Deep Learning for Natural Language Processing

Stephen Clark et al.
University of Cambridge and DeepMind

DeepMind

UNIVERSITY OF CAMBRIDGE

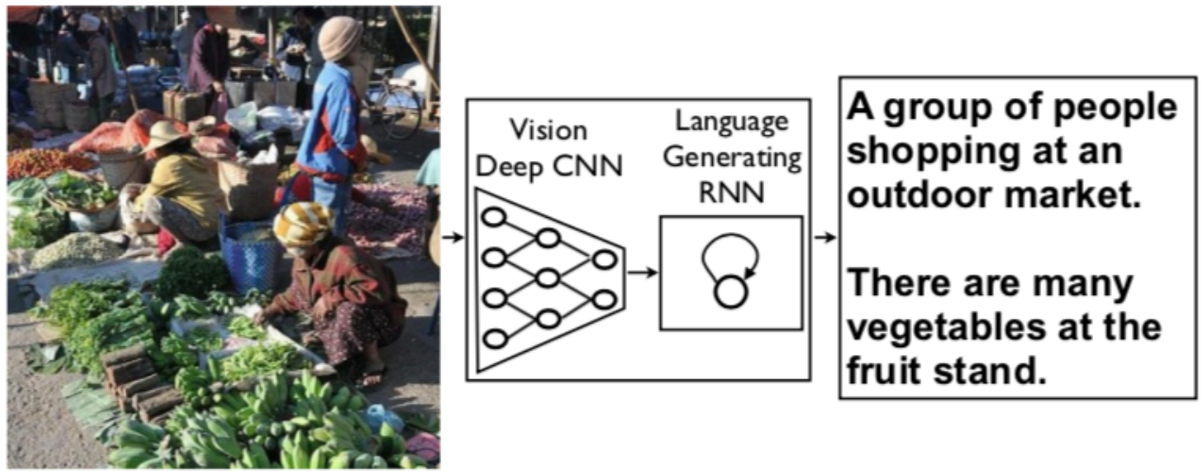
14. Image Captioning

Stephen Clark
University of Cambridge and DeepMind

DeepMind

UNIVERSITY OF CAMBRIDGE

Image Captioning



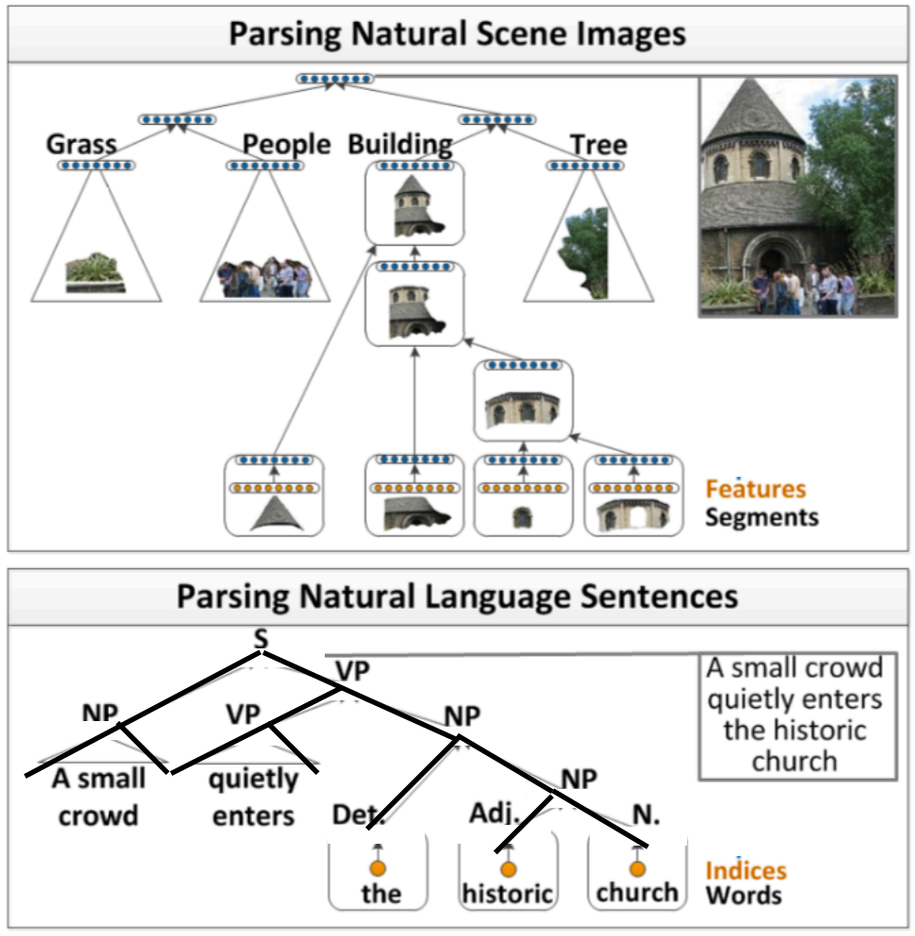
Vision **Language**

Taken from Vinyals et al. 2015

DeepMind

UNIVERSITY OF CAMBRIDGE

Image Captioning

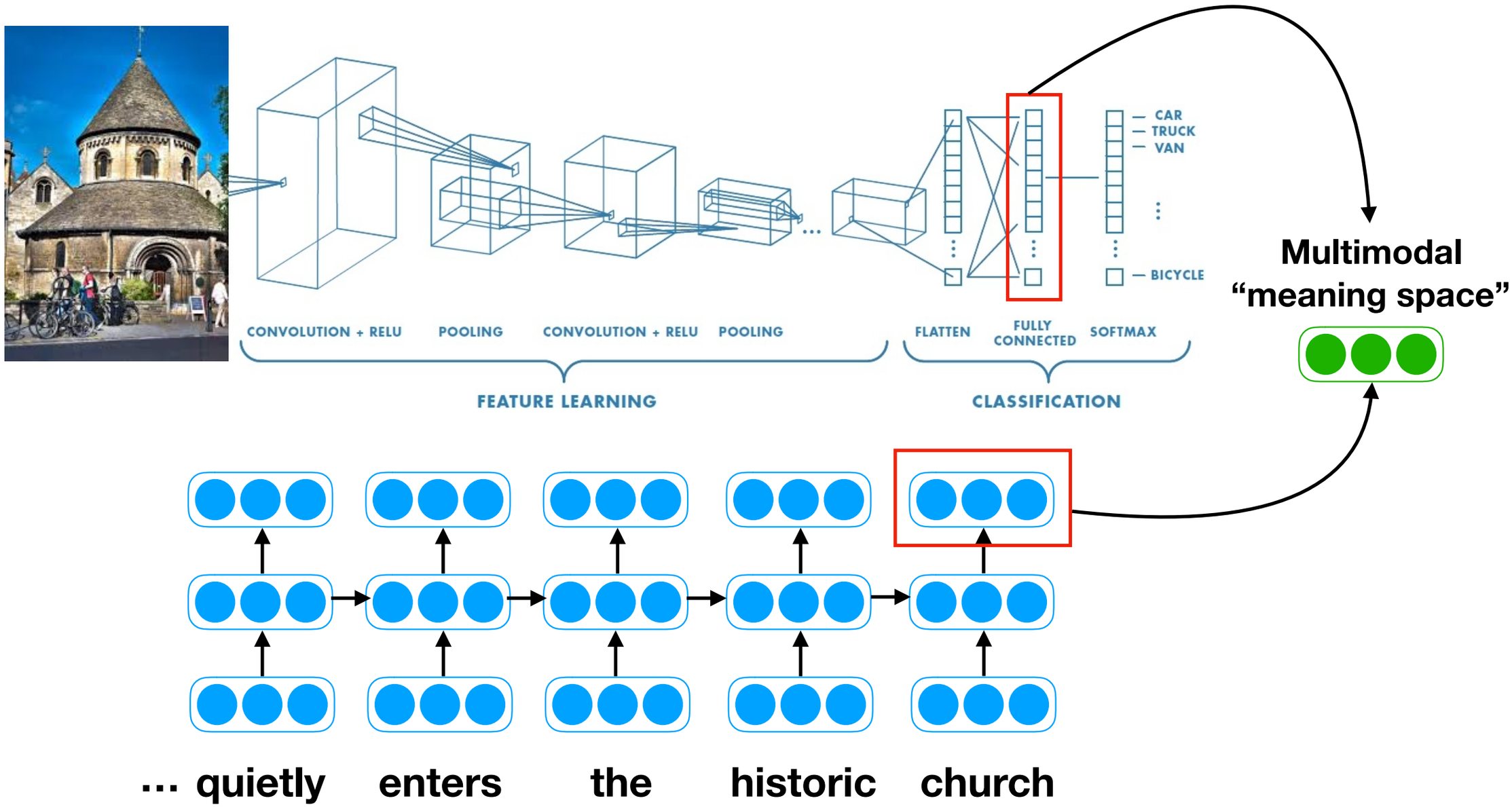


Modified from Socher et al. 2011

DeepMind

UNIVERSITY OF CAMBRIDGE

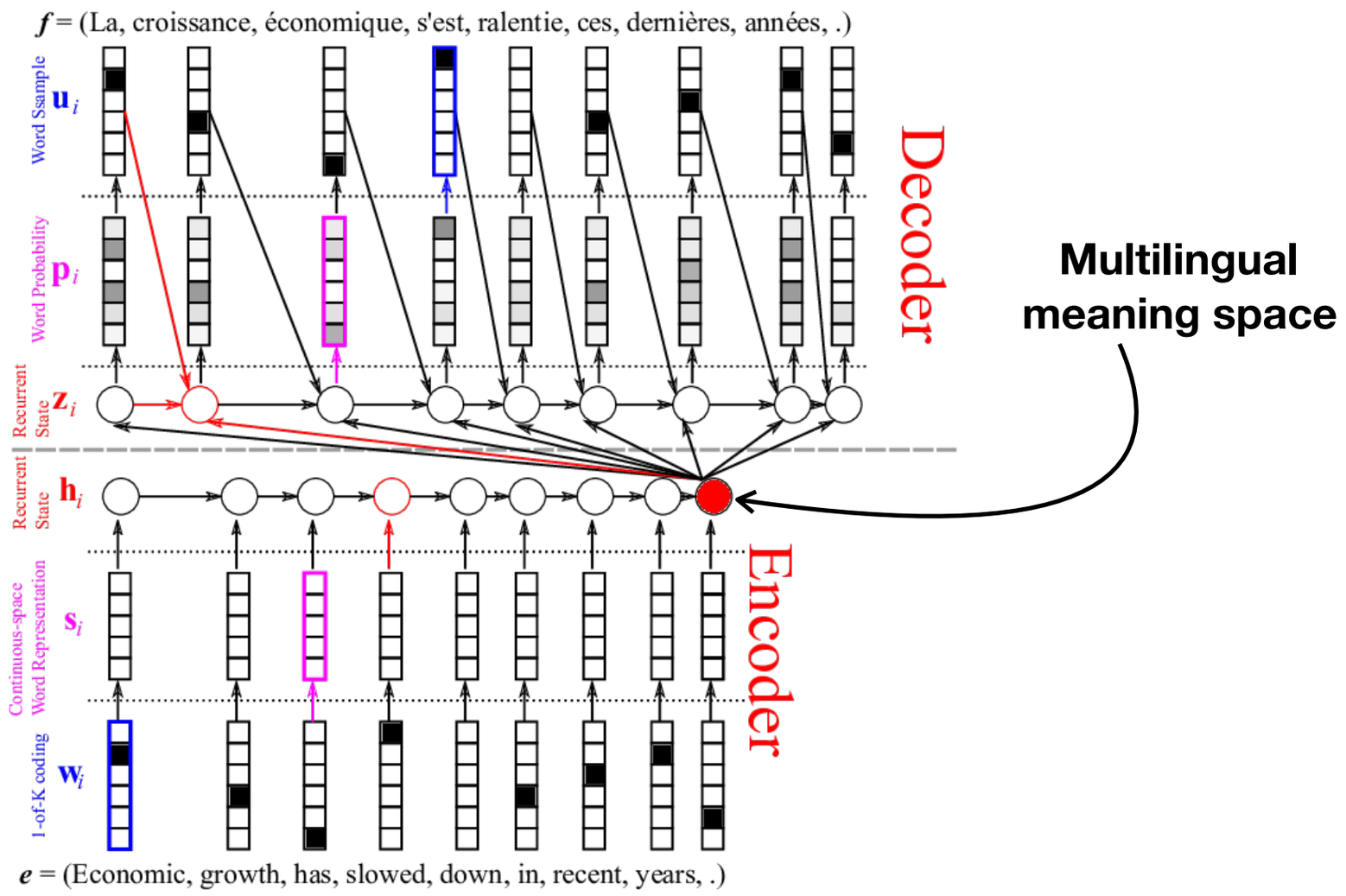
Image Captioning



DeepMind

UNIVERSITY OF CAMBRIDGE

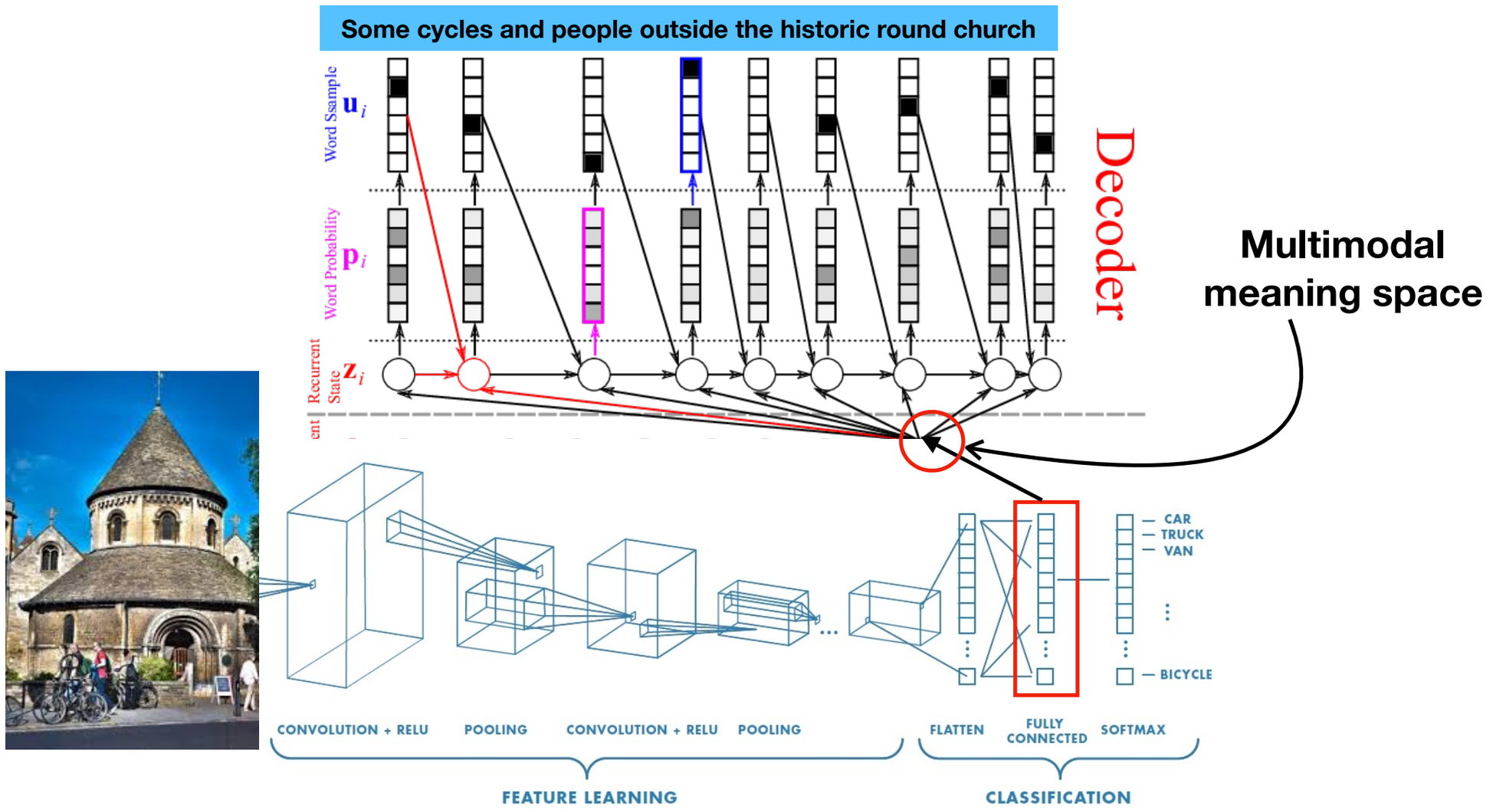
Image “Translation”



DeepMind

UNIVERSITY OF CAMBRIDGE

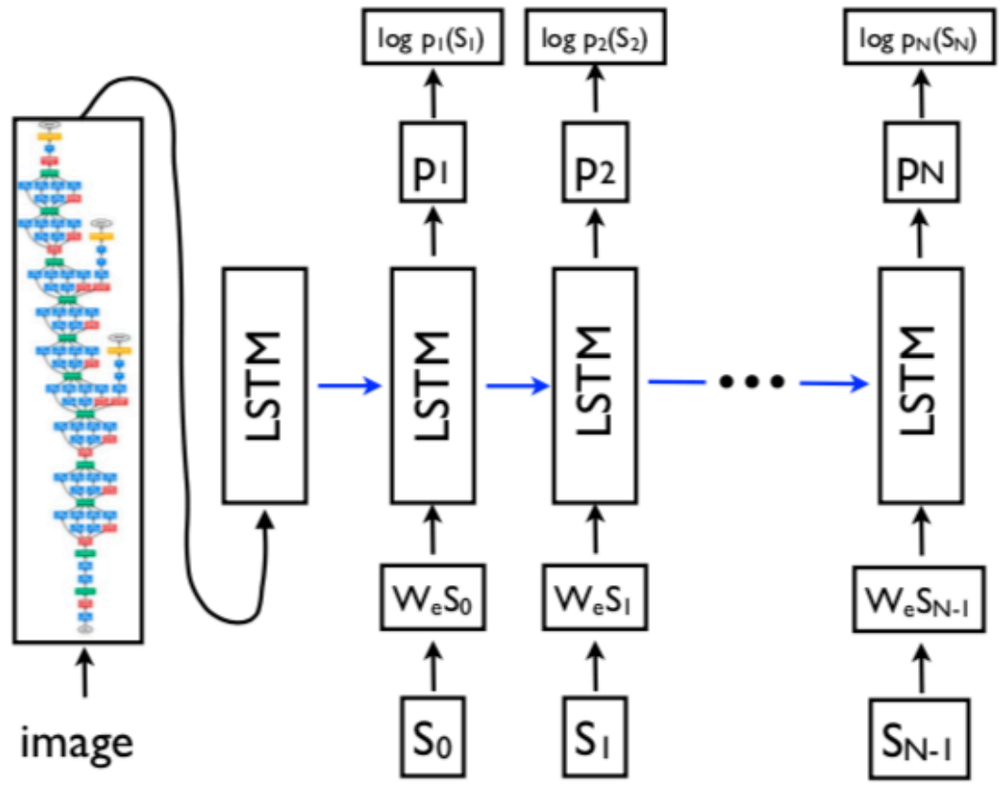
Image “Translation”



DeepMind

UNIVERSITY OF CAMBRIDGE

Caption Model



Taken from Vinyals et al. 2015: *Show and Tell: A Neural Image Caption Generator*

DeepMind

UNIVERSITY OF CAMBRIDGE

Model Optimization

$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta) \tag{1}$$

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1}) \tag{2}$$

Optimized using stochastic gradient descent

Taken from Vinyals et al. 2015

DeepMind

UNIVERSITY OF CAMBRIDGE

Is That It?! Yup, Pretty Much

State-of-the-art CNN for object classification; LSTM for the sentence generation

Image is input only once to the LSTM at the beginning

LSTM trained from scratch, only top layer of CNN retrained

CNN pre-trained on object classification; no pre-training of word embeddings

Beam search (20) used to perform the argmax at test time (better than greedy)

$$S = \arg\max_{S'} p(S'|I).$$

SGD with fixed learning rate and no momentum

Dropout and ensembles used to combat overfitting

DeepMind

UNIVERSITY OF CAMBRIDGE

Datasets

The statistics of the datasets are as follows:

Dataset name	size		
	train	valid.	test
Pascal VOC 2008 [6]	-	-	1000
Flickr8k [26]	6000	1000	1000
Flickr30k [33]	28000	1000	1000
MSCOCO [20]	82783	40504	40775
SBU [24]	1M	-	-

With the exception of SBU, each image has been annotated by labelers with 5 sentences that are relatively visual and unbiased. SBU consists of descriptions given by image owners when they uploaded them to Flickr. As such they are not guaranteed to be visual or unbiased and thus this dataset has more noise.

Taken from Vinyals et al. 2015

DeepMind

UNIVERSITY OF CAMBRIDGE

Results

Metric	BLEU-4	METEOR	CIDER
NIC	**27.7**	**23.7**	**85.5**
Random	4.6	9.0	5.1
Nearest Neighbor	9.9	15.7	36.5
Human	21.7	25.2	85.4

Table 1. Scores on the MSCOCO development set.

Taken from Vinyals et al. 2015

DeepMind

UNIVERSITY OF CAMBRIDGE

Results

Approach	PASCAL (xfer)	Flickr 30k	Flickr 8k	SBU
Im2Text [24]				11
TreeTalk [18]				19
BabyTalk [16]	25			
Tri5Sem [11]			48	
m-RNN [21]		55	58	
MNLM [14][5]		56	51	
SOTA	25	56	58	19
NIC	**59**	**66**	**63**	**28**
Human	69	68	70	

Table 2. BLEU-1 scores. We only report previous work results when available. SOTA stands for the current state-of-the-art.

Taken from Vinyals et al. 2015

DeepMind

UNIVERSITY OF CAMBRIDGE

Generation Diversity

A man throwing a frisbee in a park.
A man holding a frisbee in his hand.
A man standing in the grass with a frisbee.
A close up of a sandwich on a plate.
A close up of a plate of food with french fries.
A white plate topped with a cut in half sandwich.
A display case filled with lots of donuts.
A display case filled with lots of cakes.
A bakery display case filled with lots of donuts.

Table 3. N-best examples from the MSCOCO test set. Bold lines indicate a novel sentence not present in the training set.

Taken from Vinyals et al. 2015

DeepMind

UNIVERSITY OF CAMBRIDGE

Word Embeddings

Word	Neighbors
car	van, cab, suv, vehicule, jeep
boy	toddler, gentleman, daughter, son
street	road, streets, highway, freeway
horse	pony, donkey, pig, goat, mule
computer	computers, pc, crt, chip, compute

Table 6. Nearest neighbors of a few example words

Taken from Vinyals et al. 2015

DeepMind

UNIVERSITY OF CAMBRIDGE

Example Output

Figure 5. A selection of evaluation results, grouped by human rating.

Taken from Vinyals et al. 2015

DeepMind

UNIVERSITY OF CAMBRIDGE

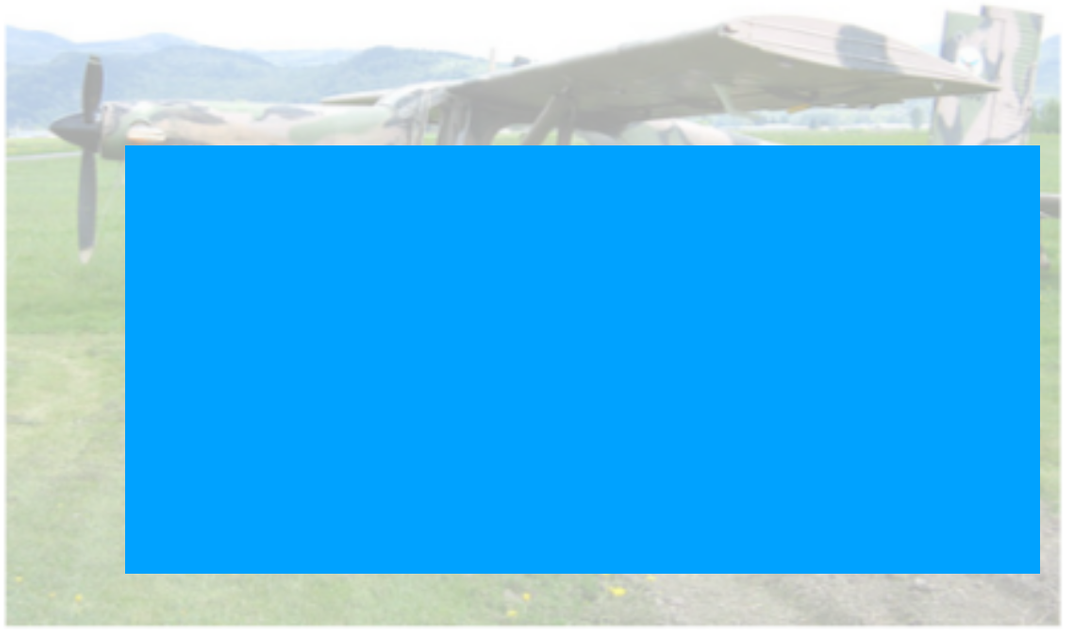
Captions with Attention

- *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,* Xu et al. 2015
- Nice demo at *http://kelvinxu.github.io/projects/capgen.html*

DeepMind

UNIVERSITY OF CAMBRIDGE

How Hard is the Task?



DeepMind

UNIVERSITY OF CAMBRIDGE

How Hard is the Task?

- A camouflaged plane sitting on the green grass.
- A plane painted in camouflage in a grassy field
- A small camouflaged airplane parked in the grass.
- Camouflage airplane sitting on grassy field.
- Parked camouflage high wing aircraft.

These examples from:
http://www.cs.toronto.edu/~fidler/slides/2017/CSC2539/Kaustav_slides.pdf

DeepMind

UNIVERSITY OF CAMBRIDGE

How Hard is the Task?

DeepMind

UNIVERSITY OF CAMBRIDGE

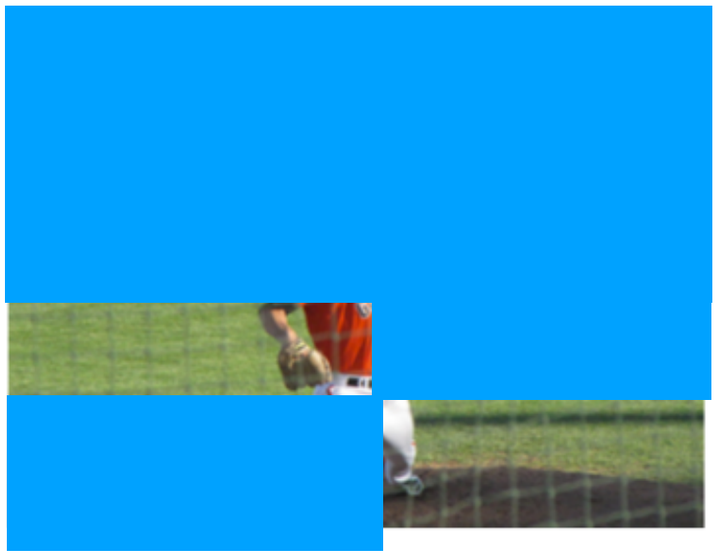
How Hard is the Task?

- A biker in red rides in the countryside.
- A biker on a dirt path.
- A person rides a bike off the top of a hill and is airborne.
- A person riding a bmx bike on a dirt course.
- The person on the bicycle is wearing red.

DeepMind

UNIVERSITY OF CAMBRIDGE

How Hard is the Task?



DeepMind

UNIVERSITY OF CAMBRIDGE

How Hard is the Task?

- A baseball winds up to pitch the ball.
- A pitcher throwing the ball in a baseball game.
- A pitcher throwing a baseball on the mound.
- A baseball player pitching a ball on the mound.
- A left-handed pitcher throwing for the San Francisco giants.

DeepMind

UNIVERSITY OF CAMBRIDGE

How Hard is the Task?

DeepMind

UNIVERSITY OF CAMBRIDGE

How Hard is the Task?

Sentences

1) A girl is eating donuts with a boy in a restaurant
2) A boy and girl sitting at a table with doughnuts.
3) Two kids sitting a coffee shop eating some frosted donuts
4) Two children sitting at a table eating donuts.
5) Two children eat doughnuts at a restaurant table.

DeepMind

UNIVERSITY OF CAMBRIDGE

Visual Question Answering

What Does the Sign Say?

DeepMind

UNIVERSITY OF CAMBRIDGE

Visual Question Answering

What does the sign say?	stop stop stop	stop stop yield
What shape is this sign?	octagon octagon octagon	diamond octagon round

What Does the Sign Say?

Taken from Agrawal et al. 2016

DeepMind

UNIVERSITY OF CAMBRIDGE

Visual Question Answering

What is the mustache made of?

DeepMind

UNIVERSITY OF CAMBRIDGE

Visual Question Answering

What color are her eyes?
What is the mustache made of?

What is the mustache made of?

Taken from Agrawal et al. 2016

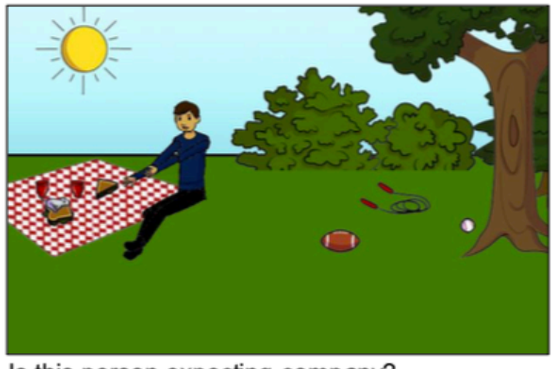
DeepMind

UNIVERSITY OF CAMBRIDGE

Visual Question Answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

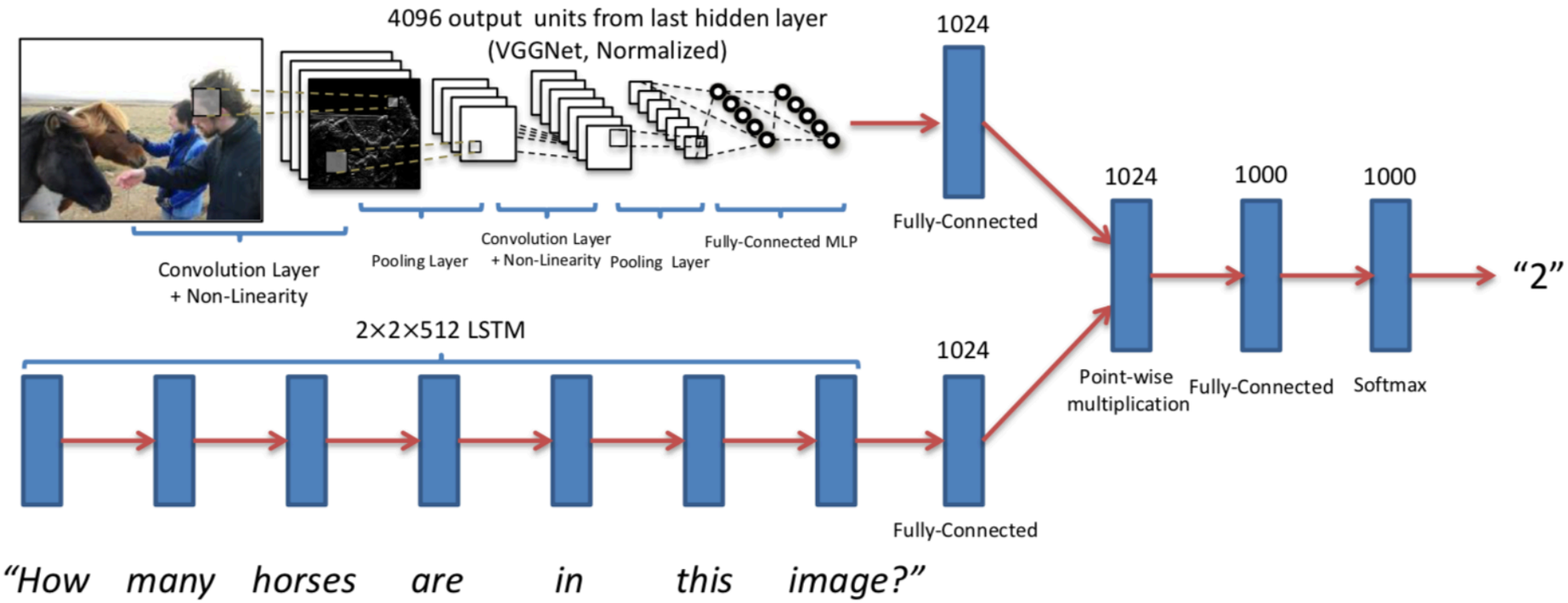
Fig. 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.

Taken from Agrawal et al. 2016

DeepMind

UNIVERSITY OF CAMBRIDGE

Visual Question Answering



Taken from Agrawal et al. 2016

DeepMind

UNIVERSITY OF CAMBRIDGE

CLEVR Dataset

Abstract

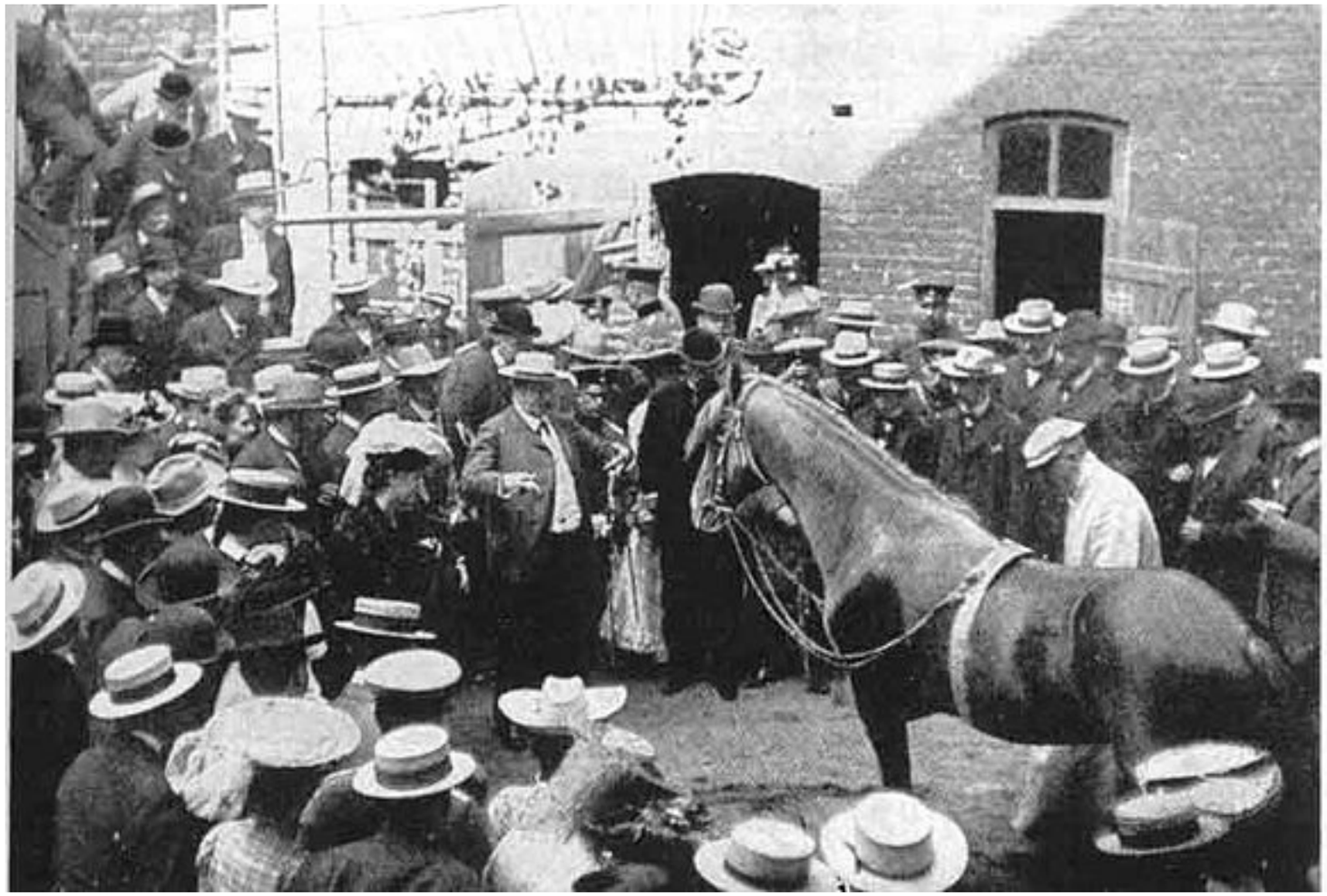
When building artificial intelligence systems that can reason and answer questions about visual data, we need diagnostic tests to analyze our progress and discover shortcomings. Existing benchmarks for visual question answering can help, but have strong biases that models can exploit to correctly answer questions without reasoning. They also conflate multiple sources of error, making it hard to pinpoint model weaknesses. We present a diagnostic dataset that tests a range of visual reasoning abilities. It contains minimal biases and has detailed annotations describing the kind of reasoning each question requires. We use this dataset to analyze a variety of modern visual reasoning systems, providing novel insights into their abilities and limitations.

Taken from Johnson et al. 2016: CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning

DeepMind

UNIVERSITY OF CAMBRIDGE

Clever Hans

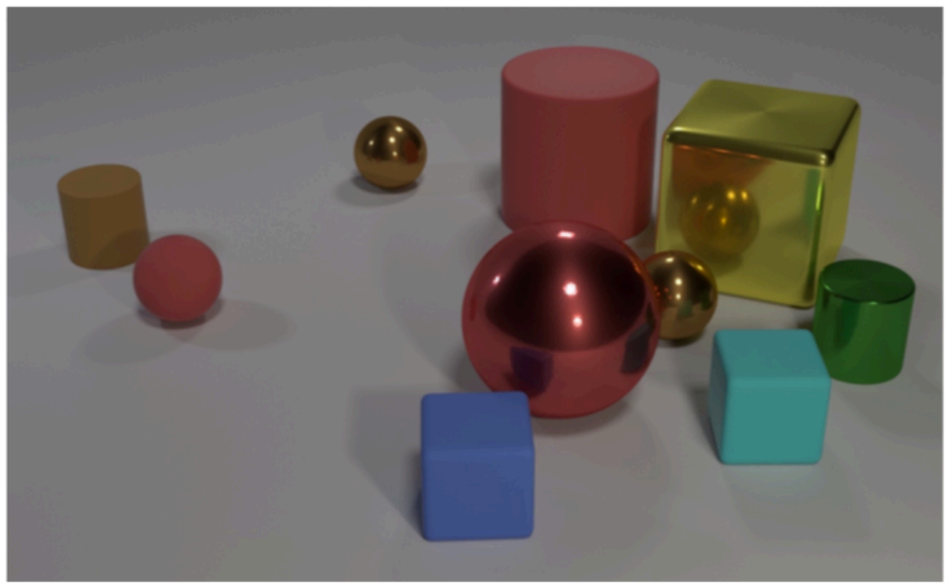


Taken from: https://en.wikipedia.org/wiki/Clever_Hans

DeepMind

UNIVERSITY OF CAMBRIDGE

CLEVR



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?
Q: How many objects are either small cylinders or metal things?

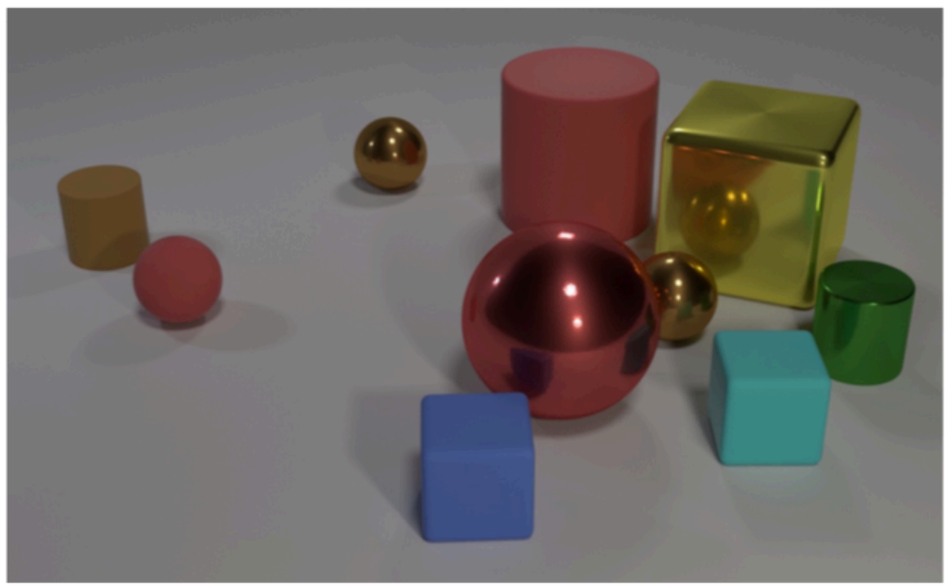
Figure 1. A sample image and questions from CLEVR. Questions test aspects of visual reasoning such as attribute identification, counting, comparison, multiple attention, and logical operations.

Taken from Johnson et al. 2016

DeepMind

UNIVERSITY OF CAMBRIDGE

CLEVR



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?
Q: How many objects are either small cylinders or metal things?

Figure 1. A sample image and questions from CLEVR. Questions test aspects of visual reasoning such as attribute identification, counting, comparison, multiple attention, and logical operations.

A functional program is used to generate the questions and answers, given a randomly generated image; see Johnson et al. 2016 for details

DeepMind

UNIVERSITY OF CAMBRIDGE

Systems Tested

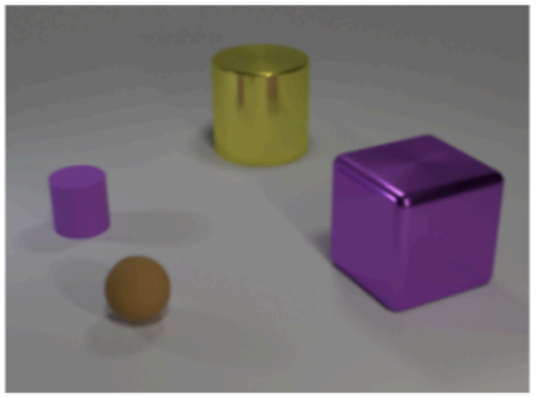
- Question LSTM without looking at the image **(46.8% acc)**
- CNN (image) + Bag-of-words (question) **(48.4%)**
- CNN + LSTM **(52.3%)**
- CNN + LSTM + sophisticated pooling **(51.4%)**
- CNN + LSTM + spatial attention **(68.5%)**

DeepMind

UNIVERSITY OF CAMBRIDGE

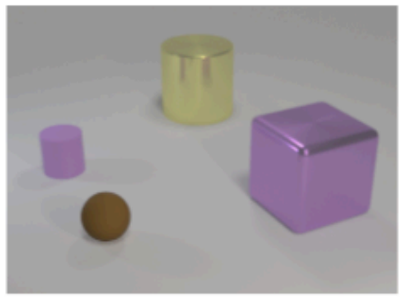
Relation Networks

Original Image:



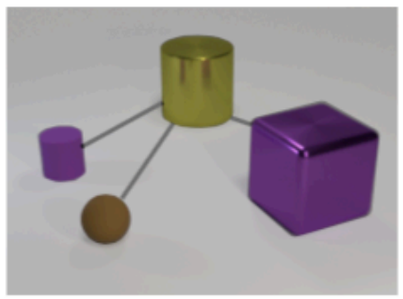
Non-relational question:

What is the size of the brown sphere?



Relational question:

Are there any rubber things that have the same size as the yellow metallic cylinder?



Taken from: Santoro, Raposo et al. 2017:
A simple neural network module for relational reasoning

DeepMind

UNIVERSITY OF CAMBRIDGE

Relation Networks

In its simplest form the RN is a composite function:

$$\mathrm{RN}(O) = f_\phi \left(\sum_{i,j} g_\theta(o_i, o_j) \right), \tag{1}$$

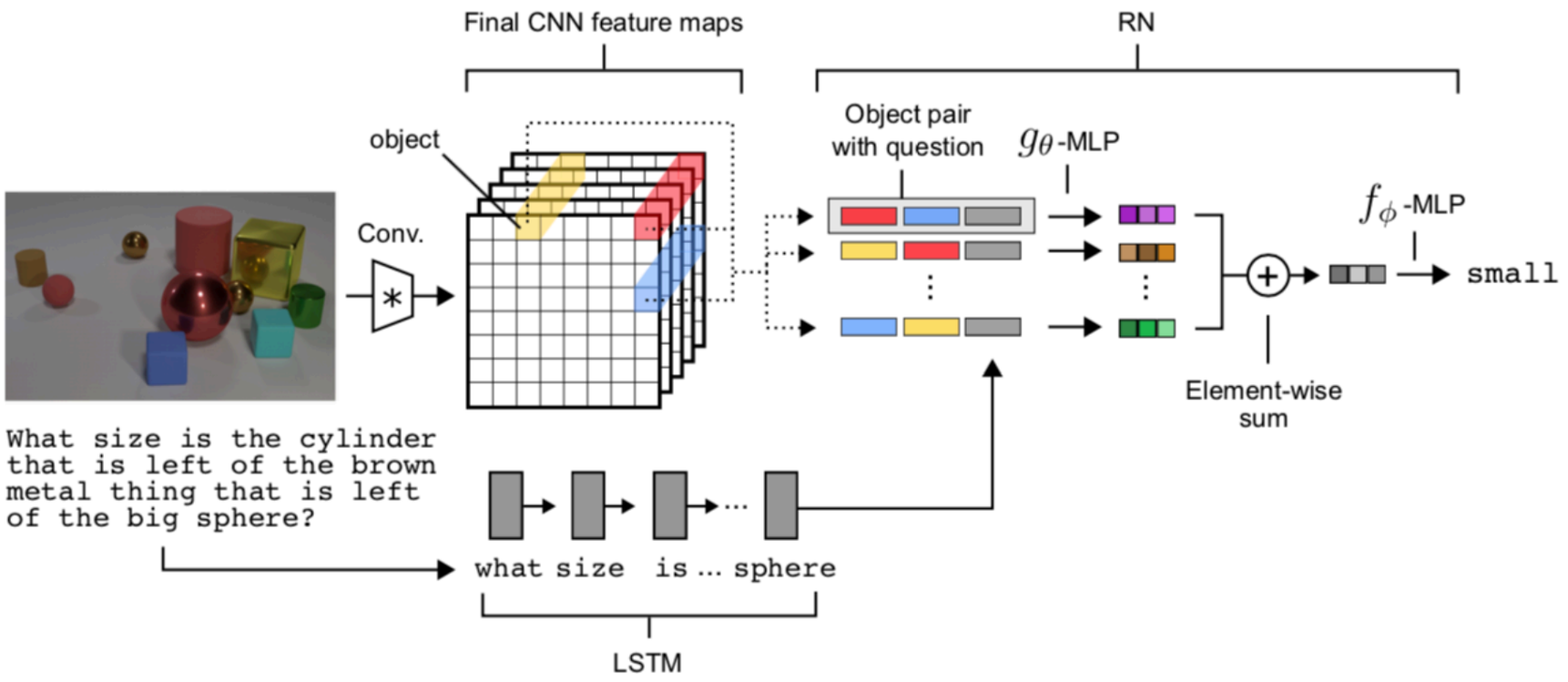
where the input is a set of "objects" $O = \{o_1, o_2, ..., o_n\}$, $o_i \in \mathbb{R}^m$ is the i^{th} object, and f_ϕ and g_θ are functions with parameters ϕ and θ, respectively. For our purposes, f_ϕ and g_θ are MLPs, and the

Taken from: Santoro, Raposo et al. 2017:
A simple neural network module for relational reasoning

DeepMind

UNIVERSITY OF CAMBRIDGE

Relation Networks



Taken from: Santoro, Raposo et al. 2017:
A simple neural network module for relational reasoning

DeepMind

UNIVERSITY OF CAMBRIDGE

Superhuman Performance!

Model	**Overall**	Count	Exist	Compare Numbers	Query Attribute	Compare Attribute
Human	92.6	86.7	96.6	86.5	95.0	96.0
Q-type baseline	41.8	34.6	50.2	51.0	36.0	51.3
LSTM	46.8	41.7	61.1	69.8	36.8	51.8
CNN+LSTM	52.3	43.7	65.2	67.1	49.3	53.0
CNN+LSTM+SA	68.5	52.2	71.1	73.5	85.3	52.3
CNN+LSTM+SA*	76.6	64.4	82.7	77.4	82.6	75.4
CNN+LSTM+RN	**95.5**	**90.1**	**97.8**	**93.6**	**97.9**	**97.1**

* Our implementation, with optimized hyperparameters and trained fully end-to-end.

Table 1: **Results on CLEVR from pixels.** Performances of our model (RN) and previously reported models [16], measured as accuracy on the test set and broken down by question category.

Taken from: Santoro, Raposo et al. 2017:
A simple neural network module for relational reasoning

DeepMind

UNIVERSITY OF CAMBRIDGE

References

- *Show and Tell: A Neural Image Caption Generator,* Vinyals et al. 2015

- *Visual Question Answering,* Agrawal et al. 2016

- *CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning,* Johnson et al. 2016

- *A Simple Neural Network Module for Relational Reasoning,* Santoro, Raposo, et al. 2017

DeepMind

UNIVERSITY OF CAMBRIDGE