Insight
Centre for Data Analytics

DEEP LEARNING WORKSHOP

Dublin City University
21-22 May 2018

DCU

Day 2 Lecture 12

Language & Vision

Amaia Salvador
amaia.salvador@upc.edu

PhD Candidate
Universitat Politècnica de Catalunya

UPC

Language & Vision

Caption this picture:

Language & Vision

Caption this picture:

“a bed decorated in black and white bedding”
“a large white bed sitting under two framed pictures”
“a bedroom scene with a bed an television on the wall”

Language & Vision

"Two children riding a horse in front of their home"

"a group of sheep trailing one another in a line"

Language & Vision

"Two children riding a horse in front of their home"

"a group of sheep trailing one another in a line"

Language & Vision

What is the mustache made of?

How many slices of pizza are there?

Outline

1. Image Captioning
2. Visual Question Answering
3. Cross-Modal Embeddings
4. Image Generation from Text

Outline

1. **Image Captioning**
2. Visual Question Answering
3. Cross-Modal Embeddings
4. Image Generation from Text

Image Captioning

Lorem ipsum dolor
sit amet, consectetur adipiscing elit,
sed do eiusmod tempor incididunt

eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Lorem ipsum dolor sit amet, consectetur adipiscing

aliquip ex ea commodo consequat.

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Lorem

Text (English)

Neural Machine Translation

Lorem ipsum dolor
sit amet, consectetur adipiscing elit,
sed do eiusmod tempor incididunt

eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Lorem ipsum dolor sit amet, consectetur adipiscing

aliquip ex ea commodo consequat.

Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. Lorem ipsum sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Lorem

Text (French)

Image Captioning

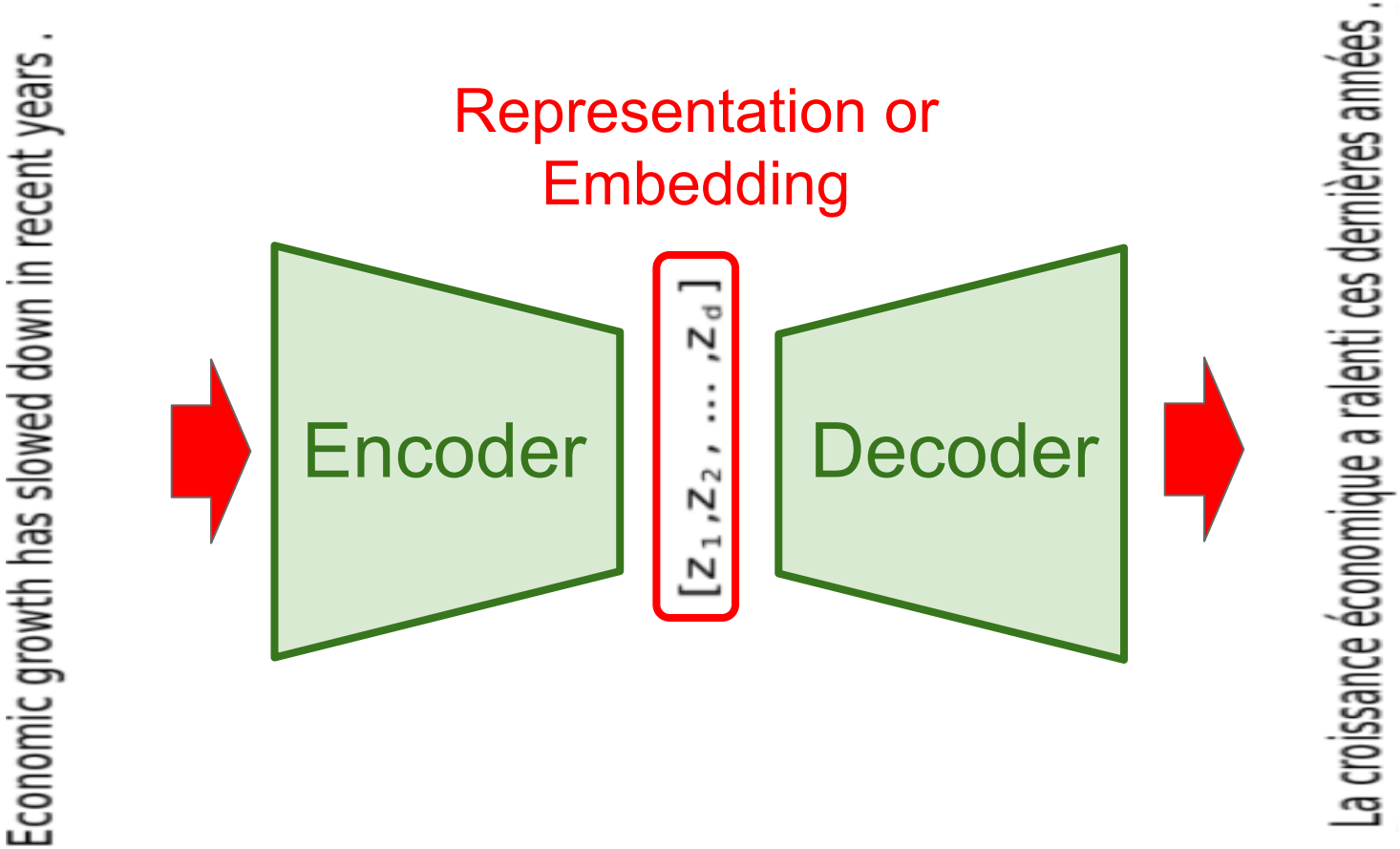
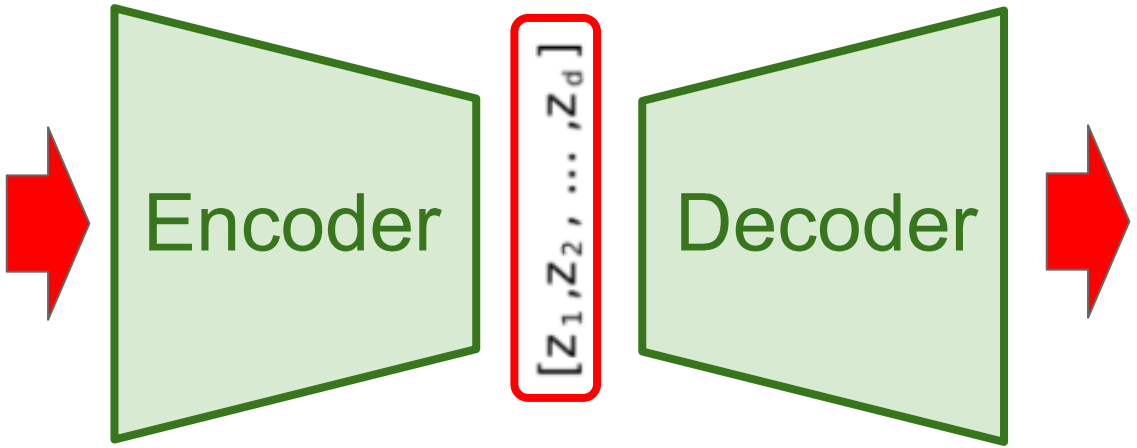


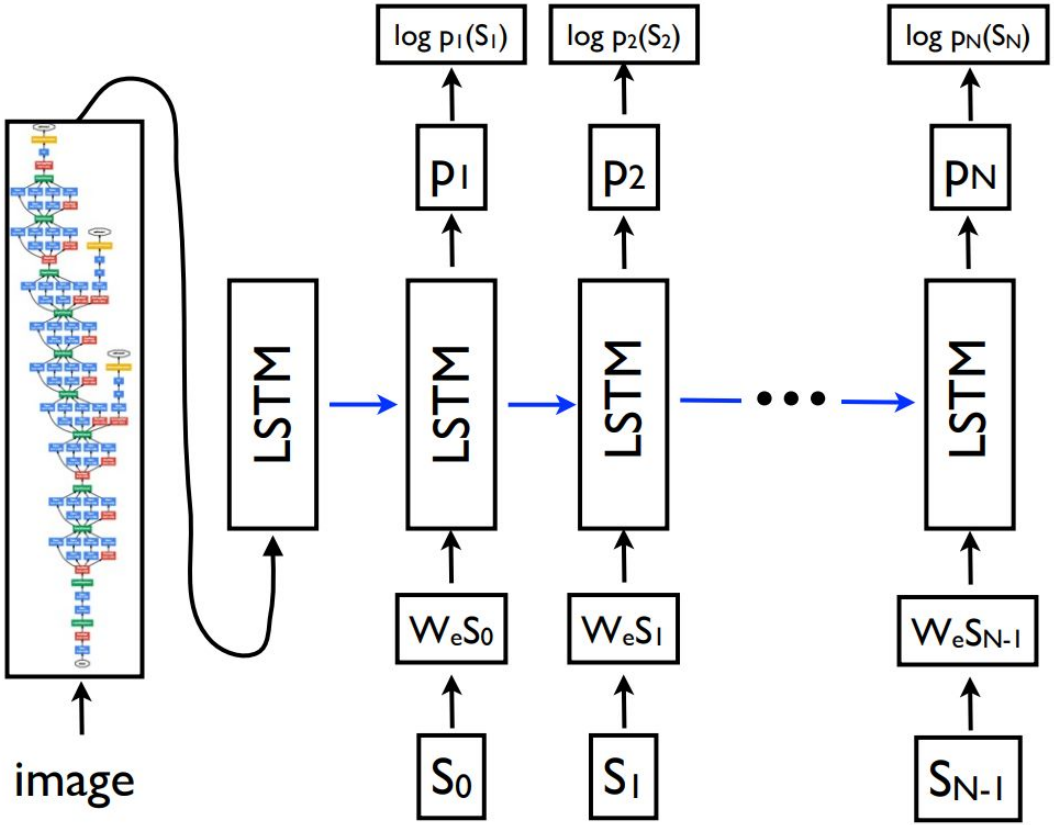
Image Captioning

Representation or Embedding



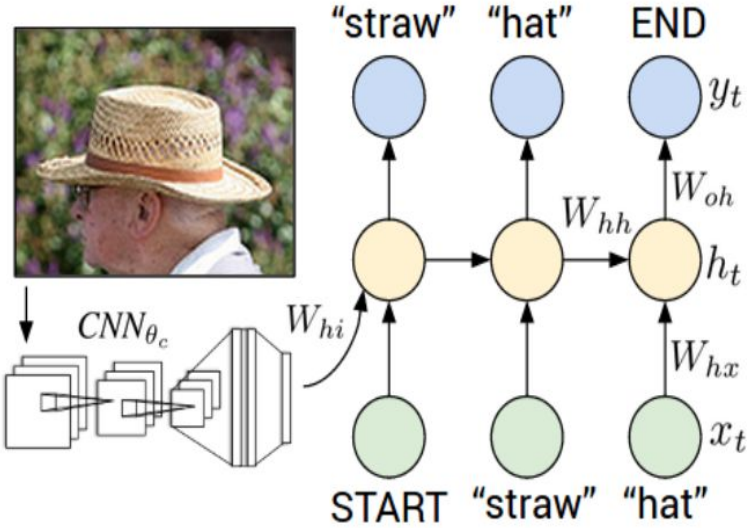
a bed decorated in black and white bedding

Image Captioning



Vinyals et al. "Show and tell: A neural image caption generator." CVPR 2015.

Image Captioning

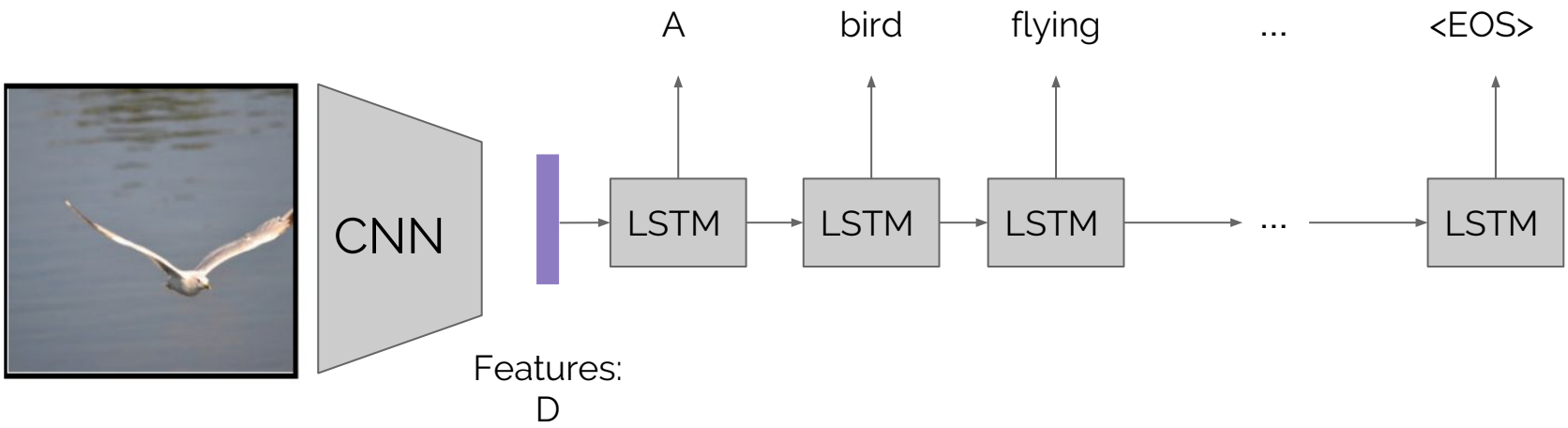


only takes into account image features in the first time step

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

Karpathy et al. "Deep visual-semantic alignments for generating image descriptions." CVPR 2015

Image Captioning

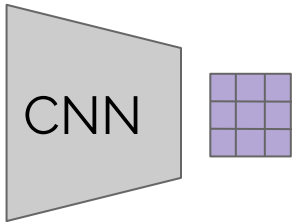


Limitation:

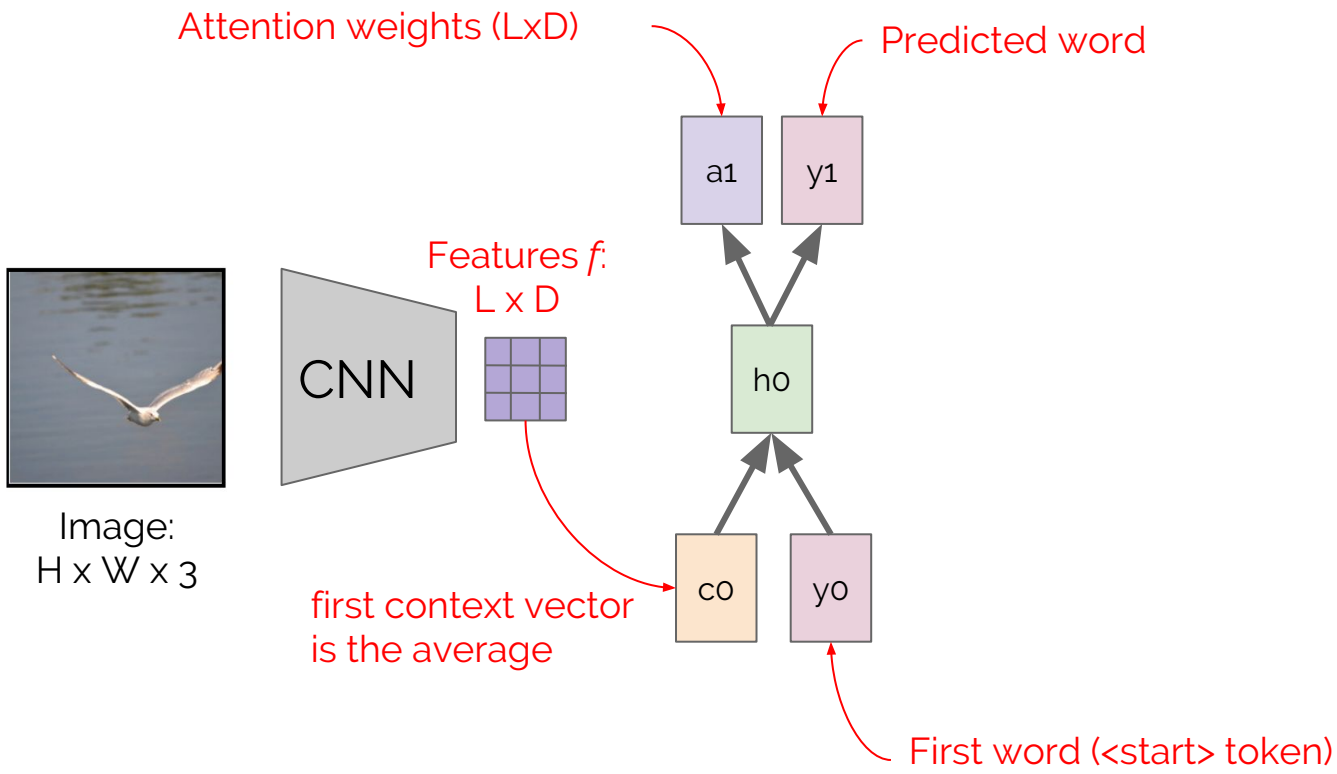
All output predictions are based on the **final and static** output of the encoder

Visual Attention for Image Captioning

Image:
H x W x 3



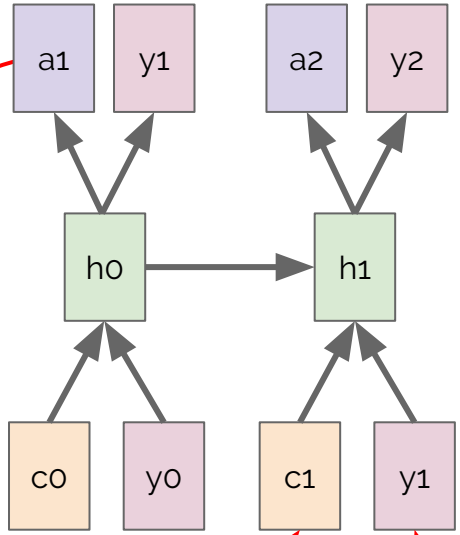
Visual Attention for Image Captioning



Visual Attention for Image Captioning

Image:
H x W x 3

CNN



$$c_i = \sum_{j=1}^{LD} f_j a_{ij}$$

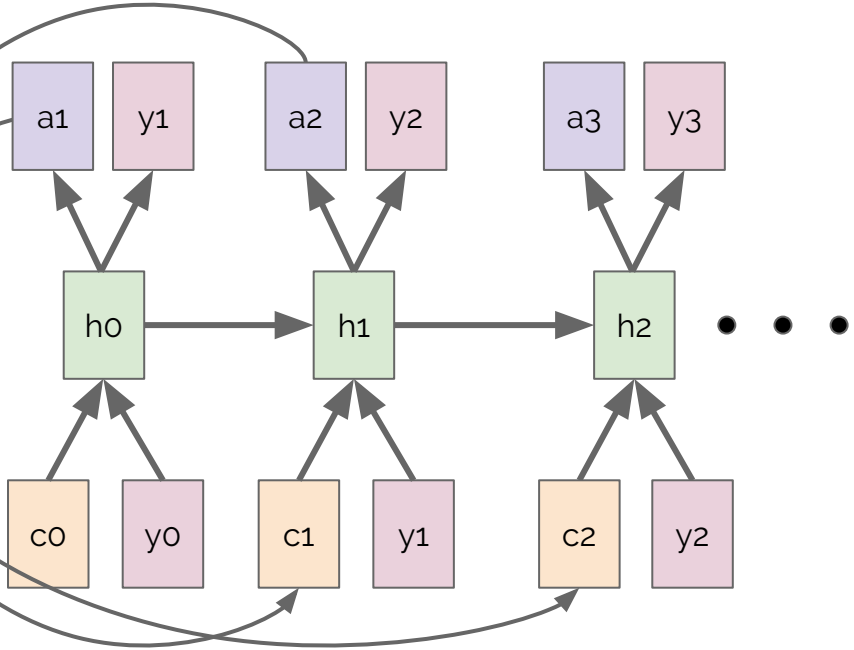
Visual features weighted with attention give the next context vector

Predicted word in previous timestep

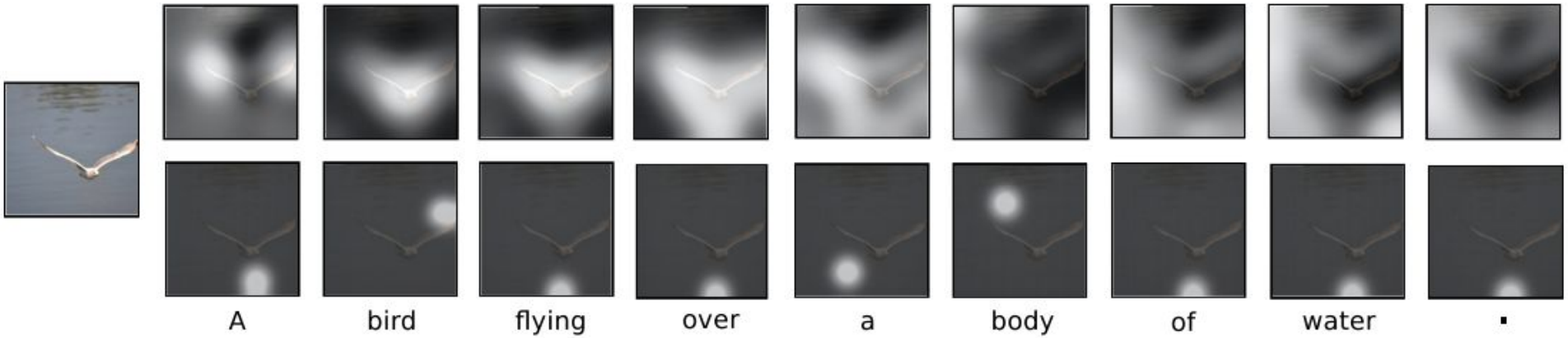
Visual Attention for Image Captioning

Image:
H x W x 3

CNN



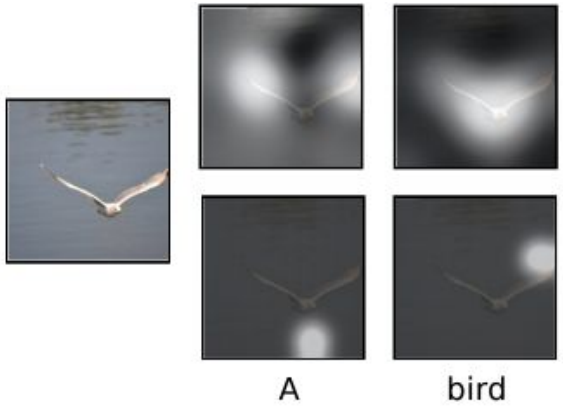
Visual Attention for Image Captioning



Xu et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML 2015

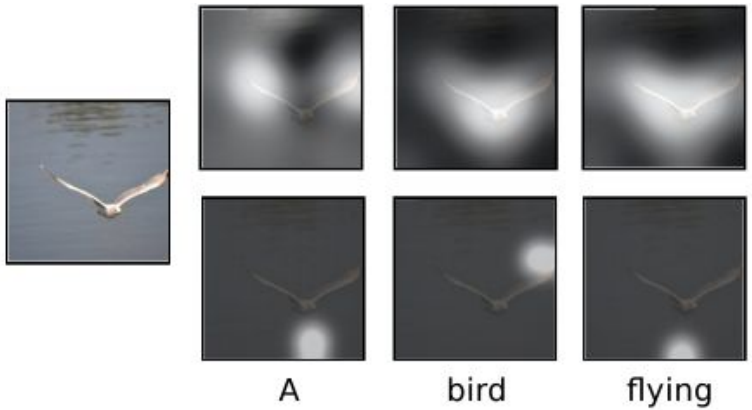
Visual Attention for Image Captioning

Some outputs can probably be predicted without looking at the image...



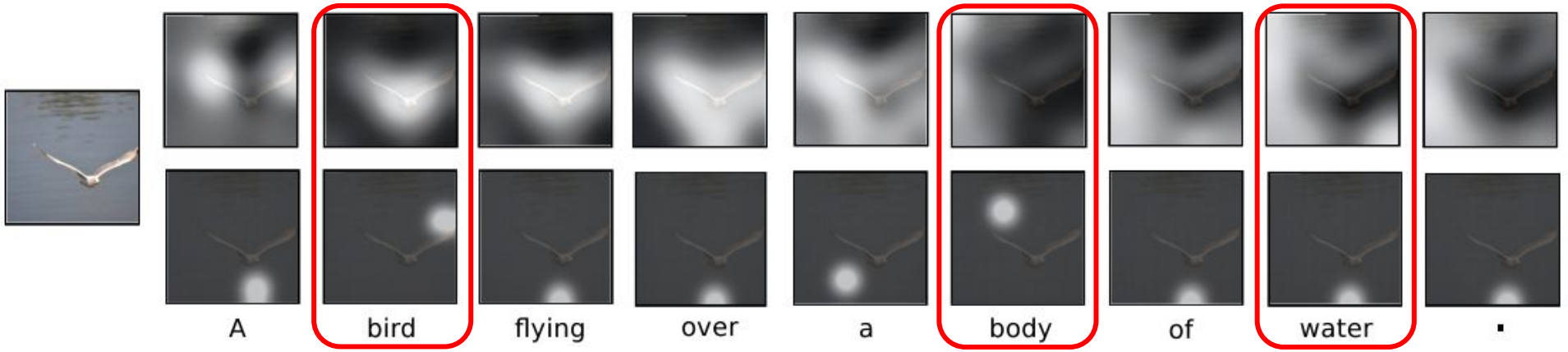
Visual Attention for Image Captioning

Some outputs can probably be predicted without looking at the image...



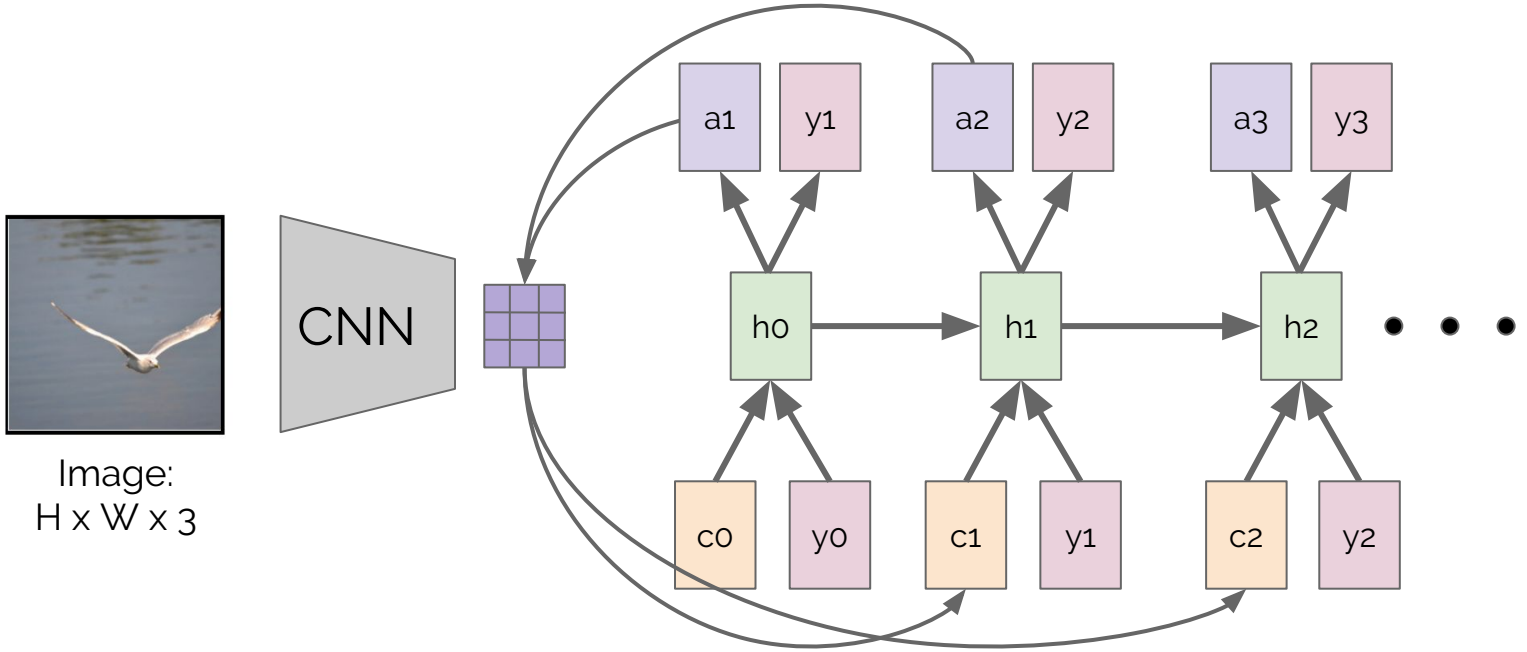
Visual Attention for Image Captioning

Can we focus on the image only when necessary?



Visual Attention for Image Captioning

"Regular" spatial attention



Visual Attention for Image Captioning

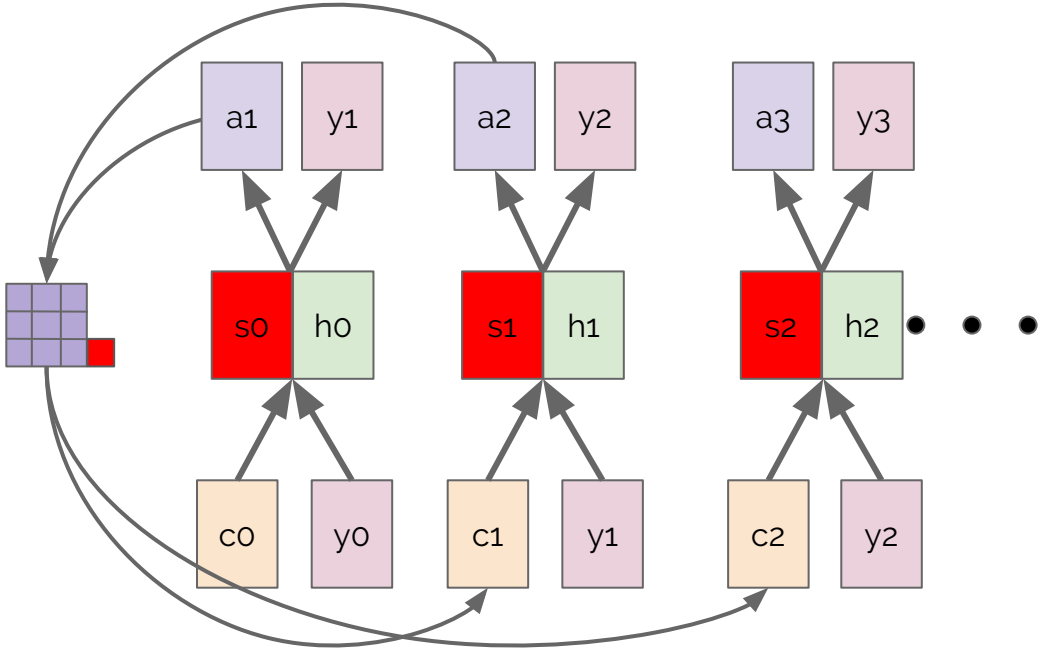
Attention with sentinel: LSTM is modified to output a “non-visual” feature to attend to

$$f = [f; s_i]$$

$$c_i = \sum_{j=1}^{LD+1} f_j a_{ij}$$

Image:
H x W x 3

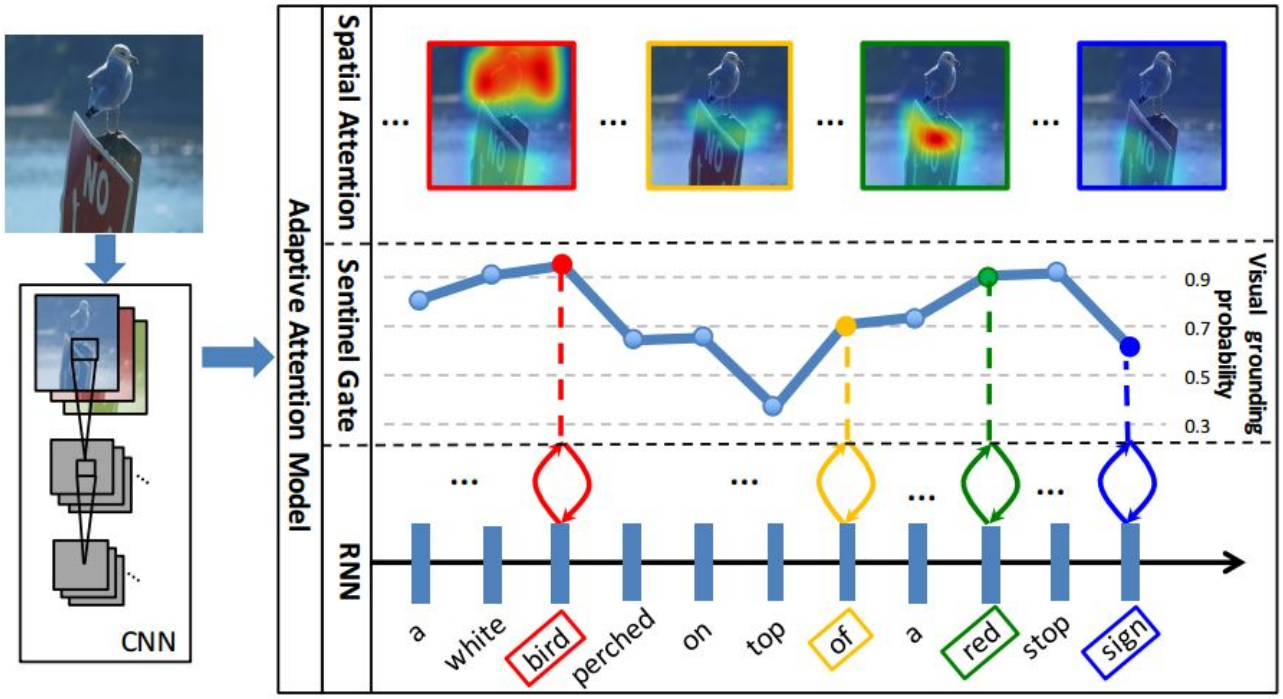
CNN



Lu et al. [Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning.](#) CVPR 2017

Visual Attention for Image Captioning

Attention weights indicate when it's more important to look at the image features, and when it's better to rely on the current LSTM state



Lu et al. [Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning.](#) CVPR 2017

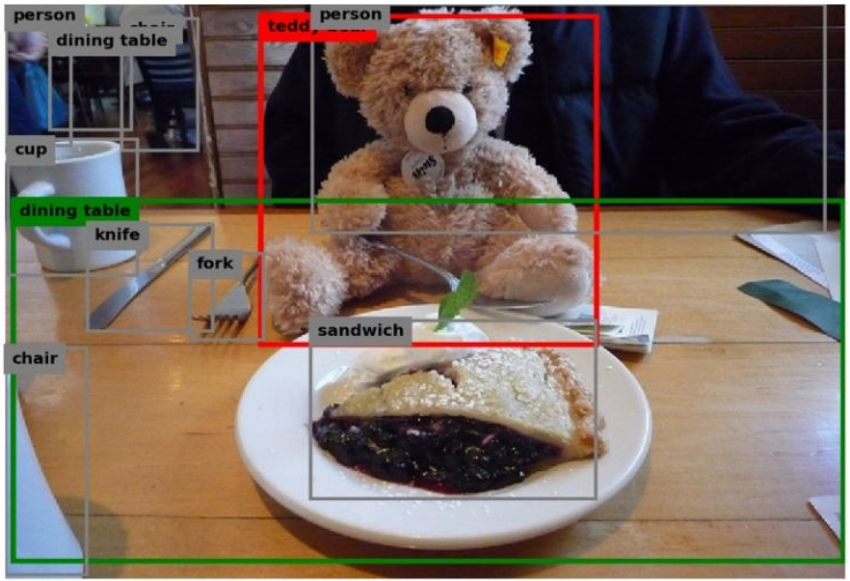
Grounded Image Captioning

Typical captioning models tend to produce plausible generic captions based on image features

A close up of a stuffed animal on a plate.

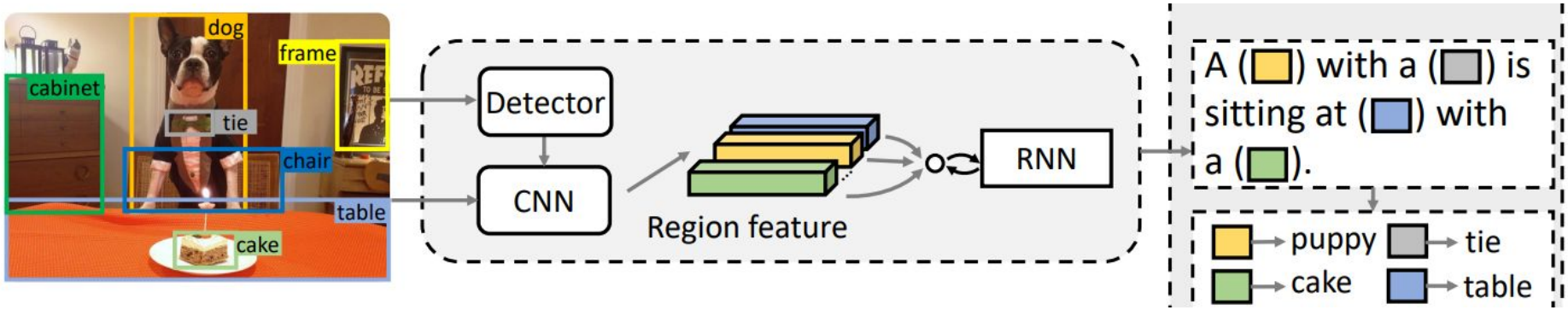
Grounded Image Captioning

The outputs of an object detector can assist the language model to give more accurate descriptions



A teddy bear sitting on a table with a plate of food.

Grounded Image Captioning

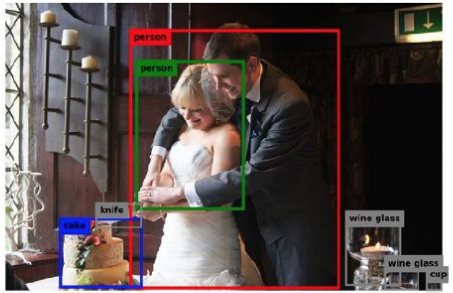


Slot-filling approach: generating a sentence template with empty slots to be filled using the outputs of an object detection model

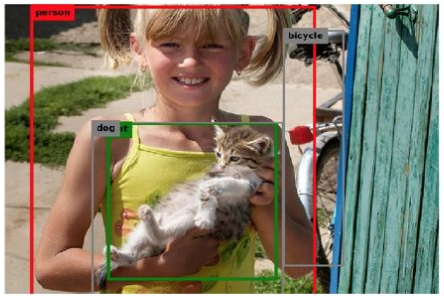
Lu et al. "Neural Baby Talk" CVPR 2018

Grounded Image Captioning

A dog is laying in the grass with a Frisbee.



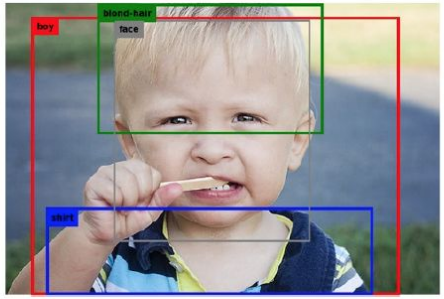
A bride and groom cutting a cake together.



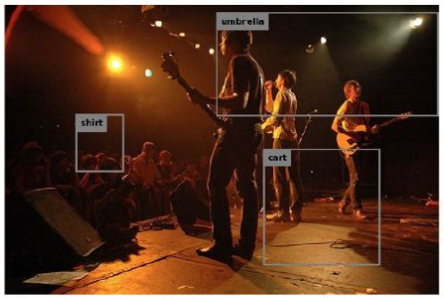
A little girl holding a cat in her hand.

A woman sitting on a boat in the water.

A cat is standing on a sign that says “UNK”.



A young boy with blond-hair and a blue shirt is eating a chocolate



A band is performing on a stage.

Two people are sitting on a boat in the water.

Lu et al. [“Neural Baby Talk”](#) CVPR 2018

Image Captioning Evaluation

Table 1: A summary of the evaluation metrics considered in this study.

Metric	Proposed to evaluate	Underlying idea
BLEU (Papineni et al., 2002)	Machine translation	n-gram precision
ROUGE (Lin, 2004)	Document summarization	n-gram recall
METEOR (Banerjee and Lavie, 2005)	Machine translation	n-gram with synonym matching
CIDEr (Vedantam et al., 2015)	Image description generation	*tf-idf* weighted n-gram similarity
SPICE (Anderson et al., 2016)	Image description generation	Scene-graph synonym matching
WMD (Kusner et al., 2015)	Document similarity	Earth Mover Distance on *word2vec*

Table 2: Drawbacks of automatic evaluation metrics for image captioning. See text for details.

	Description	BLEU	METEOR	ROUGE	CIDEr	SPICE	WMD
original	a man wearing a red life jacket is sitting in a canoe on a lake	1	1	1	10	1	1
candidate	a man wearing a life jacket is in a small boat on a lake	0.45	0.28	0.67	2.19	0.40	0.19
synonyms	a **guy** wearing a life **vest** is in a small boat on a lake	**0.20**	**0.17**	0.57	**0.65**	**0.00**	**0.10**
redundancy	a man wearing a life jacket is in a small boat on a lake **at sunset**	0.45	0.28	0.66	2.01	0.36	0.18
word order	**in a small boat on a lake** a man is wearing a life jacket	**0.26**	0.26	**0.38**	**1.32**	0.40	0.19

Kilickaya et al. Re-evaluating Automatic Metrics for Image Captioning. arXiv preprint arXiv 2016.

Outline

1. Image Captioning
2. **Visual Question Answering**
3. Cross-Modal Embeddings
4. Image Generation from Text

Visual Question Answering (VQA)

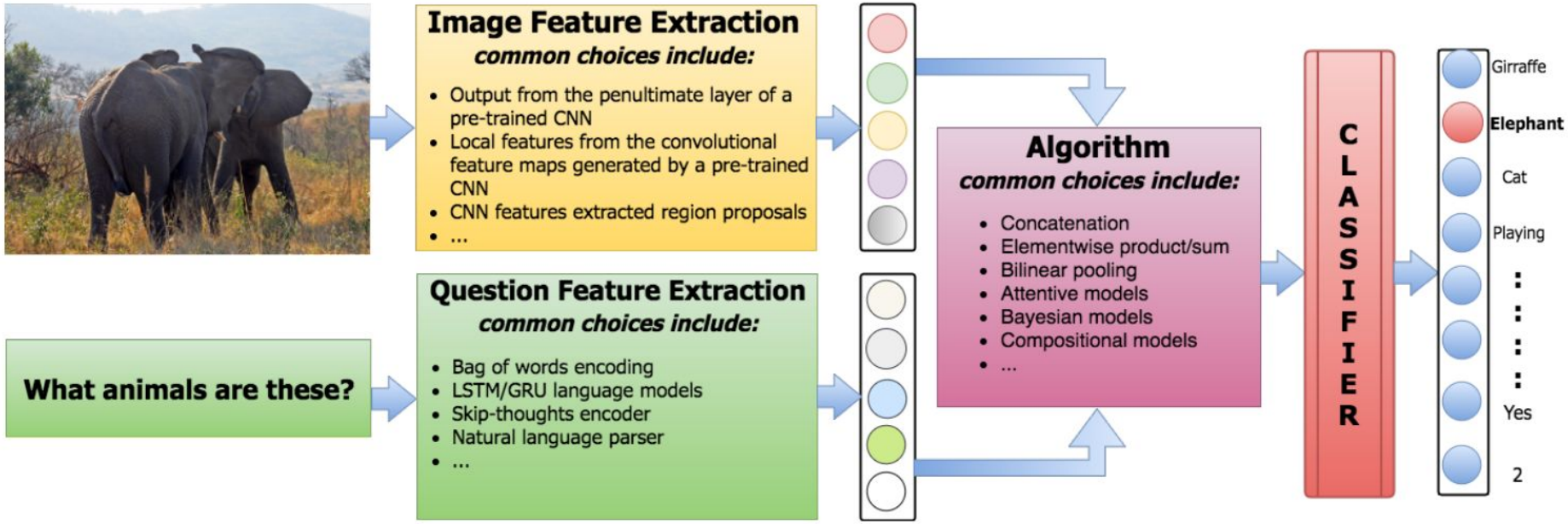
What is the mustache made of?

AI System

bananas

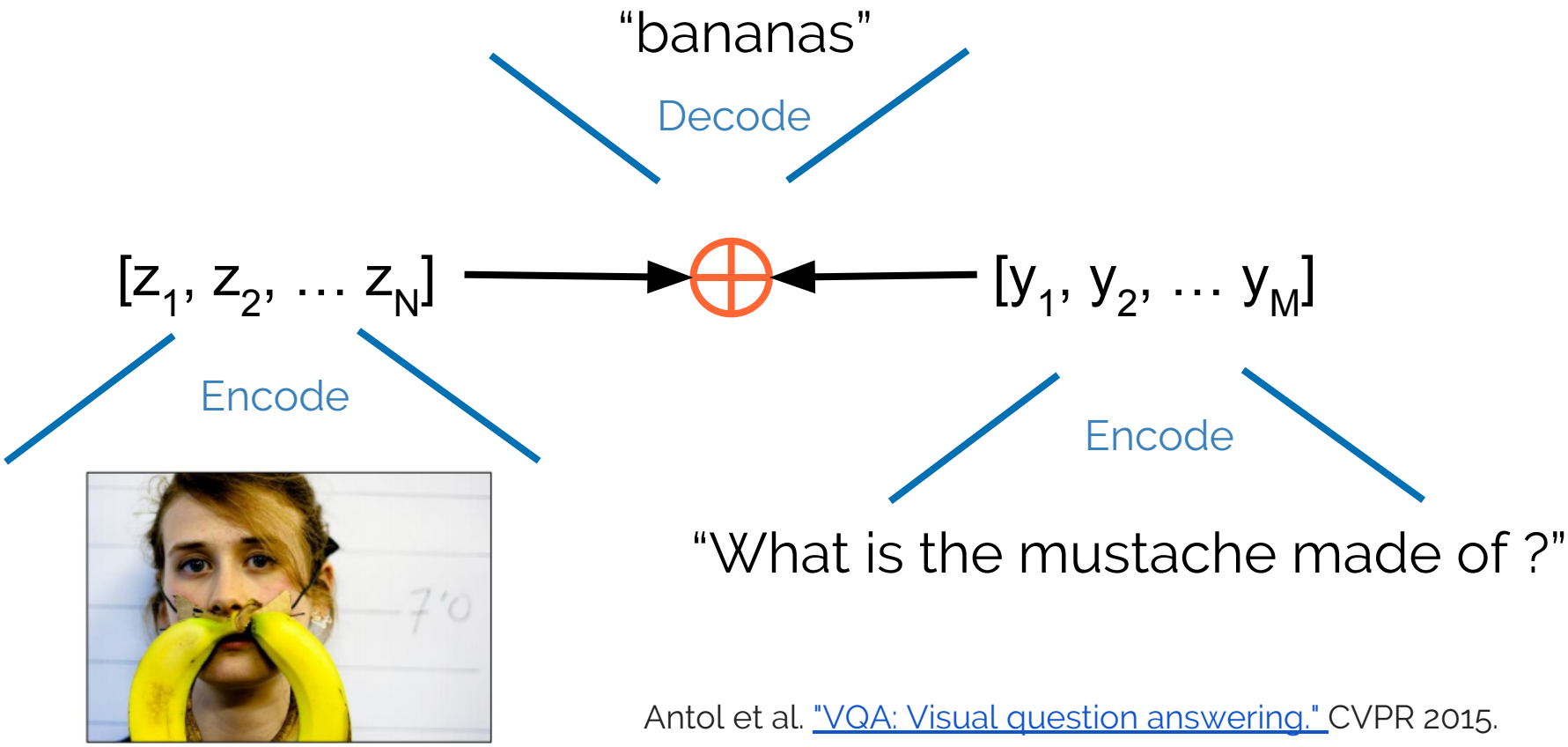
Antol et al. "VQA: Visual question answering." CVPR 2015.

Visual Question Answering (VQA)



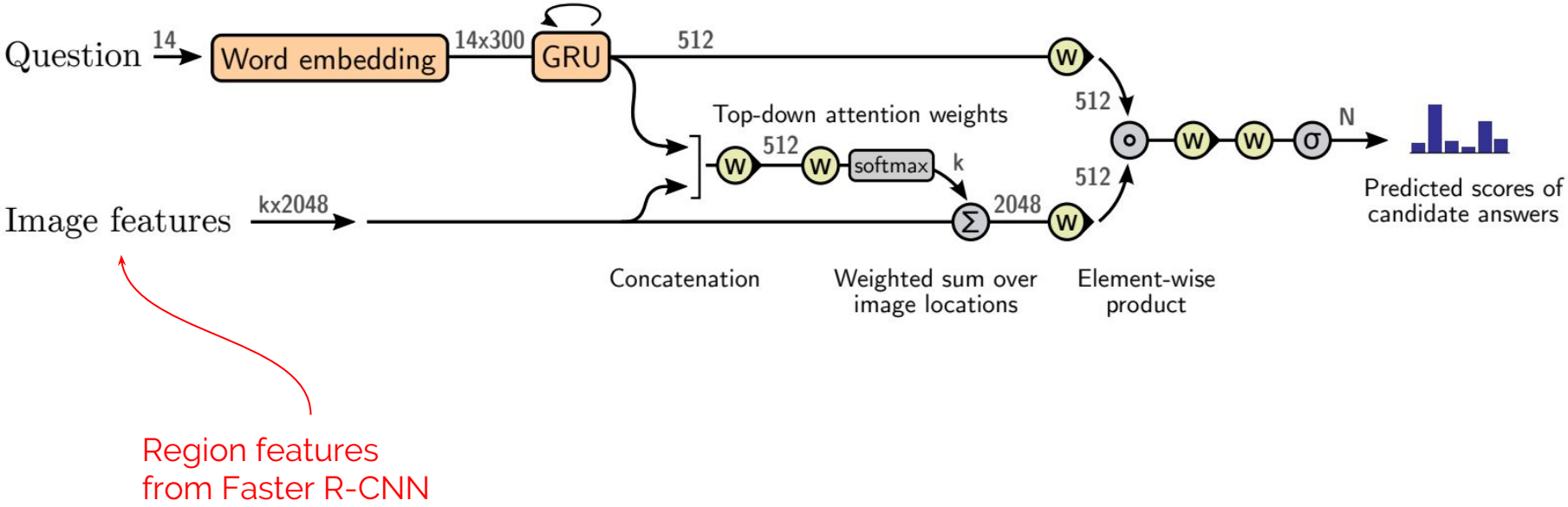
Kafle & Kanan. Visual Question Answering: Datasets, Algorithms, and Future Challenges. In Computer Vision and Image Understanding 2017

Visual Question Answering (VQA)



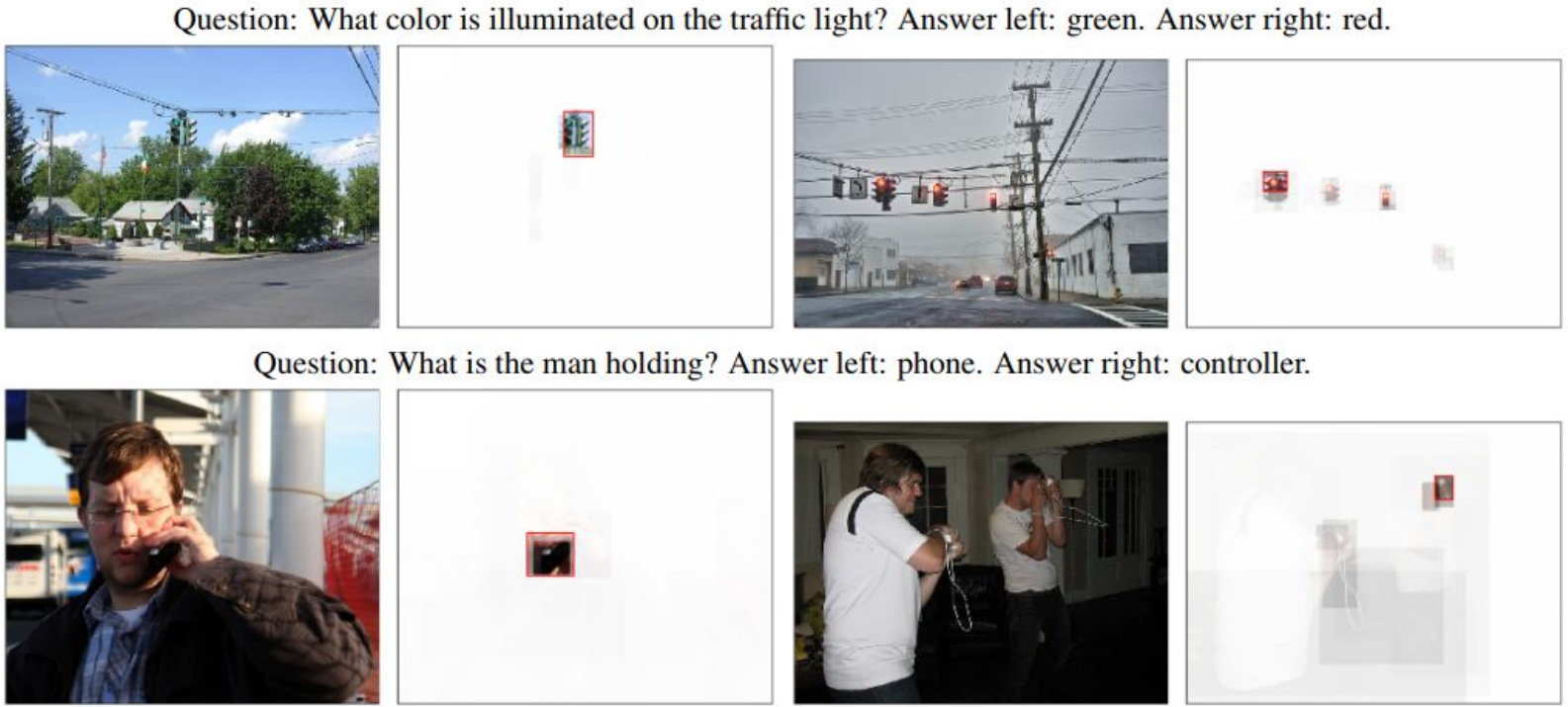
Antol et al. "VQA: Visual question answering." CVPR 2015.

Visual Question Answering (VQA)



Anderson et al. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. CVPR 2018

Visual Question Answering (VQA)



Anderson et al. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. CVPR 2018

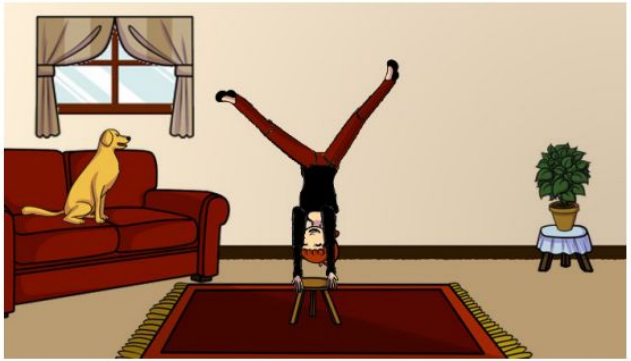
VQA Evaluation

$$\text{Acc}(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\}$$

Dataset bias in VQA

(a) Q: Would you like to fly in that? GT: yes (4x), no (6x). The VQA Dataset contains subjective questions that are prone to cause disagreement between annotators and also clearly lack a single objectively correct answer.

(b) Q: What color are the trees? GT: green. There are 73 total questions in the dataset asking this question. For 70 of those questions, the majority answer is green. Such questions can be often answered without information from the image.



(c) Q: Why would you say this woman is strong? GT: yes (5x), can lift up on arms, headstand, handstand, can stand on her head, she is standing upside down on stool. Questions seeking descriptive or explanatory answers can pose significant difficulty in evaluation.

Kafle & Kanan. Visual Question Answering: Datasets, Algorithms, and Future Challenges. In Computer Vision and Image Understanding 2017

Dataset bias in VQA

Who is wearing glasses?

man

woman

Where is the child sitting?

fridge

arms

Is the umbrella upside down?

yes

no

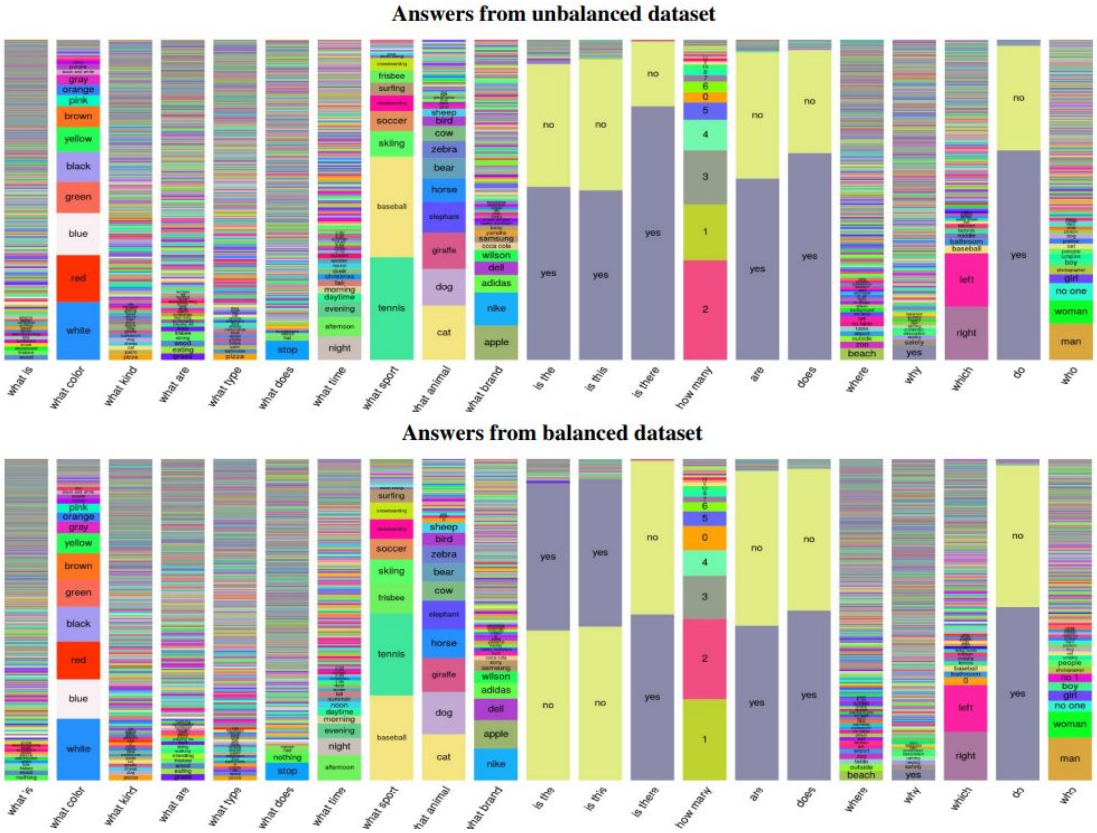
How many children are in the bed?

2

1

Goyal et al. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017

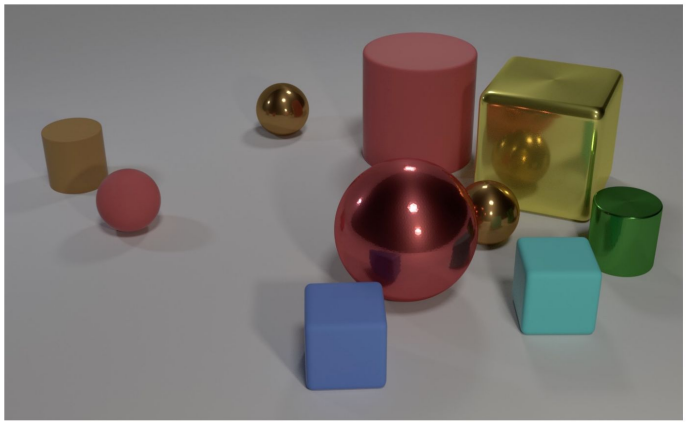
Dataset bias in VQA



Goyal et al. [Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering](). CVPR 2017

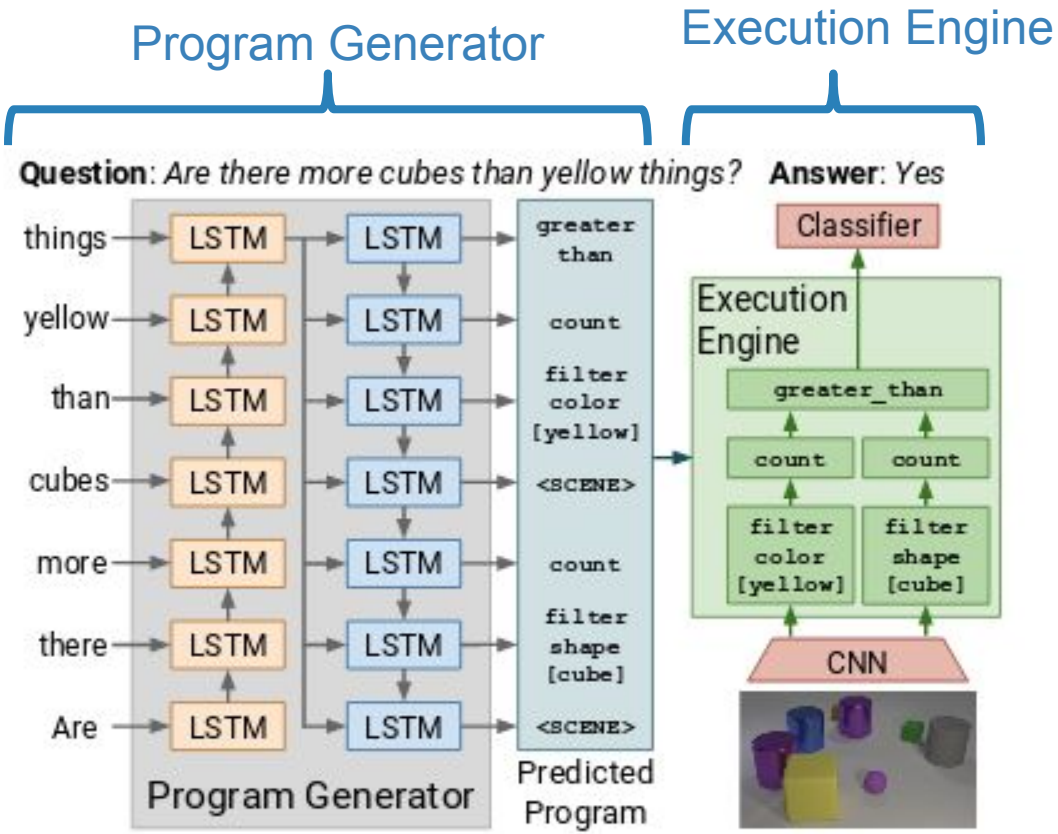
Visual Reasoning

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 et al. "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning." CVPR 2017

Visual Reasoning



Johnson et al. [“Inferring and Executing Programs for Visual Reasoning”](). ICCV 2017

Outline

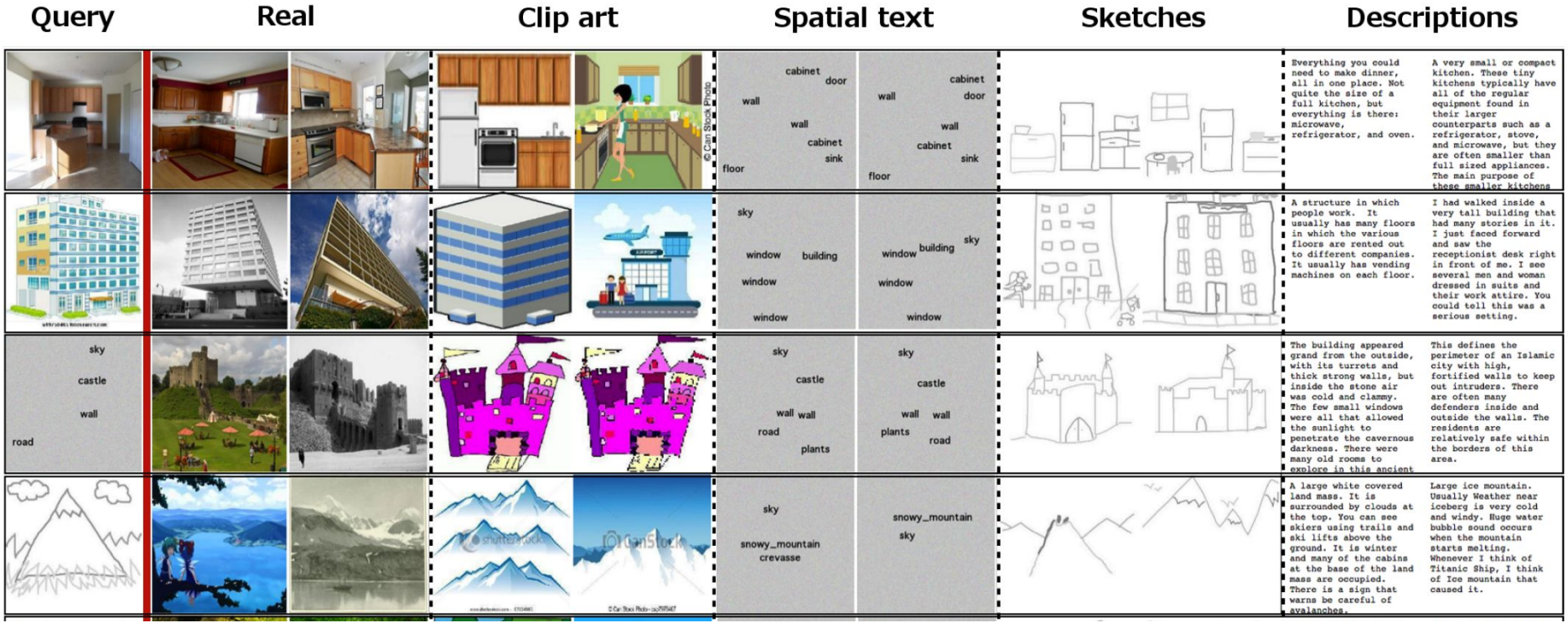
1. Image Captioning
2. Visual Question Answering
3. **Cross-Modal Embeddings**
4. Image Generation from Text

Cross-Modal Representations

"Two children riding a horse in front of their home"

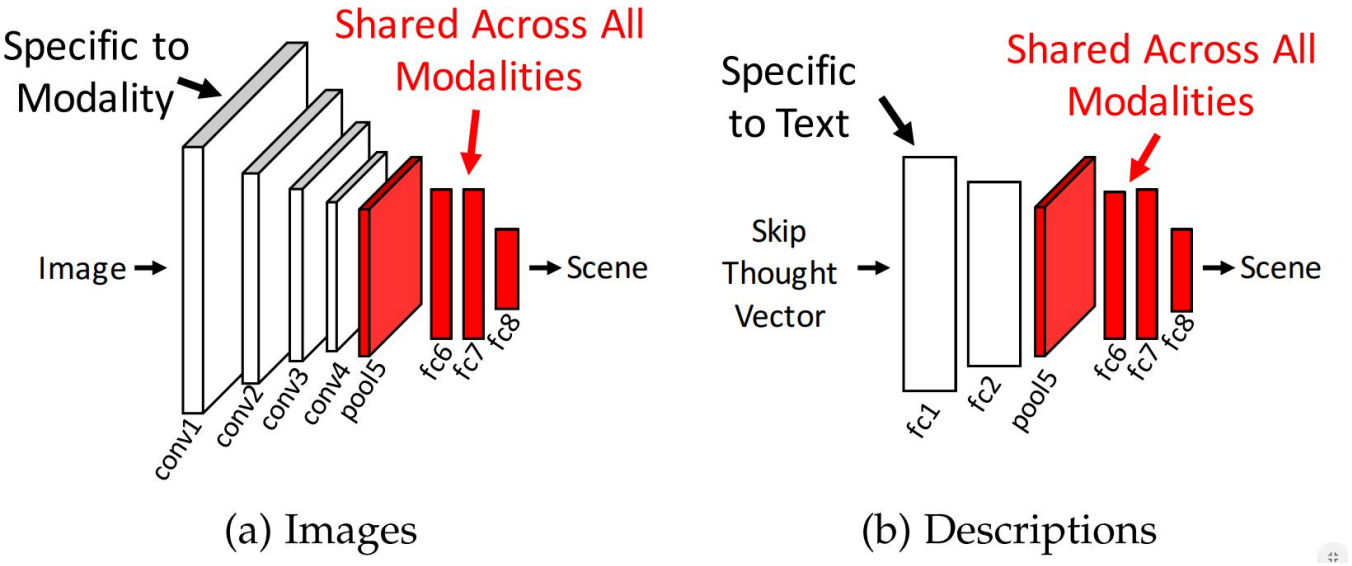
"a group of sheep trailing one another in a line"

Cross-Modal Representations



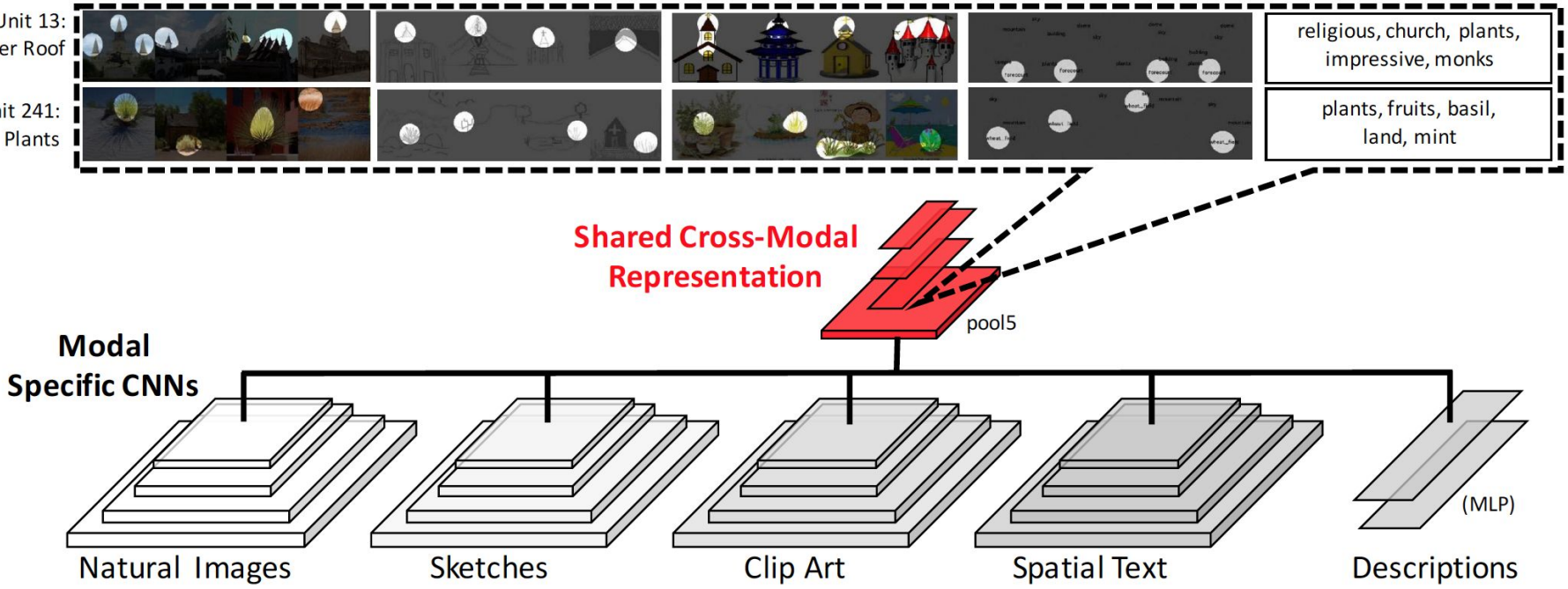
Castrejon et al. [Learning Aligned Cross-Modal Representations from Weakly Aligned Data.](#) CVPR 2016.

Cross-Modal Representations



Castrejon et al. [Learning Aligned Cross-Modal Representations from Weakly Aligned Data.](#) CVPR 2016.

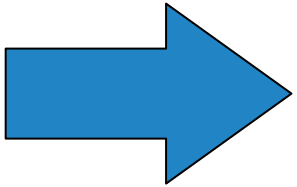
Cross-Modal Representations



Castrejon et al. [Learning Aligned Cross-Modal Representations from Weakly Aligned Data.](#) CVPR 2016.

Cross-Modal Representations

Image and text retrieval with joint neural embeddings



Ingredients

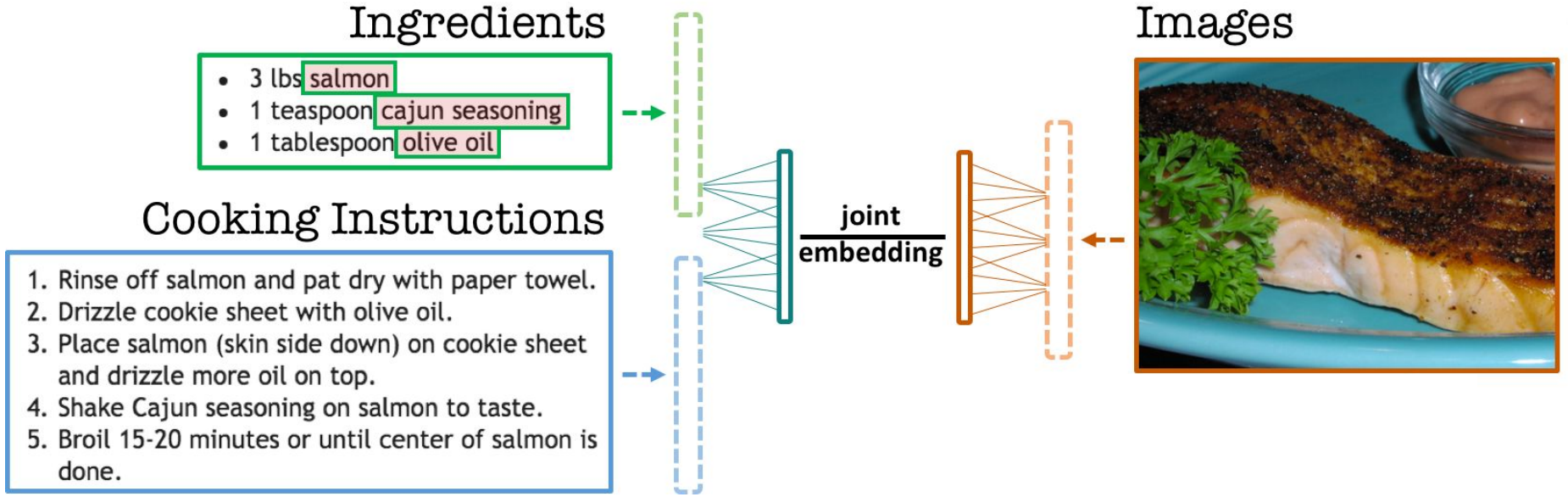
- 3 lbs salmon
- 1 teaspoon cajun seasoning
- 1 tablespoon olive oil

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.

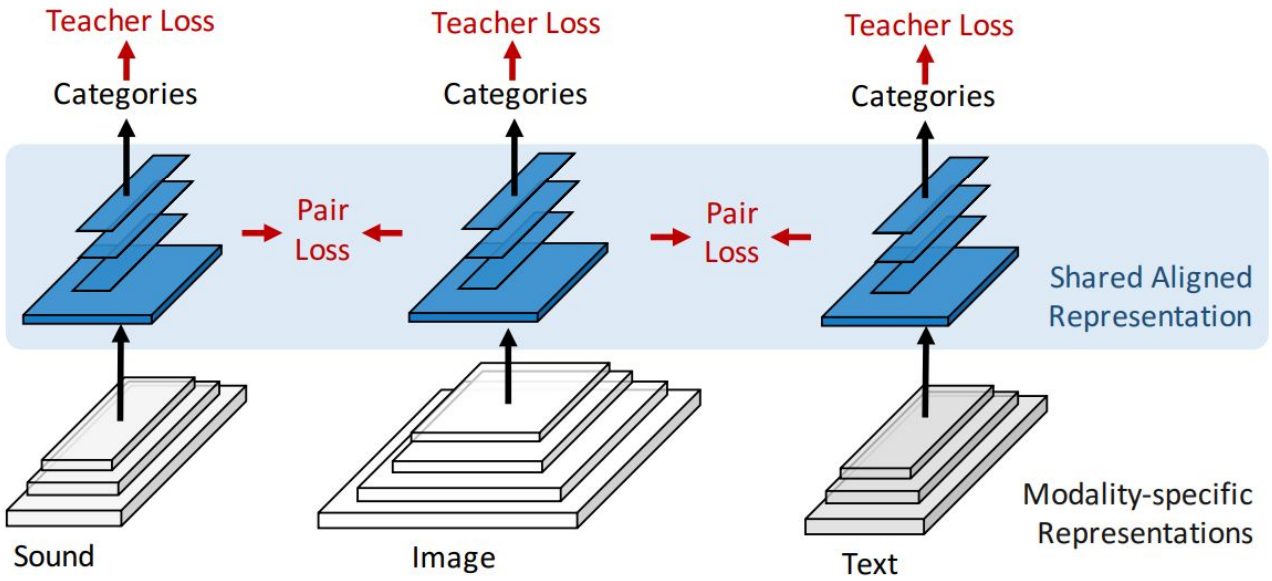
Salvador et al. "Learning Cross-modal Embeddings for Cooking Recipes and Food Images". CVPR 2017

Cross-Modal Representations



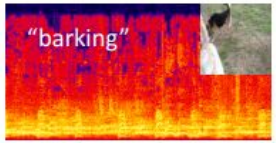
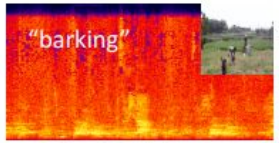
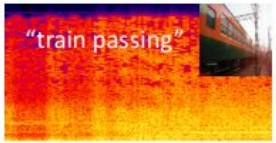
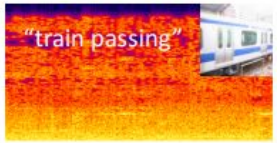
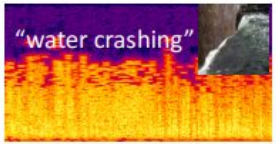
Salvador et al. "Learning Cross-modal Embeddings for Cooking Recipes and Food Images". CVPR 2017

Cross-Modal Representations



Aytar et al. "See, Hear and Read: Deep Aligned Representations". arXiv June 2017

Cross-Modal Representations

Input Query	Sound Retrievals	Text Retrievals	Image Retrievals
	 	- A dog lying down on the beach - The dog belongs to the homeowner	 
	 	- Steel tracks under the train. - The train platform	 
The choppy water the man is riding	 	- A person stands on water skis in the water - A couple of kayakers paddling through water	 

Aytar et al. "See, Hear and Read: Deep Aligned Representations". arXiv June 2017

Outline

1. Image Captioning
2. Visual Question Answering
3. Cross-Modal Embeddings
4. **Image Generation from Text**

Image Generation from Text

"Mark Zuckerberg wearing a flowery shirt"

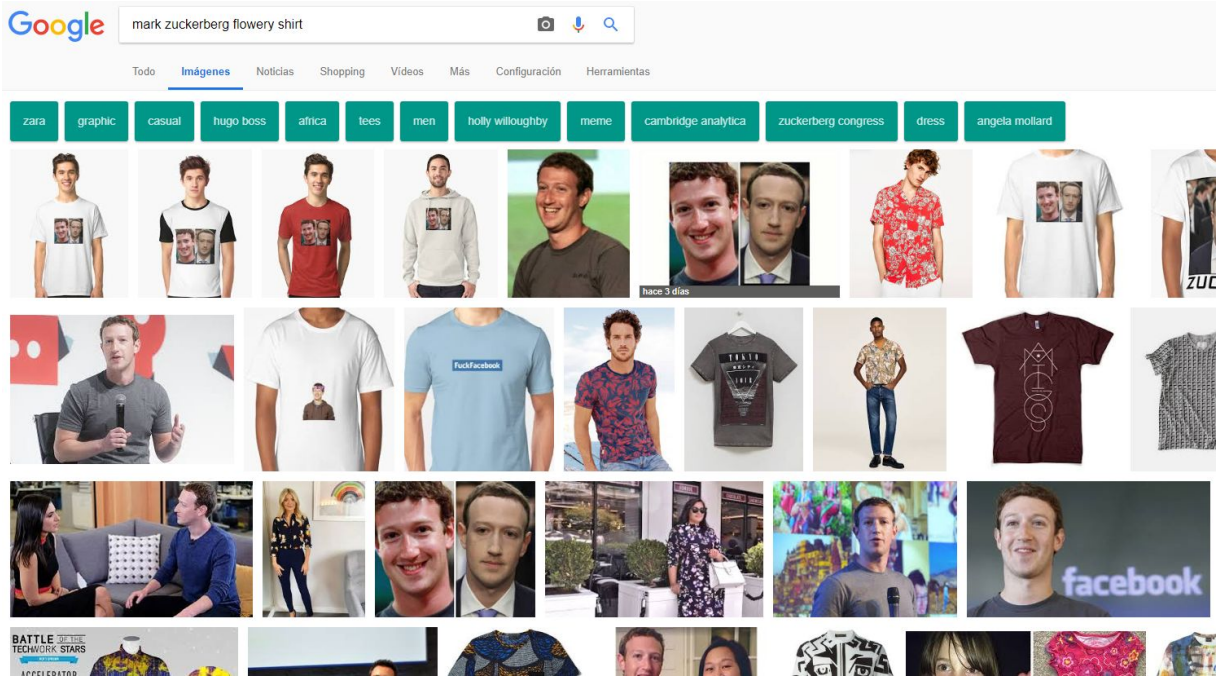
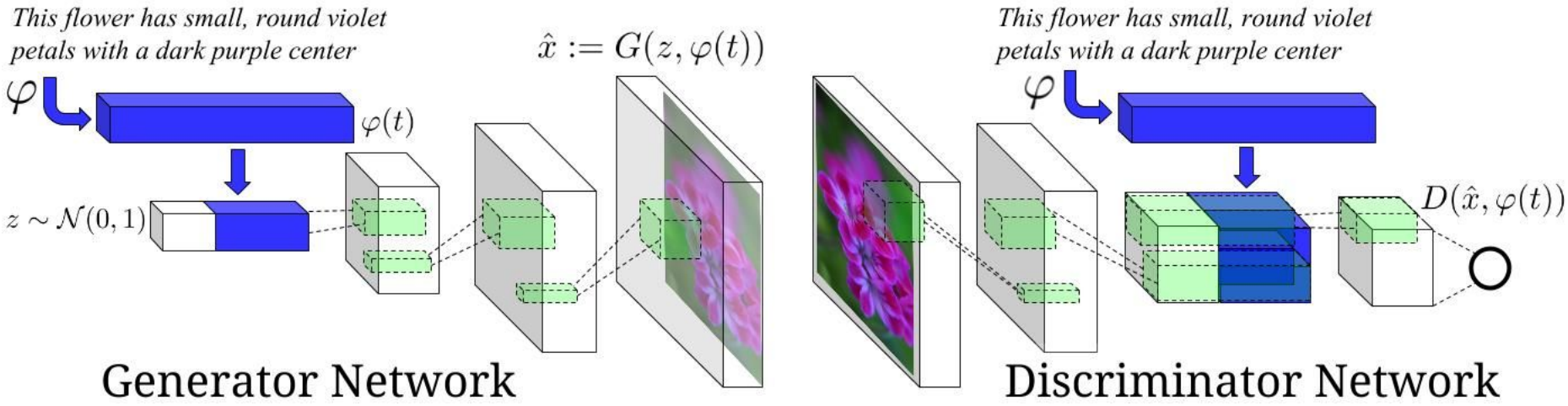


Image Generation from Text

Reed et al. "Generative adversarial text to image synthesis." ICML 2016.

Image Generation from Text

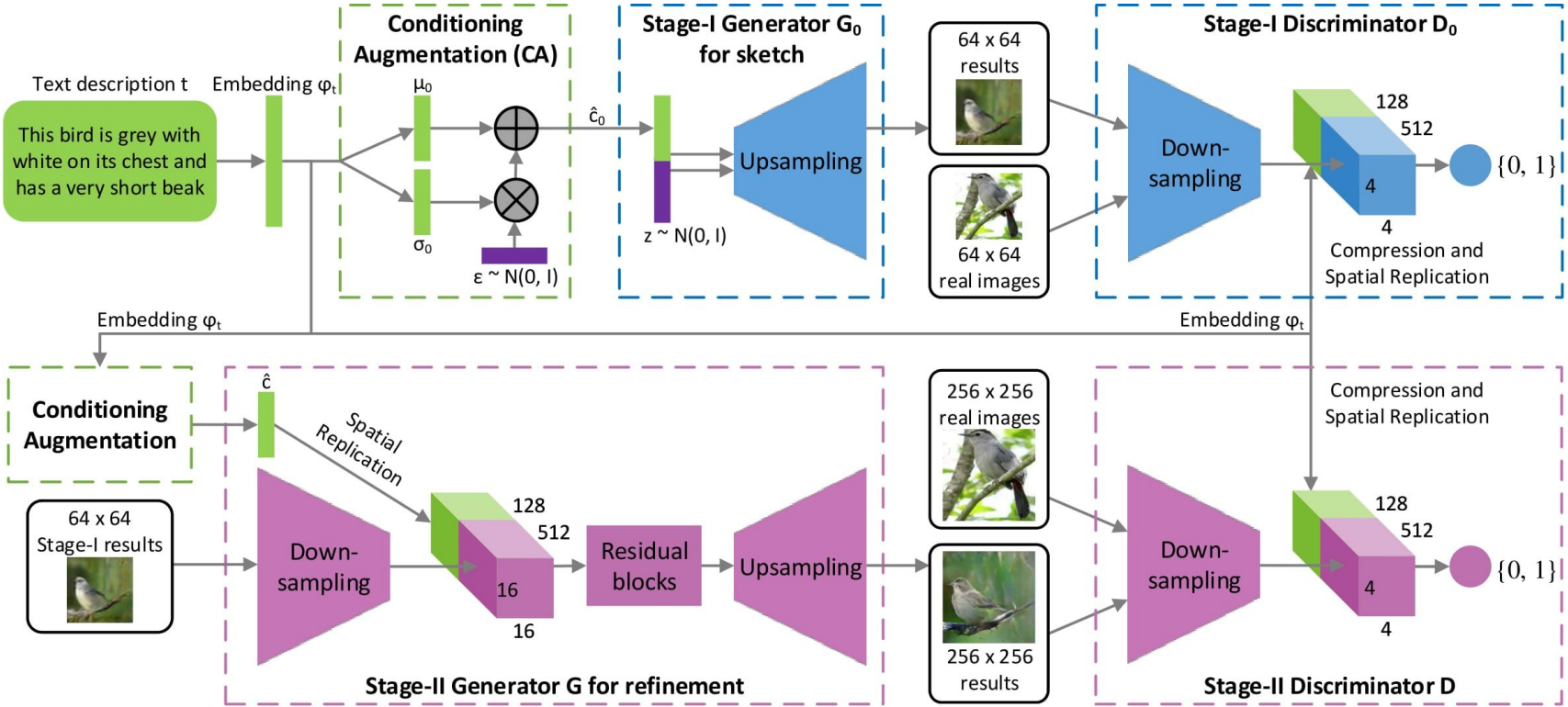


Reed et al. ["Generative adversarial text to image synthesis."](#) ICML 2016.

Image Generation from Text

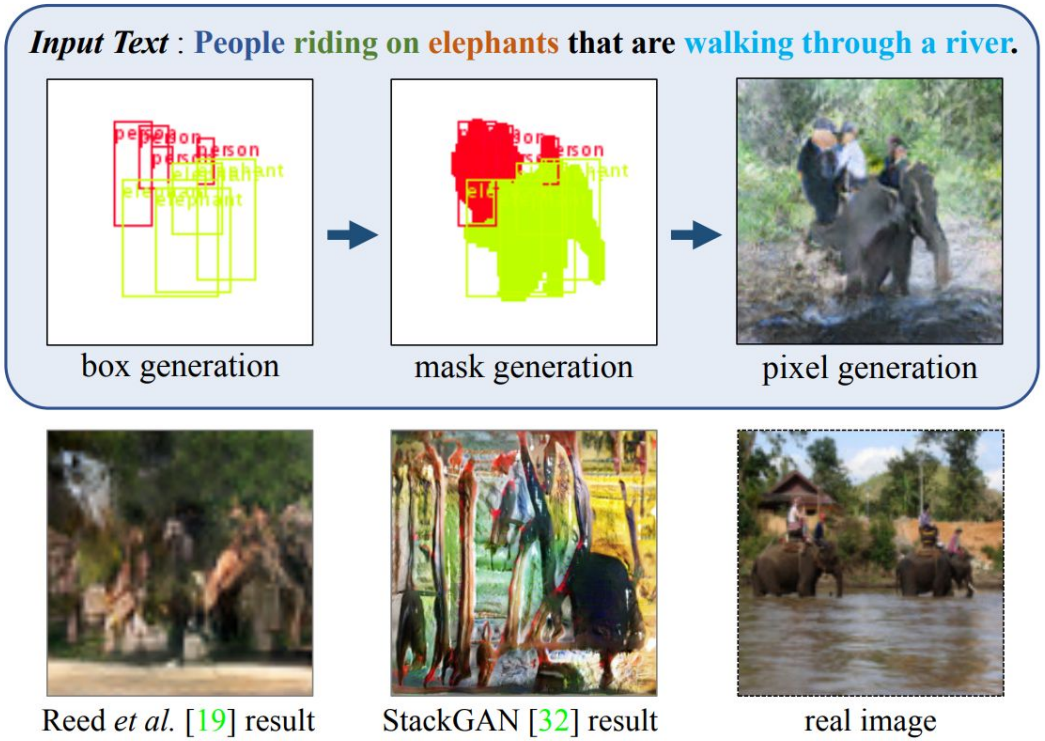
Zhang et al. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks." ICCV 2017

Image Generation from Text



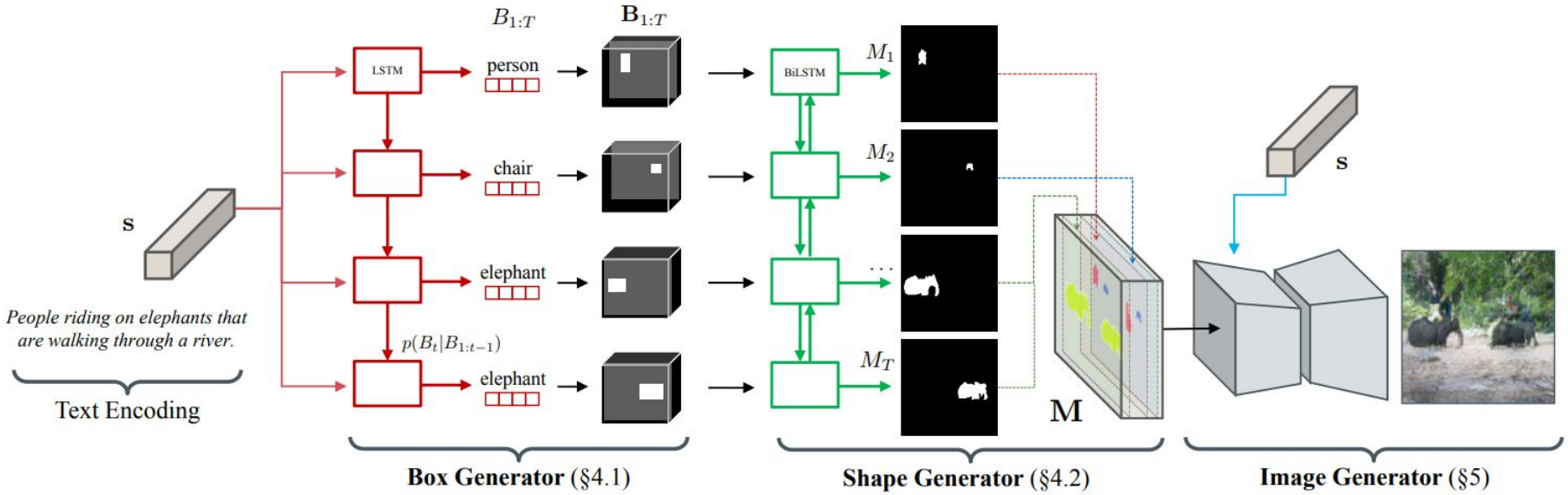
Zhang et al. ["Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks."](#) ICCV 2017

Image Generation from Text



Hong et al. Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis. CVPR 2018

Image Generation from Text



Hong et al. Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis. CVPR 2018

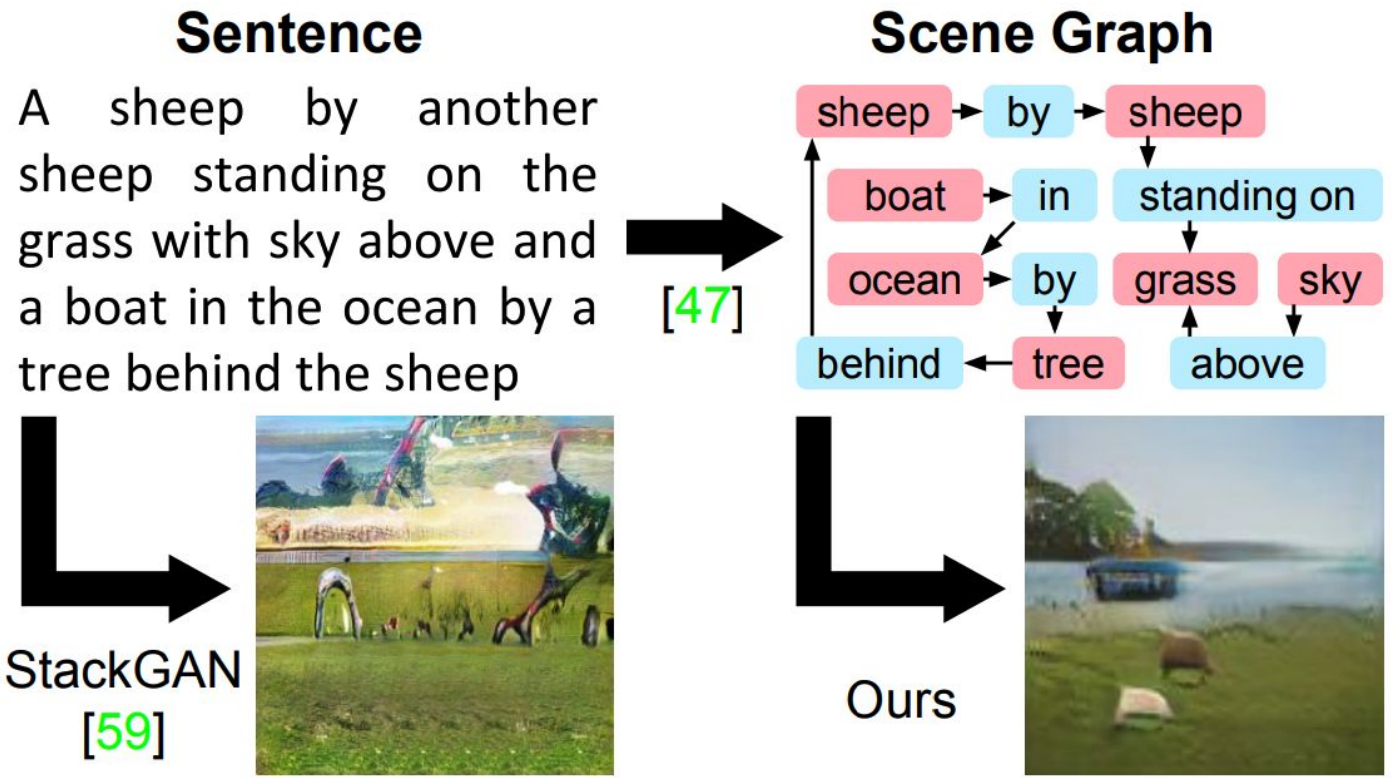
Image Generation from Text

Ground Truth		**(GT)** ***A kid in wet-suit on surfboard in the ocean.***		**(GT)** ***a lady that is on some skies on some snow***
	generated image and caption		generated image and caption	
StackGAN 256x256		a person flying a kite on a beach .		a man is walking on a beach with a surfboard .
Reed *et al.* 64x64		a man is flying a kite in the sky		a person is riding a snowboard on a snowy slope .
Ours 128x128		a man is surfing in the ocean with a surfboard .		a man is skiing down a hill with a snowboard .

		**(GT)** ***A young man playing frisbee while people watch.***		**(GT)** ***A bus that is sitting in the street.***
	generated image and caption		generated image and caption	
		a man is standing next to a cow .		a city street with a traffic light and a green light .
		a group of people standing around a field with kites .		a large boat is in the water near a city .
		a man is playing with a frisbee in a field .		a red and white bus parked on a city street .

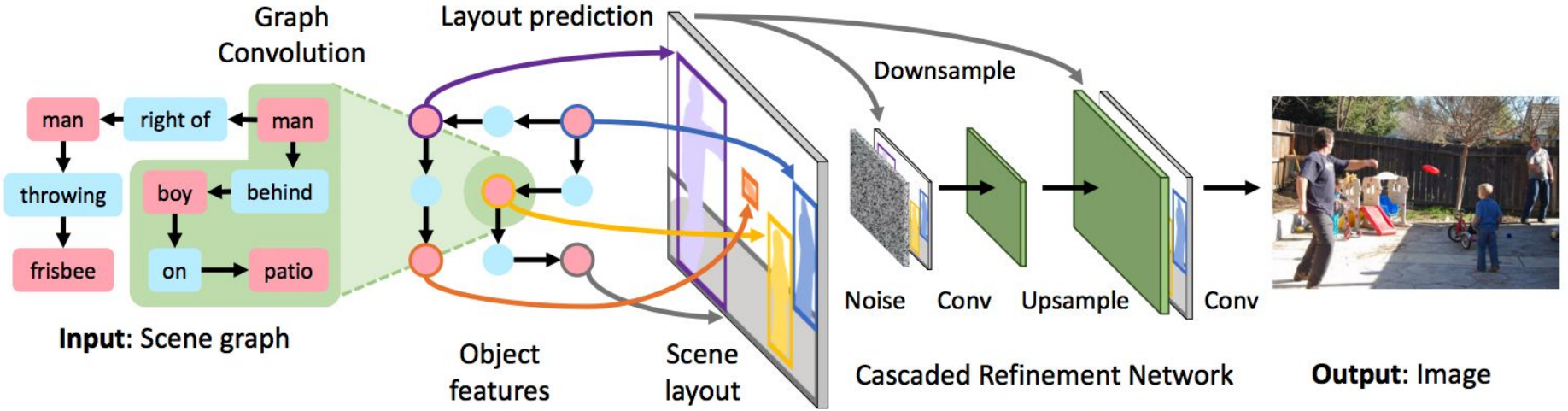
Hong et al. [Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis.](#) CVPR 2018

Image Generation from Text



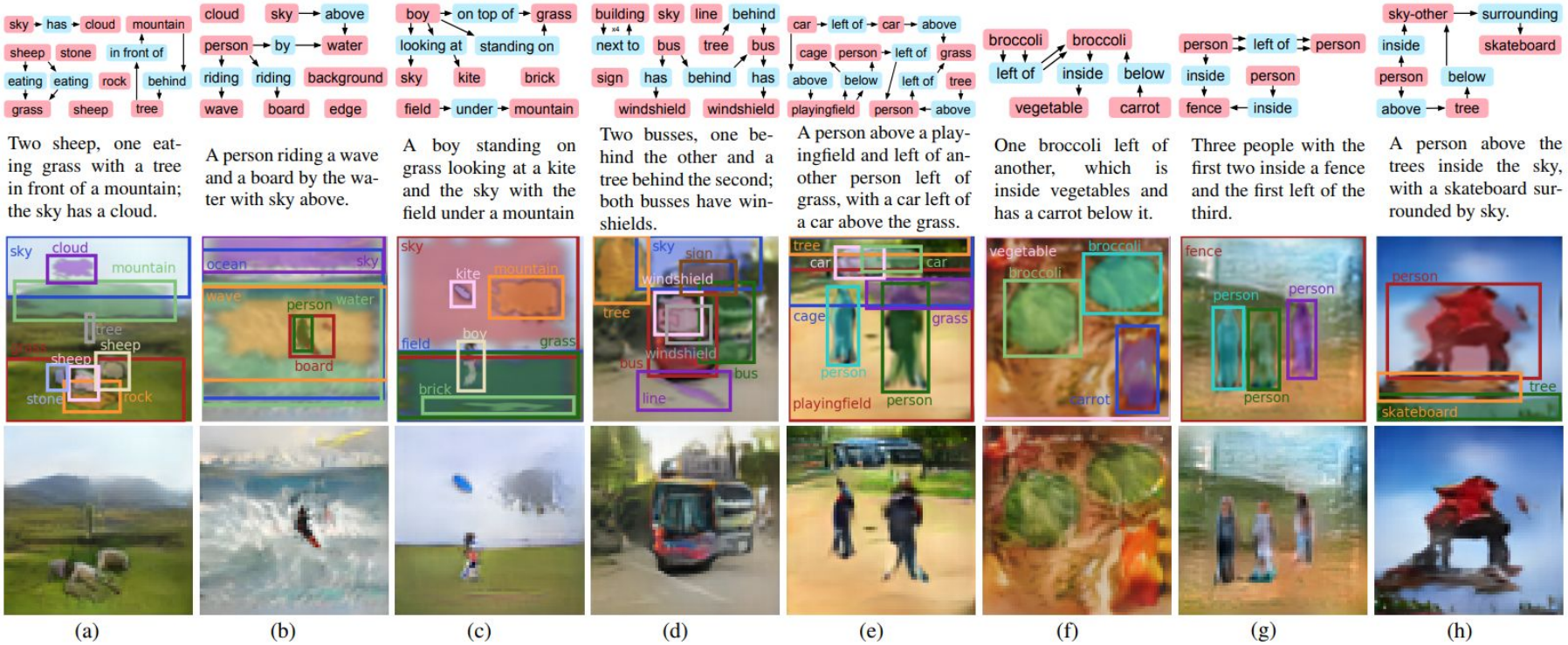
Johnson et al. "Image Generation from Scene Graphs" CVPR 2018

Image Generation from Text



Johnson et al. "Image Generation from Scene Graphs" CVPR 2018

Image Generation from Text

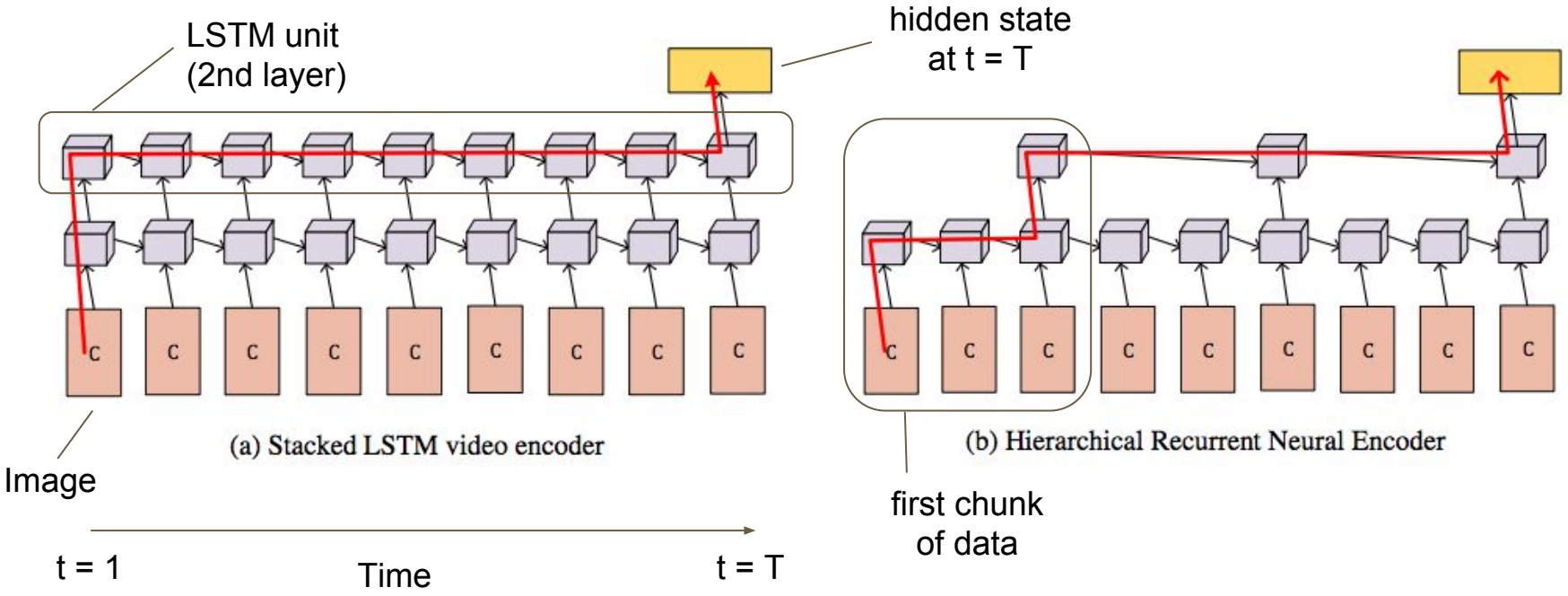


Johnson et al. "Image Generation from Scene Graphs" CVPR 2018

Questions?

1. Image Captioning
2. Visual Question Answering
3. Cross-Modal Embeddings
4. Image Generation from Text

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

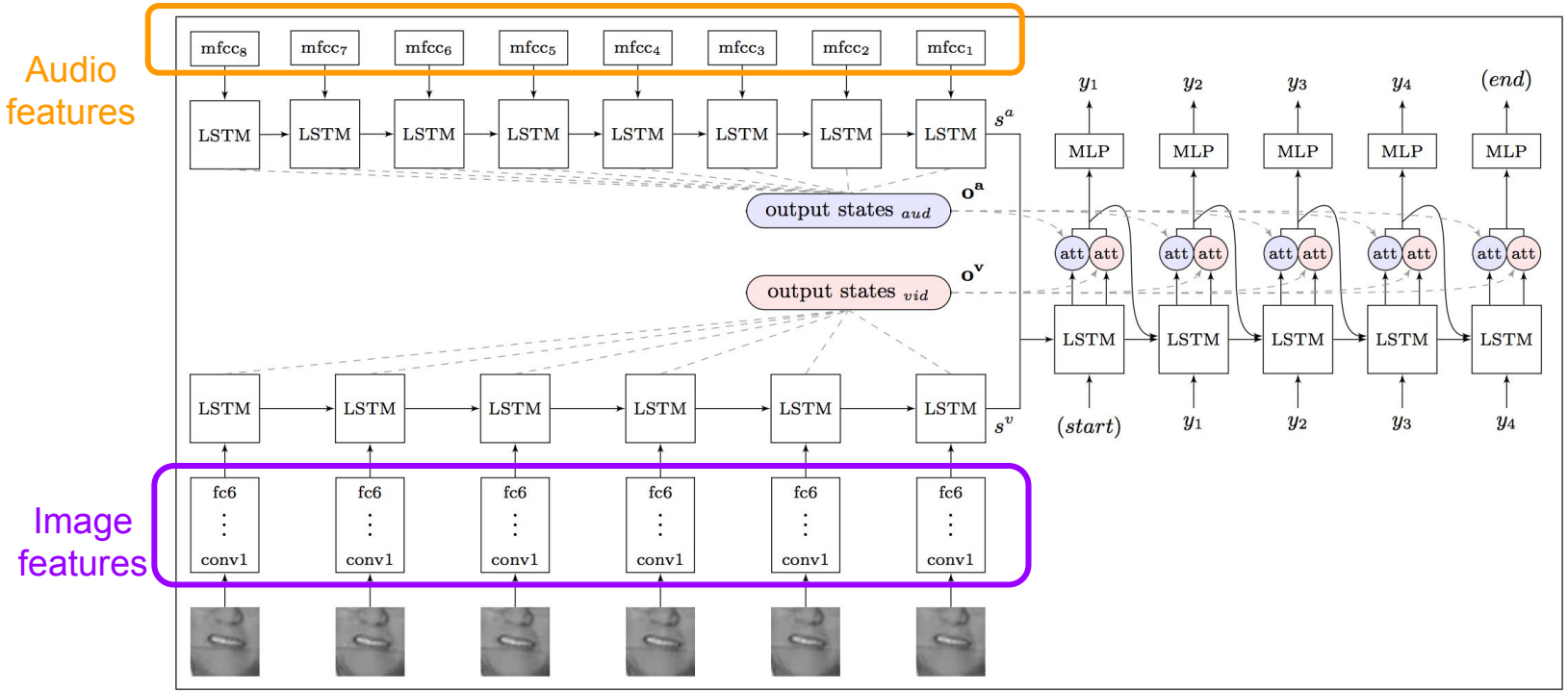


Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character y_i, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

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

Lipreading: Watch, Listen, Attend & Spell

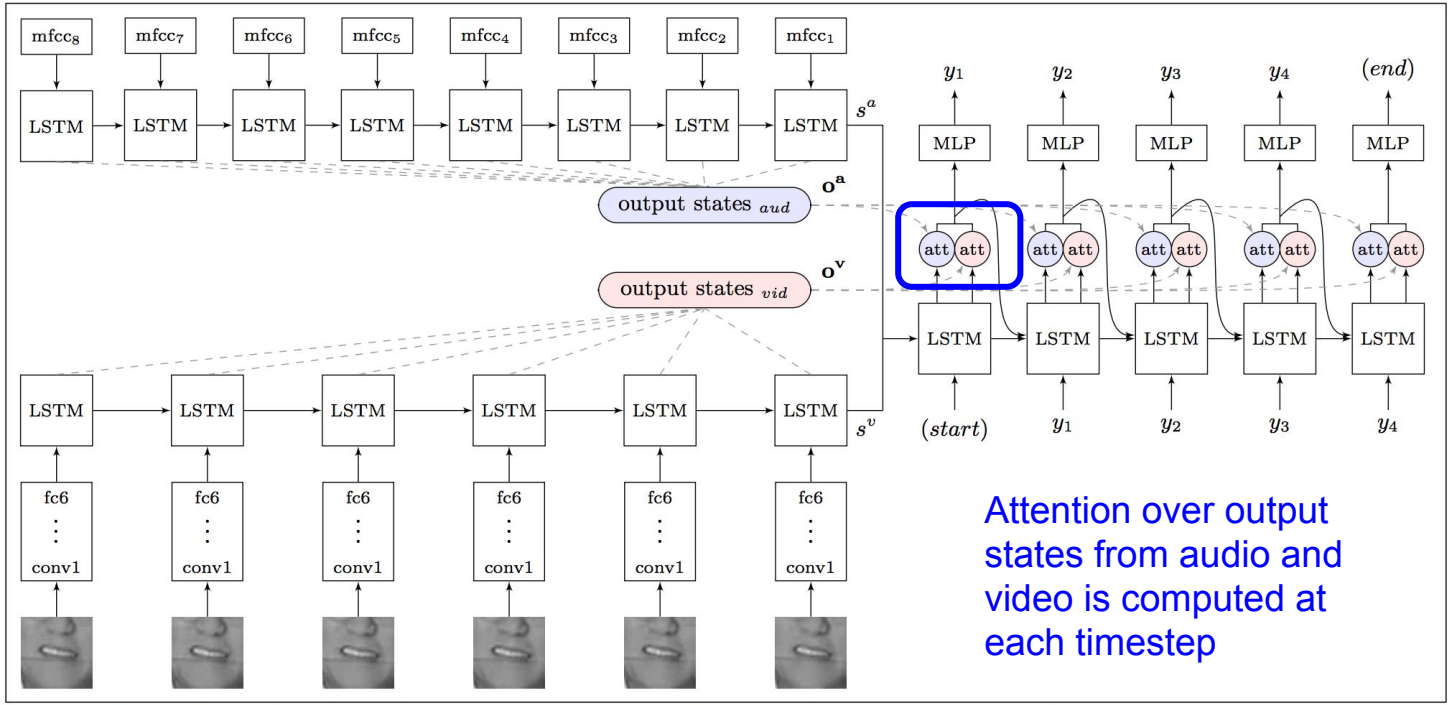
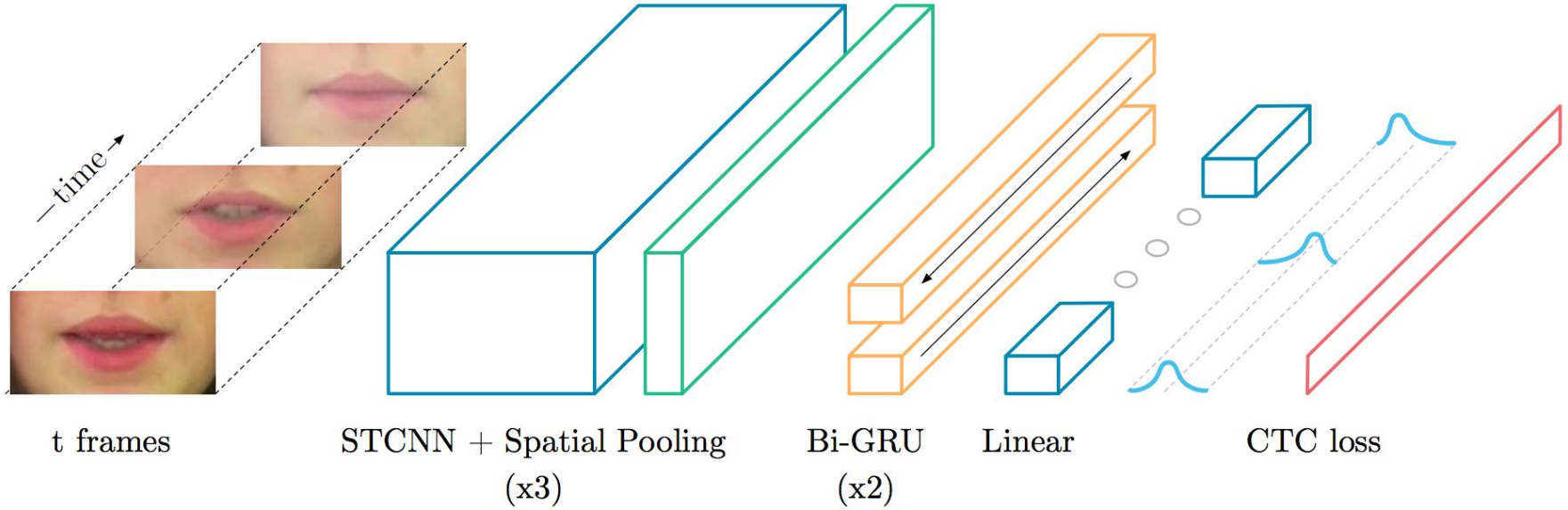


Figure 1. *Watch, Listen, Attend and Spell* architecture. At each time step, the decoder outputs a character y_i, as well as two attention vectors. The attention vectors are used to select the appropriate period of the input visual and audio sequences.

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

Lip Reading: LipNet

Input (video frames) and output (sentence) sequences are not aligned

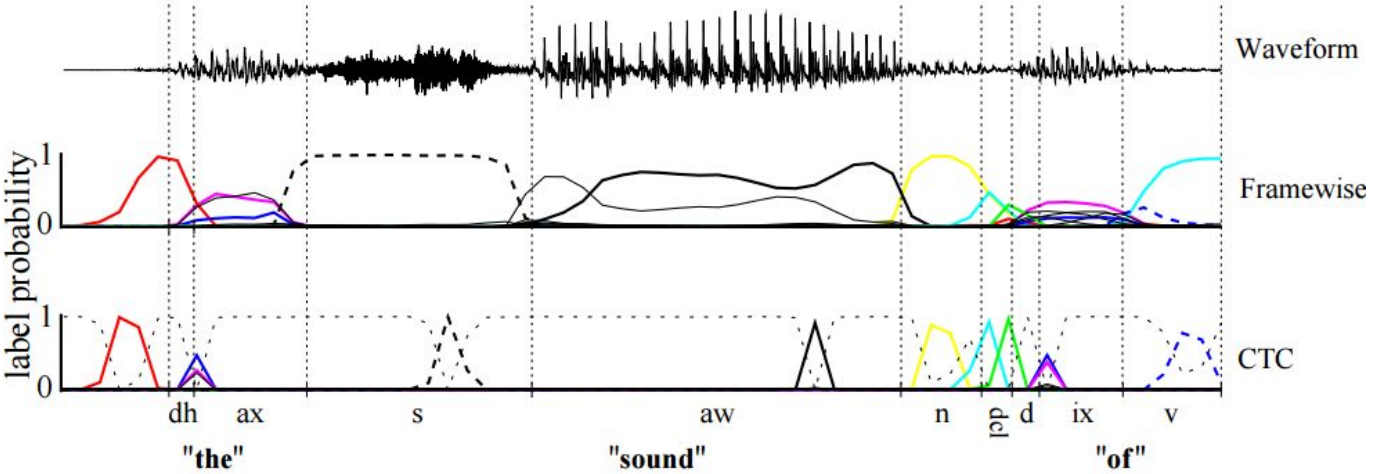


Assael, Yannis M., Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. "LipNet: End-to-End Sentence-level Lipreading." (2016).

Lip Reading: LipNet

CTC Loss: Connectionist temporal classification

- Avoiding the need for alignment between input and output sequence by predicting an additional “_” blank word
- Before computing the loss, repeated words and blank tokens are removed
- “a _ a b _ ” == “_ a a _ _ b b” == “a a b”

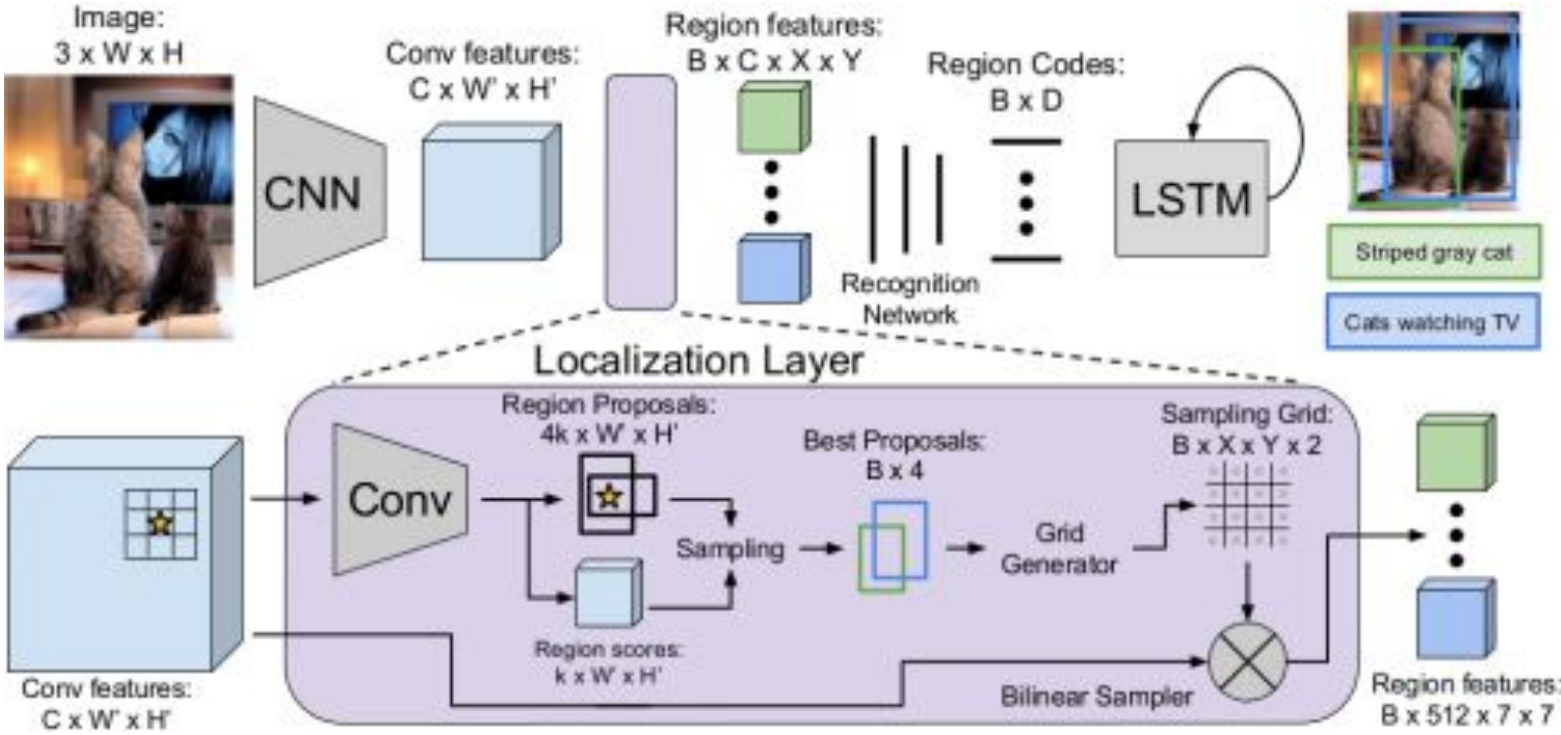


Graves et al. [Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks](). ICML 2006

Lip Reading: LipNet

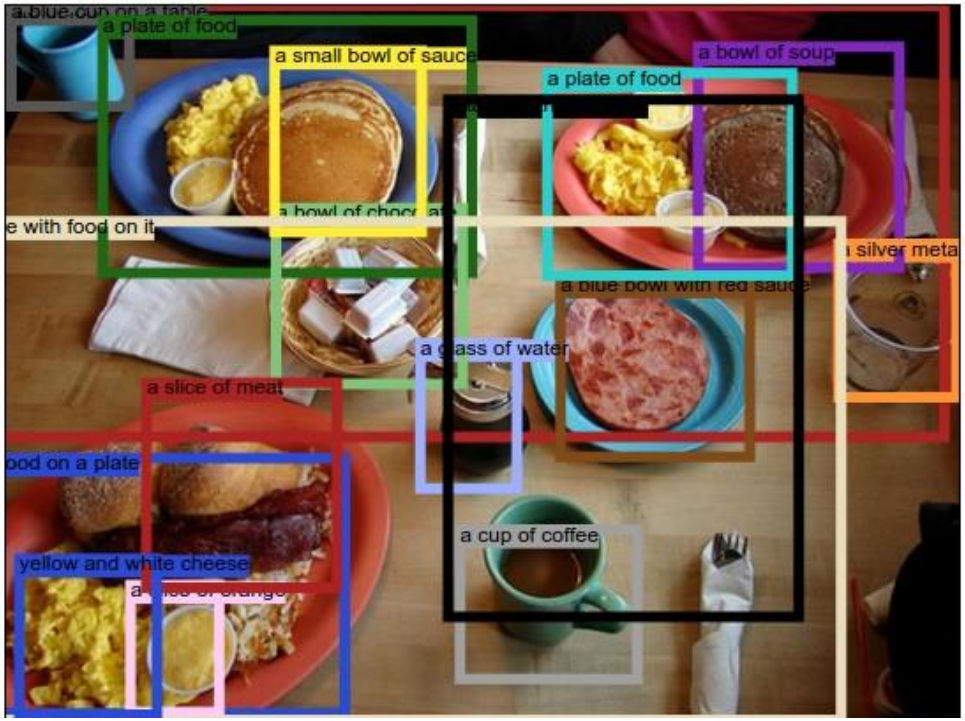
Assael et al. [LipNet: Sentence-level Lipreading.](#) *arXiv Nov 2016*

Captioning (+ Detection): DenseCap



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

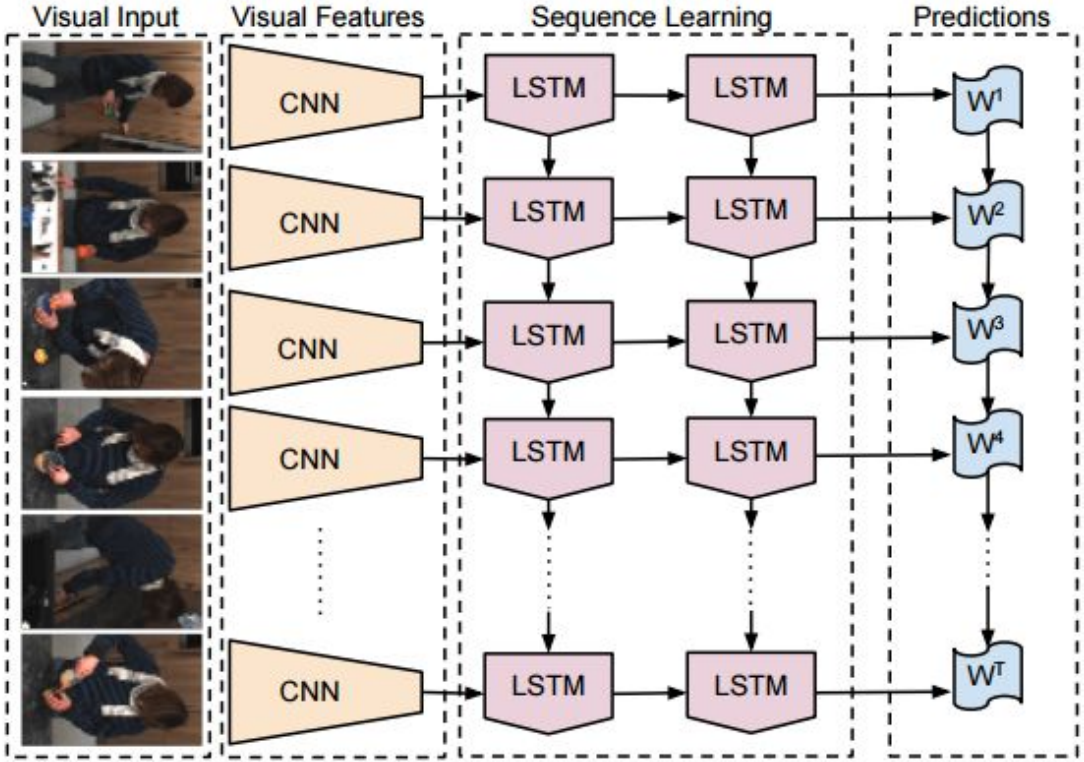
Captioning (+ Detection): DenseCap



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.

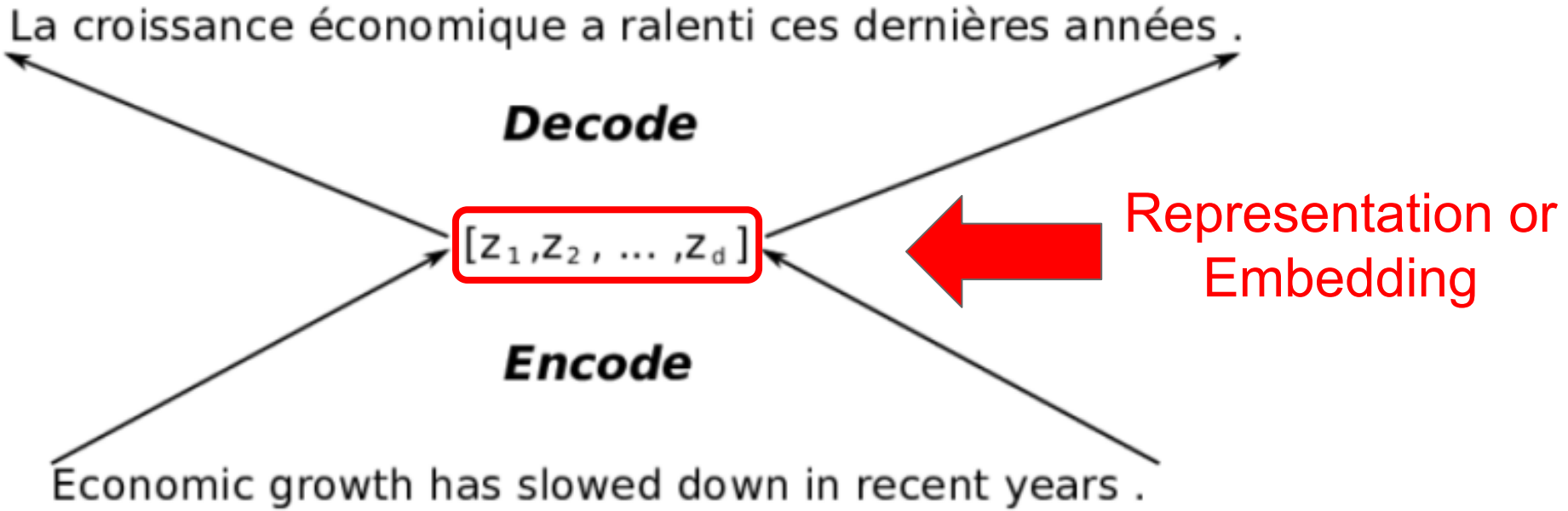
Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning." *CVPR 2016*

Captioning: 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

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 .

Word Embeddings

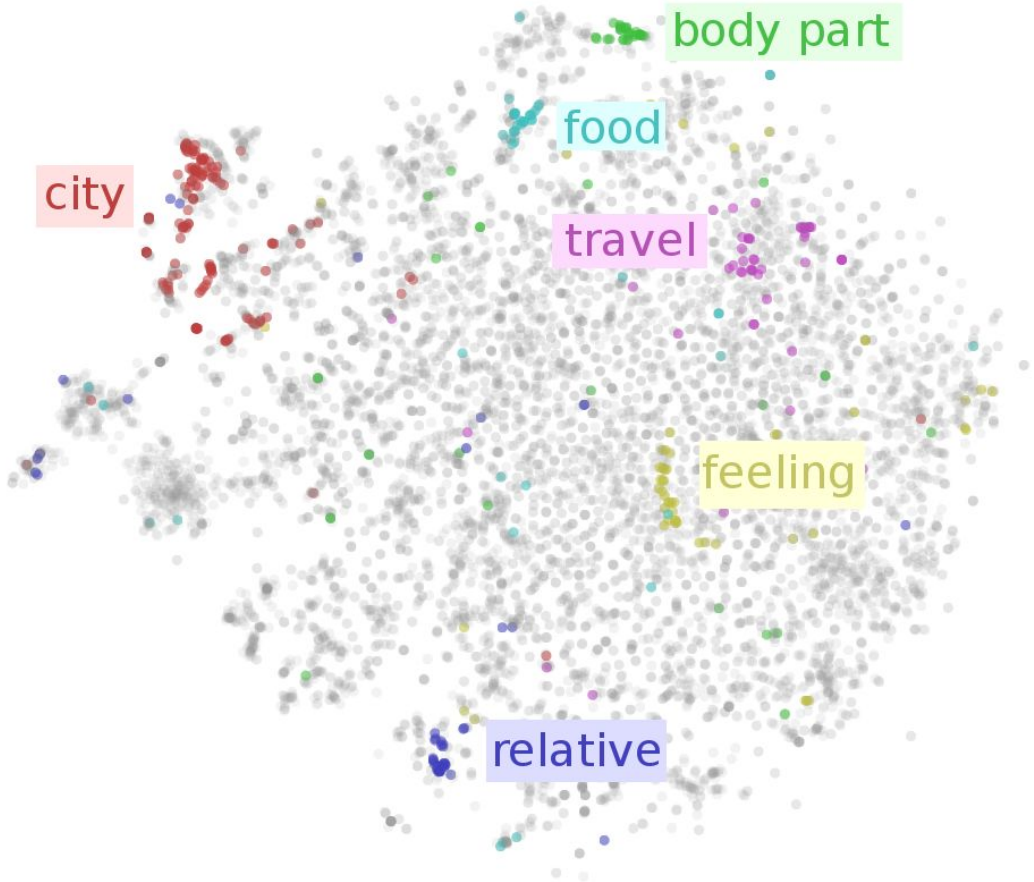
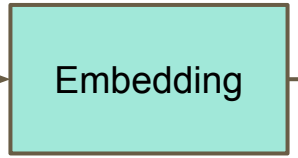


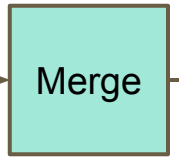
Figure:
Christopher Olah
Visualizing Representations

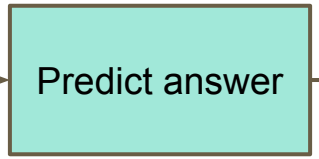
Visual Question Answering (VQA)

Question

What object is flying?





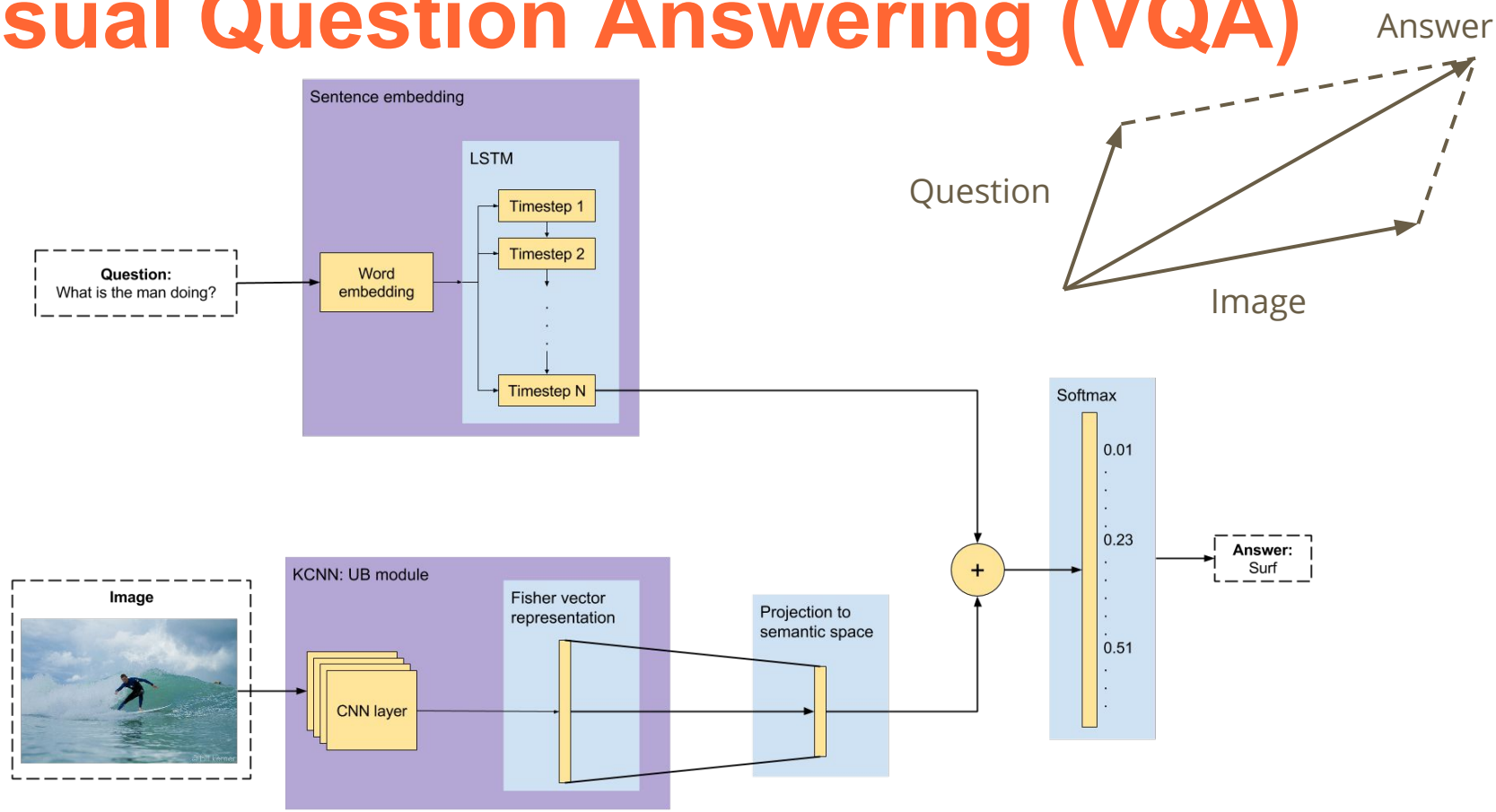


Answer

Kite

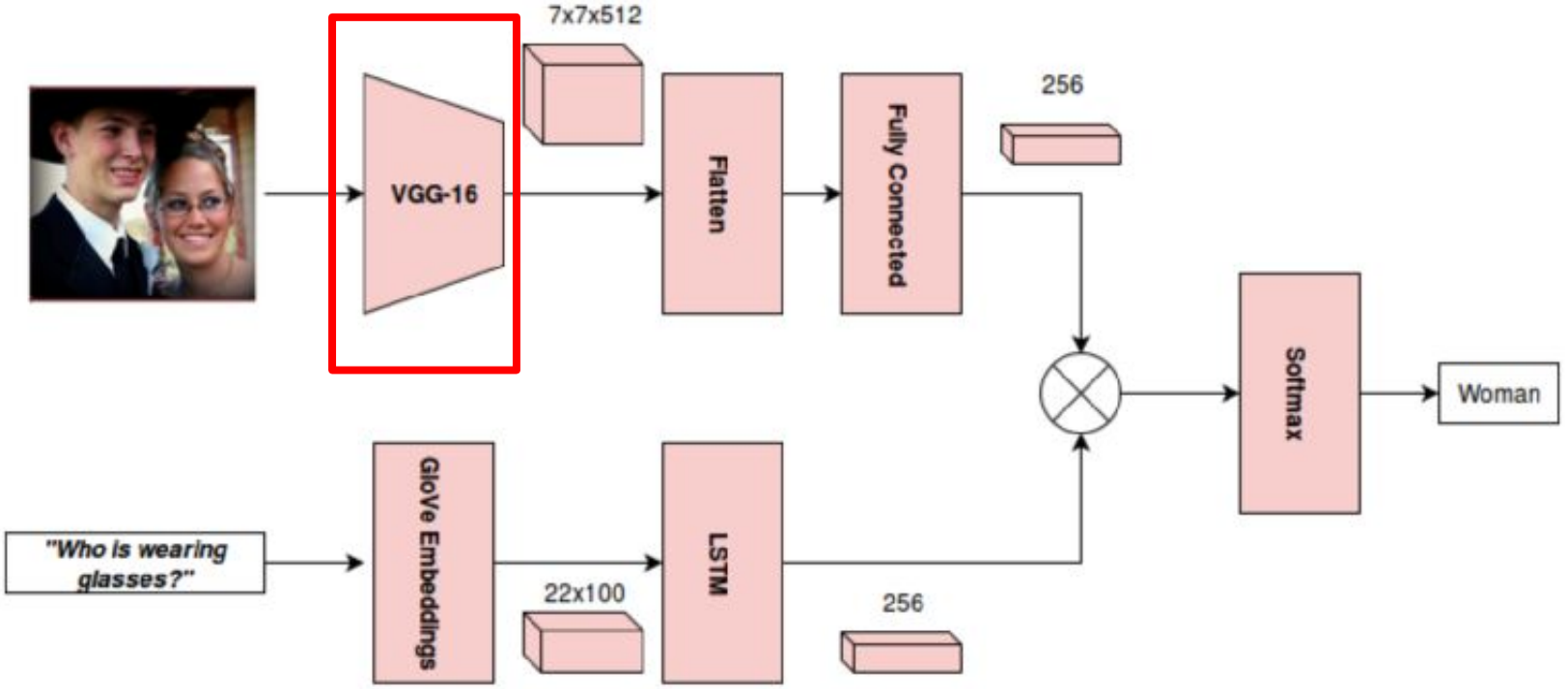
Slide credit: Issey Masuda

Visual Question Answering (VQA)



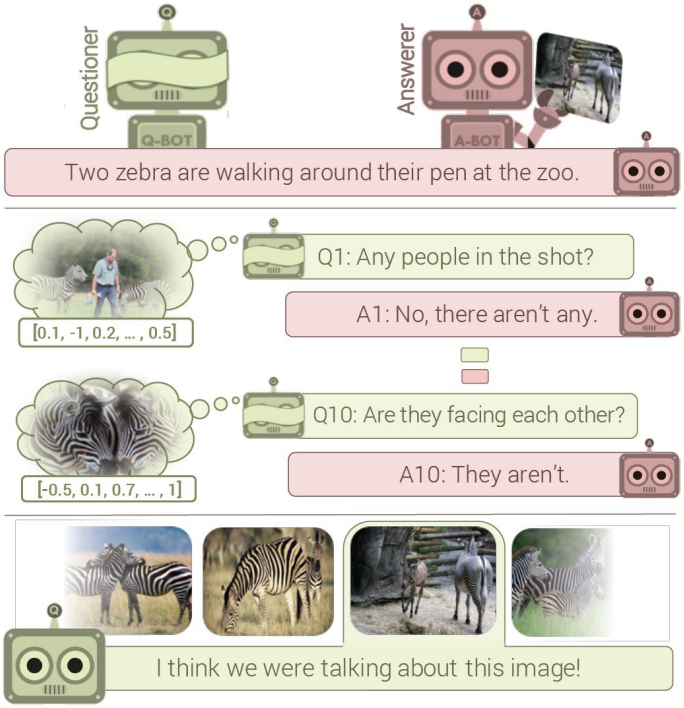
Masuda, Issey, Santiago Pascual de la Puente, and Xavier Giro-i-Nieto. "Open-Ended Visual Question-Answering." ETSETB UPC TelecomBCN (2016).

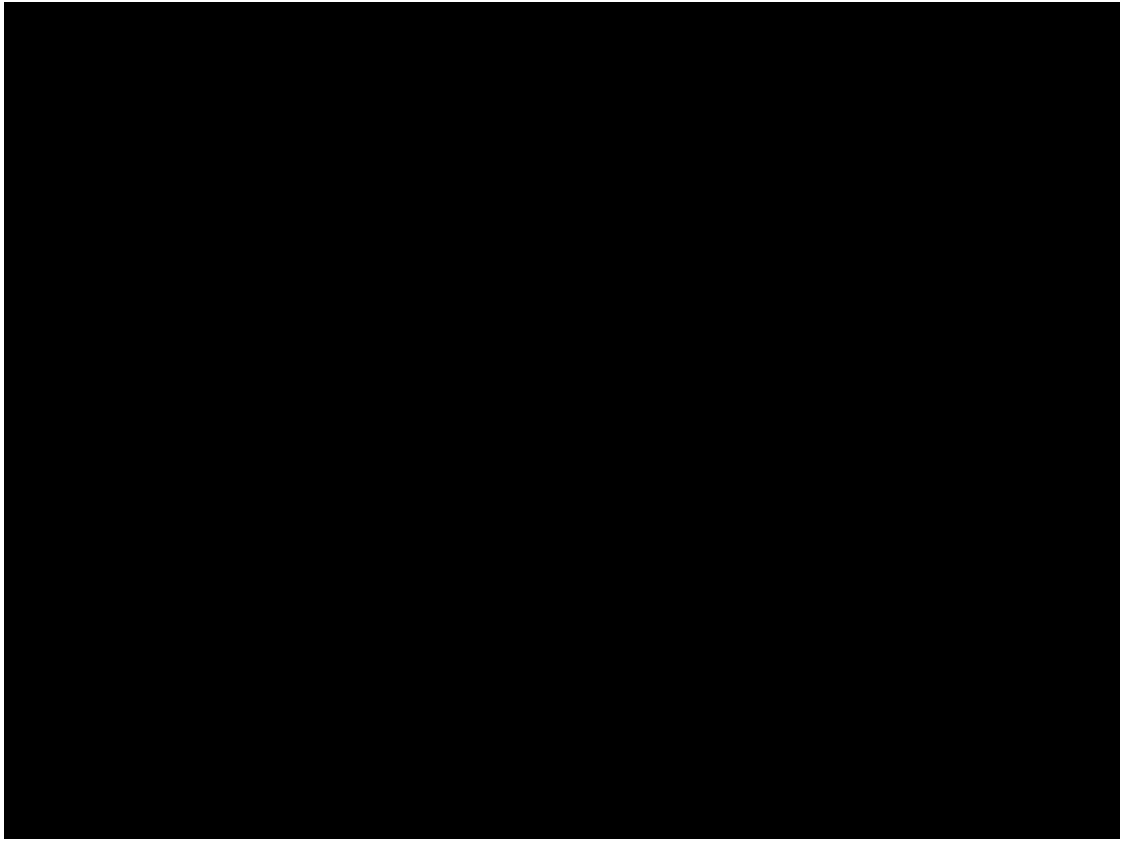
Visual Question Answering (VQA)



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

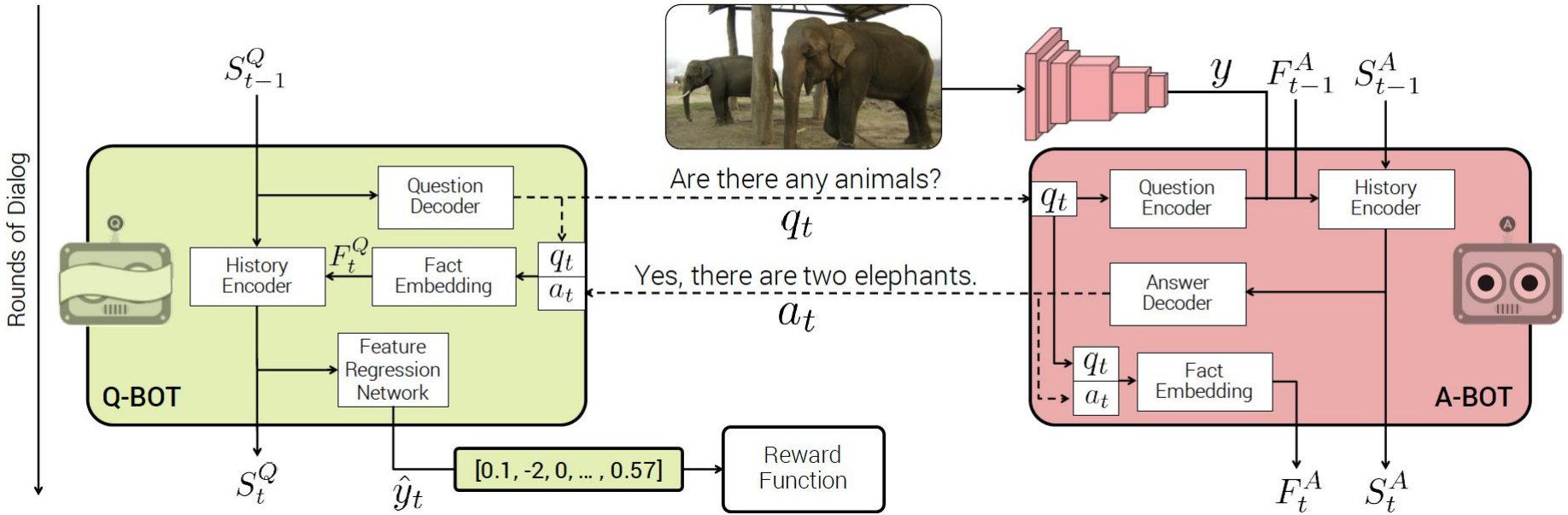
Visual Dialog (Image Guessing Game)





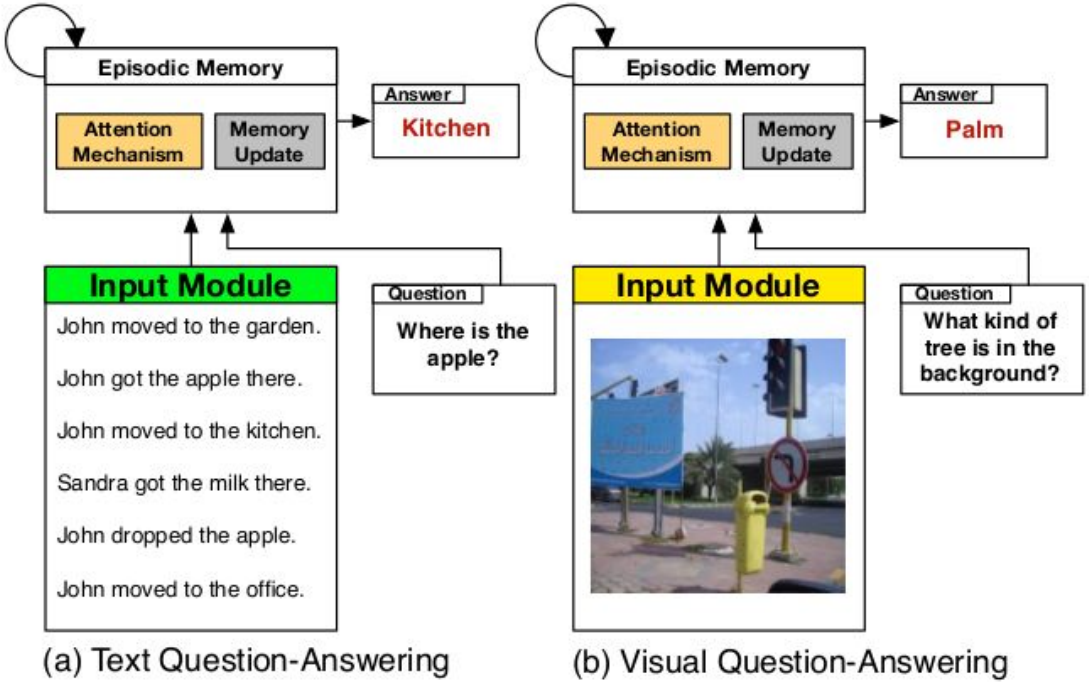
Das et al. "Visual Dialog." CVPR 2017

Visual Dialog (Image Guessing Game)



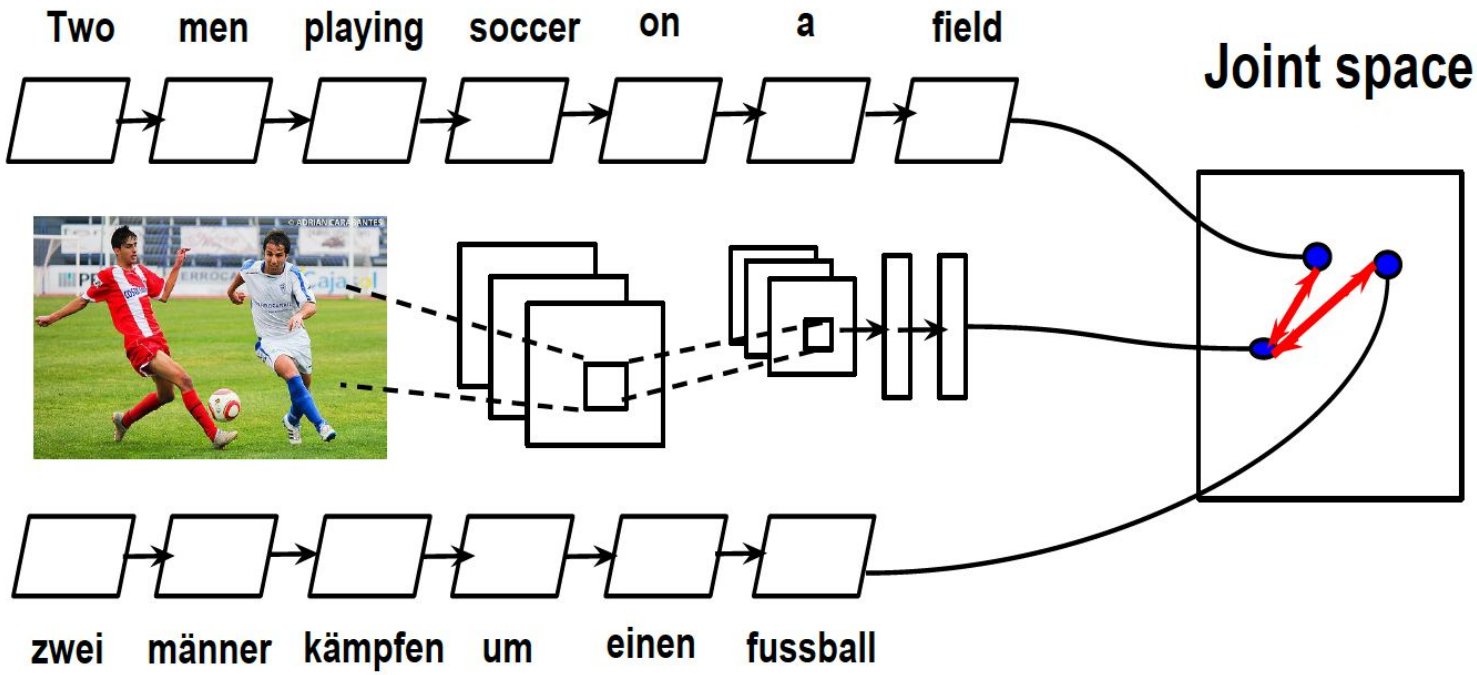
Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. "Visual Dialog." CVPR 2017

Visual Question Answering: Dynamic



Xiong et al. "Dynamic Memory Networks for Visual and Textual Question Answering." ICML 2016

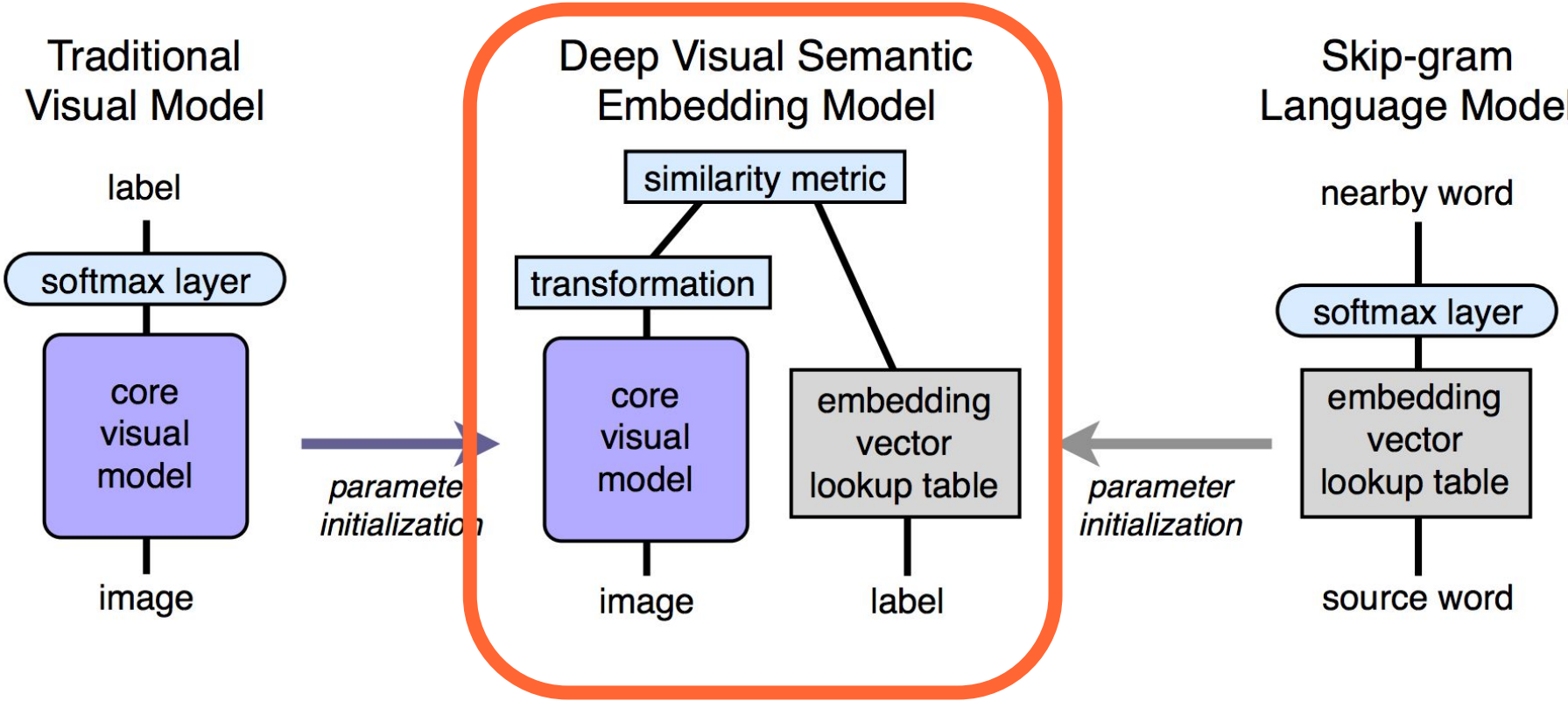
Multilingual & Multimodal Embeddings



Gella, Spandana, Rico Sennrich, Frank Keller, and Mirella Lapata. "Image Pivoting for Learning Multilingual Multimodal Representations." arXiv preprint arXiv:1707.07601 (2017).

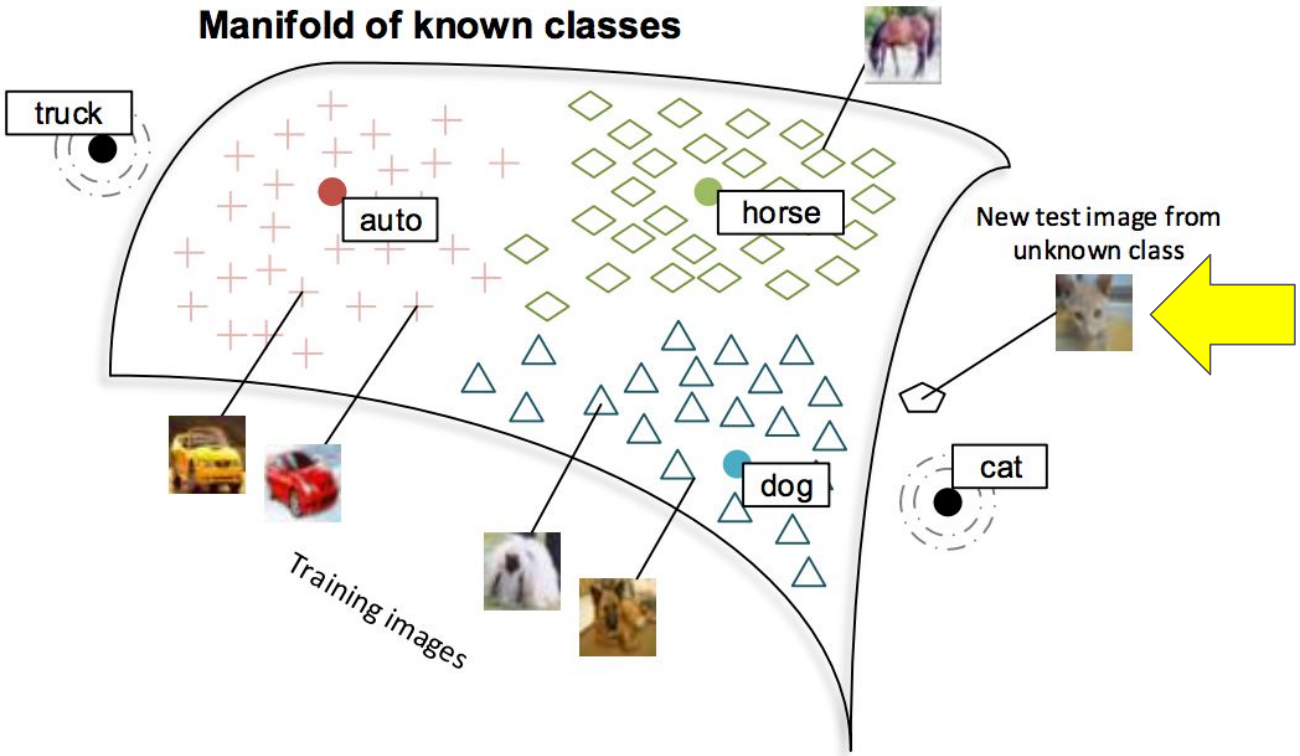
Janarthanan Rajendran, Mitesh M Khapra, Sarath Chandar, Balaraman Ravindran, Bridge Correlational Neural Networks for Multilingual Multimodal Representation Learning NAACL, 2016

Cross-Modal Embeddings



Frome et al. "Devise: A deep visual-semantic embedding model." NIPS 2013

Joint Neural Embeddings

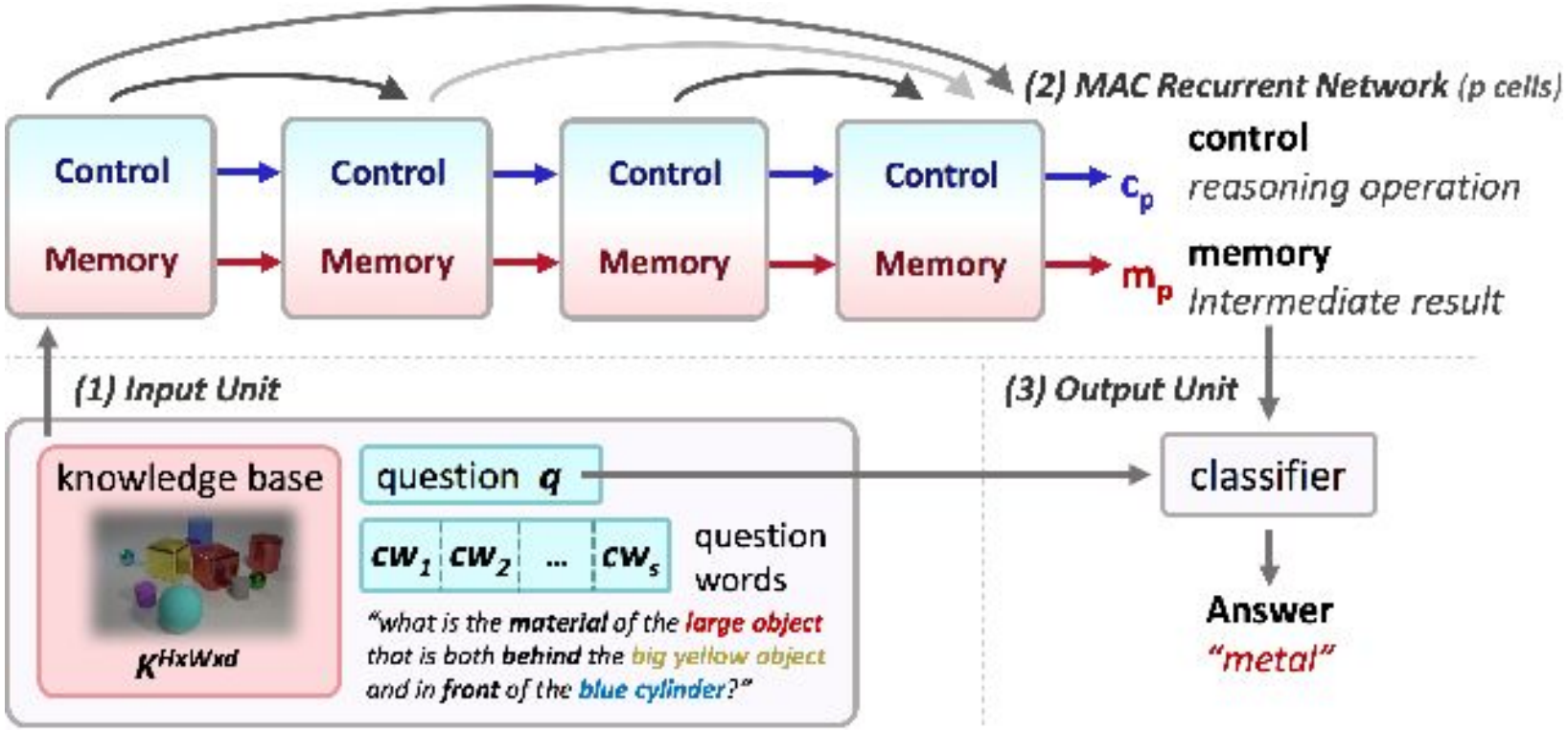


Zero-shot learning:
a class not present in the training set of images can be predicted

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

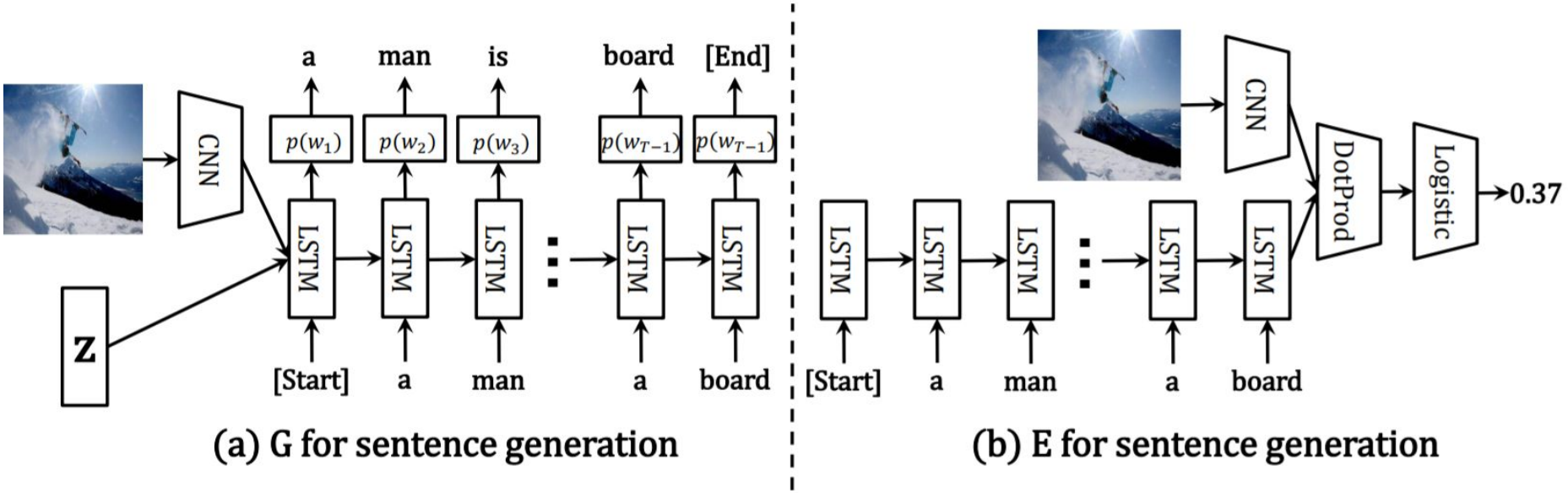
Socher et al. Zero-shot learning through cross-modal transfer. NIPS 2013

Reasoning: MAC



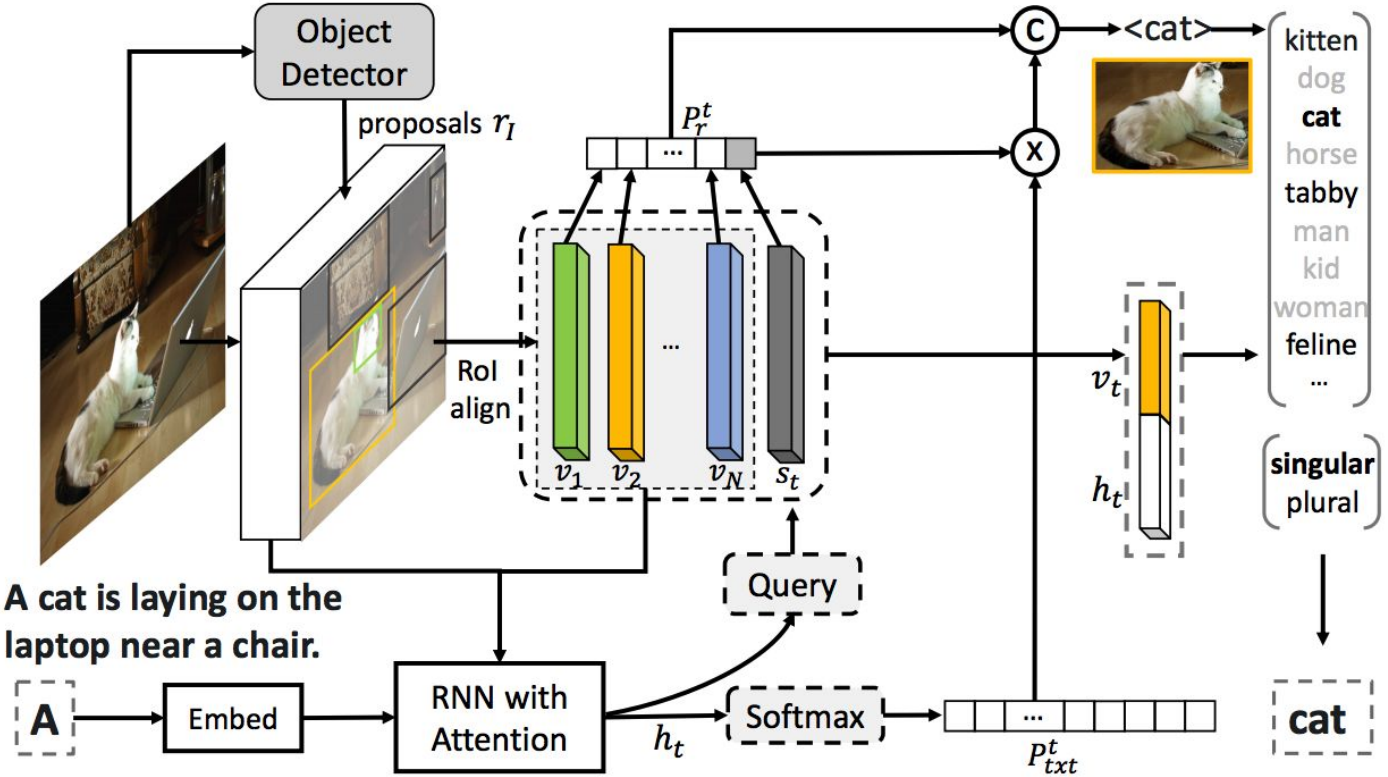
Hudson et al. "Compositional attention networks for machine reasoning." arXiv preprint arXiv:1803.03067 (2018).

GANs for Image Captioning



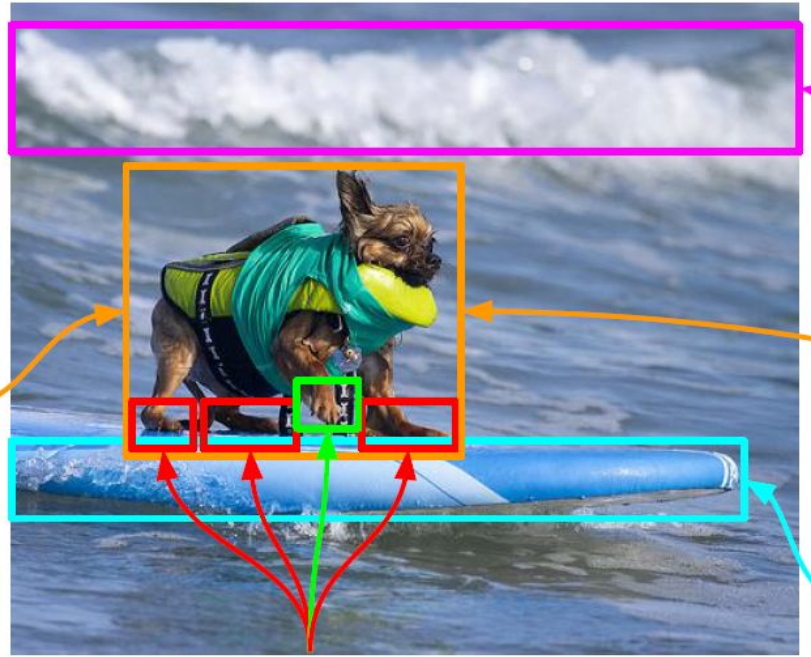
Dai et al. [Towards Diverse and Natural Image Descriptions via a Conditional GAN](). ICCV 2017

Grounded Image Captioning



Lu et al. "Neural Baby Talk" CVPR 2018

Grounded VQA



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.

Zhu et al."Visual7W: Grounded Question Answering in Images." CVPR 2016.