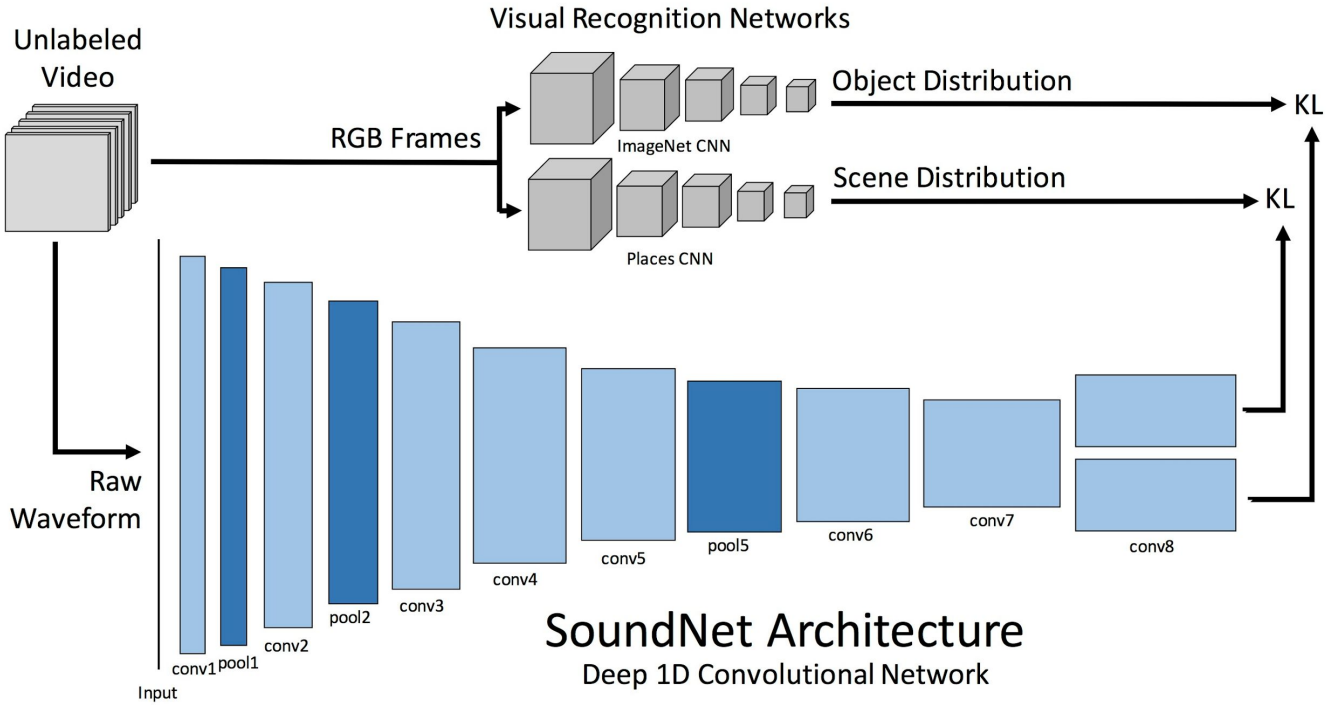
Audio

Vision

and ratings, geolocation,
time stamps...

Audio and Video: Soundnet

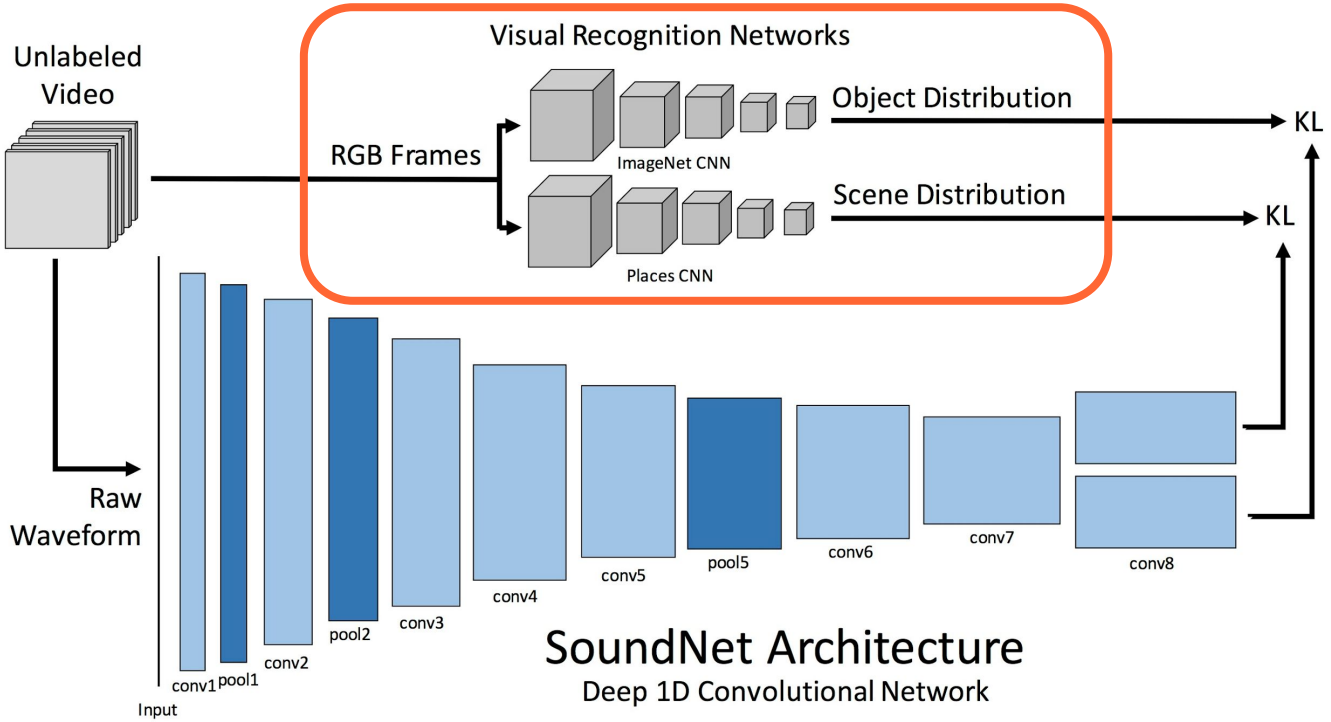
Object & Scenes recognition in videos by analysing the audio track (only).



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." In *Advances in Neural Information Processing Systems*, pp. 892-900. 2016.

Audio and Video: Soundnet

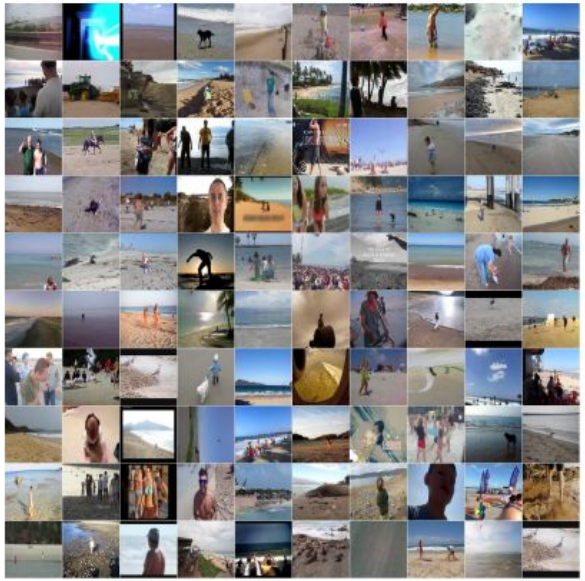
Videos for training are unlabeled. Relies on CNNs trained on labeled images.



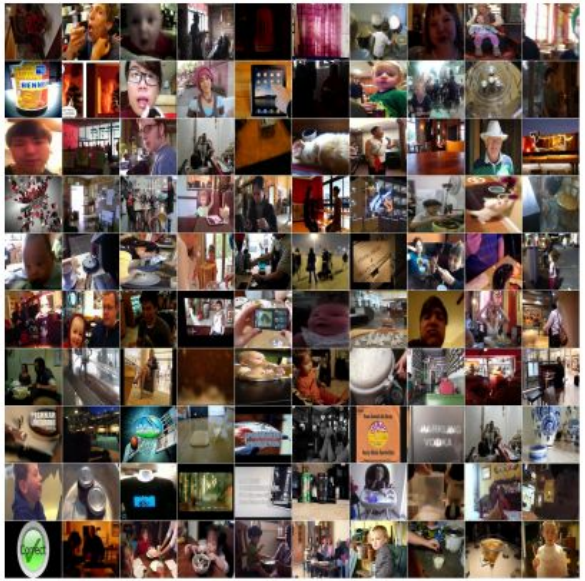
Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016.

Audio and Video: Soundnet

Videos for training are unlabeled. Relies on CNNs trained on labeled images.



Beach



Classroom

Construction

Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016.

Audio and Video: Soundnet

Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." In *Advances in Neural Information Processing Systems*, pp. 892-900. 2016.

Audio and Video: Soundnet

Hidden layers of Soundnet are used to train a standard SVM classifier that outperforms state of the art.

Method	Accuracy
RG [29]	69%
LTT [21]	72%
RNH [30]	77%
Ensemble [34]	78%
SoundNet	**88%**

Table 3: **Acoustic Scene Classification on DCASE:** We evaluate classification accuracy on the DCASE dataset. By leveraging large amounts of unlabeled video, SoundNet generally outperforms hand-crafted features by 10%.

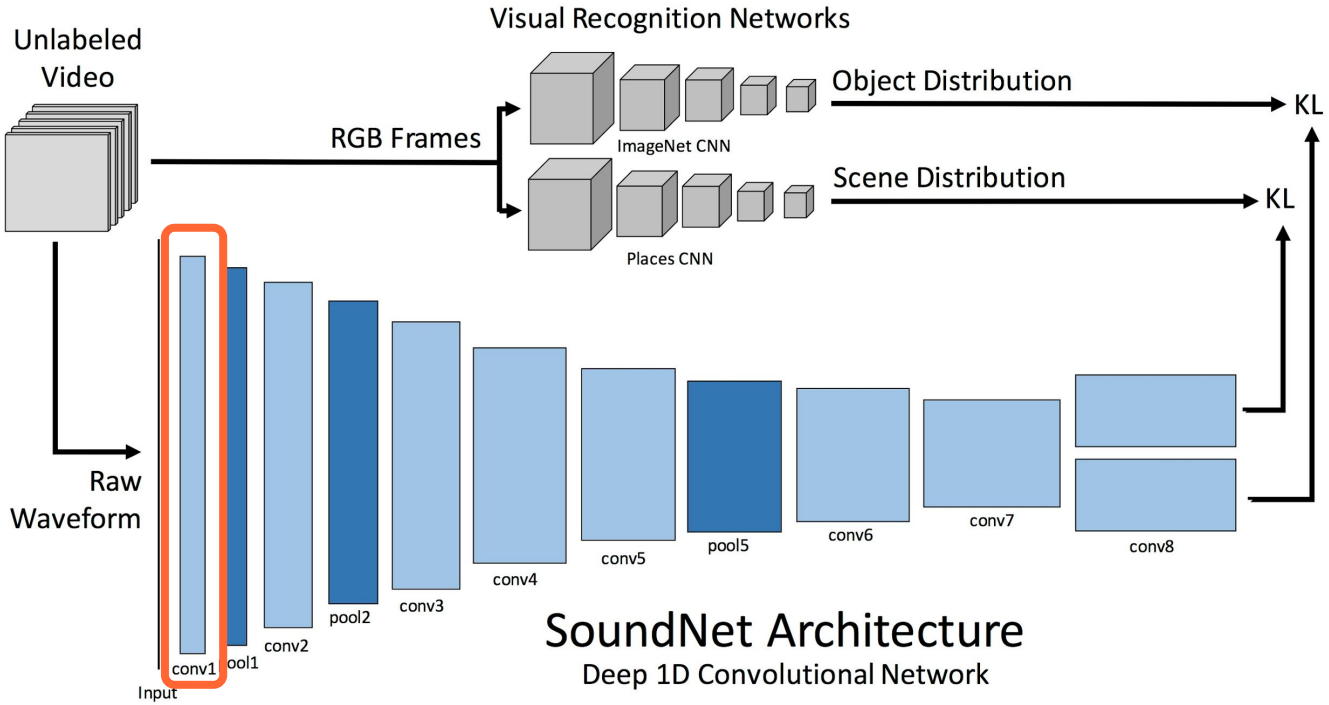
Method	Accuracy on ESC-50	Accuracy on ESC-10
SVM-MFCC [28]	39.6%	67.5%
Convolutional Autoencoder	39.9%	74.3%
Random Forest [28]	44.3%	72.7%
Piczak ConvNet [27]	64.5%	81.0%
SoundNet	**74.2%**	**92.2%**
Human Performance [28]	81.3%	95.7%

Table 4: **Acoustic Scene Classification on ESC-50 and ESC-10:** We evaluate classification accuracy on the ESC datasets. Results suggest that deep convolutional sound networks trained with visual supervision on unlabeled data outperforms baselines.

Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016.

Audio and Video: Soundnet

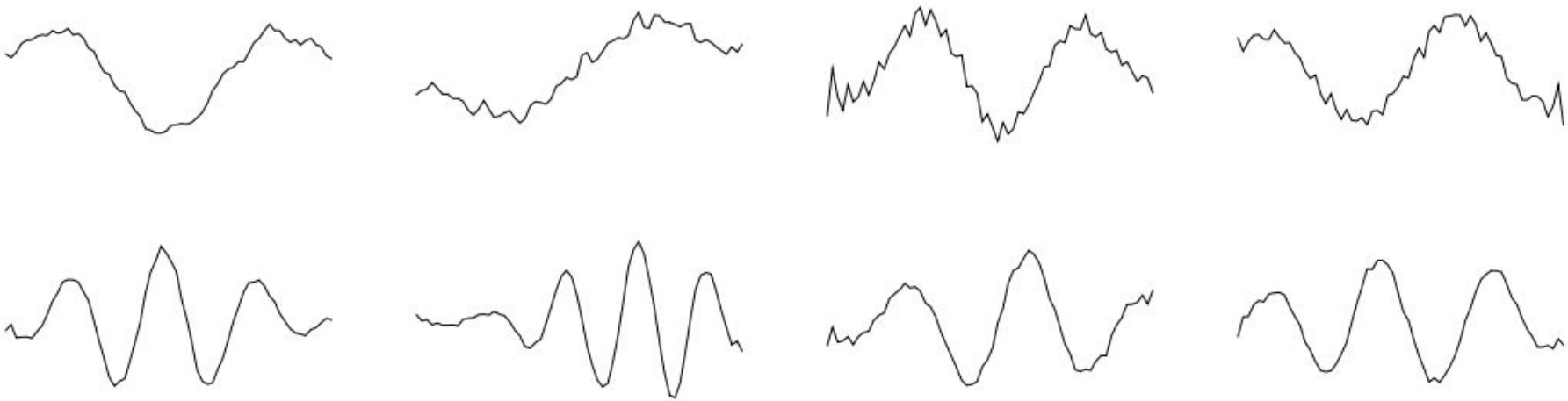
Visualization of the 1D filters over raw audio in conv1.



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016.

Audio and Video: Soundnet

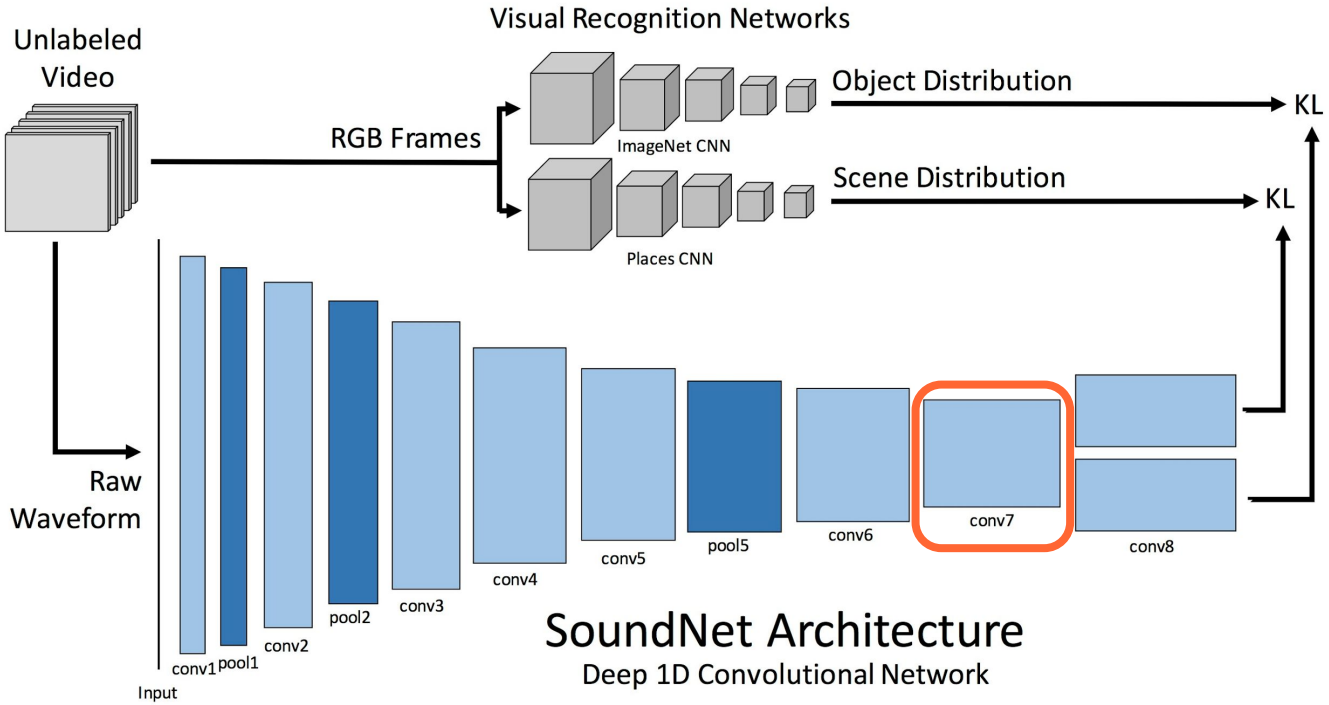
Visualization of the 1D filters over raw audio in conv1.



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016.

Audio and Video: Soundnet

Visualization of the 1D filters over raw audio in conv1.



Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." NIPS 2016.

Audio and Video: Soundnet

Visualization of the video frames associated to the sounds that activate some of the last hidden units (conv7):

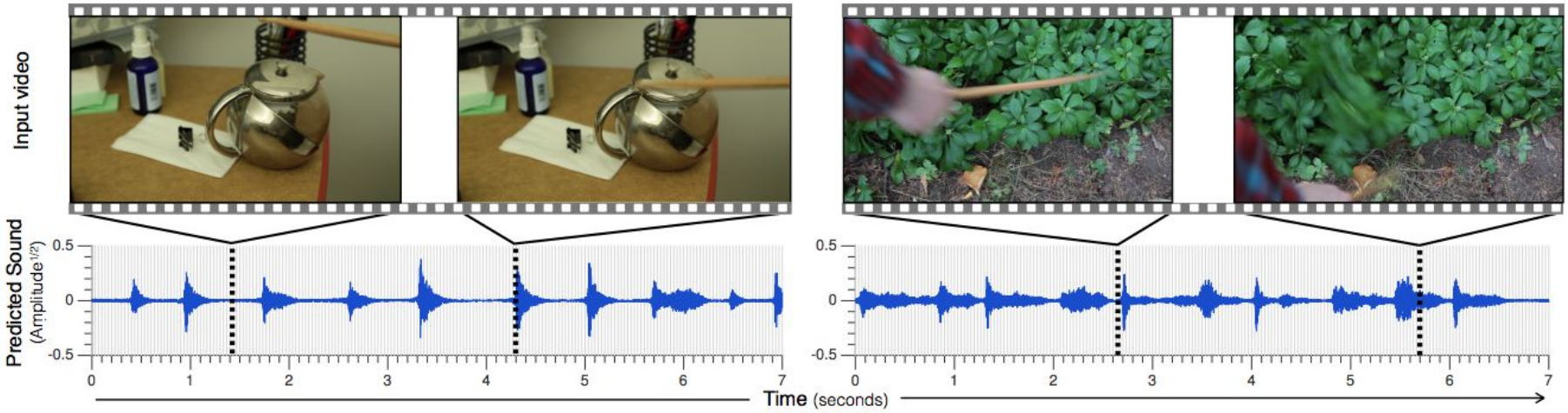
Baby Talk

Bubbles

Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." In *Advances in Neural Information Processing Systems*, pp. 892-900. 2016.

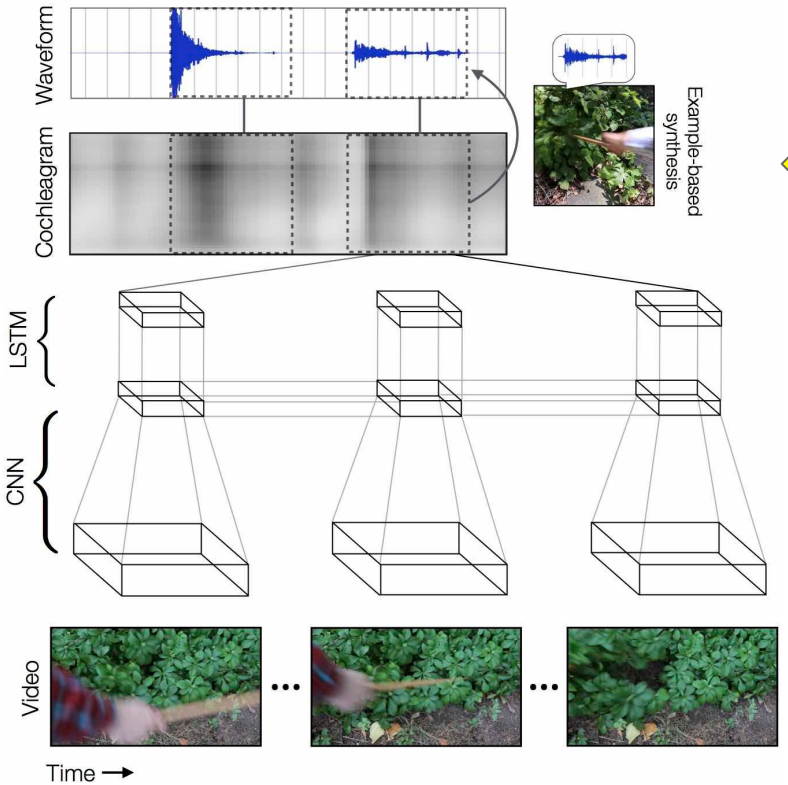
Audio and Video: Sonorizaton

Learn synthesized sounds from videos of people hitting objects with a drumstick.



Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. "Visually indicated sounds." CVPR 2016.

Audio and Video: Visual Sounds

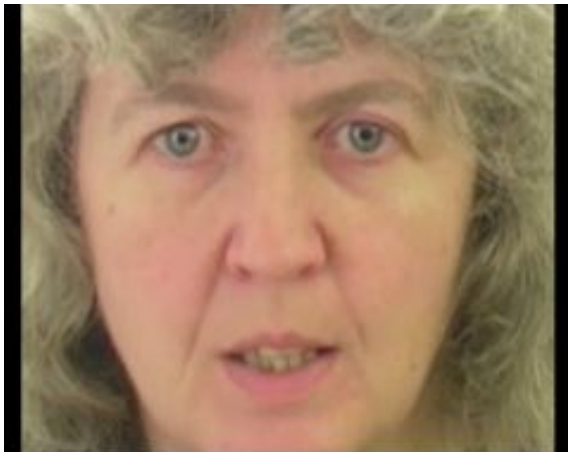


Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. "Visually indicated sounds." CVPR 2016.

Audio and Video: Visual Sounds

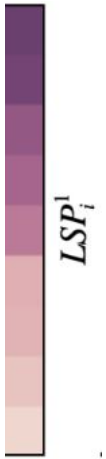
Owens, Andrew, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. "Visually indicated sounds." CVPR 2016.

Speech and Video: Vid2Speech



Frame from a slient video

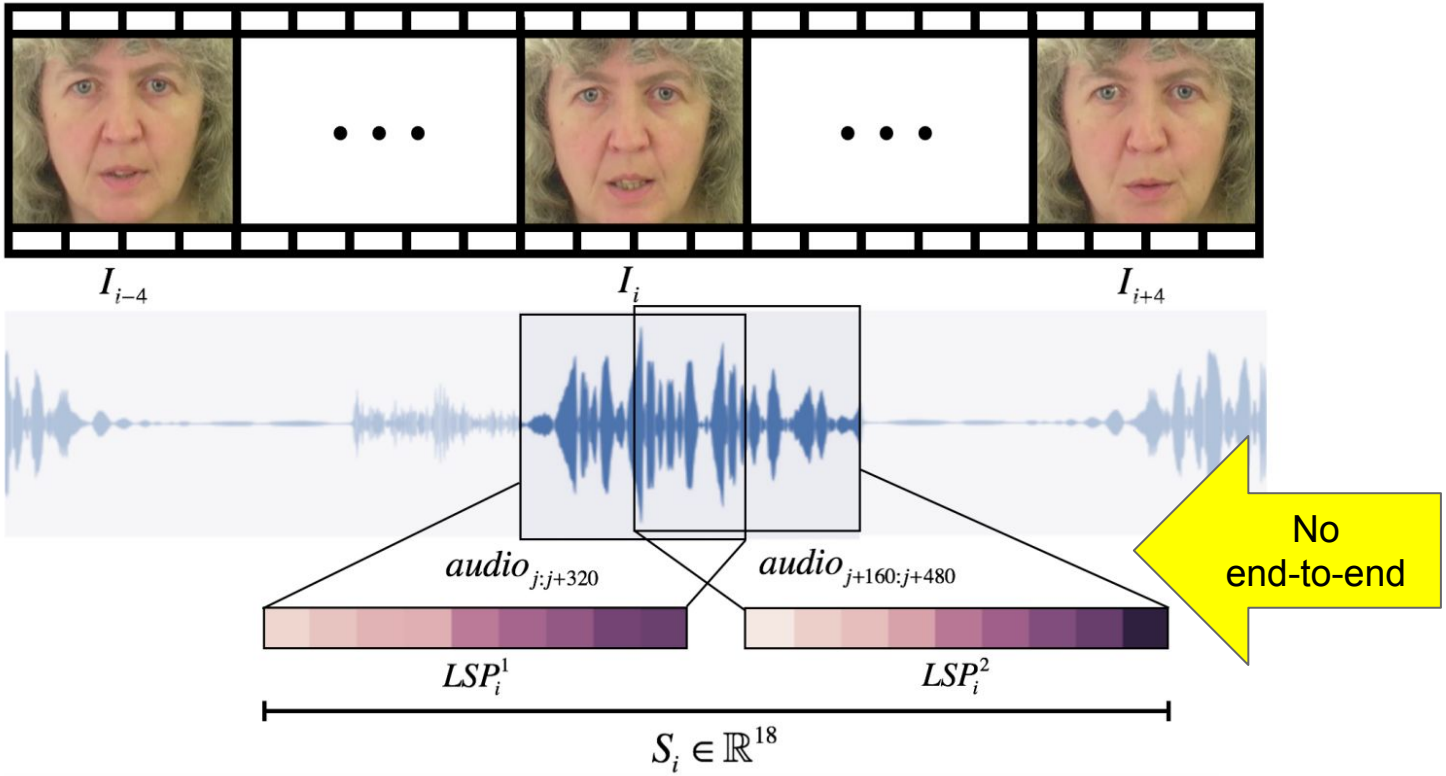
CNN (VGG)



Audio feature

Ephrat, Ariel, and Shmuel Peleg. "Vid2speech: Speech Reconstruction from Silent Video." *ICASSP 2017*

Speech and Video: Vid2Speech



Ephrat, Ariel, and Shmuel Peleg. "Vid2speech: Speech Reconstruction from Silent Video." *ICASSP 2017*

Speech and Video: Vid2Speech

Ephrat, Ariel, and Shmuel Peleg. "Vid2speech: Speech Reconstruction from Silent Video." *ICASSP 2017*

Speech and Video: LipNet

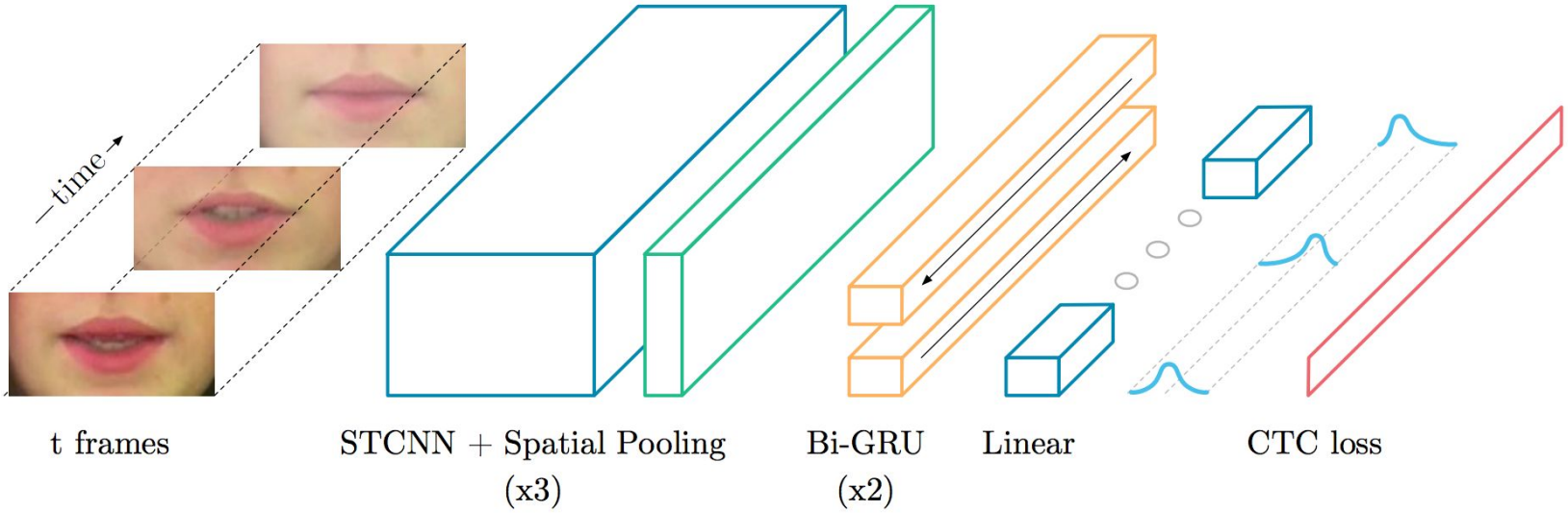


Figure 1: LipNet architecture. A sequence of T frames is used as input, and is processed by 3 layers of STCNN, each followed by a spatial max-pooling layer. The features extracted are processed by 2 Bi-GRUs; each time-step of the GRU output is processed by a linear layer and a softmax. This end-to-end model is trained with CTC.

Assael, Yannis M., Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. "LipNet: Sentence-level Lipreading." *arXiv preprint arXiv:1611.01599* (2016).

Speech and Video: LipNet

Assael, Yannis M., Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. "LipNet: Sentence-level Lipreading." *arXiv preprint arXiv:1611.01599* (2016).

Speech and Video: Watch, Listen, Attend & Spell

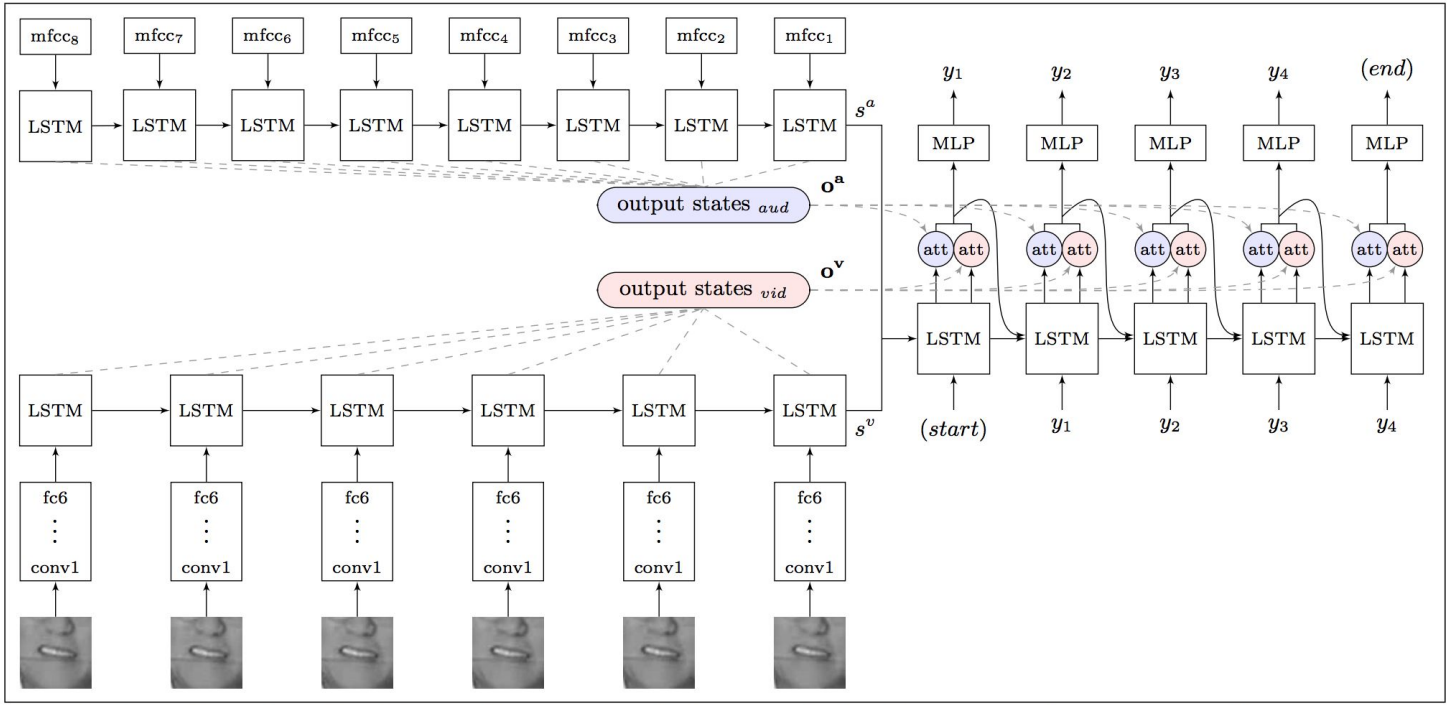


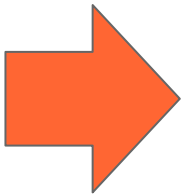
Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character y_i, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

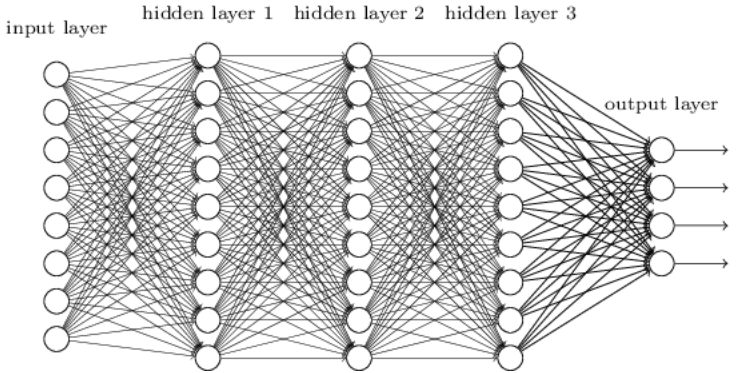
Chung, Joon Son, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. "Lip reading sentences in the wild." *arXiv preprint arXiv:1611.05358* (2016).

Speech and Video: Watch, Listen, Attend & Spell

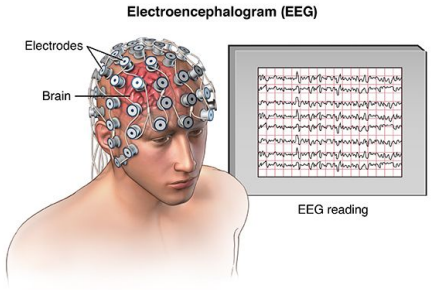
Chung, Joon Son, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. "Lip reading sentences in the wild." *arXiv preprint arXiv:1611.05358* (2016).

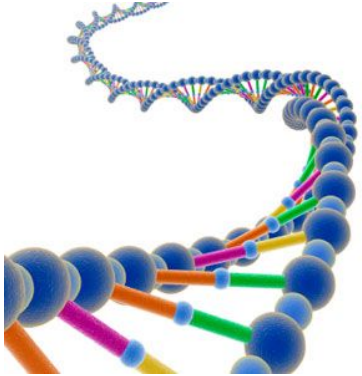
Conclusions

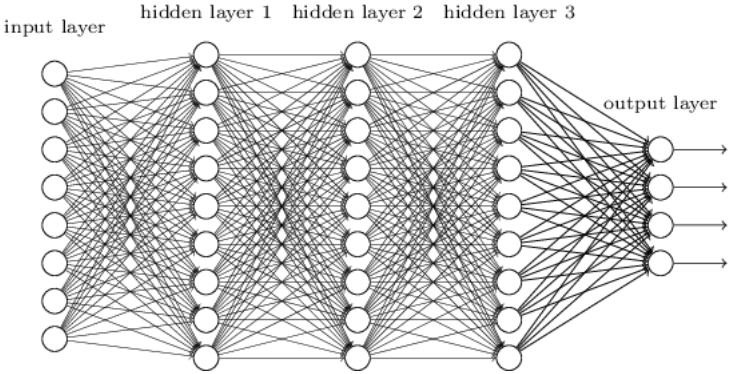




Conclusions







Conclusions

[course site]

TALP UPC

UPC

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

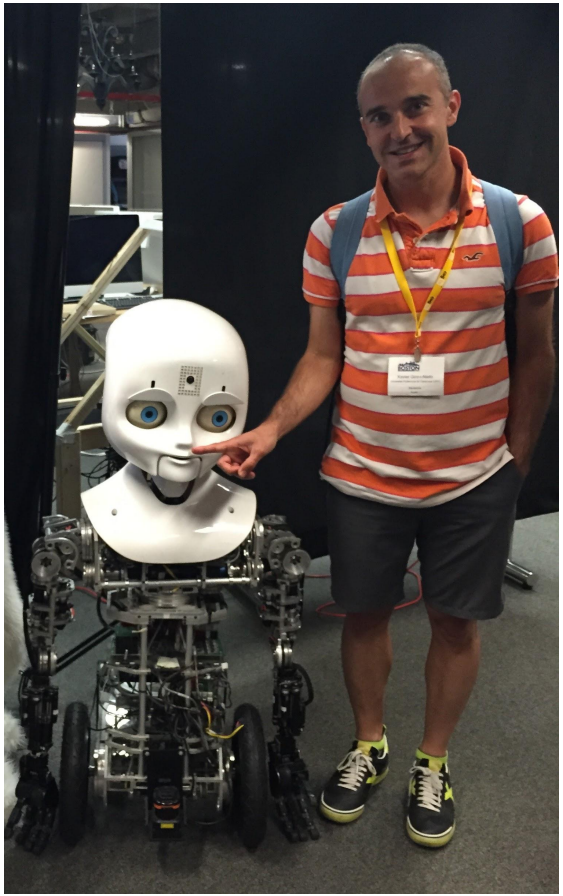
Department of Signal Theory and Communications

Image Processing Group

Learn more

Ruus Salakhutdinov, “Multimodal Machine Learning” (NIPS 2015 Workshop)

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