

Multimodality, NLP + Vision

COMP3361 – Week 12

Lingpeng Kong

Department of Computer Science, The University of Hong Kong

Language and Vision

First become a “track” in EMNLP in 2015.



Language and Vision

Bag of words (BoW)

Very good drama although it appeared to have a few blank areas leaving the viewers to fill in the action for themselves. I can imagine life being this way for someone who can neither read nor write. This film simply smacked of the real world: the wife who is suddenly the sole supporter, the live-in relatives and their quarrels, the troubled child who gets knocked up and then, typically, drops out of school, a jackass husband who takes the nest egg and buys beer with it. 2 thumbs up... very very very good movie.



('the', 8),
(',', 5),
('very', 4),
('.', 4),
('who', 4),
('and', 3),
('good', 2),
('it', 2),
('to', 2),
('a', 2),
('for', 2),
('can', 2),
('this', 2),
('of', 2),
('drama', 1),
('although', 1),
('appeared', 1),
('have', 1),
('few', 1),
('blank', 1)

■ ■ ■

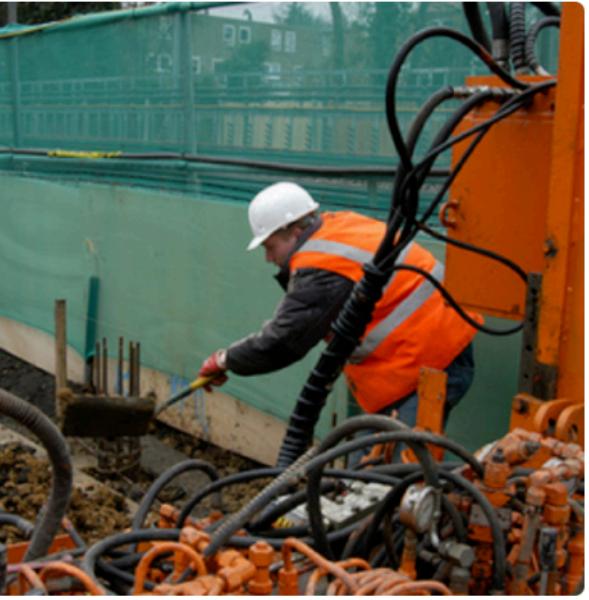


Bag of Words (BoW) in NLP and Vision

Language and Vision



"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."

Two dogs play with an orange toy in tall grass.

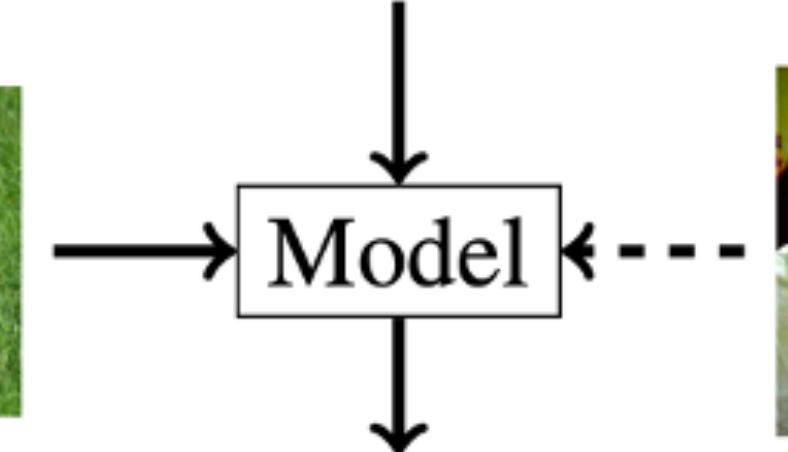
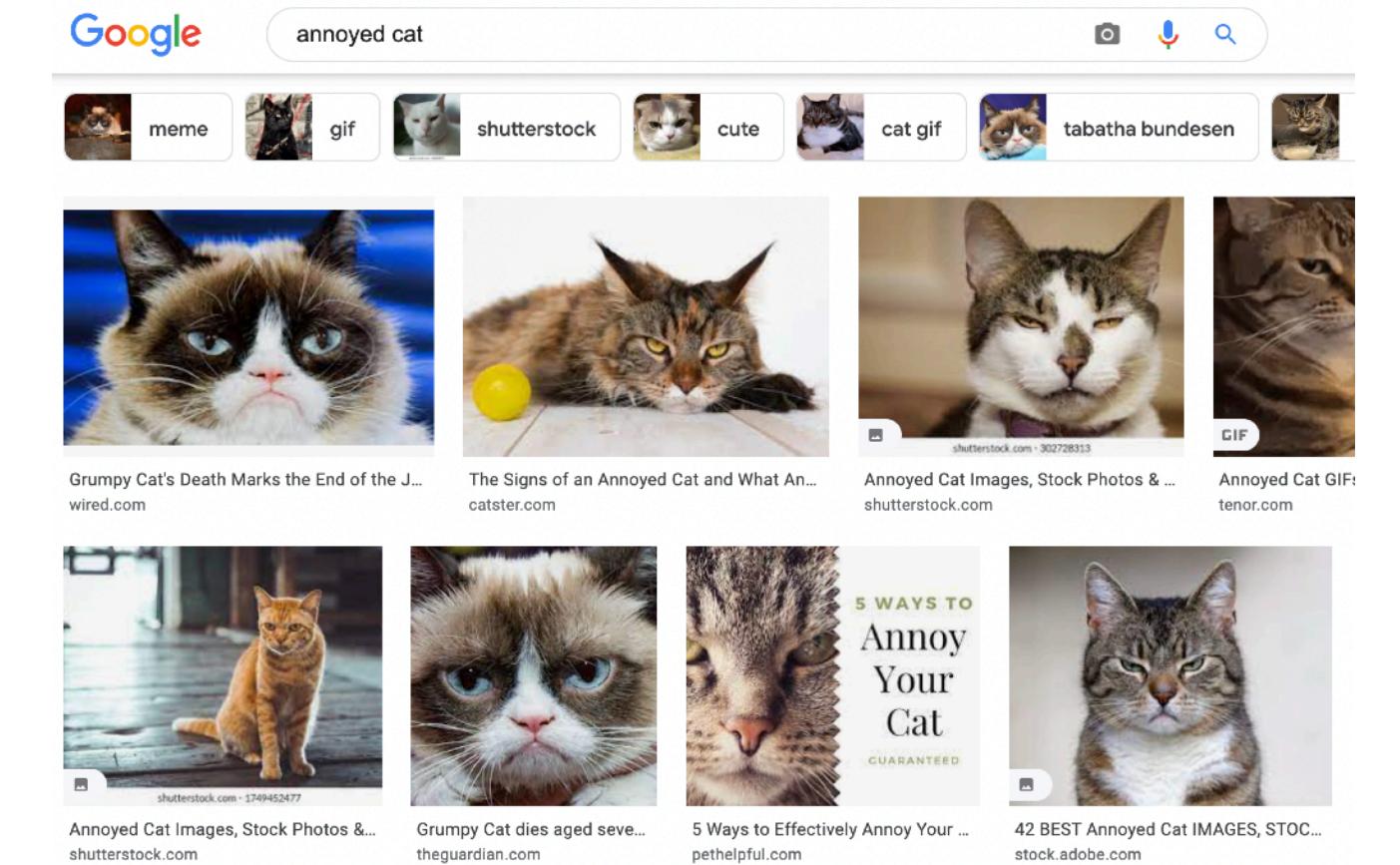


Image Captioning



Multimodal Machine Translation



Visual Question Answering

Image Search (retrieval)

Image Captioning



A stop sign is on a road with a mountain in the background.

“Scene understanding”

Sequence to Sequence Model

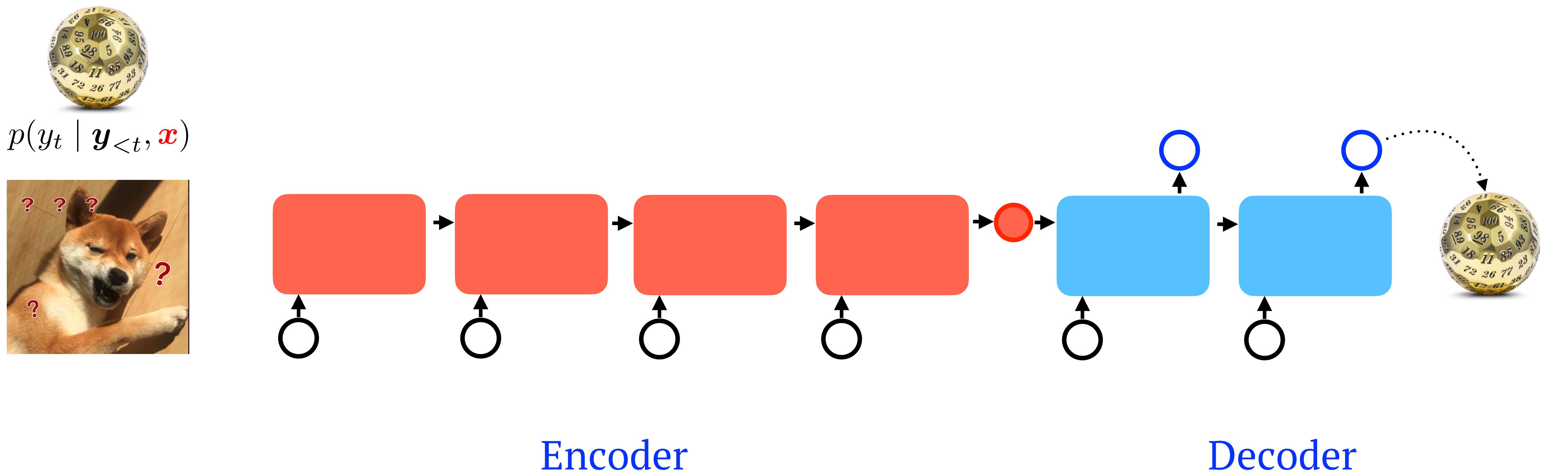
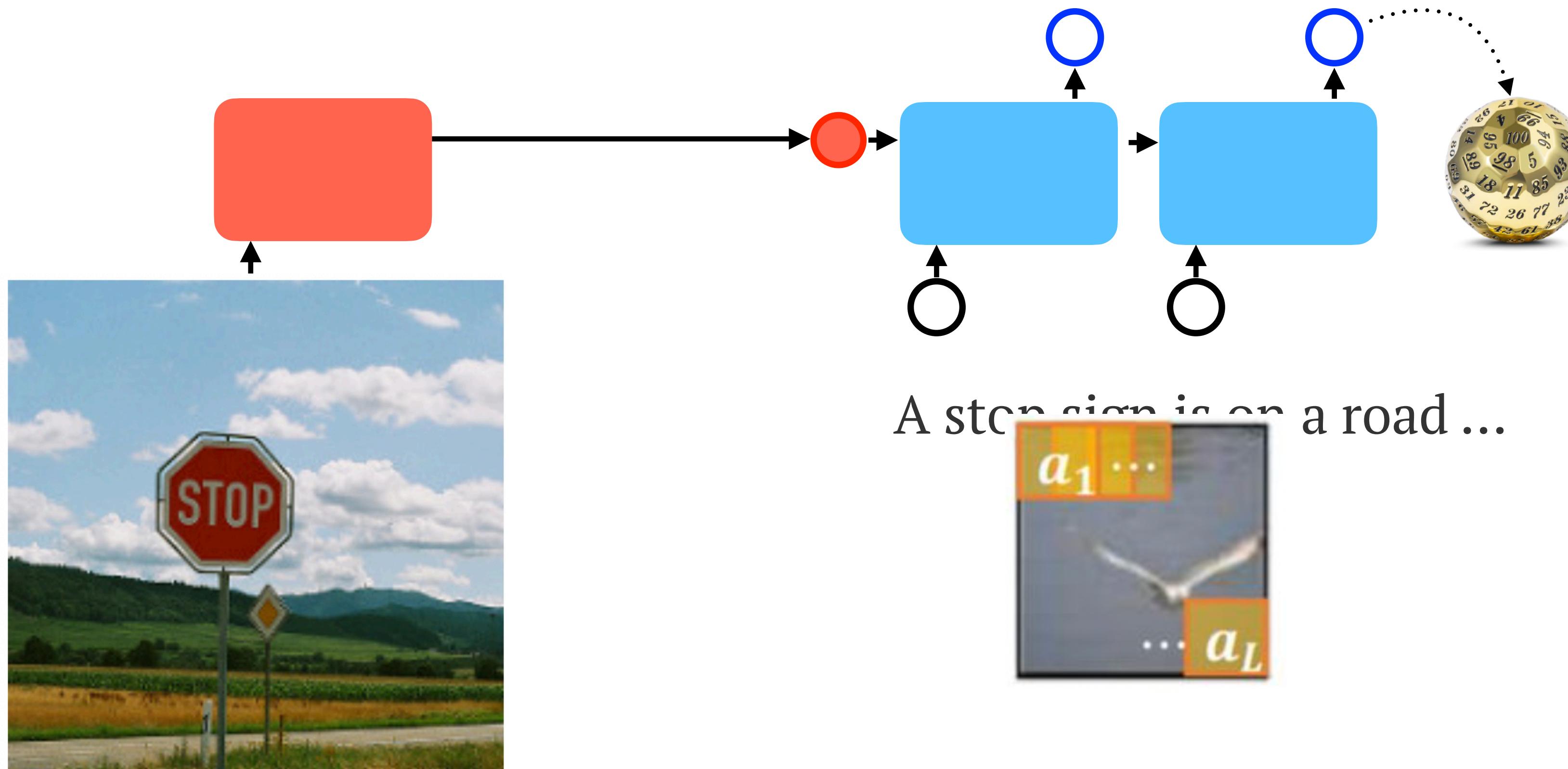
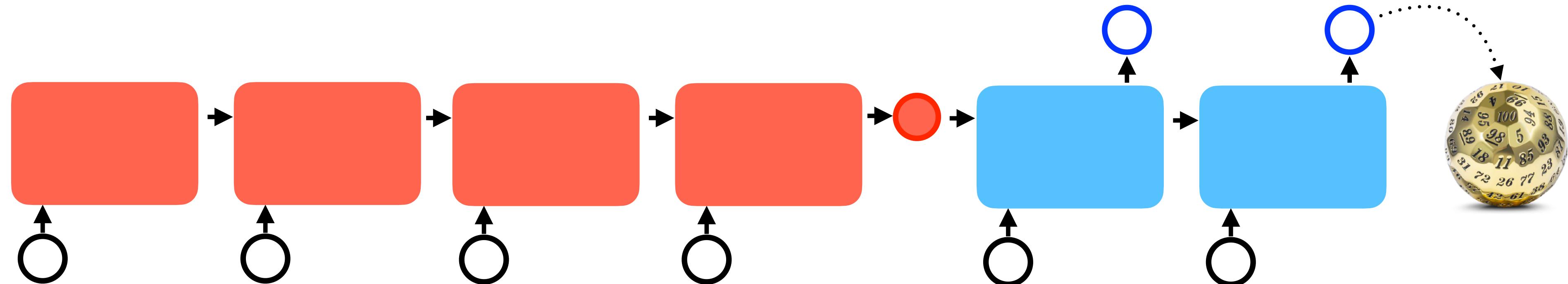


Image Captioning

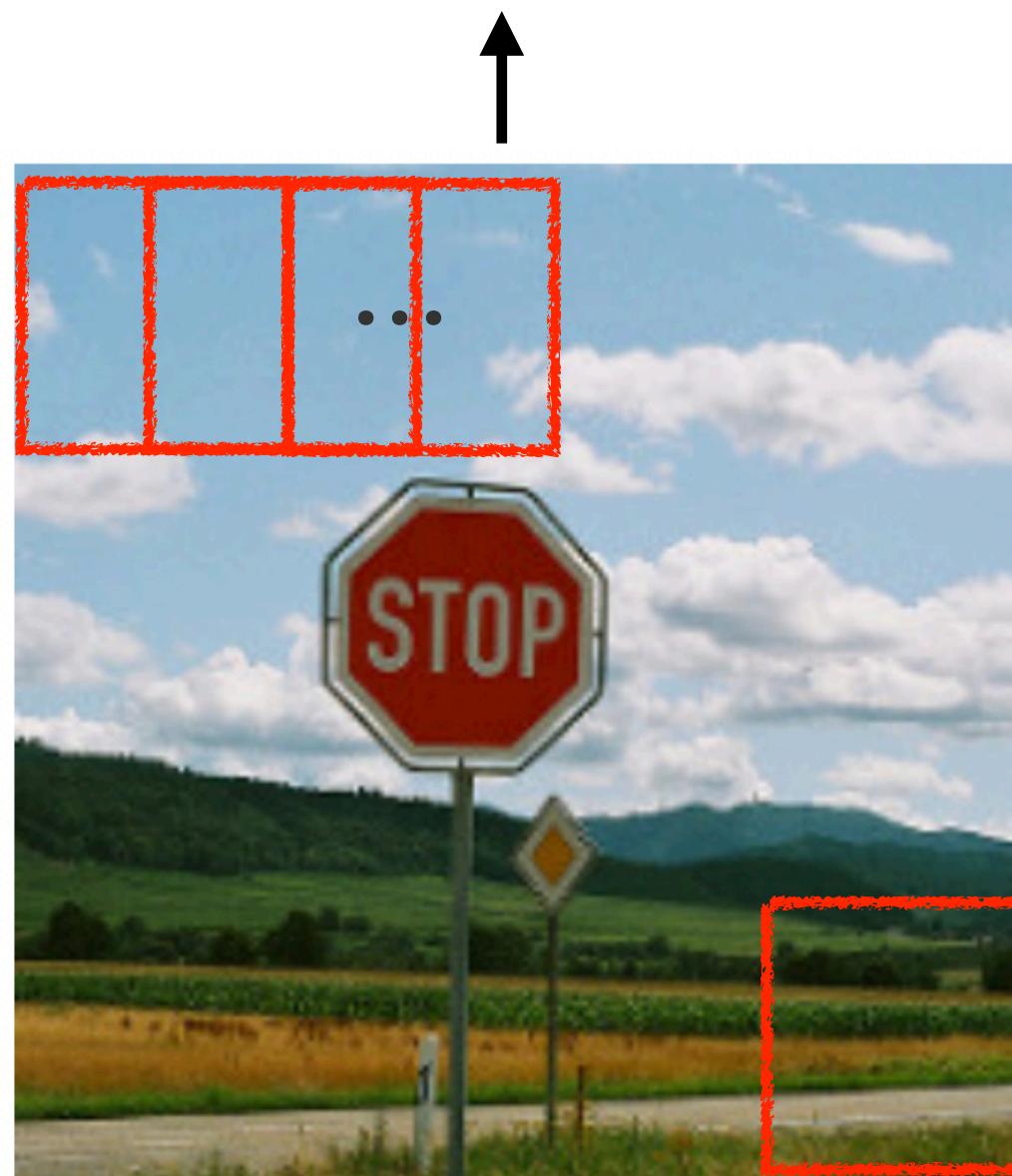


(Xu et al., 2015)

Image Captioning



A stop sign is on a road ...



(Xu et al., 2015)

Image Captioning



A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



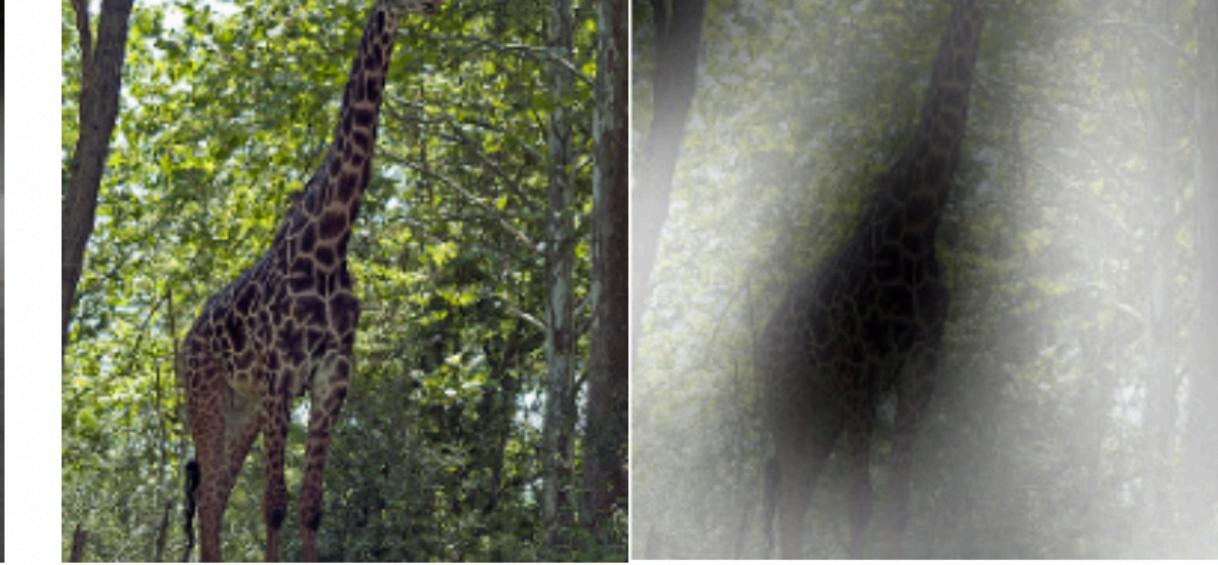
A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.

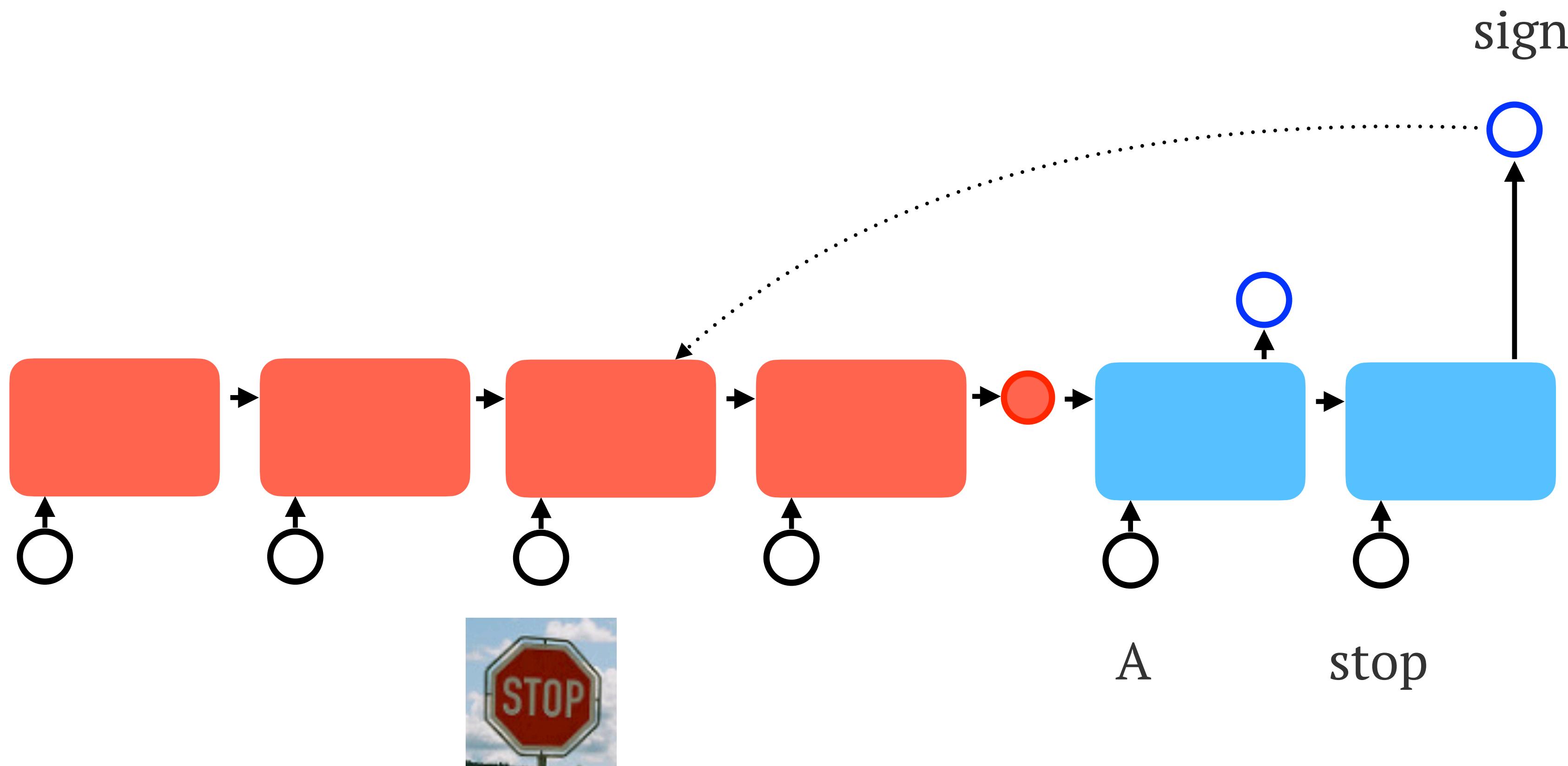


A giraffe standing in a forest with trees in the background.

To allow salient features to dynamically come to the forefront as needed.

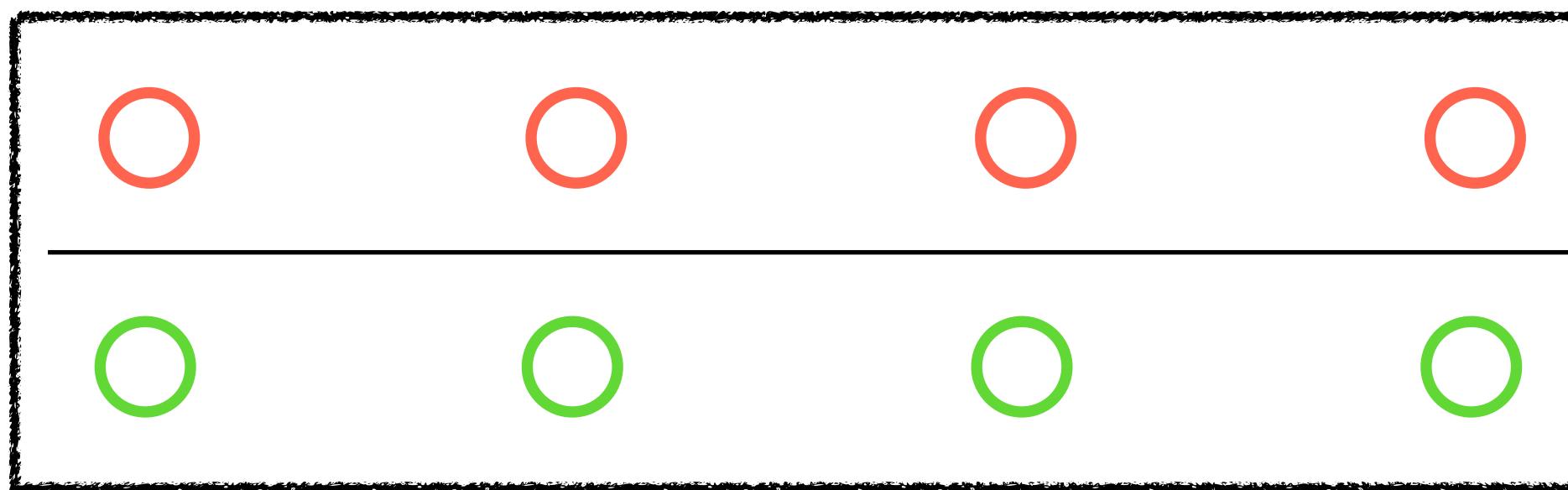
(Xu et al., 2015)

Image Captioning



(Xu et al., 2015)

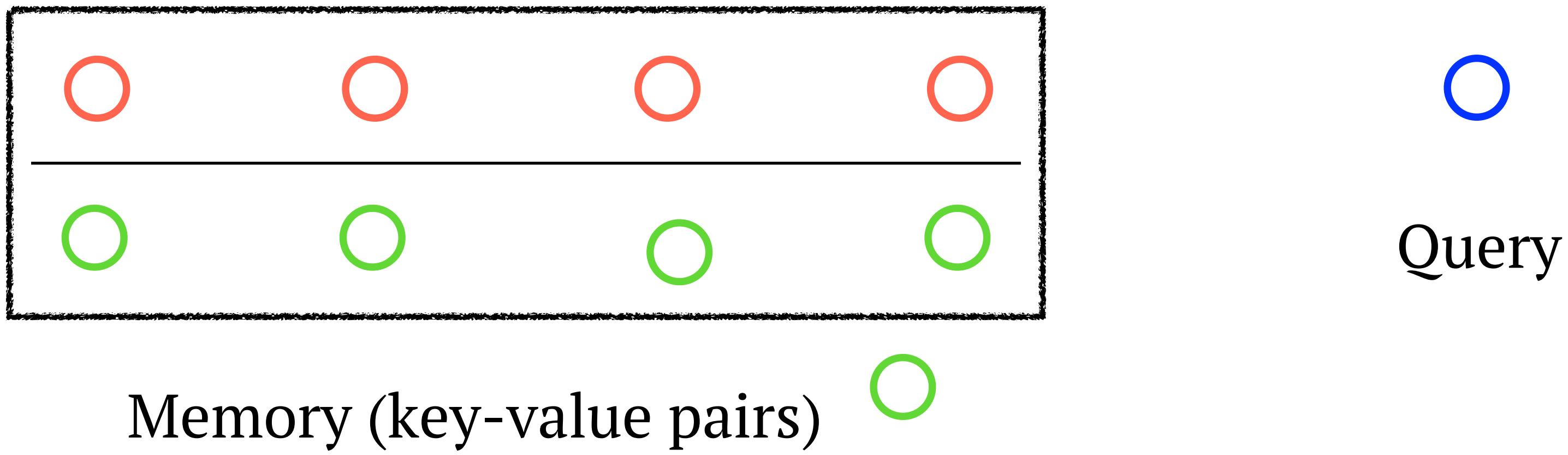
“Soft” Attention



Memory (key-value pairs)

$$\begin{array}{ll} \text{blue circle} \text{ } \text{red circle} & q \cdot k_1 \\ \text{blue circle} \text{ } \text{red circle} & q \cdot k_2 \\ \text{blue circle} \text{ } \text{red circle} & q \cdot k_3 \\ \text{blue circle} \text{ } \text{red circle} & q \cdot k_4 \end{array} \quad \begin{array}{l} q \cdot k_1 \\ q \cdot k_2 \\ q \cdot k_3 \\ q \cdot k_4 \end{array} \quad \text{softmax} \left(\begin{array}{l} q \cdot k_1 \\ q \cdot k_2 \\ q \cdot k_3 \\ q \cdot k_4 \end{array} \right) \rightarrow \begin{bmatrix} 0.6 \\ 0.1 \\ 0.2 \\ 0.1 \end{bmatrix} \quad \begin{array}{l} \text{green circle} \\ \text{green circle} \\ \text{green circle} \\ \text{green circle} \end{array} \quad \rightarrow \quad \begin{array}{l} 0.6 \text{ } \text{green circle} + 0.1 \text{ } \text{green circle} + 0.2 \text{ } \text{green circle} + 0.1 \text{ } \text{green circle} \\ = \text{ } \text{green circle} \text{ context vector } \mathbf{c} \end{array}$$

“Hard” Attention



$$\text{blue circle} \text{ } \text{red circle} \quad q \cdot k_1$$

$$\text{blue circle} \text{ } \text{red circle} \quad q \cdot k_2$$

$$\text{blue circle} \text{ } \text{red circle} \quad q \cdot k_3$$

$$\text{blue circle} \text{ } \text{red circle} \quad q \cdot k_4$$

$$q \cdot k_1$$

$$q \cdot k_2$$

$$q \cdot k_3$$

$$q \cdot k_4$$

$$\text{HardMax} \left(\begin{array}{c} q \cdot k_1 \\ q \cdot k_2 \\ q \cdot k_3 \\ q \cdot k_4 \end{array} \right) = \text{green circle} \quad \text{context vector } c$$

Image Captioning



Extract visual features

Image Captioning

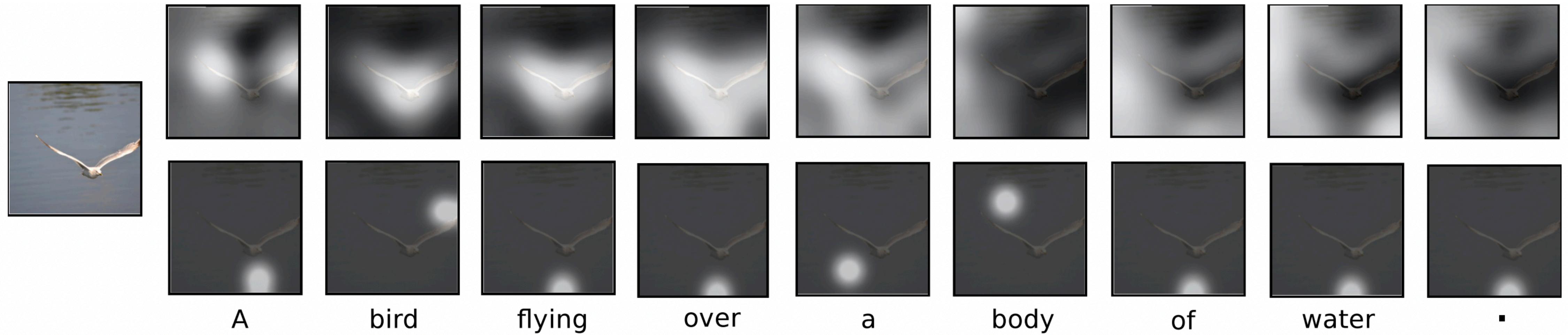
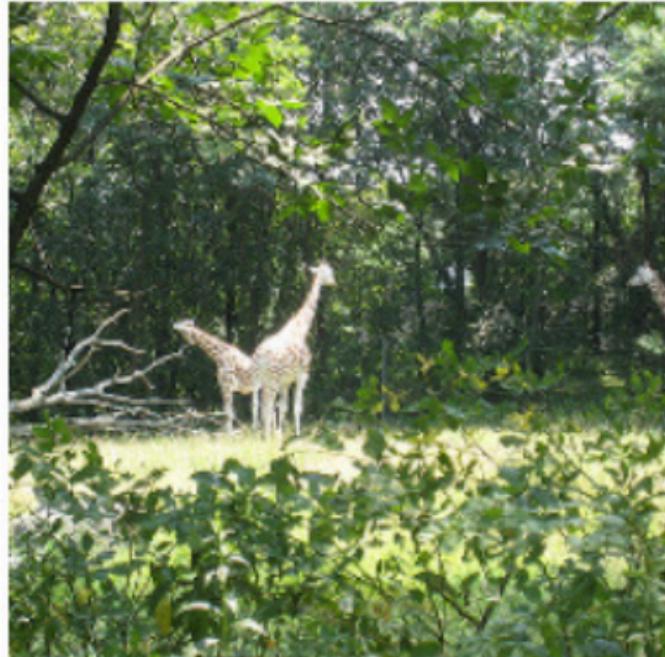


Image Captioning

Dataset	Model	BLEU				METEOR
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	
Flickr8k	Google NIC(Vinyals et al., 2014) $^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a) $^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	67	44.8	29.9	19.5	18.93
	Hard-Attention	67	45.7	31.4	21.3	20.30
Flickr30k	Google NIC $^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	18.49
	Hard-Attention	66.9	43.9	29.6	19.9	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014) a	—	—	—	—	20.41
	MS Research (Fang et al., 2014) $^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014) $^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC $^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear $^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	23.90
	Hard-Attention	71.8	50.4	35.7	25.0	23.04

(Xu et al., 2015)

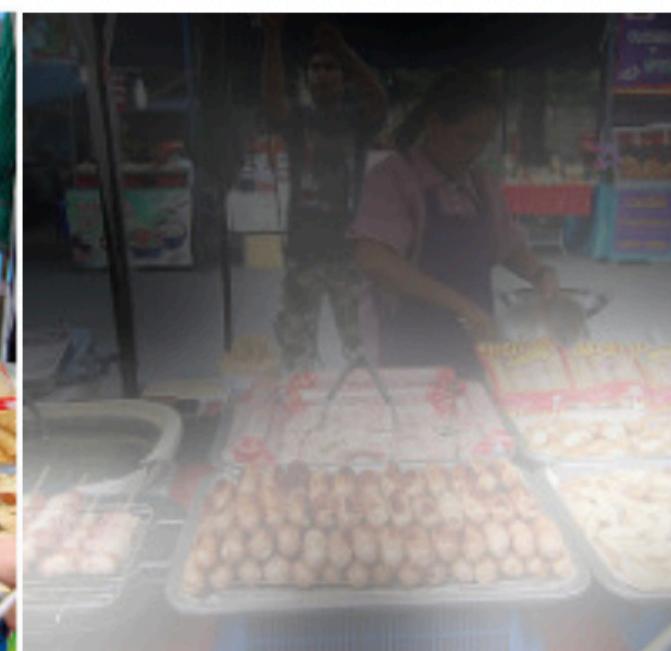
Image Captioning



A large white bird standing in a forest.

A woman holding a clock in her hand.

A man wearing a hat and a hat on a skateboard.



A person is standing on a beach with a surfboard.

A woman is sitting at a table with a large pizza.

A man is talking on his cell phone while another man watches.

Mistakes of the model.

(Xu et al., 2015)

Multimodal Machine Translation

A boy stand near the bank.



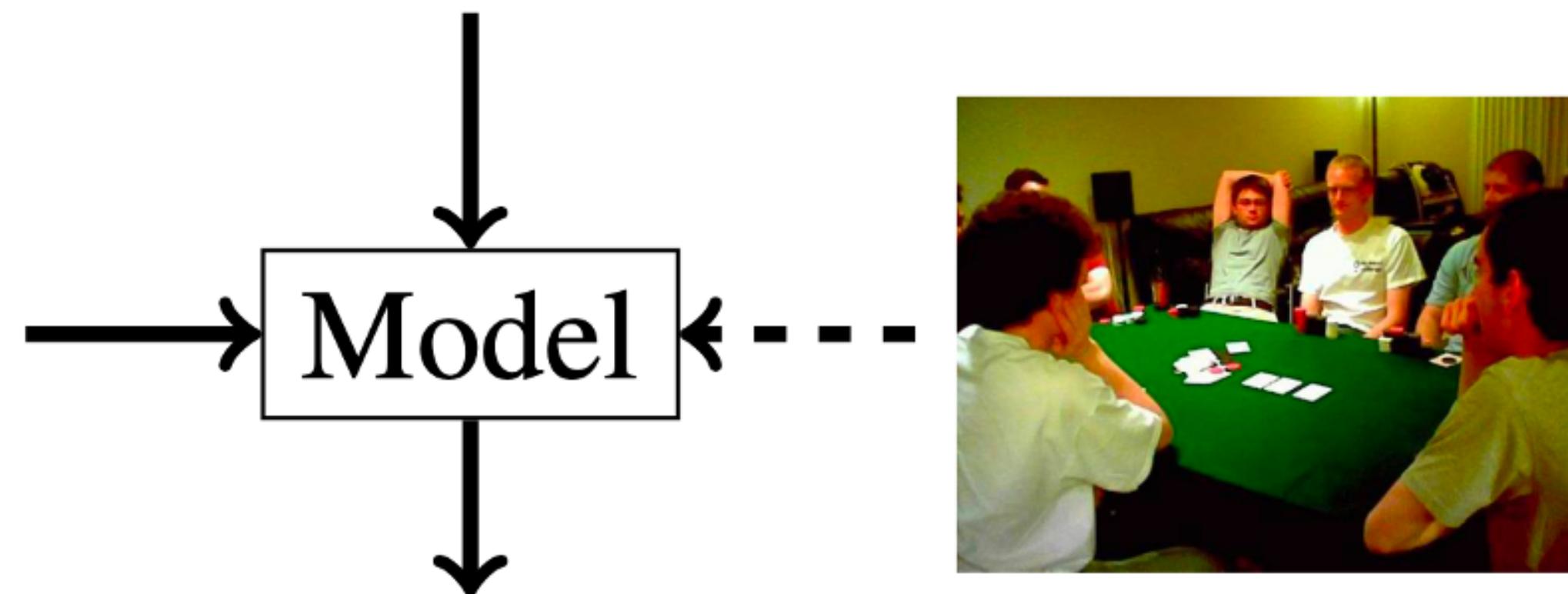
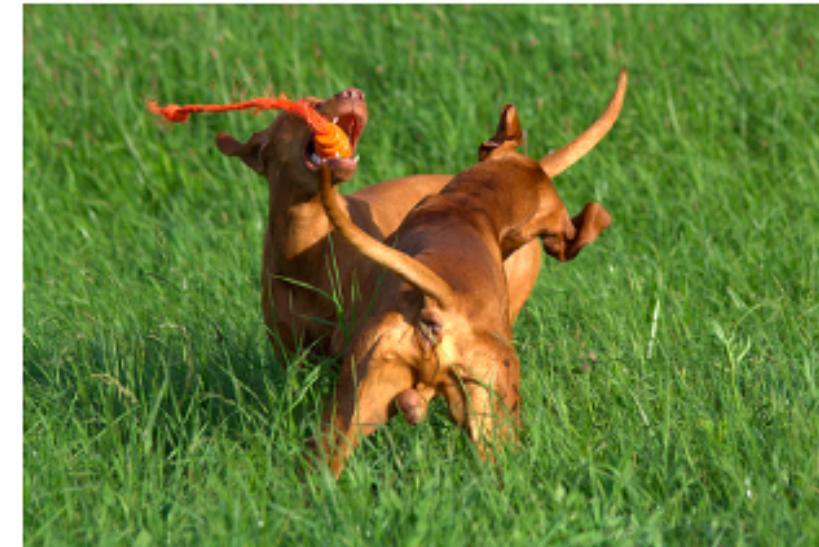
MMT System

一个男孩站在银行边。

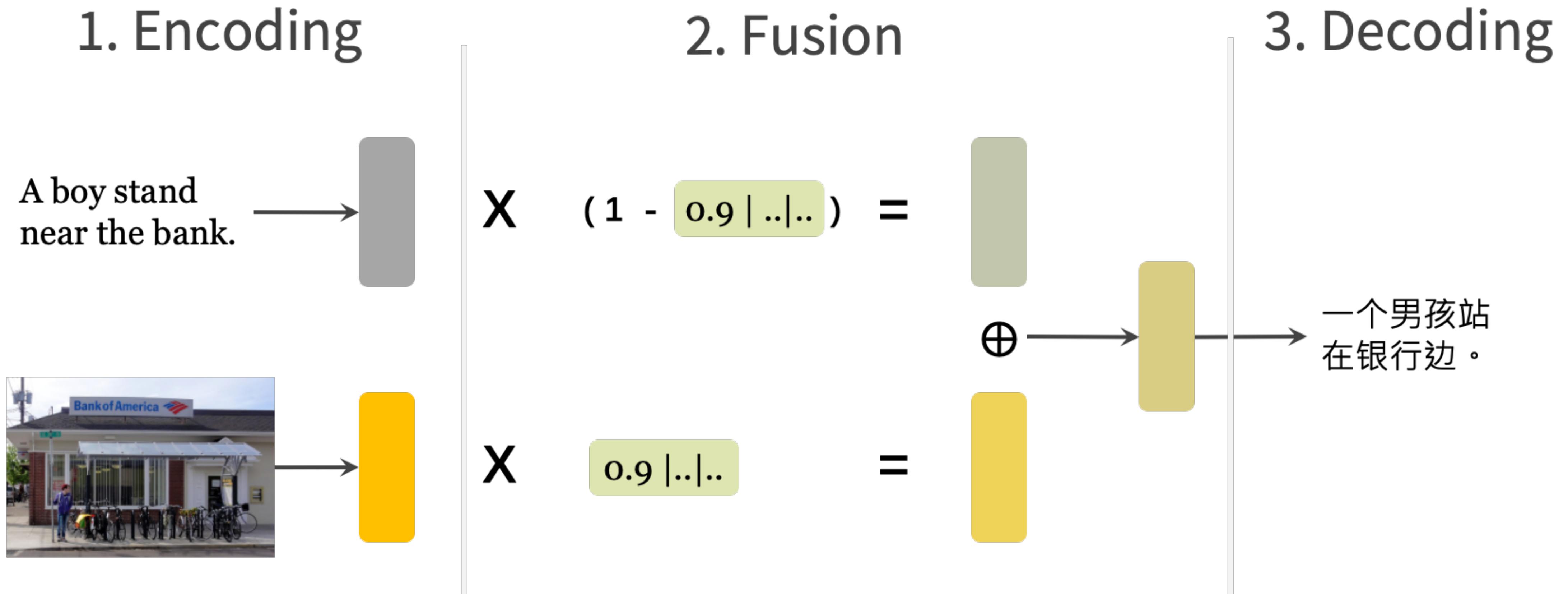
Multimodal Machine Translation

- Improvement is not evident (Elliott et al., 2017; Barrault et al., 2018)
- MMT models are insensitive to visual input (Elliott 2018; Gronroos et al., 2018)

Two dogs play with an
orange toy in tall grass.



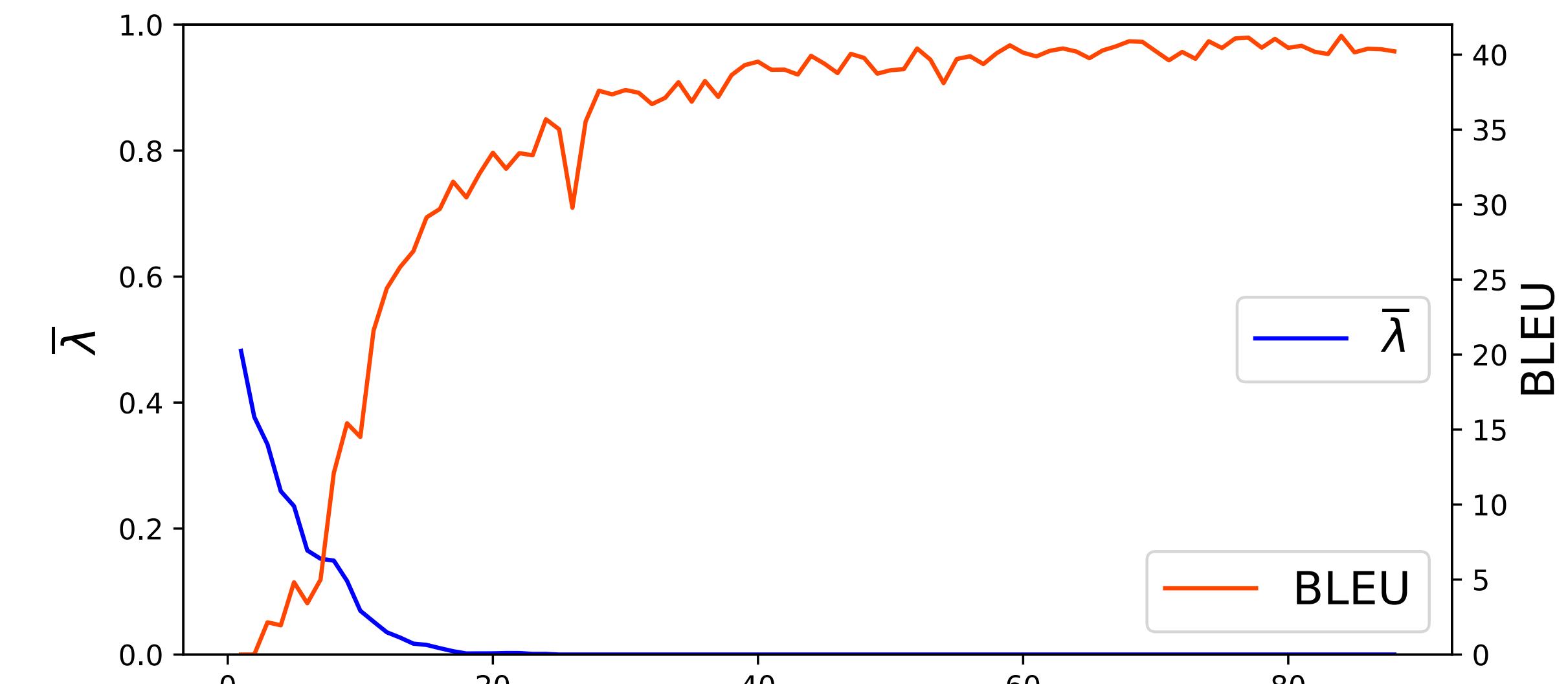
Multimodal Machine Translation



(Wu et al., 2021)

Multimodal Machine Translation

Multi30k	Gated Fusion
En→De	
Test2016	4.5E-21
Test2017	7.0E-17
MSCOCO	9.7E-21
En→Fr	
Test2016	1.6E-18
Test2017	7.2E-15
MSCOCO	2.3E-18

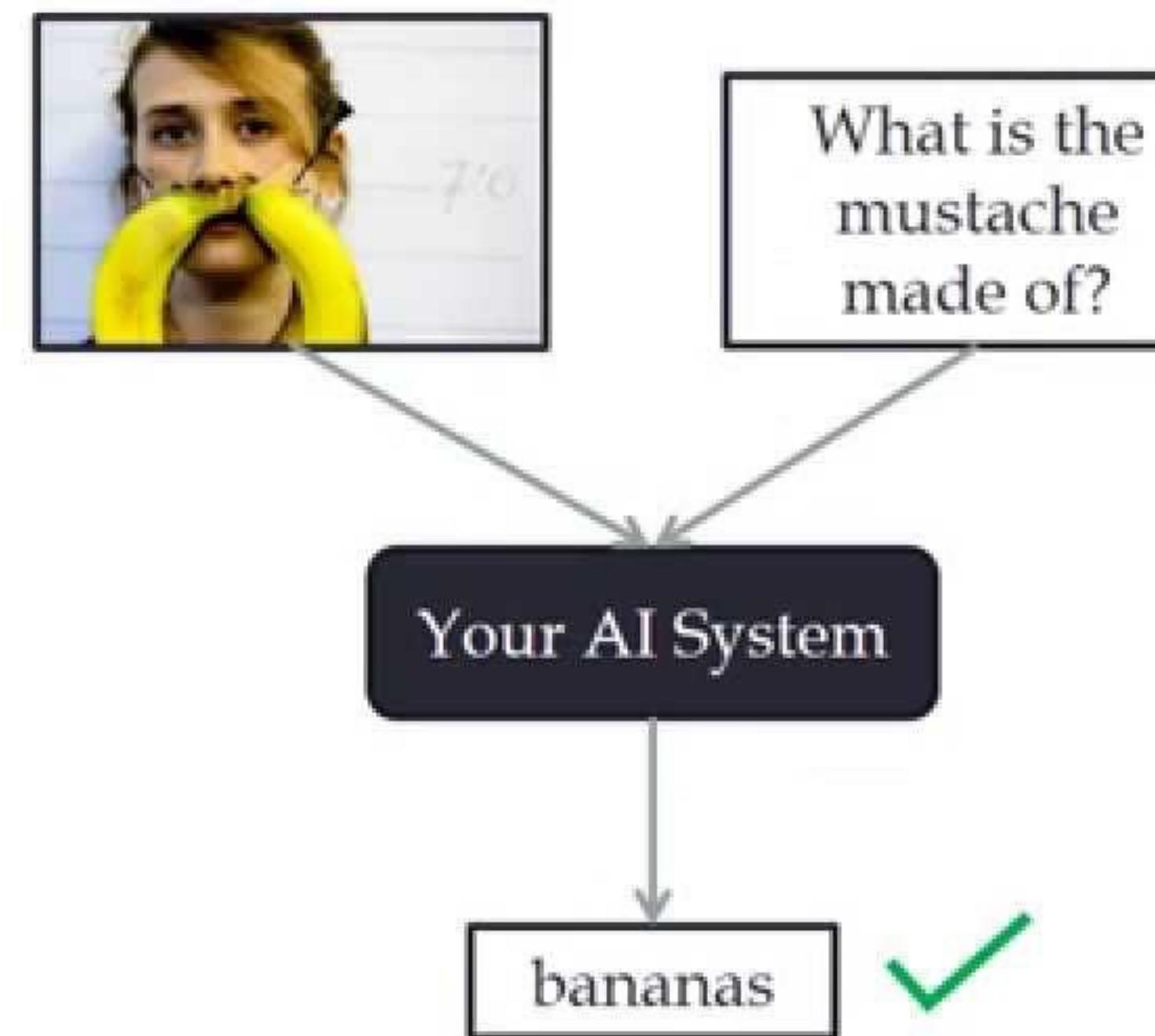


- After convergence: negligibly small averaged gating weight => both models learn to discard visual context during inference
- During training: the importance of visual context keep decreasing

Visual Question Answering (VQA)

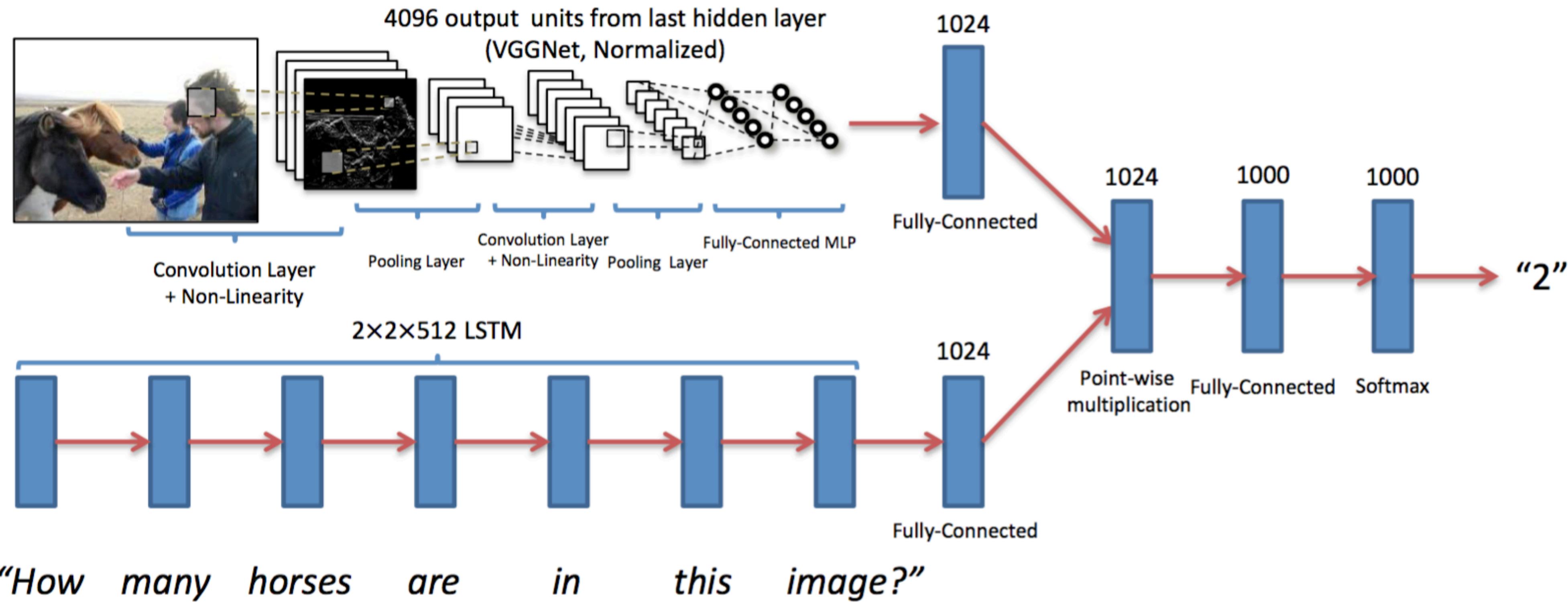
www.visualqa.org

VQA Challenges on www.codalab.org



Challenge Type	Last Updated	Participants
VQA Real Image Dev Evaluation (Multiple-Choice)	Oct 06, 2015 - No available	1 participant
VQA Real Image Dev Evaluation (Open-Ended)	Oct 06, 2015 - No available	1 participant
VQA Real Image Challenge (Multiple-Choice)	Oct 06, 2015 - No available	1 participant
VQA Real Image Challenge (Open-Ended)	Oct 06, 2015 - No available	1 participant
VQA Abstract Scene Challenge (Multiple-Choice)	Oct 06, 2015 - No available	1 participant
VQA Abstract Scene Challenge (Open-Ended)	Oct 06, 2015 - No available	1 participant

Visual Question Answering (VQA)



Visual Question Answering (VQA)



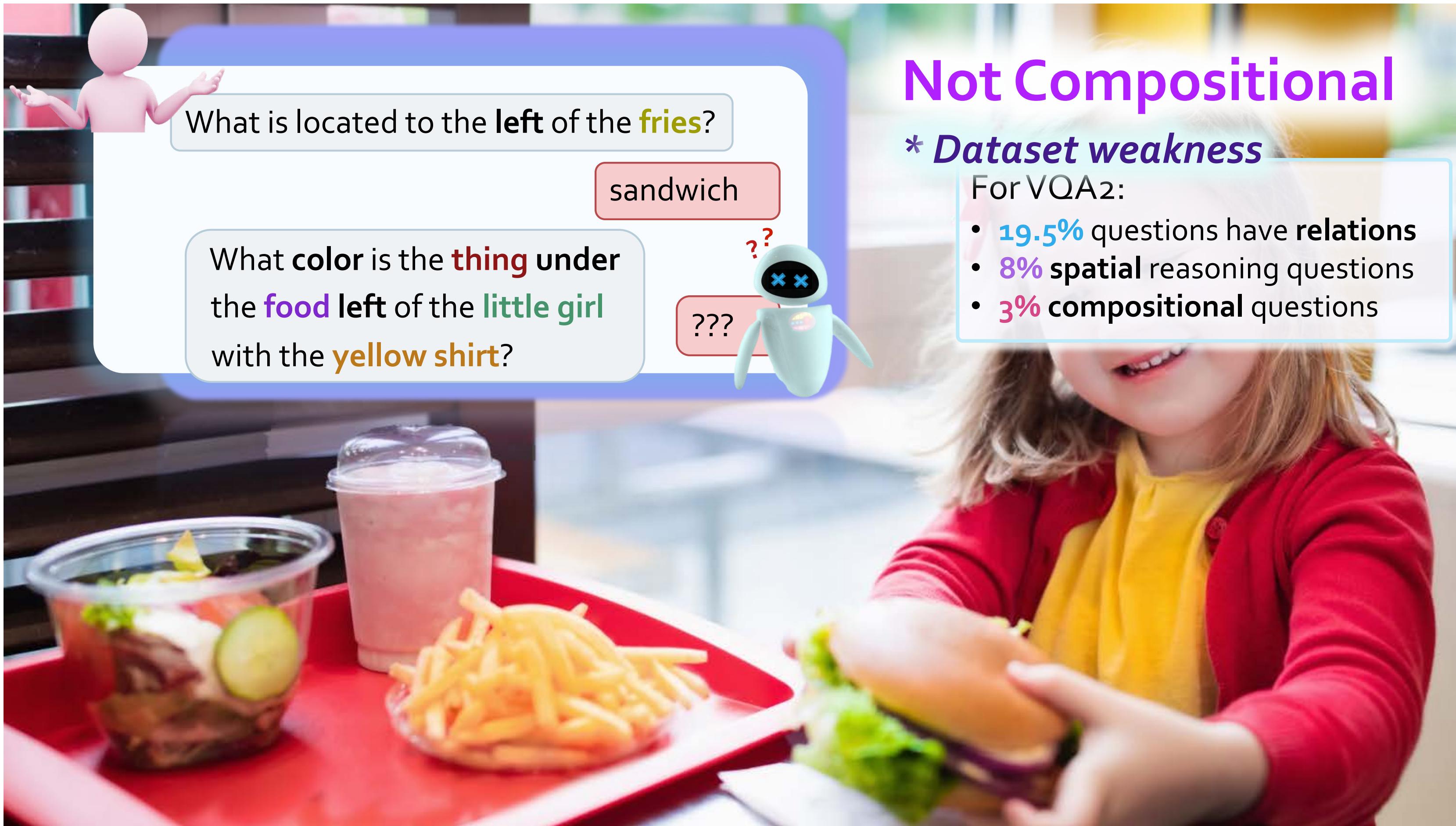
(Hudson and Manning, 2019)

Visual Question Answering (vQA)



(Hudson and Manning, 2019)

Visual Question Answering (VQA)



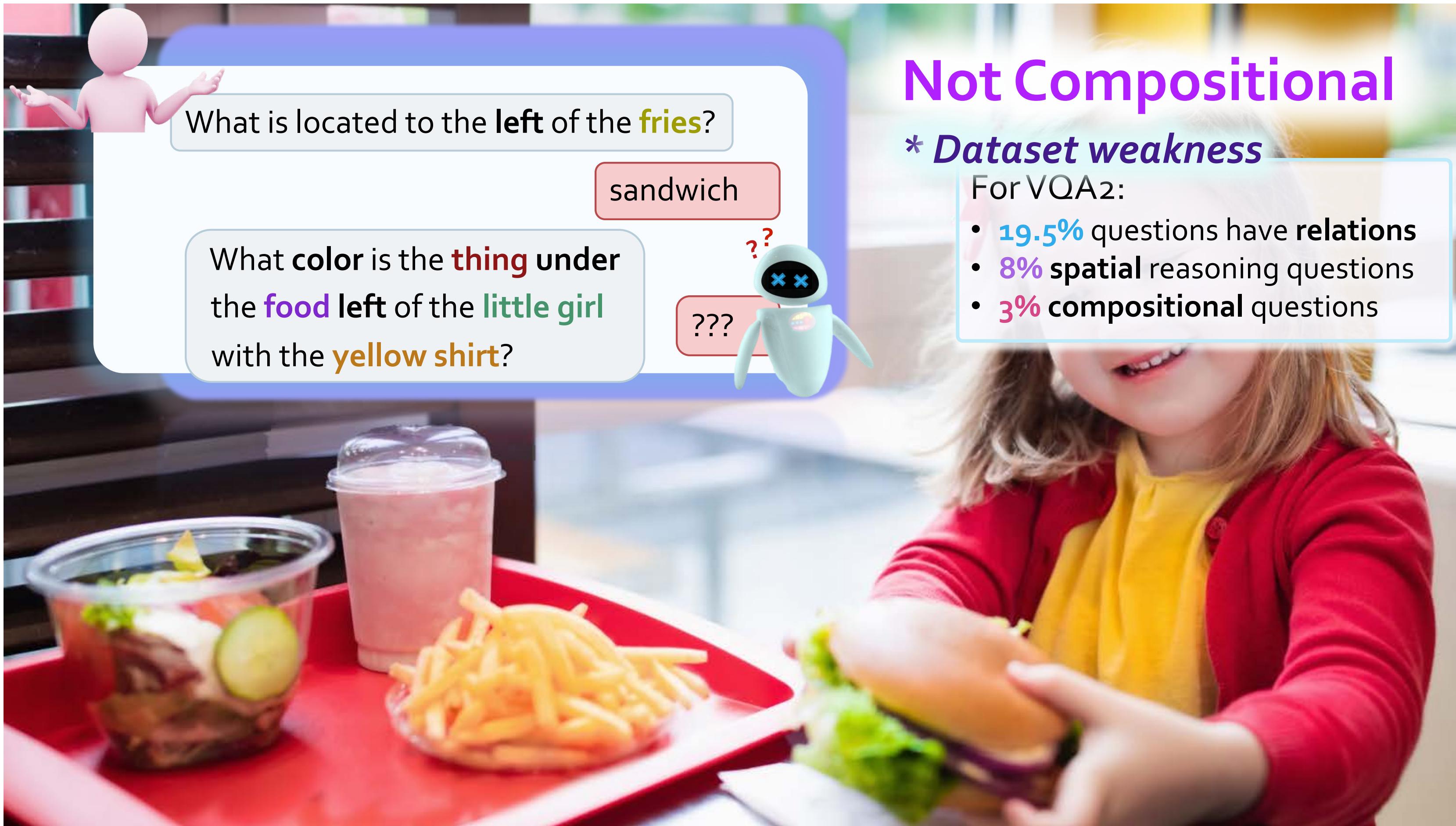
Not Compositional

* *Dataset weakness*

For VQA2:

- **19.5%** questions have **relations**
- **8%** **spatial reasoning** questions
- **3%** **compositional** questions

Visual Question Answering (VQA)



Not Compositional

* *Dataset weakness*

For VQA2:

- **19.5%** questions have **relations**
- **8%** **spatial reasoning** questions
- **3%** **compositional** questions

GQA: Real-World Visual Reasoning and Compositional Question Answering

Example Questions



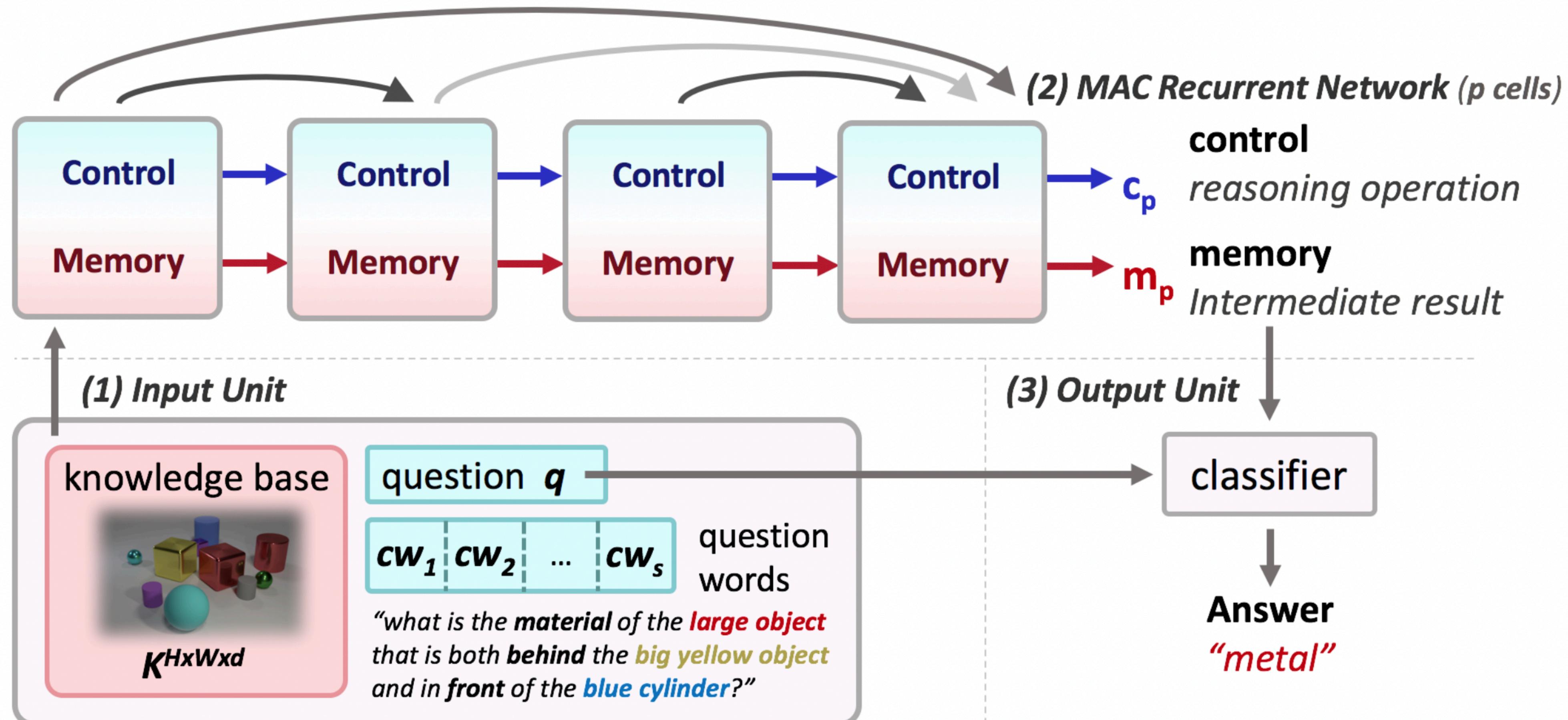
VQA

1. Does this **man** need a **haircut**?
2. What **color** is the **guy's tie**?
3. What is different about the **man's suit** that shows this is for a special occasion?

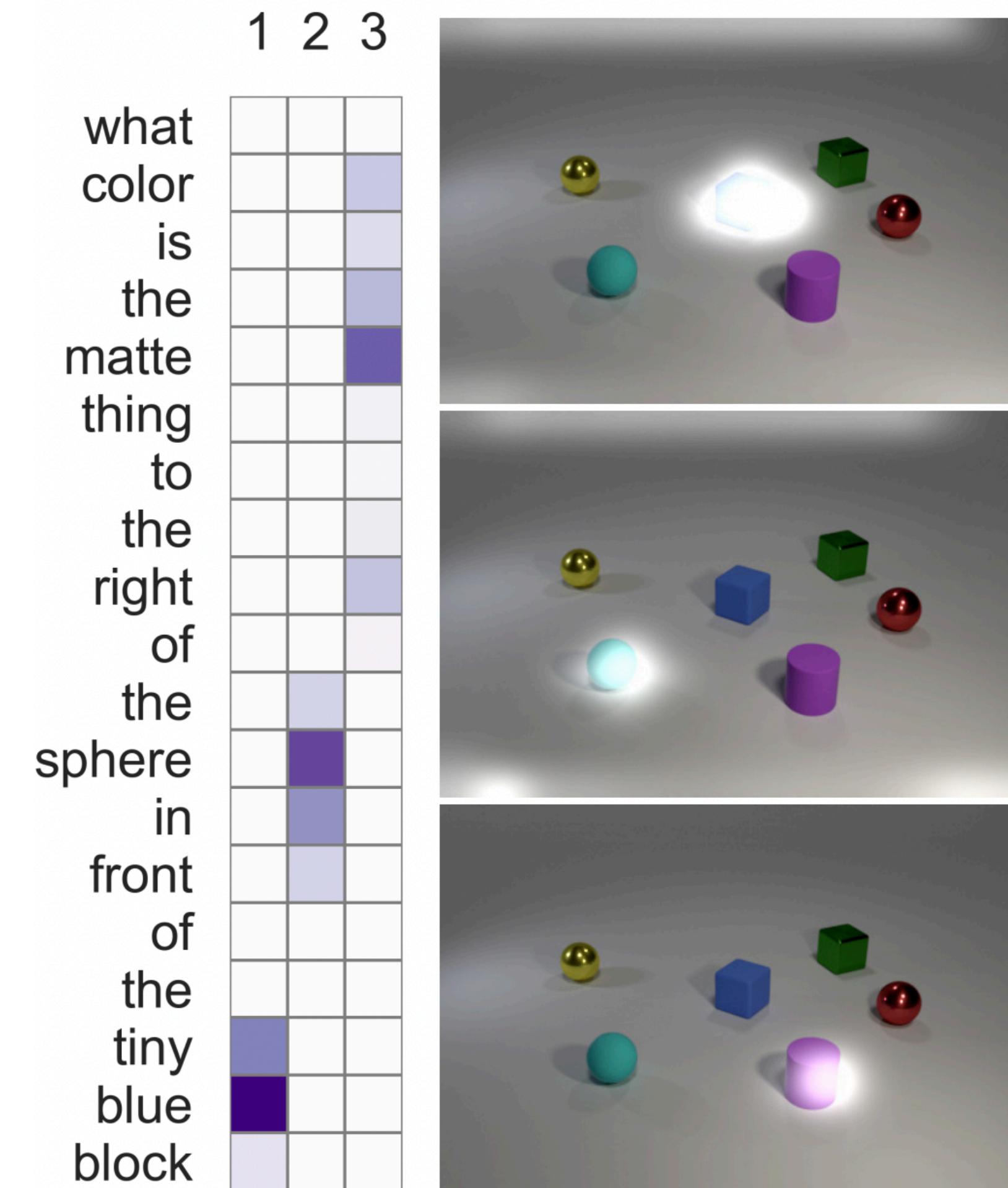
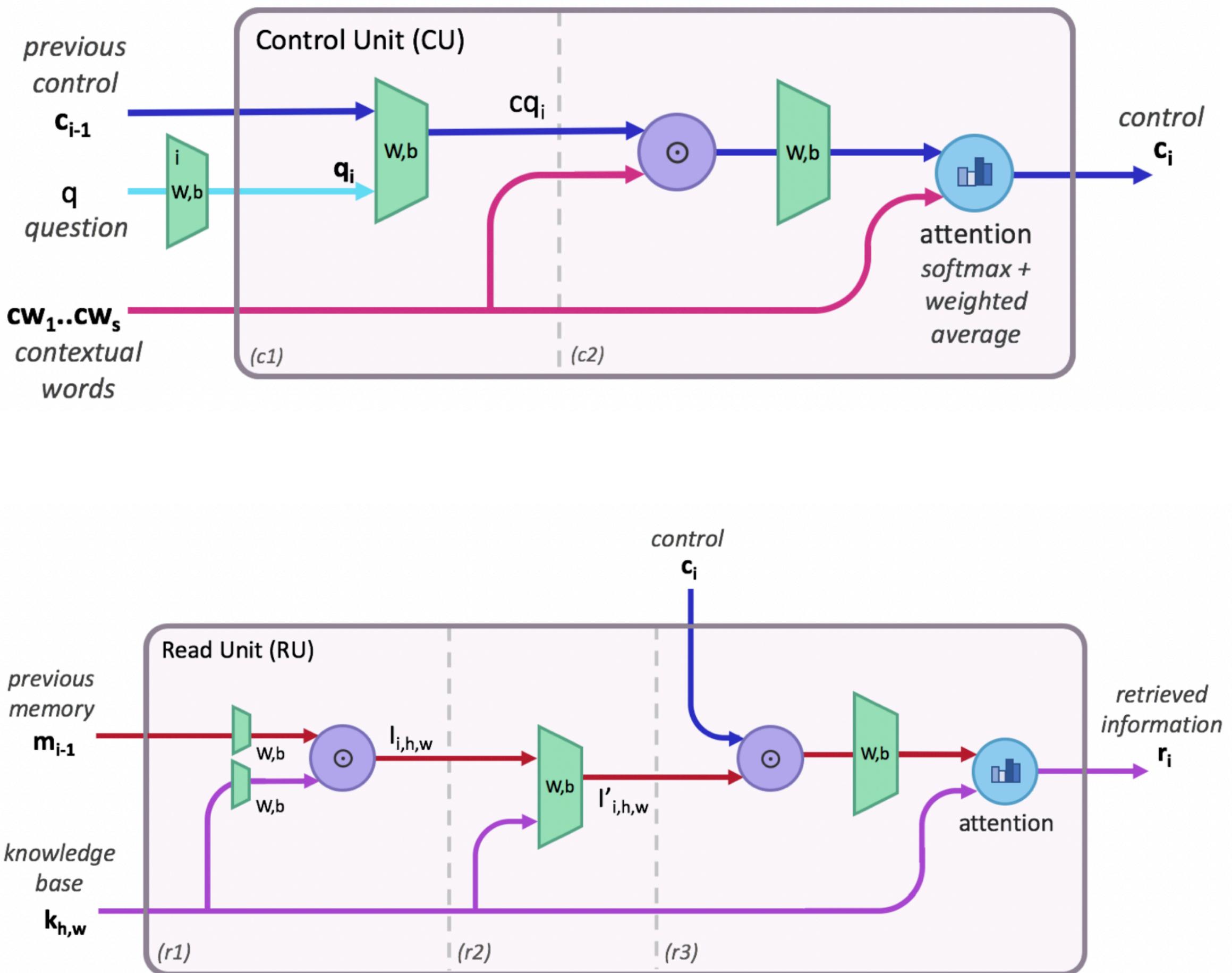
GQA

1. Is the **person's hair** long and brown?
2. What **appliance** is to the left of the **man**?
3. Who is in front of the **refrigerator** on the left?
4. Is there a **necktie** in the picture that is not red?
5. Is the color of the **vest** different than **shirt**?

Compositional Attention Networks

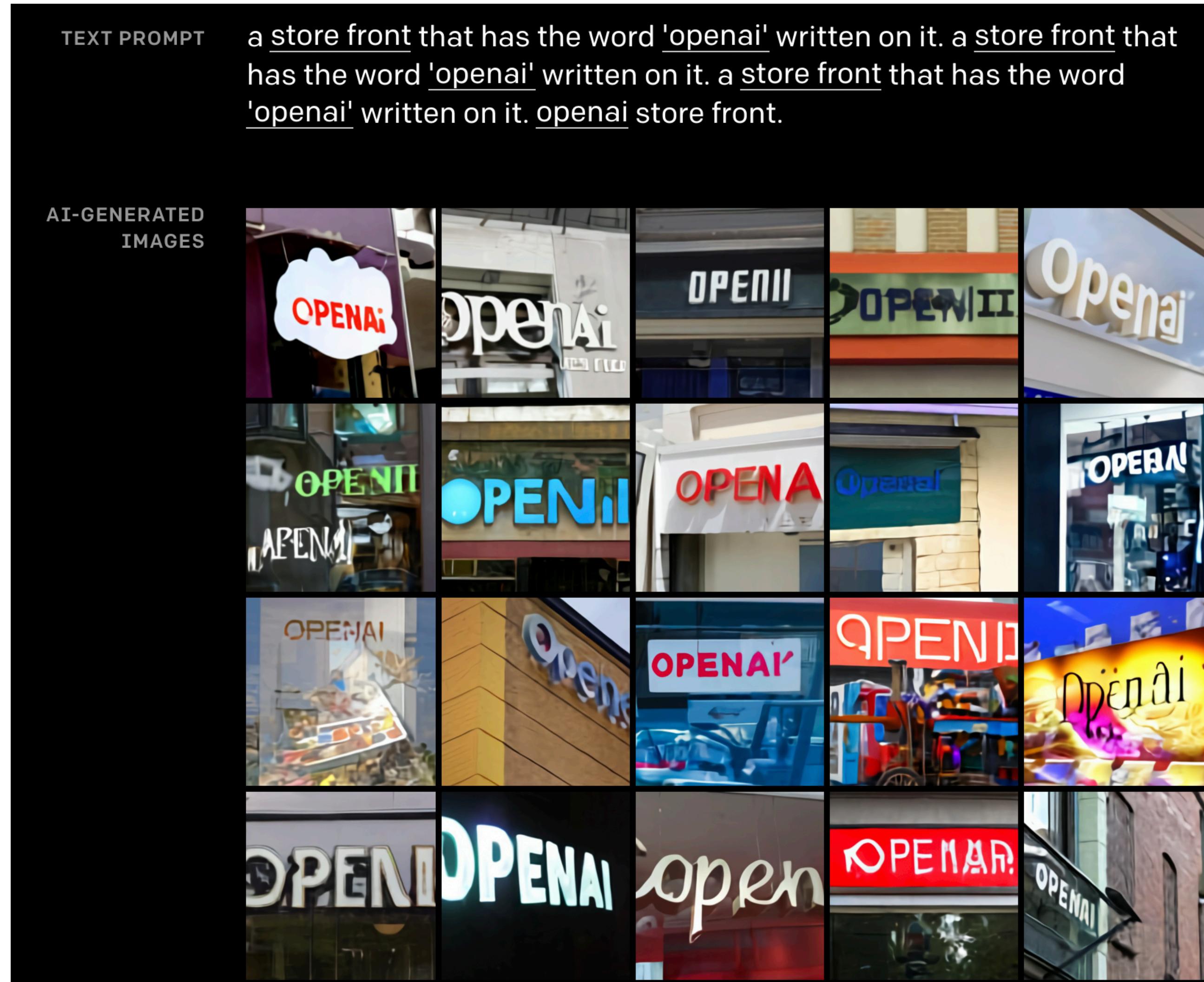


Compositional Attention Networks



(Hudson and Manning, 2018)

Image Generation



DALL-E: Creating Images from Text

Image Generation



DALL-E: Creating Images from Text

Image Generation

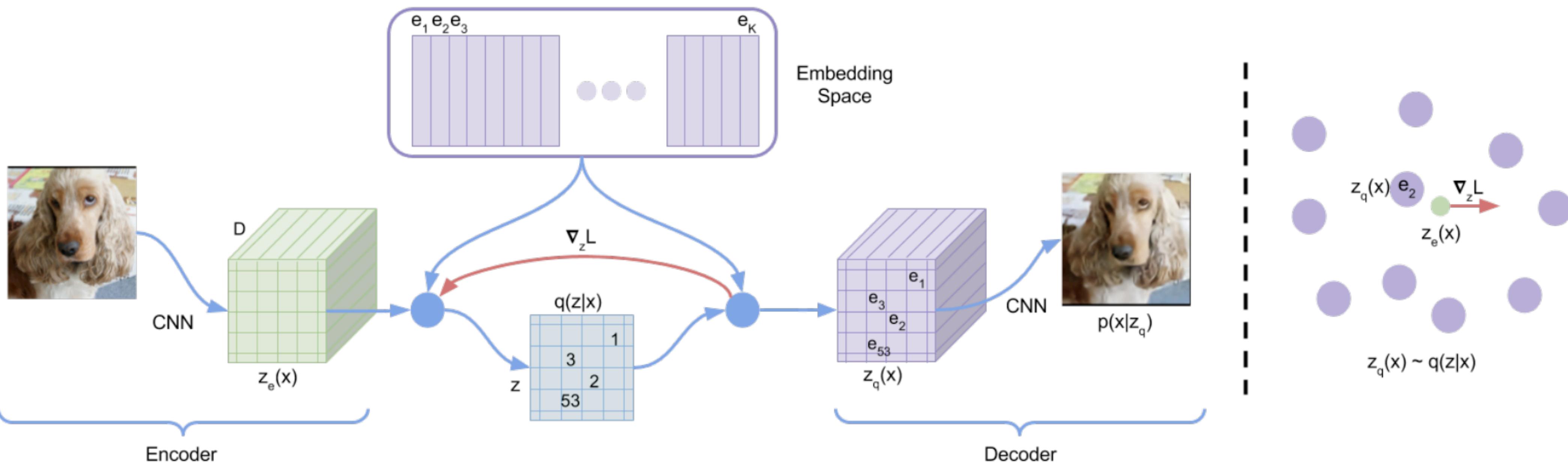


Figure 1: Left: A figure describing the VQ-VAE. Right: Visualisation of the embedding space. The output of the encoder $z(x)$ is mapped to the nearest point e_2 . The gradient $\nabla_z L$ (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass.