UNSW
THE UNIVERSITY OF NEW SOUTH WALES

9c. Vision and Language Reasoning

Never Stand Still | Faculty of Engineering | COMP9444 Week 9c

Sonit Singh
School of Computer Science and Engineering
Faculty of Engineering
The University of New South Wales, Sydney, Australia
sonit.singh@unsw.edu.au

Goal

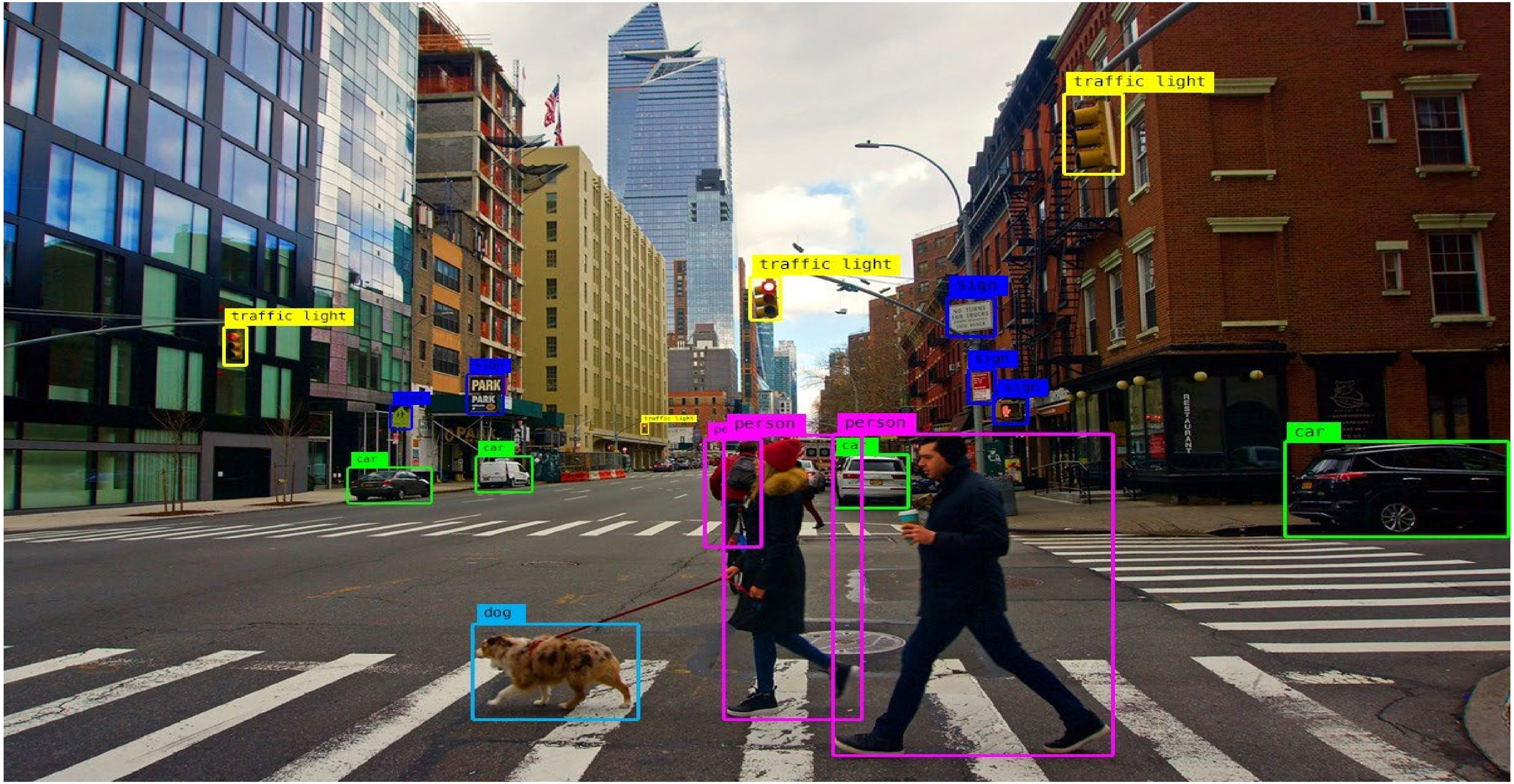
- To motivate the need for “Computer Vision + Natural Language Processing”
- After the talk, everyone can confidently say: “yeah, I know various tasks at the intersection of computer vision and natural language processing”
- Focus on high-level overview, not technical details
- Focus on static images, not videos (although they are easy to translate to videos)
- Focus on selective set of papers for various tasks, not a comprehensive literature review

Agenda

- Computer Vision
- Natural Language Processing
- Computer Vision + Natural Language Processing
- Building Blocks
  - Convolutional Neural Networks (CNNs)
  - Recurrent Neural Networks (RNNs)
  - Attention Mechanism
- Encoder-Decoder Framework
- Image Captioning
- Visual Question Answering (VQA)
- Visual Dialog (VisDial)
- Vision-Language Navigation (VLN)
- Visual Grounding
- Summary

Computer Vision

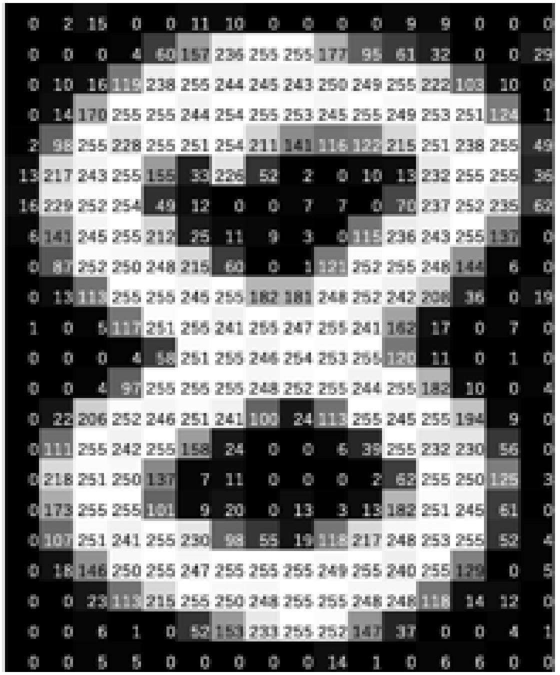
- Enabling machines to process, represent, understand, and generate visual data



Source: Computer Vision in Practice. https://levelup.gitconnected.com/deployment-computer-vision-in-practice-99b7bbb6aa7c

How computers see images?

- A matrix of numbers

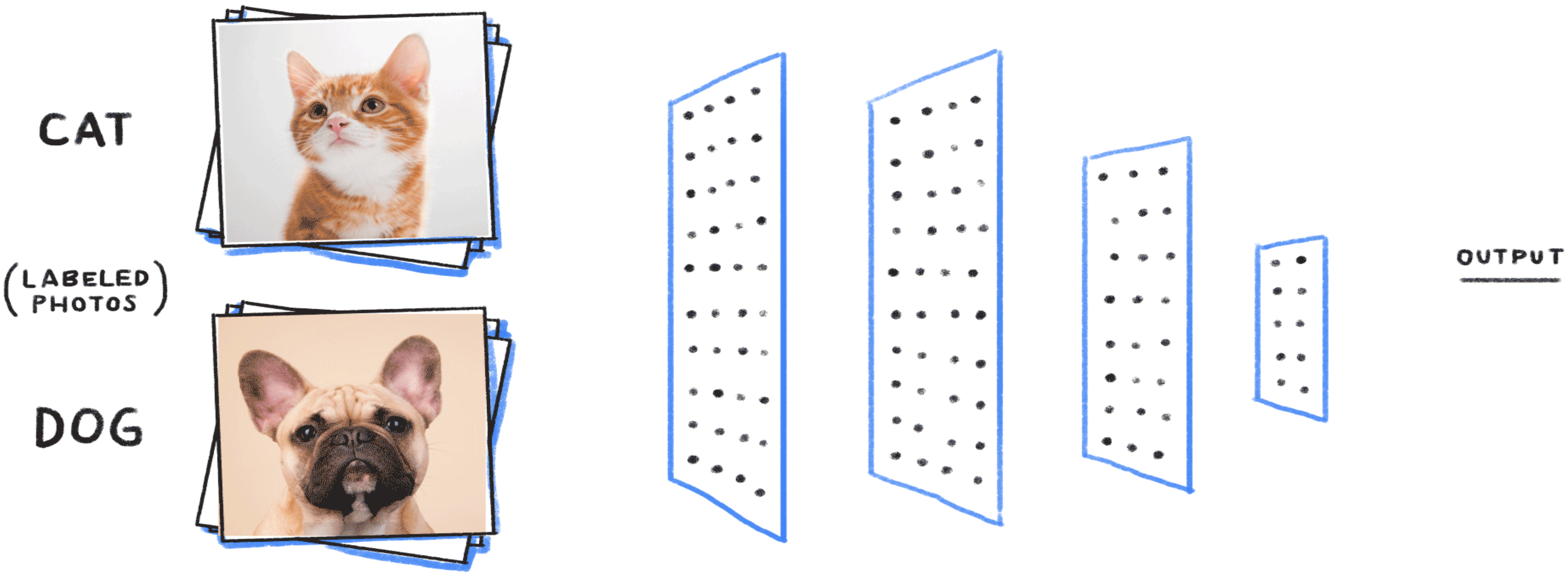


What Computer Sees

0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0

CV Applications

- Image Classification

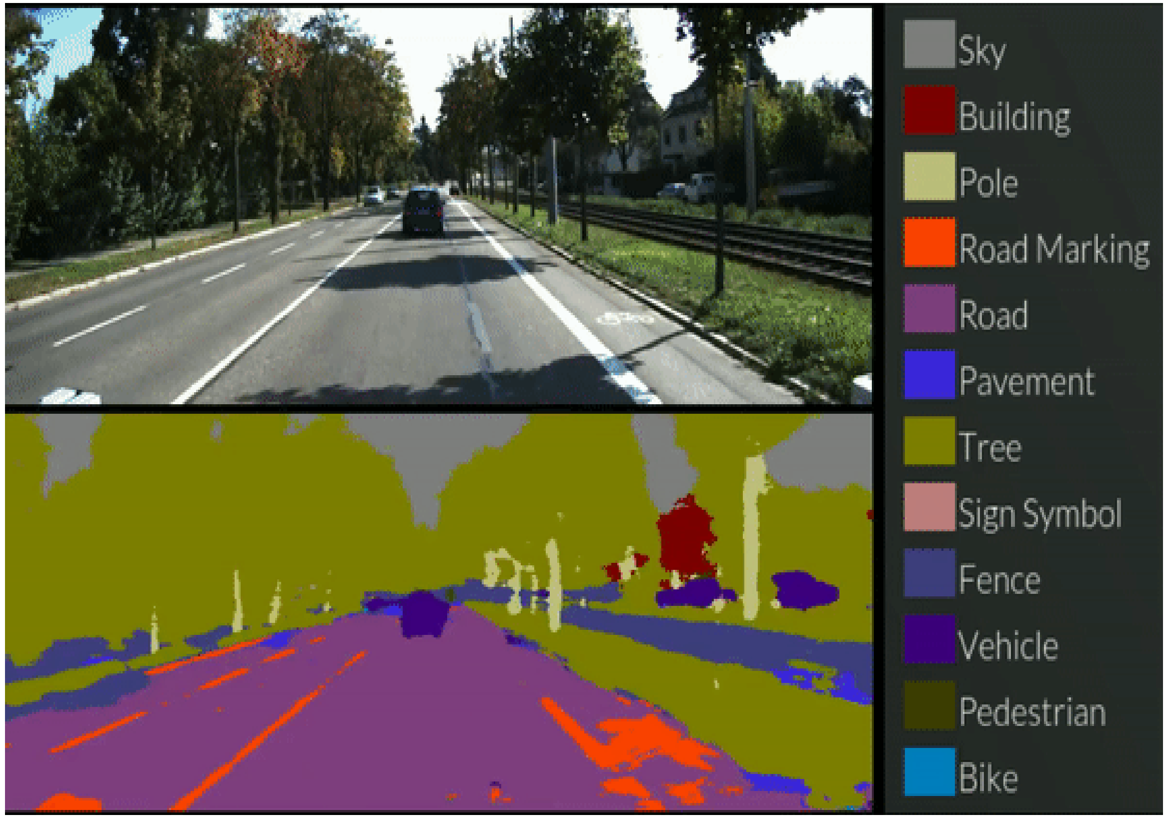


CV Applications

- Object Detection

CV Applications

➢ Segmentation



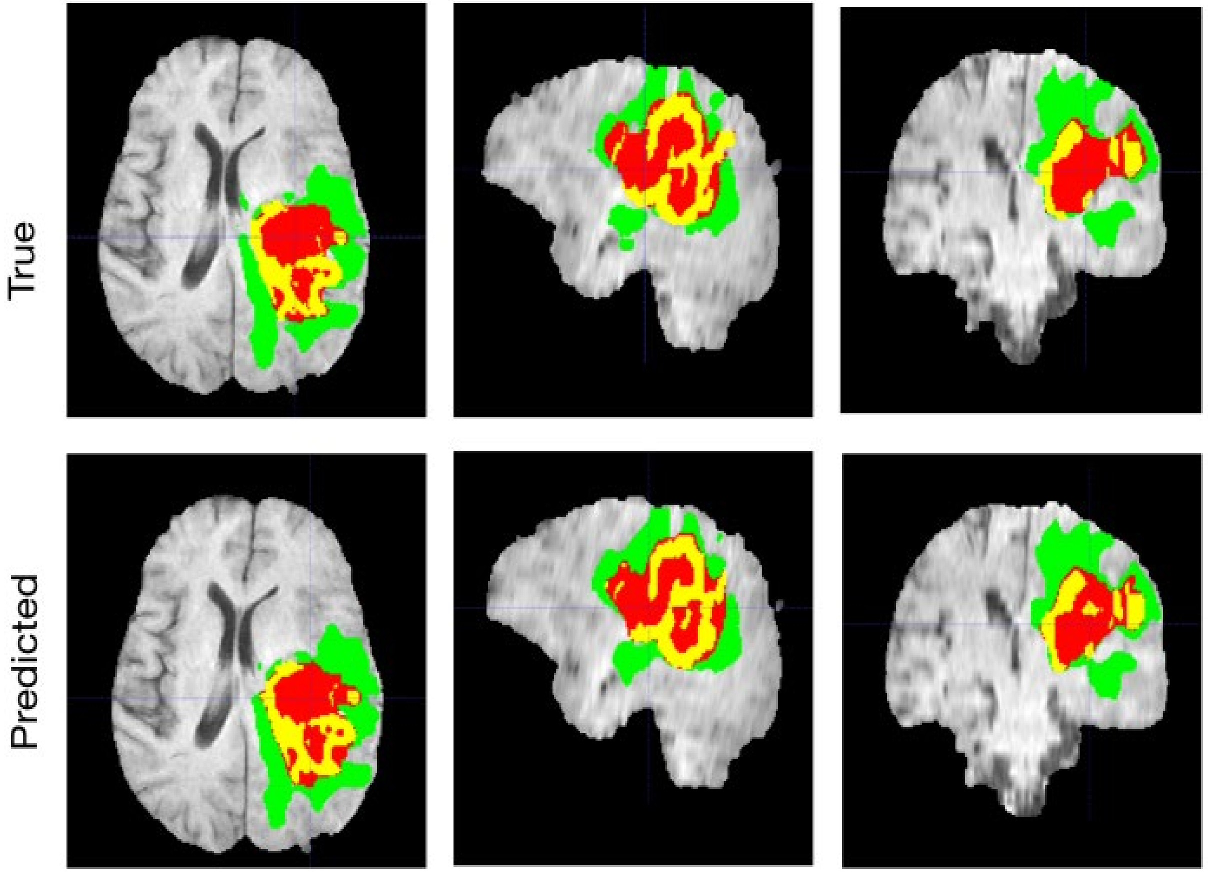


Image Credit: Analytics Vidhya: Introduction to Semantic Image Segmentation. https://medium.com/analytics-vidhya/introduction-to-semantic-image-segmentation-856cda5e5de8
Nvidia AI: Automatically segmenting brain tumours with AI. https://developer.nvidia.com/blog/automatically-segmenting-brain-tumors-with-ai/

Natural Language Processing

➢ Enabling machines to process, represent, understand, and generate languages

In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG , IBM ORG , and Microsoft ORG .

NLP Applications

- Text Understanding

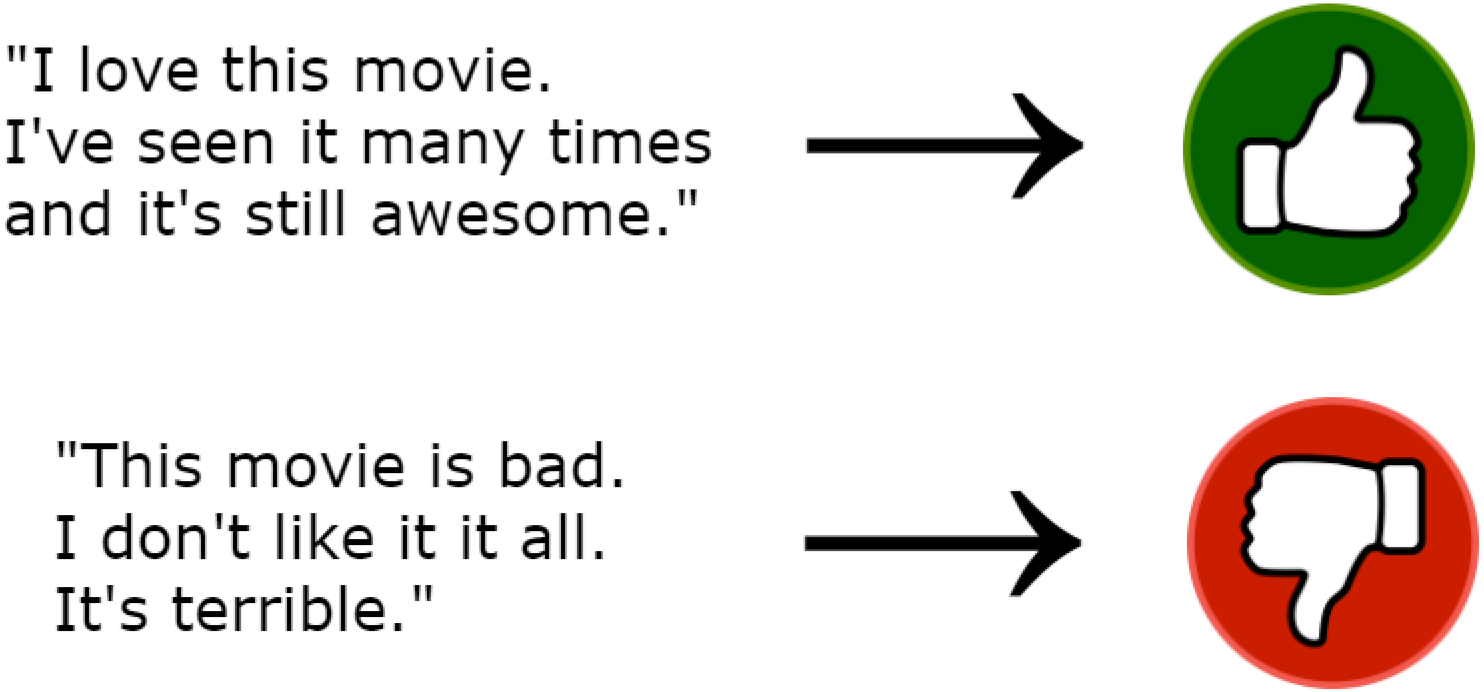
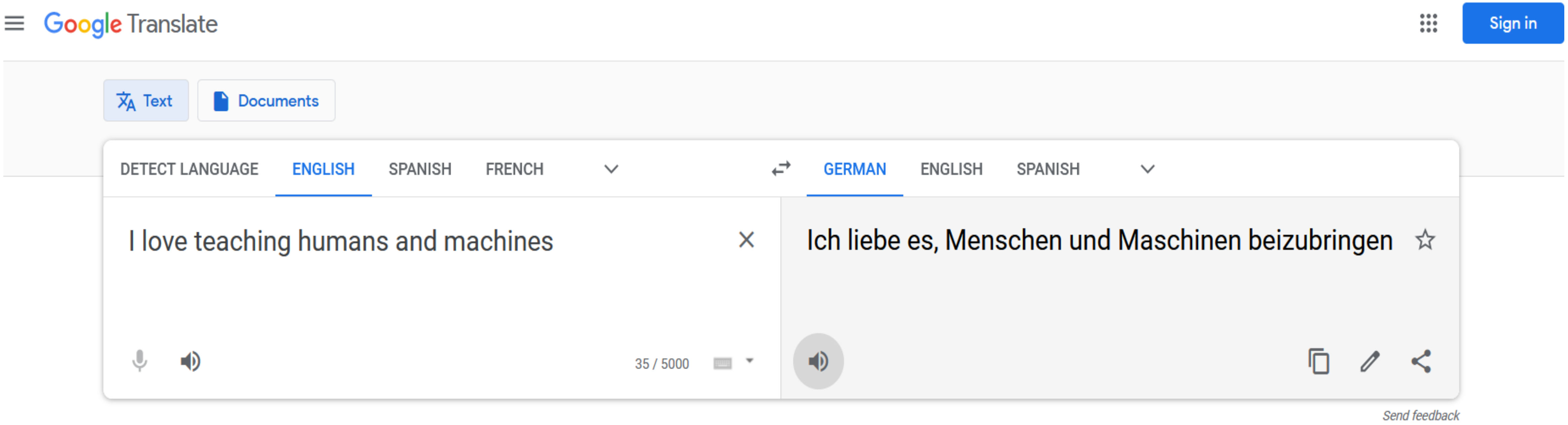


Image Credit: Sentiment Analysis. https://colab.research.google.com/github/shangeth/Google-ML-Academy/blob/master/2-Deep-Neural-Networks/2_9_ANN_Natural_Language_Processing.ipynb

NLP Applications

➢ Machine Translation



NLP Applications

➢ Question Answering/Comprehension

Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

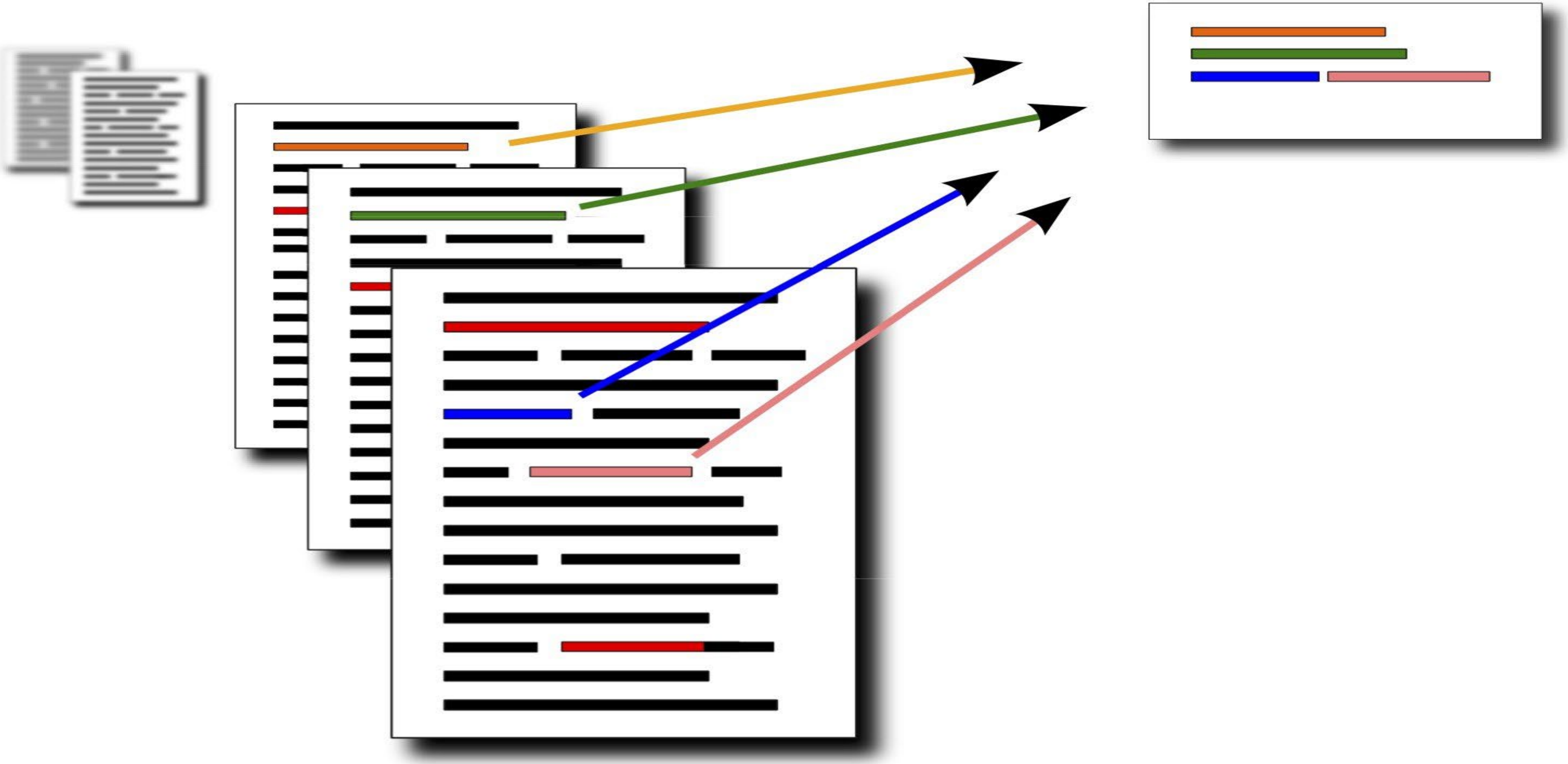
Answer Candidate

gravity

Source: The Stanford Question Answering Dataset (SQUAD). https://rajpurkar.github.io/mlx/qa-and-squad/

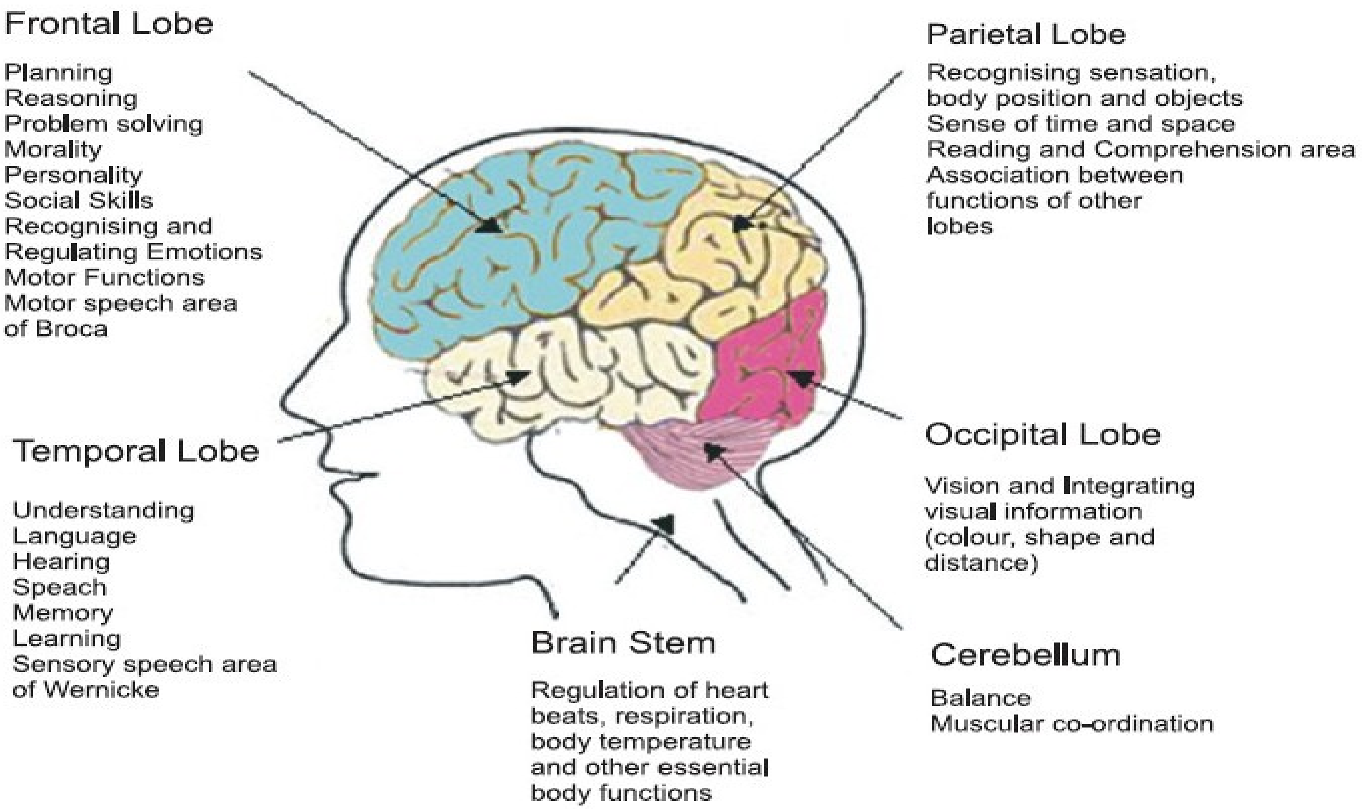
NLP Applications

- Text Summarization



Vision + Language

- Science Perspective
  - Vision is how we observe and understand the world
  - Language is how we communicate
- A move towards Artificial General Intelligence (AGI)



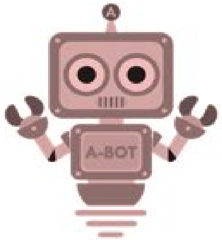
Source: Barrow Neurological Institute

Applications at the intersection of Vision and Language

- To aid “visually impaired” people

Applications at the intersection of Vision and Language

- To aid "situationally impaired" analysts



Did anyone enter this room last week?

Yes, 127 instances logged on camera

Show me images of anyone carrying a black bag.

...

Applications at the intersection of Vision and Language

- Personal Assistants

And what's next?

Add the chopped zucchini to the pan…

Hey, can you order more zucchini?

Done.

Applications at the intersection of Vision and Language

➢ Natural Language Instructions for Robots

Is there smoke in any room around you?

Yes, in one room

Go there and look for people

...

Slide Credit: Peter Anderson. Connecting Language and Vision to Actions. https://lvatutorial.github.io/

Applications at the intersection of Vision and Language

- Generating and summarizing radiology reports from medical images

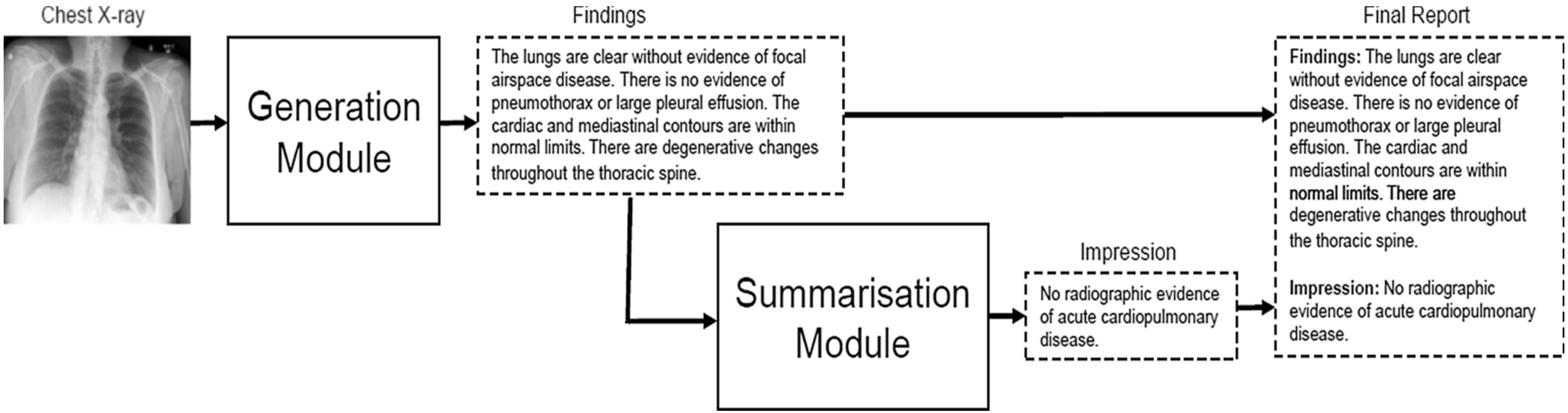
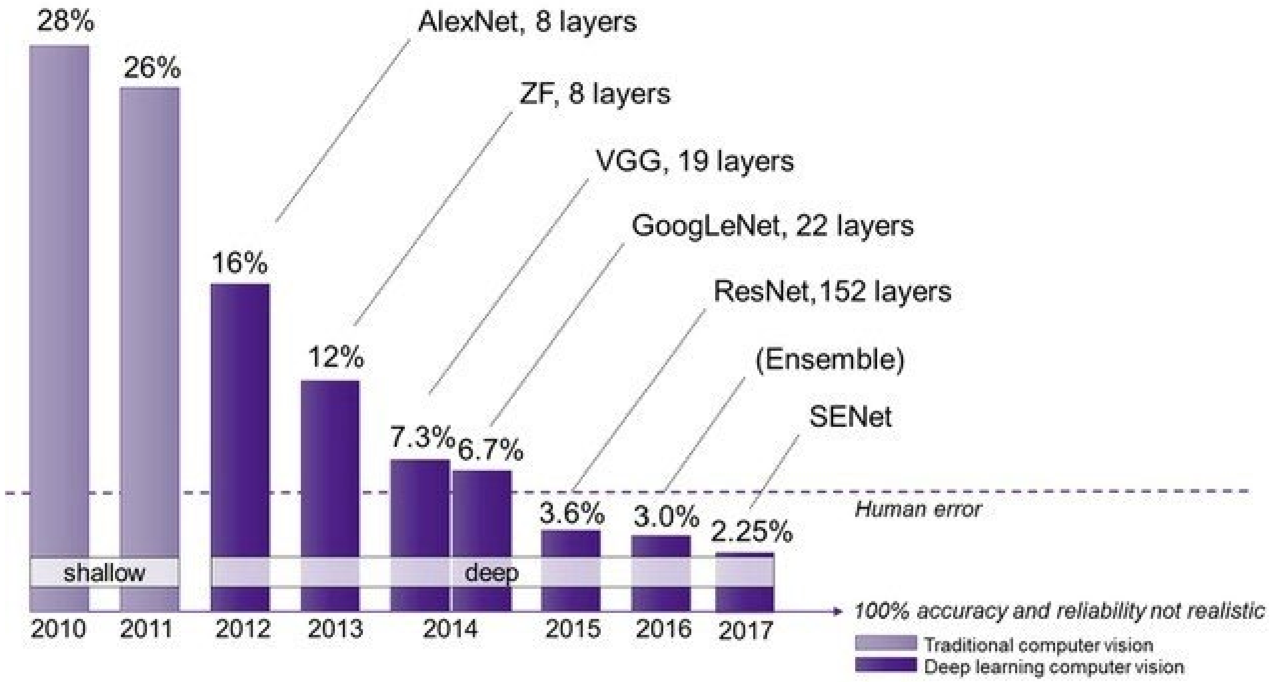


Image Credit: Singh et al., (2021). Show, tell and summarise: learning to generate and summarise radiology findings from medical images

Building Blocks: Convolutional Neural Networks (CNNs)

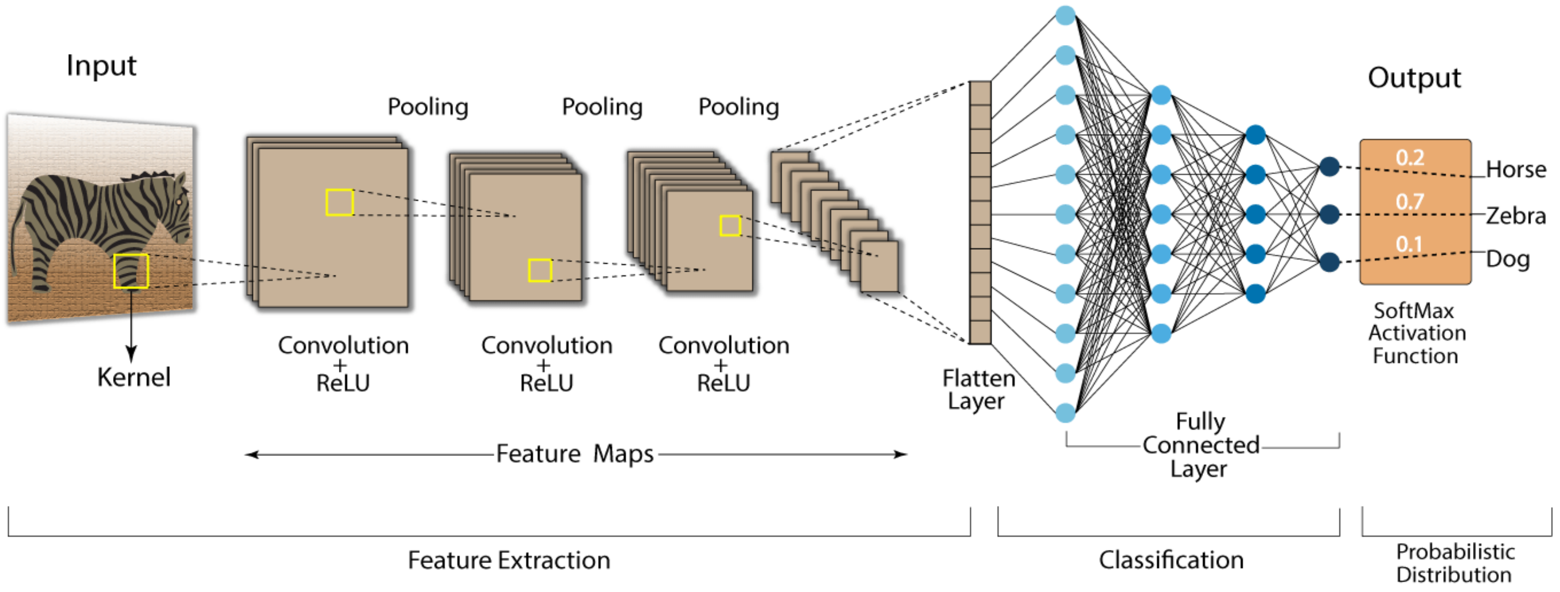
- A class of deep neural networks suitable for processing 2D/3D data. For e.g., Images and Videos
- CNNs can capture high-level representation of images/videos which can be used for end-tasks such as classification, object detection, segmentation, etc.
- A range of CNNs improving over the years



Source: Convolutional Neural Networks. https://medium.com/@rajat.k.91/convolutional-neural-networks-why-what-and-how-f8f6dbebb2f9

CNN Architecture

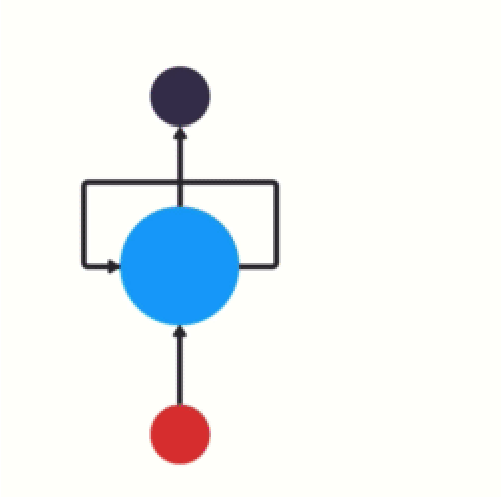
- A typical CNN architecture consists of the following layers:
  - Convolution layer
  - ReLU layer (non-linearity)
  - Pooling layer
  - Flattening
  - Fully-connected layer
  - Output layer

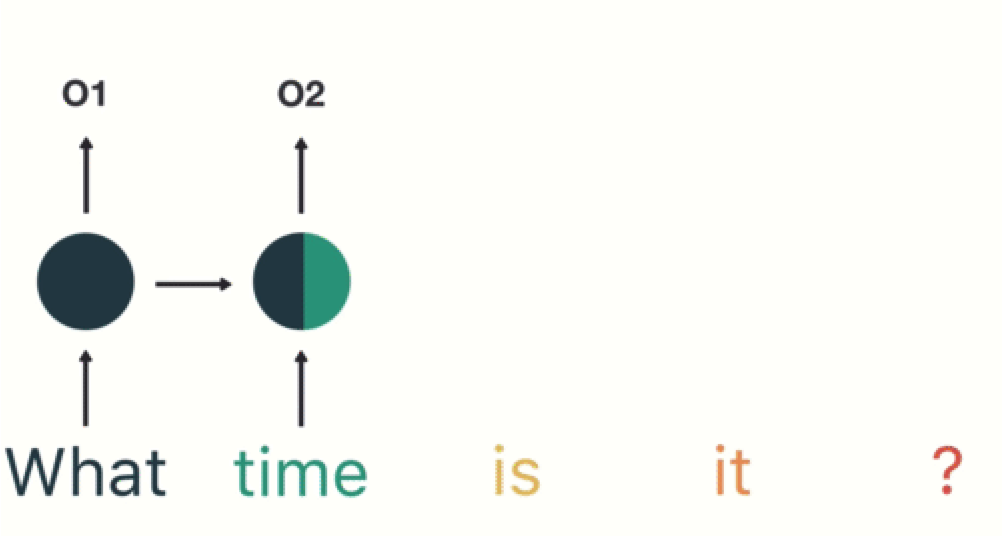


- There can be multiple steps of convolution followed by pooling, before reaching the fully connected layers.

Building Blocks: Recurrent Neural Networks (RNNs)

- A class of neural networks suitable for processing temporal or sequential data
- The basic unit of RNN is called "cell", and each cell consists of layers and a series of cells that enables the sequential processing of recurrent neural network models
- RNNs have a looping mechanism that acts as a highway to allow information to flow from one step to the next. This information is the hidden state, which is a representation of previous inputs.





Long-range Dependency problem

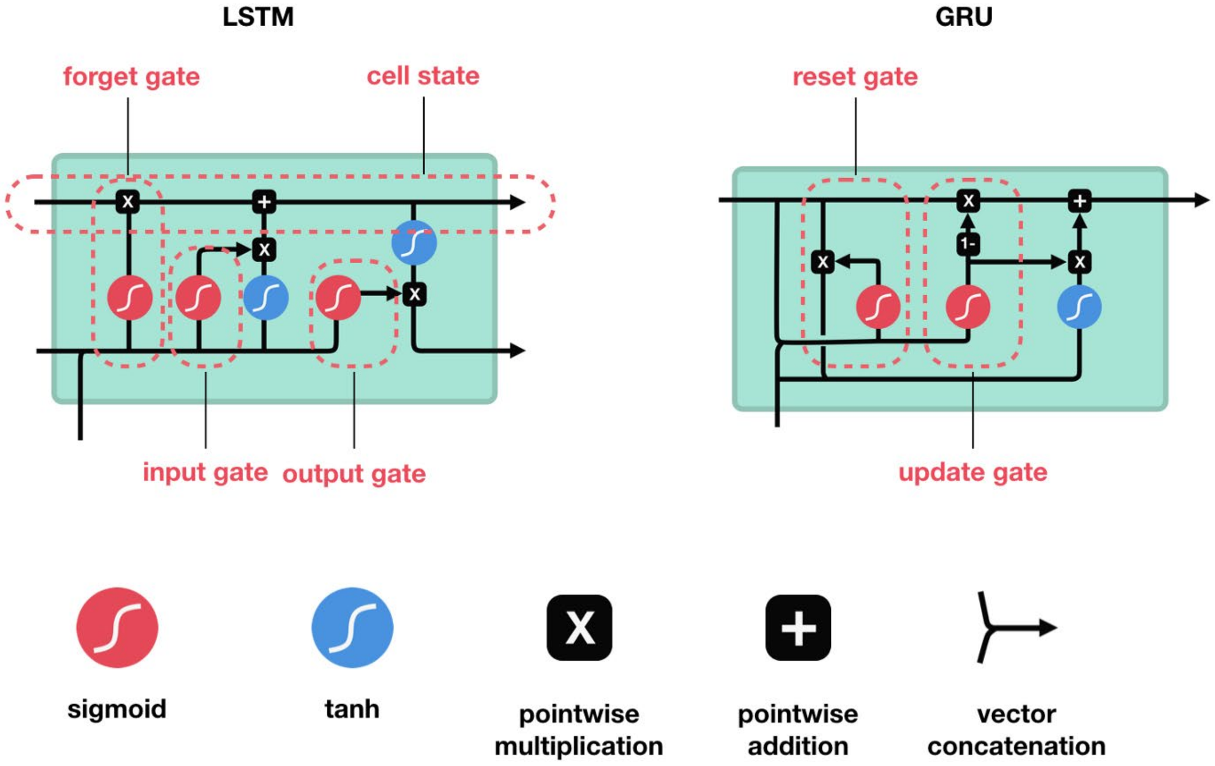
- Vanilla RNNs suffers from vanishing gradient problem
  - As the RNNs processes more steps, it has troubles retaining information from previous steps.
  - Due to back-propagation, the earlier layers fail to do any learning as the internal weights are barely being adjusted due to extremely small gradients.
  - Does not learn the long-range dependencies across time steps

"Once upon a time, there was a king who ruled a great and glorious nation. Favourite amongst his subjects was the court painter of whom he was very proud. Everybody agreed this wizzened old man painted the greatest pictures in the whole kingdom and the king would spend hours each day gazing at them in wonder. However, one day a dirty and disheveled stranger presented himself at the court claiming that in fact he was the greatest painter in the land. The indignant king decreed a competition would be held between the two artists, confident it would teach the vagabond an embarrassing lesson. Within a month they were both to produce a masterpiece that would out do the other. After thirty days of working feverishly day and night, both artists were ready. They placed their paintings, each hidden by a cloth, on easels in the great hall of the castle. As a large crowd gathered, the king ordered the cloth be pulled…"

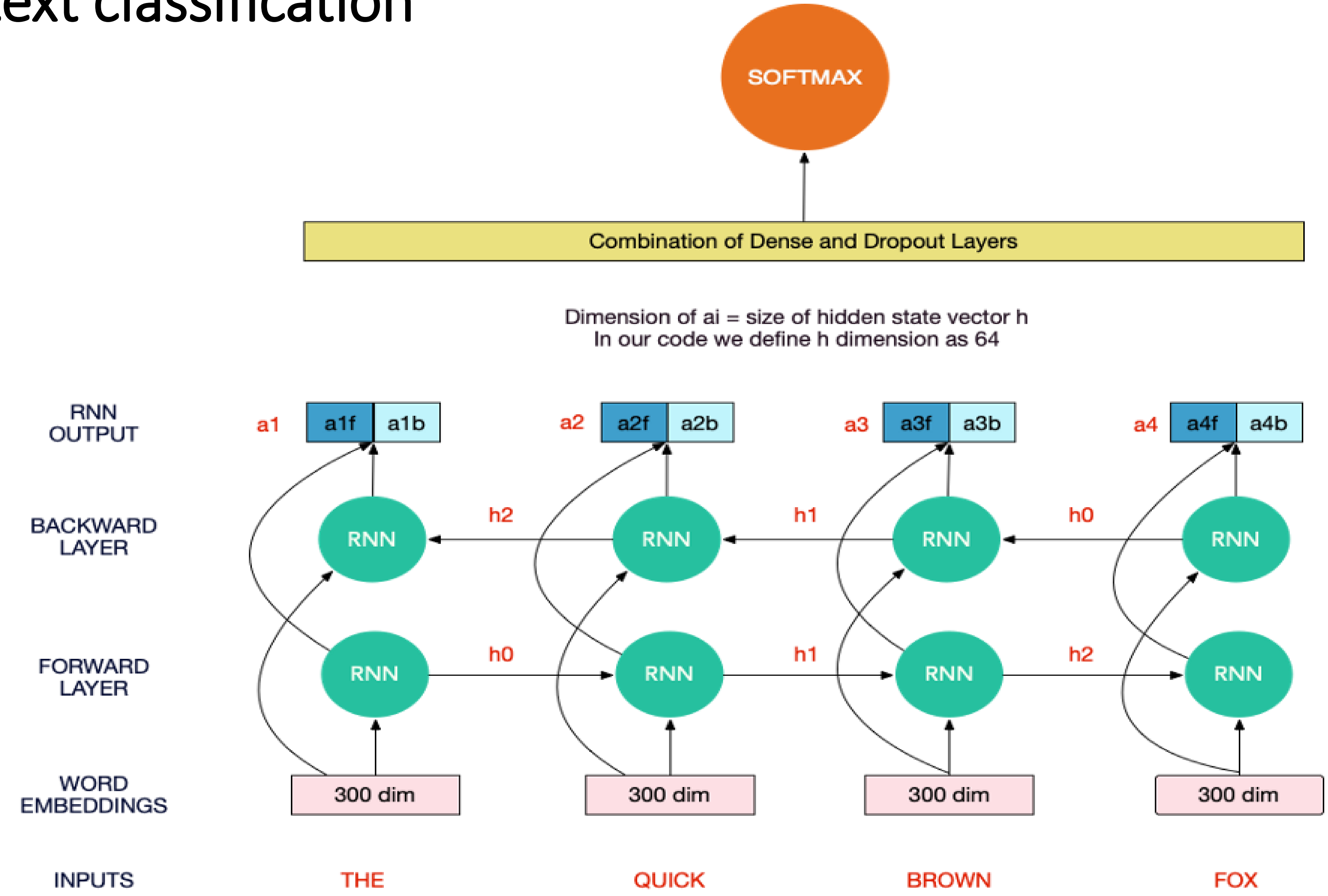
Credit: https://www.goodreads.com/quotes/51722-once-upon-a-time-there-was-a-king-who-ruled

LSTMs/GRUs

- LSTMs and GRUs are two special RNNs, capable of learning long-term dependencies using mechanisms called **gates**.
- These gates are different tensor operations that can learn what information to add or remove to the hidden state.



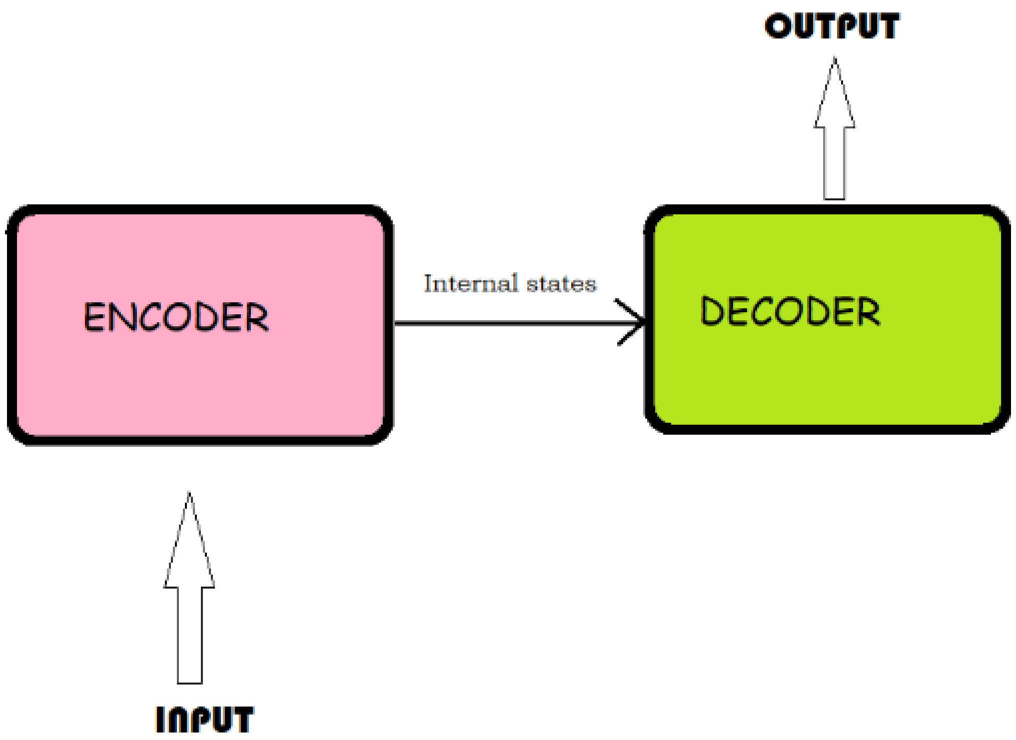
End-to-End text classification



Credit: Rahul Agarwal. Towards Data Science. https://towardsdatascience.com/using-deep-learning-for-end-to-end-multiclass-text-classification-39b46aecac81

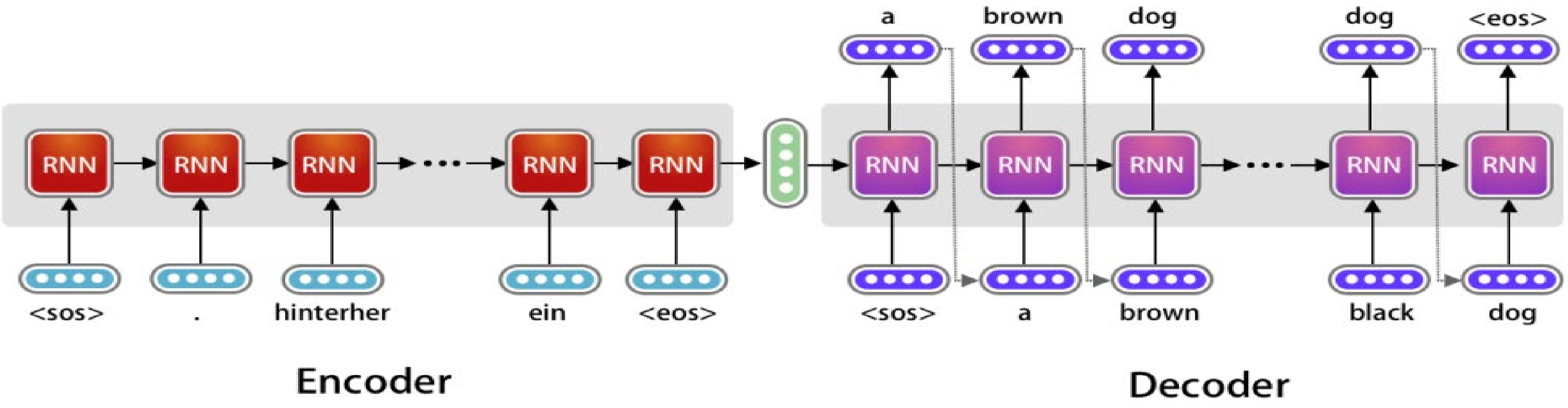
Seq2Seq model

- Model has two parts: Encoder and Decoder (Encoder-Decoder Framework)

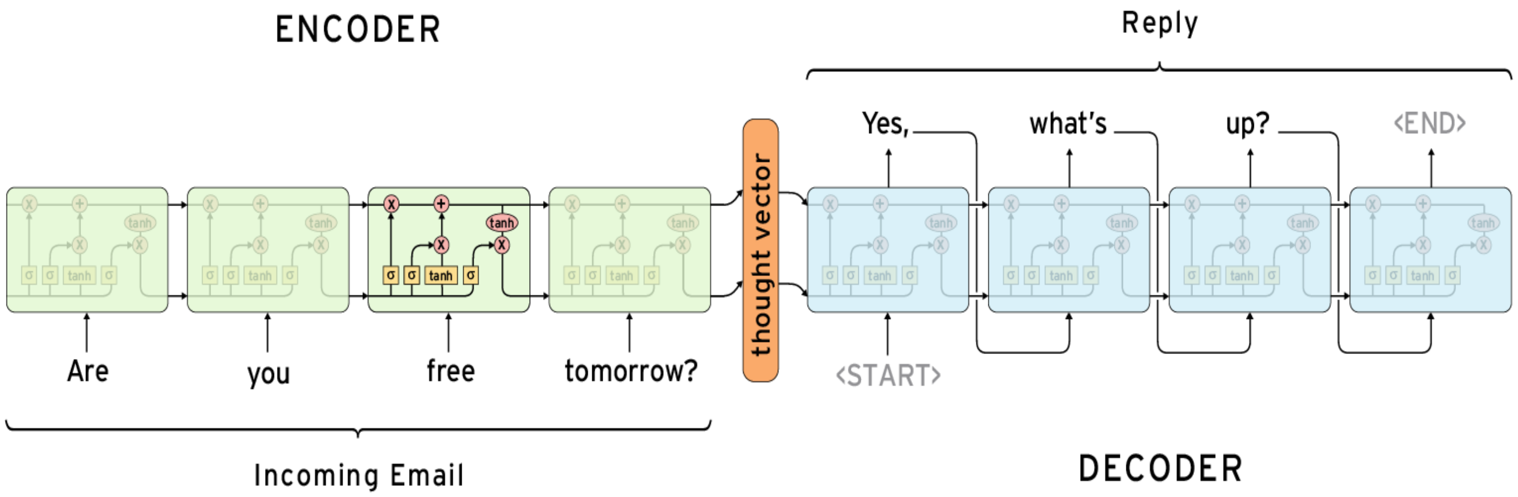


Seq2Seq used for various NLP applications

- Machine Translation

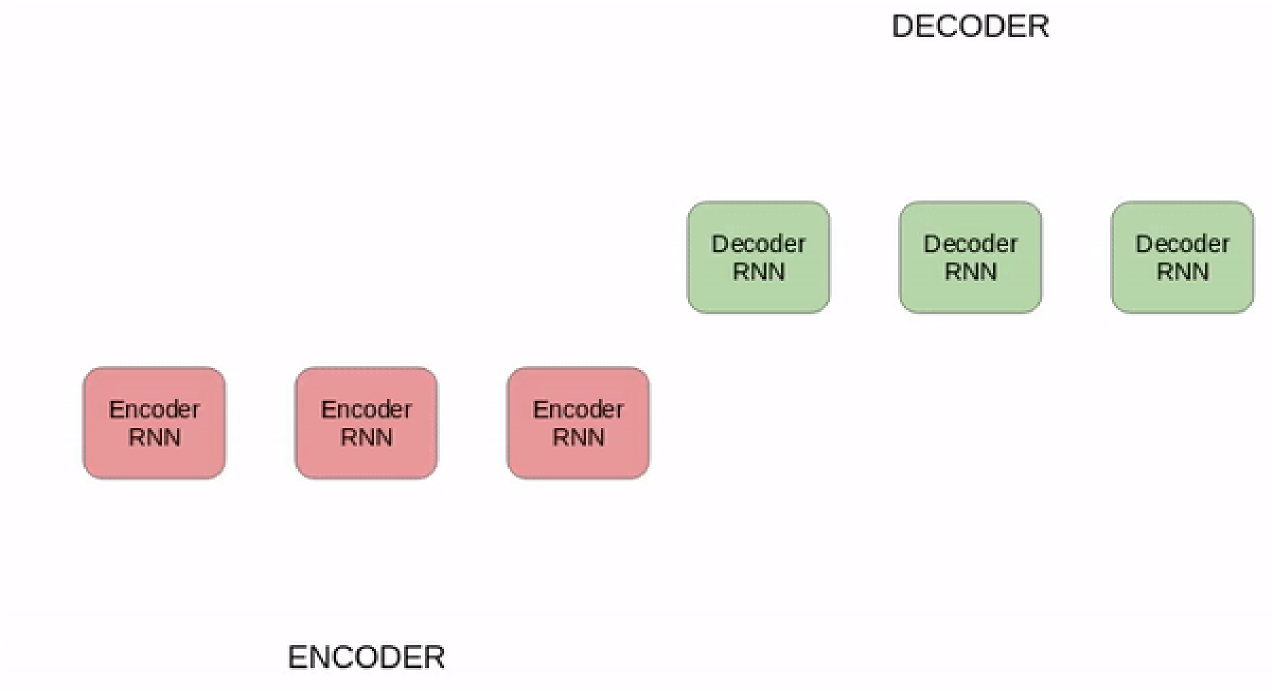


- Automatic Email Reply



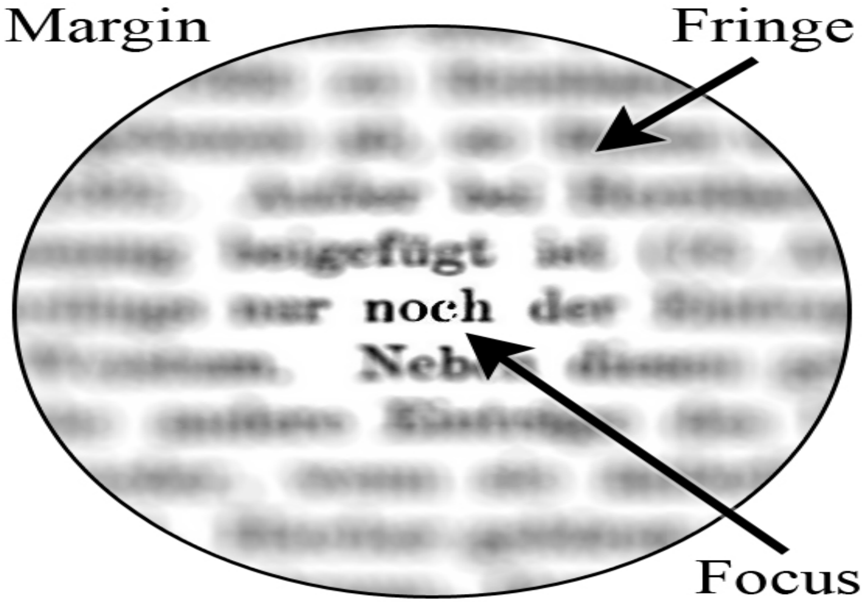
Issues with RNNs for seq2seq tasks

- Despite being very successful for various NLP tasks, there were challenges:
  - dealing with long-range dependencies
  - the sequential nature of the architecture prevents parallelization
- Attention mechanism helped to overcome first issue to certain extent



Attention Mechanism

- A set of mechanisms that limit some processing to a subset of incoming stimuli (reducing computational demands)
- Attention in neural networks
  - A mechanism that learns to focus on a subset of the input that is relevant to the task.



task context representation

attended feature $\longrightarrow$ $\hat{\boldsymbol{v}} = f(\mathbf{h}, V)$

learned attention function (neural net)

set of attention candidates

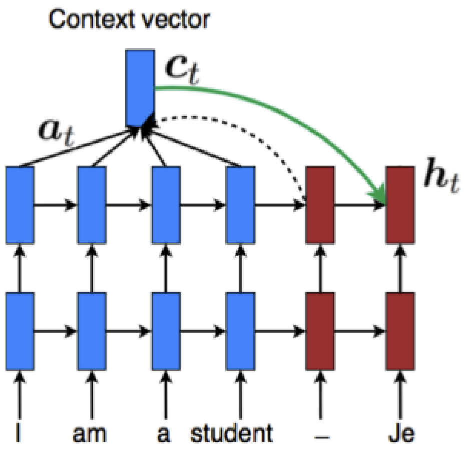
Neural Machine Translation by jointly Learning to Align and Translate

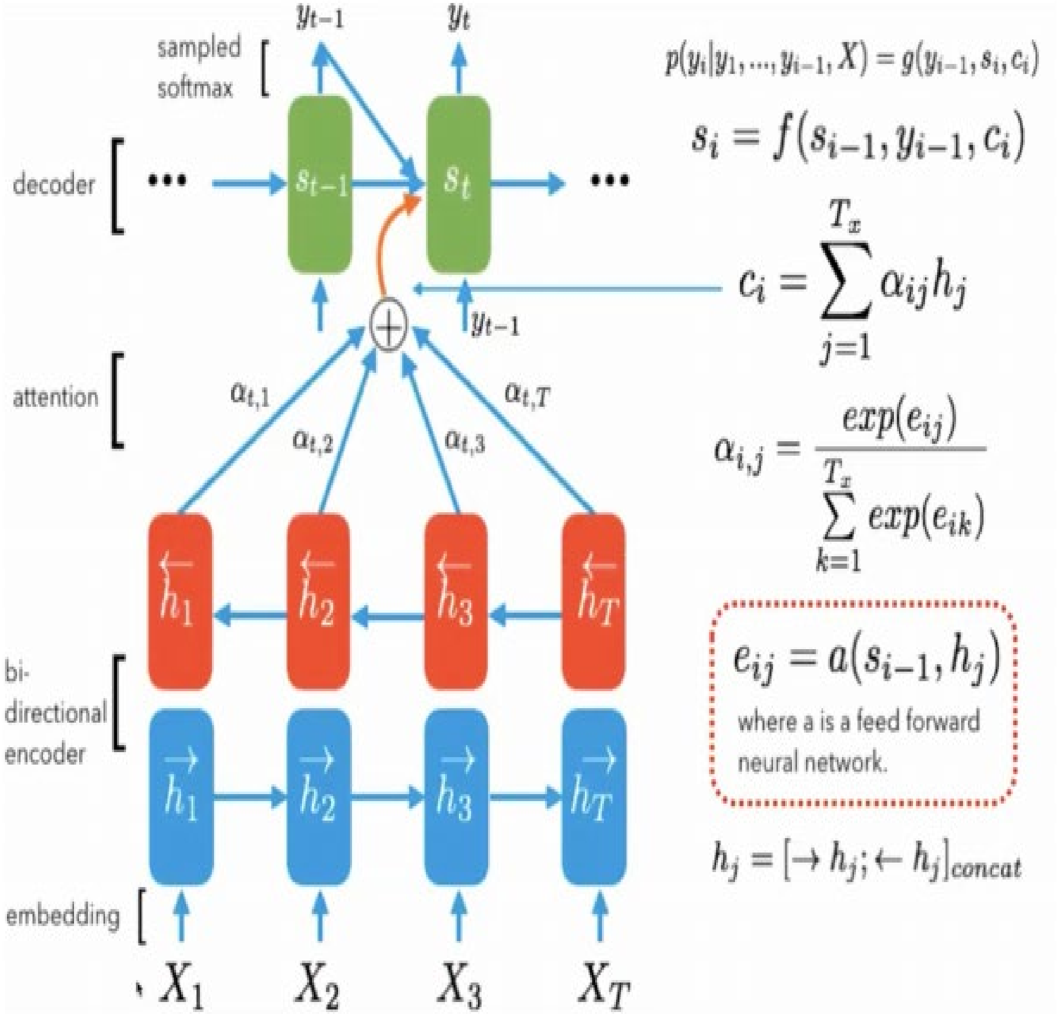
- Issue: The encoder-decoder framework compresses all the necessary information of a source sentence into a fixed-length vector.
- Solution: Enable the network to pay attention to specific areas of the input by adding new (weighted) connections

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$

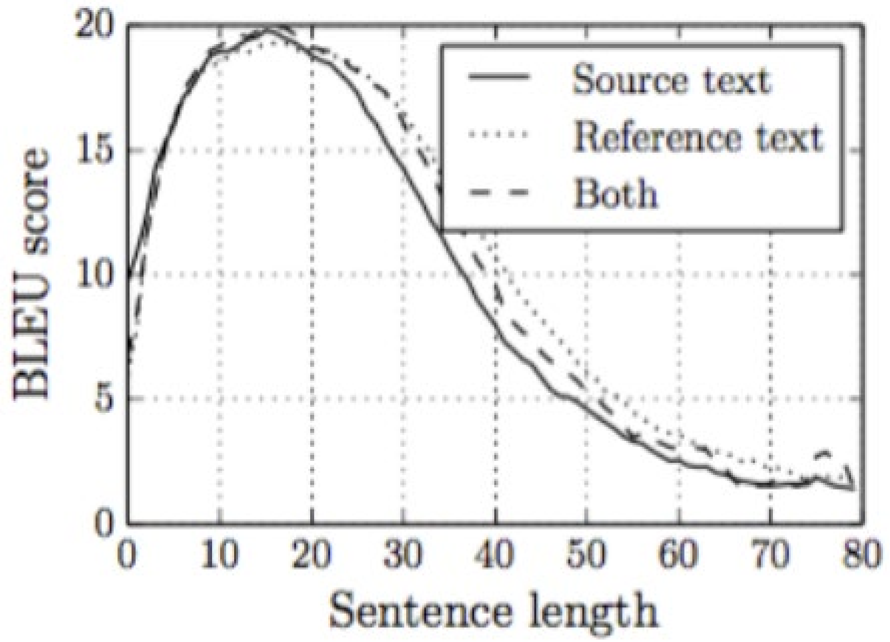
$$e_{ij} = a(s_{i-1}, h_j)$$





NMT by jointly Learning to Align and Translate

Before Attention: Long sentences are very hard as they are compressed to a fixed length vector



After Attention: The attention mechanism helps to overcome the issue

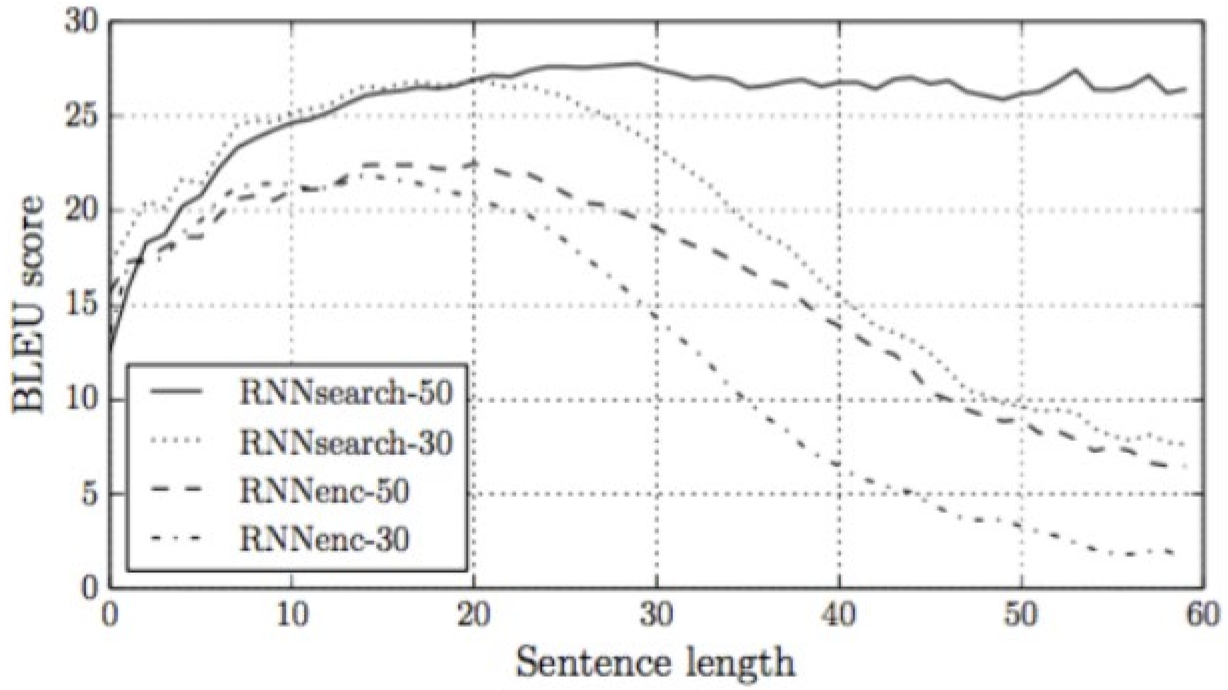


Image Captioning

- Developing models to generate textual description of an image

"man in black shirt is playing guitar."

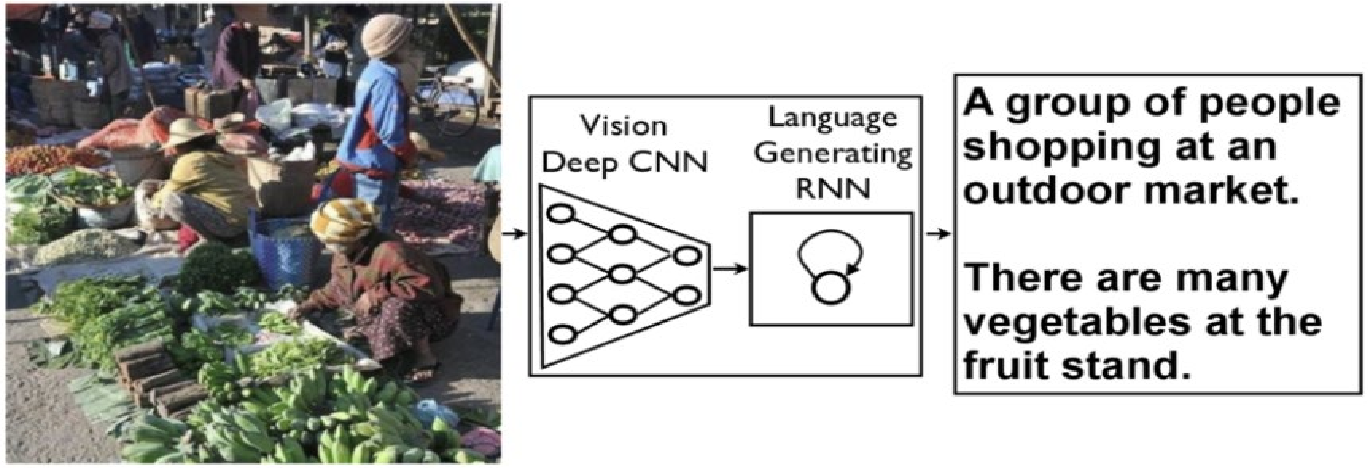
"construction worker in orange safety vest is working on road."

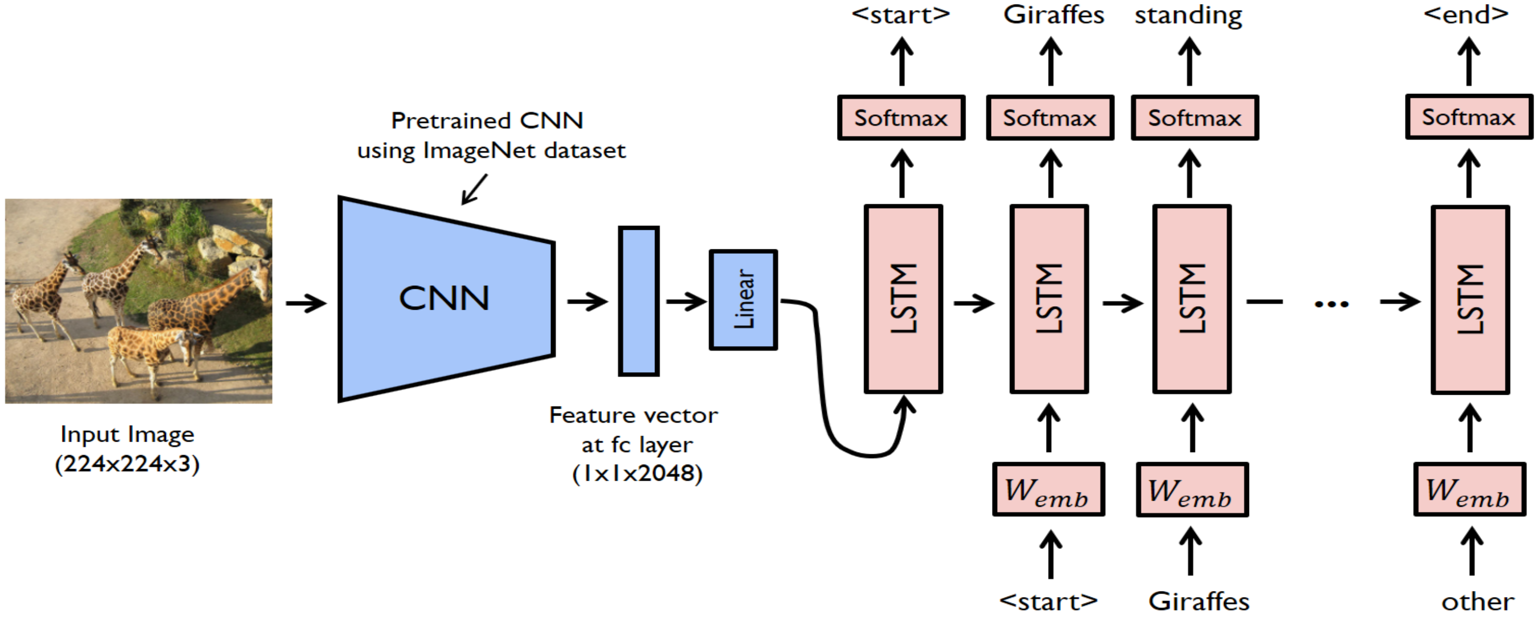
"two young girls are playing with lego toy."

Image Credit: Machine Translation using Recurrent Neural Network and PyTorch. http://www.adeveloperdiary.com/data-science/deep-learning/nlp/machine-translation-recurrent-neural-network-pytorch/

Show and Tell (Vinyals et al., 2015)

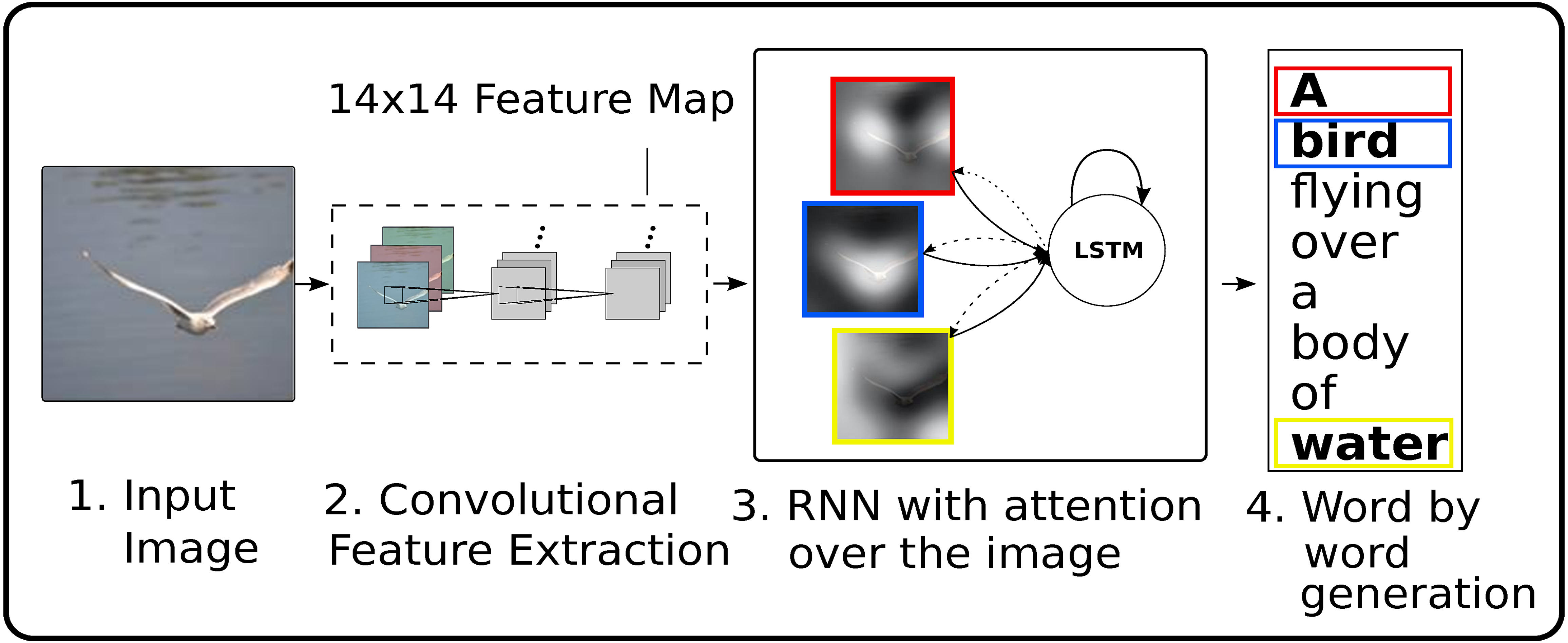
➢





Show, Attend and Tell (Xu et al., 2015)

- “Rather than compressing an entire image into static representation, attention allows for **salient features** to dynamically come to forefront as needed”



Show, Attend and Tell (Xu et al., 2015)

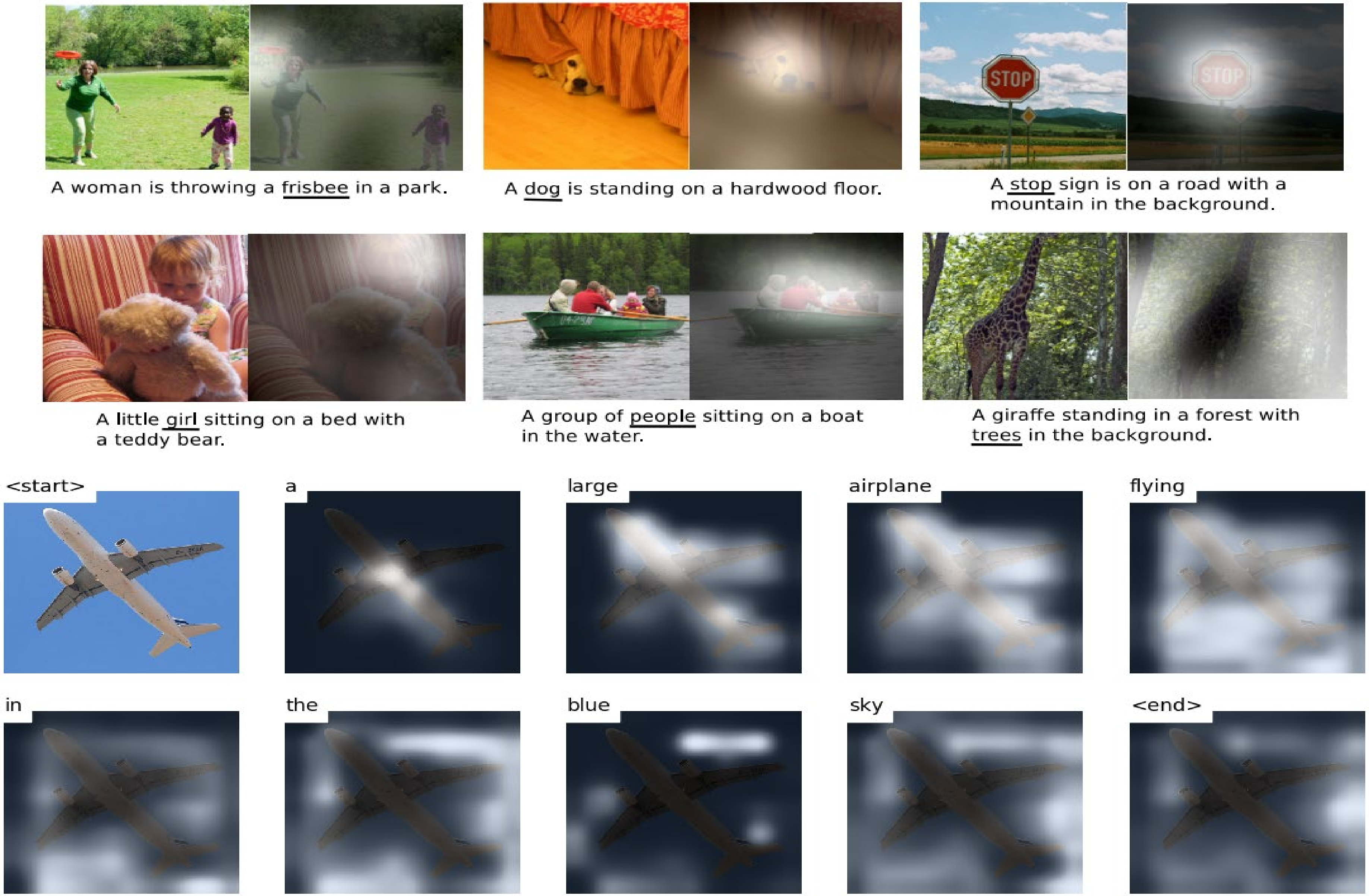


Image Captioning with Semantic Attention (You et al., 2016)

- Learns to selectively attend to semantic concept proposals and fuse them into hidden states and outputs of RNNs.

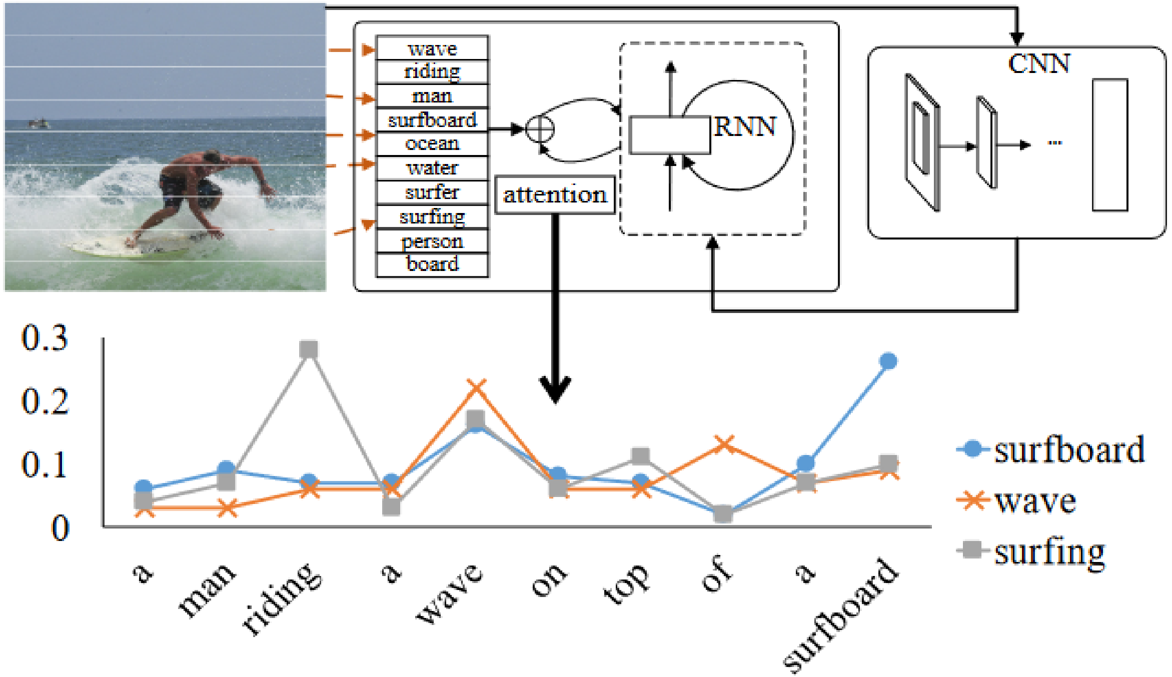


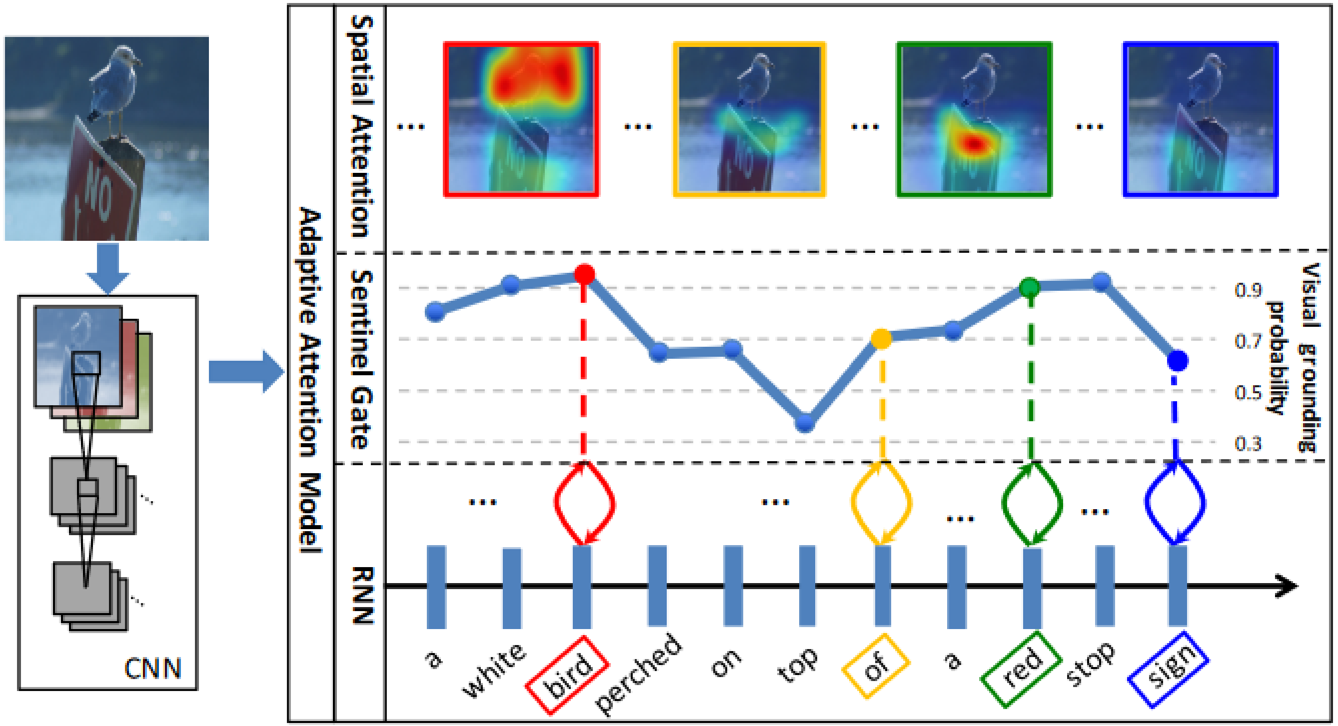
Image Captioning with Semantic Attention (You et al., 2016)

➢ Learns to selectively attend to semantic concept proposals and fuse them into hidden states and outputs of RNNs.

Source: You et al., (2016). Image Captioning with Semantic Attention

Knowing When to Look (Lu et al., 2017)

- Words such as “a”, “of”, “it” may be seen as not worth attending
- Words such as “woman”, “dog”, “traffic light” need attending to the image
- Automatically determines when to look (sentinel gate) and where to look (spatial attention) for word generation.



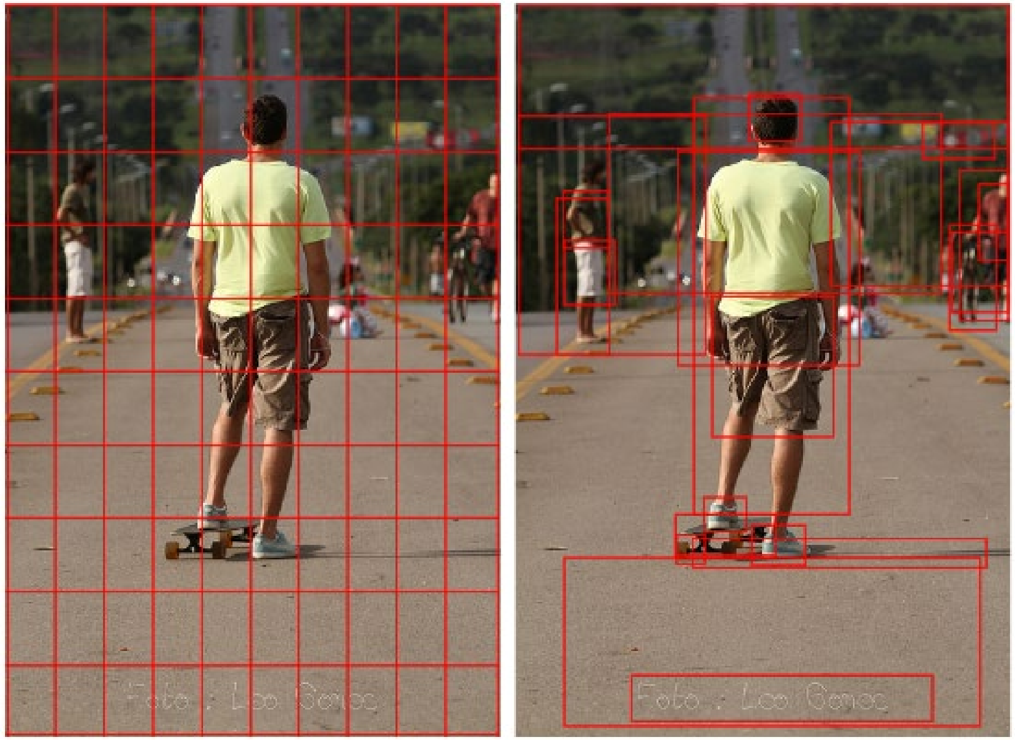
Knowing When to Look (Lu et al., 2017)

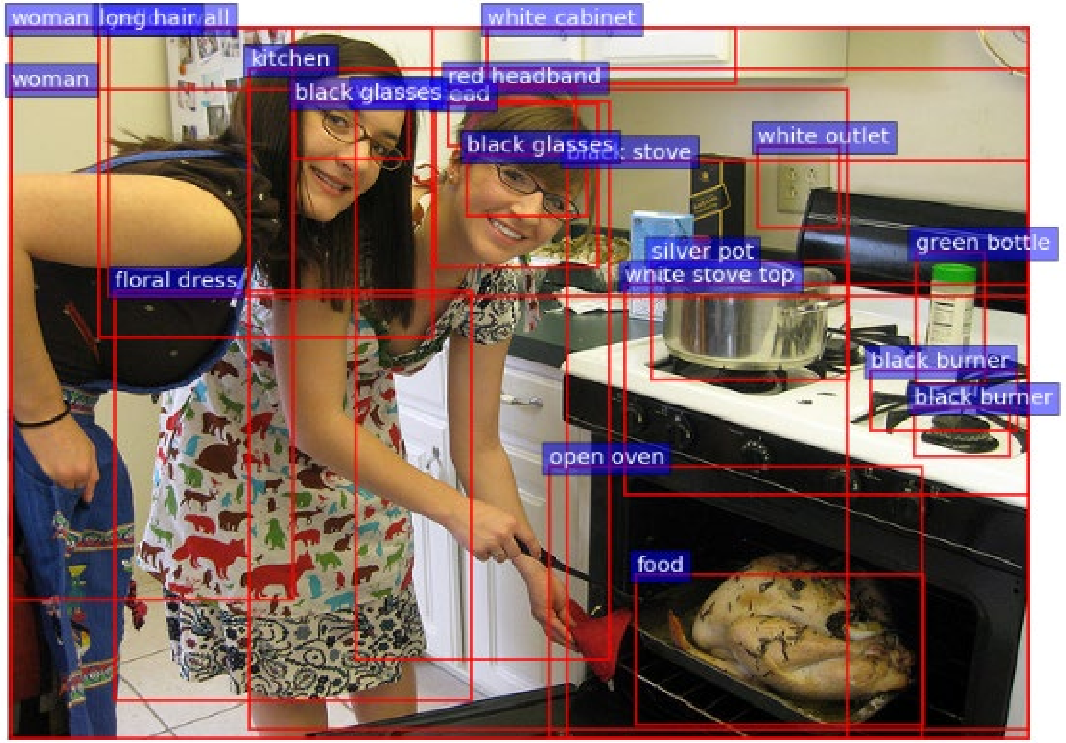
- Visualization of generated captions and image attention maps on COCO dataset

Source: Lu et al., (2017). Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning

Bottom-up and Top-down attention (Anderson et al., 2018)

- Previous attention models operate on CNN features corresponding to a uniform grid of equally-sized image regions
- Bottom-up and Top-down enables attention to be calculated at the level of objects (or salient image regions)
- Features vectors extracted from Faster R-CNN are used





Bottom-up and Top-down attention (Anderson et al., 2018)

➢

Ours: Resnet – A man sitting on a *toilet* in a bathroom.

Ours: Up-Down – A man sitting on a *couch* in a bathroom.

And many more work …

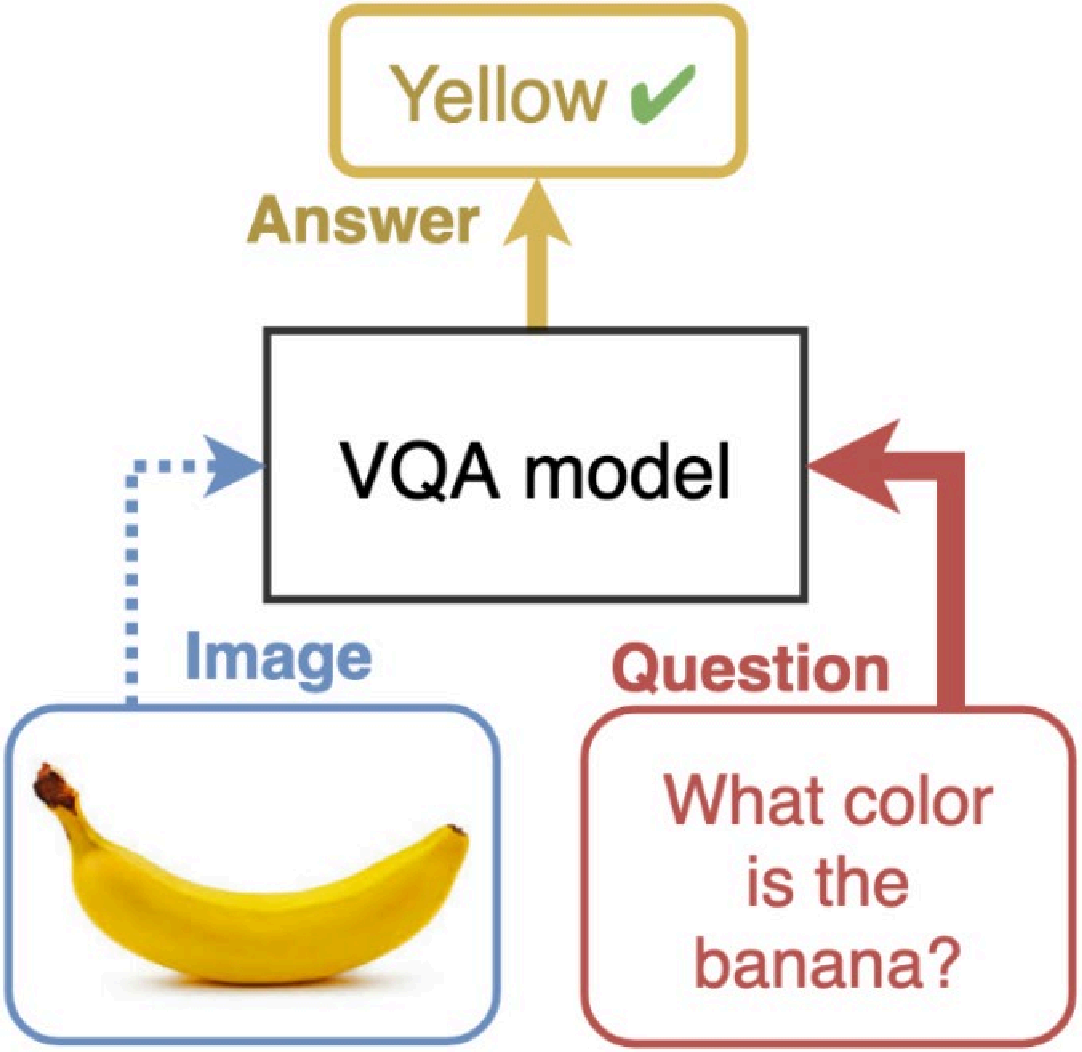
- Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions – Cornia M et al., CVPR 2019
- Pointing Novel Objects in Image Captioning, Li Y et al., CVPR 2019
- Describing like humans: on diversity in image captioning – Wang Q et al., CVPR 2019
- Length-Controllable Image Captioning – Deng C et al., ECCV 2020
- Transformer-based local-global guidance for image captioning – Parvin et al., 2023
- Show, tell and summarise: learning to generate and summarise radiology findings from medical images – Singh et al., Neural Computing & Applications, 2021.
- Medical image captioning via generative pretrained transformers – Selivanov et al. Scientific Reports, 2023

- Let's see Image captioning in action
  - https://milhidaka.github.io/chainer-image-caption/

Source: Awesome Image Captioning. https://githubmemory.com/repo/alibabadoufu/awesome-image-captioning

Visual Question Answering (VQA)

➢ Given an image, can our machine answer the corresponding questions in natural language?



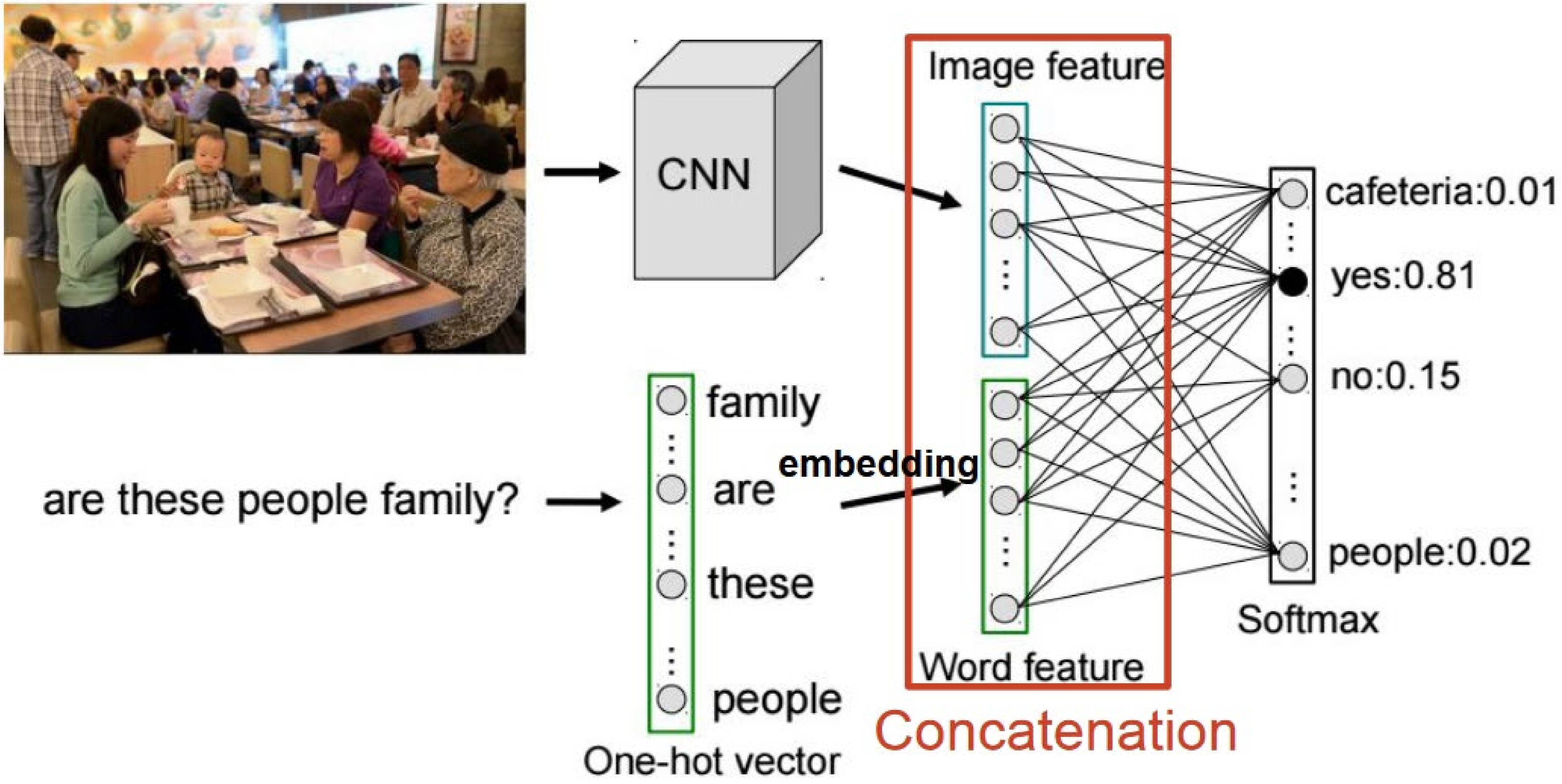
VQA Dataset

➤


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

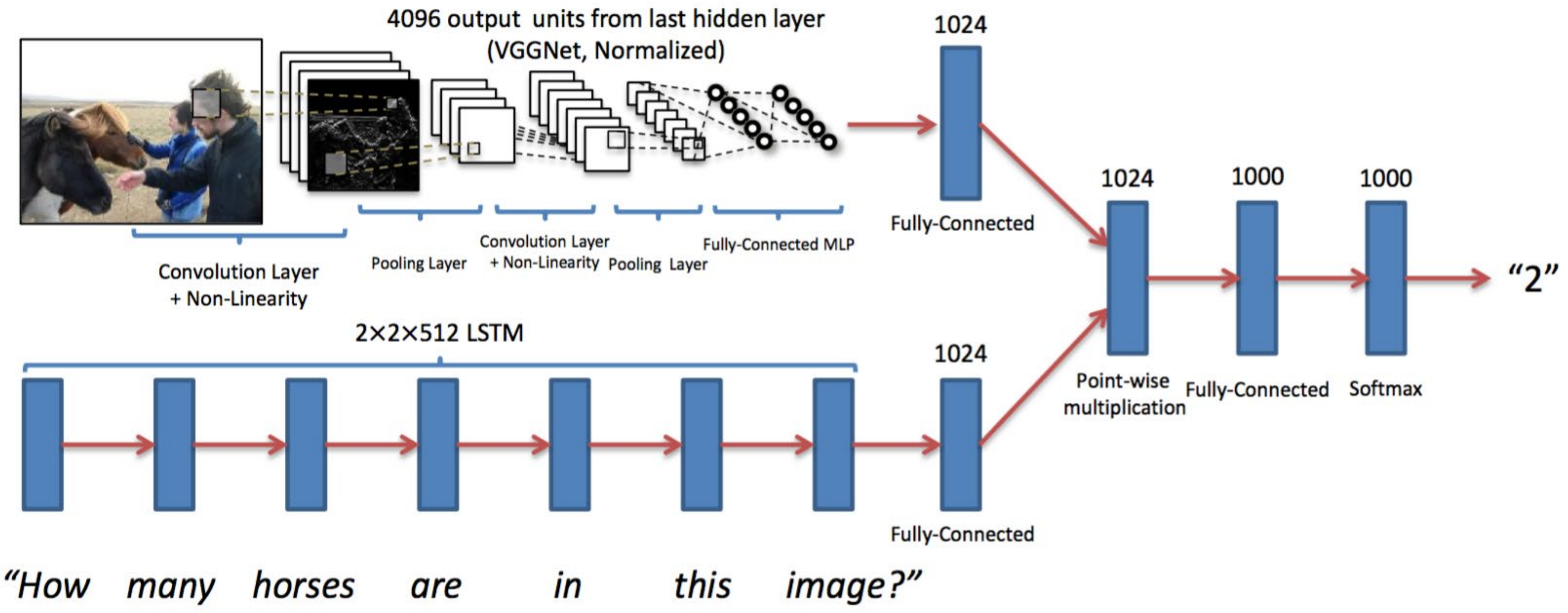
VQA Approach: Bag-of-words + Image feature (iBOWIMG)

- Combine image and word embeddings to predict answer



VQA Common Approach

- Combine CNN (for vision) and RNN (for language) to predict answer



VQA – set of problems

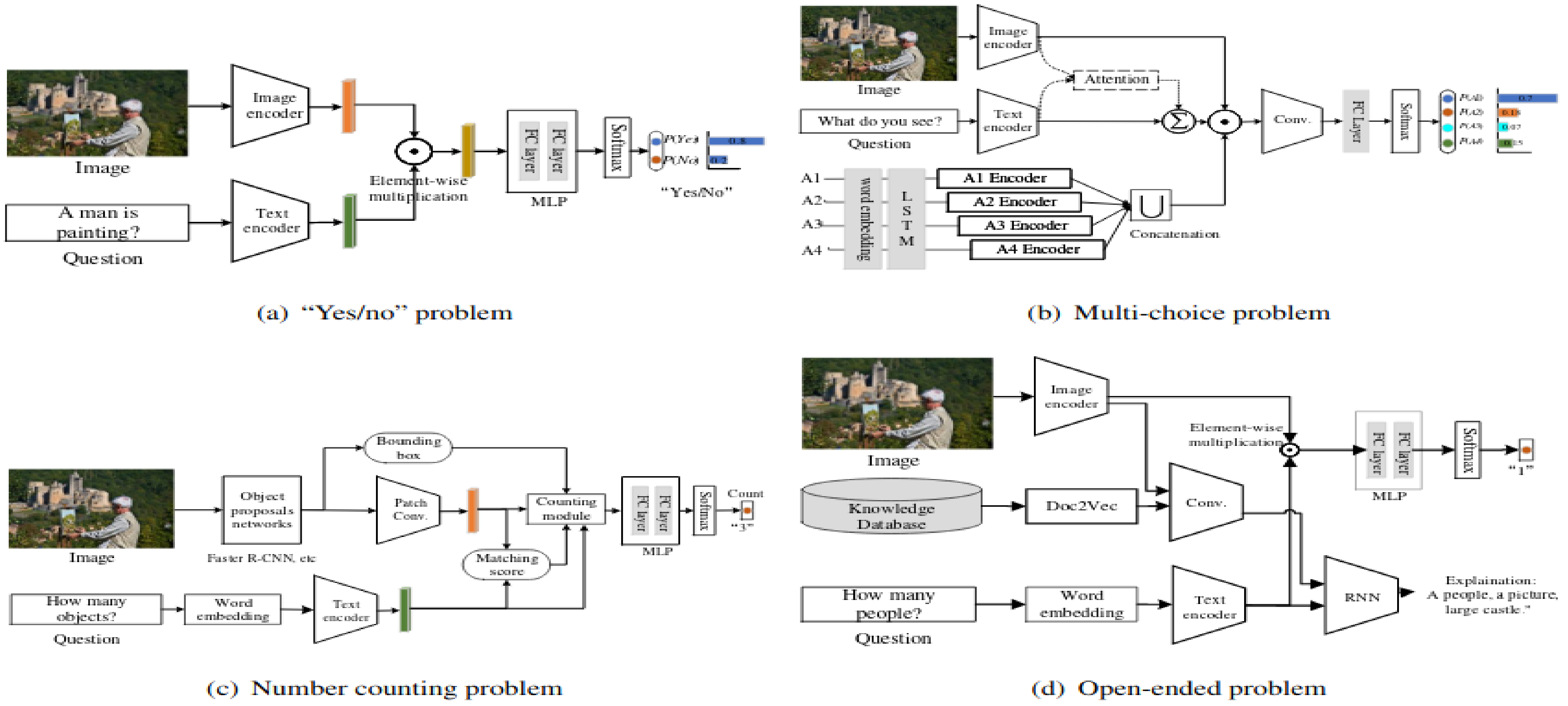
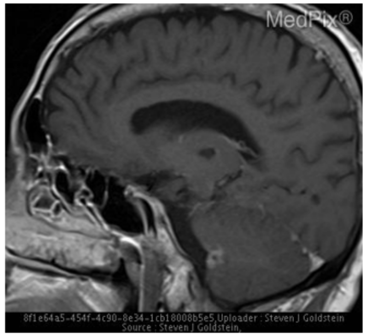


Figure 2 Common types of visual question answering. "Yes/No" problem and multi-choice problem can be regarded as a classification problem, while number counting problem and open-ended problem can be viewed as a caption generation problem.

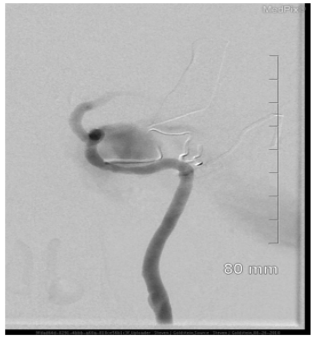
VQA in the Medical Domain

- Aims to create a system that can answer natural language questions based on a given medical image.
- VQA-Med 2019 Challenge dataset

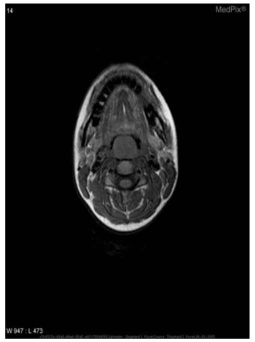


Q: What is the plane of this MRI?
A: Sagittal

Q: The CT scan shows what organ system?
A: Spine and contents



Q: With what modality is this image taken?
A: AN - angiogram

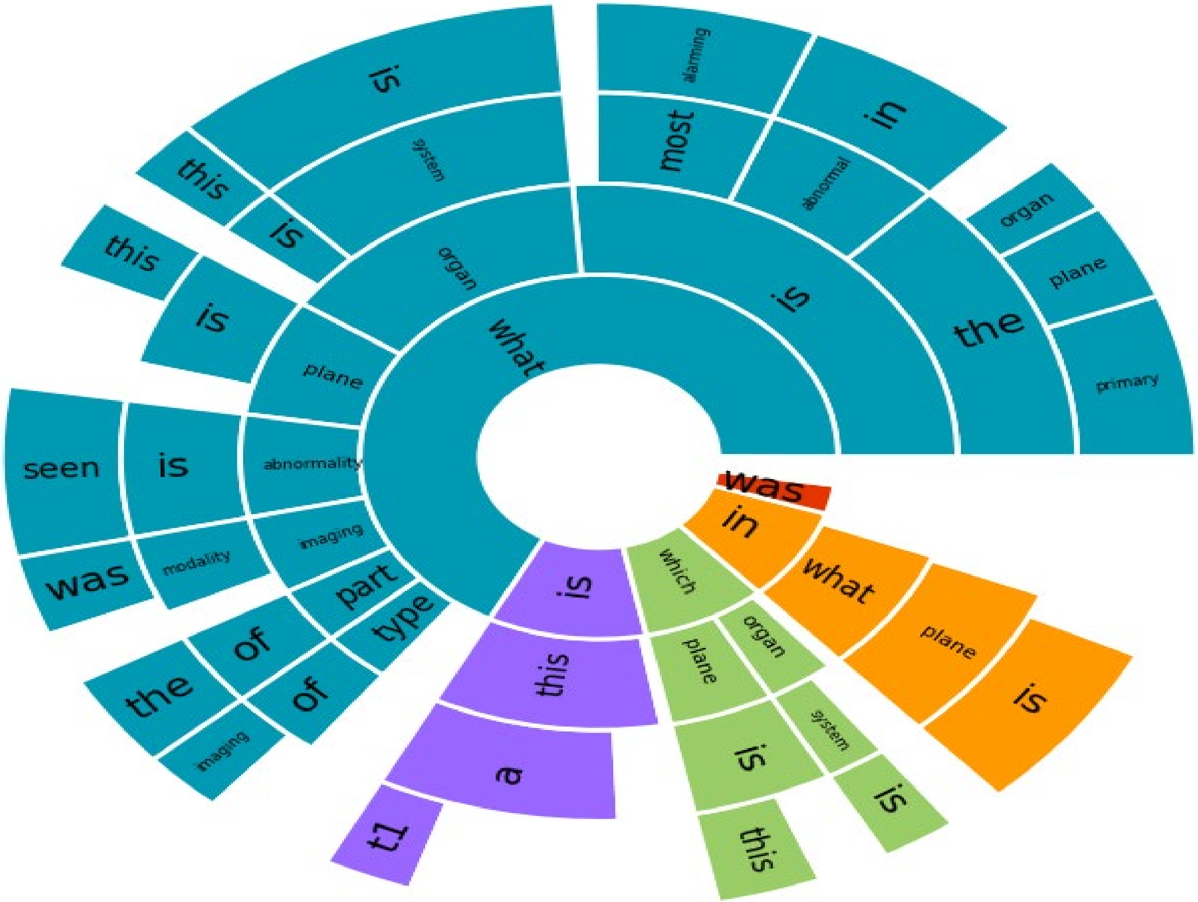


Q: What is most alarming about this MRI?
A: Schwannoma

Source: Canepa et al. (2023). Visual Question Answering in the Medical Domain. DICTA. 2023.

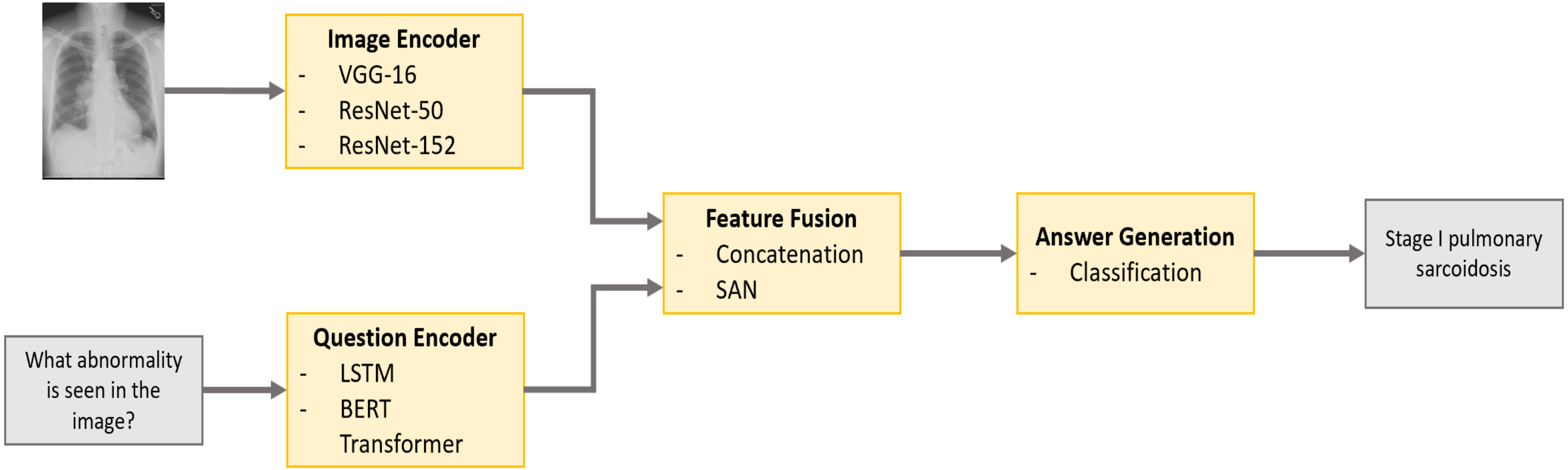
VQA-Med

- Most questions in VQA-Med 2019 dataset are "close-ended"
- More than 50% of answers consists of only one word, and more than 82% of answers have between one and three words.
- Best strategy is to do classification rather than generation



VQA-Med Methodology

- Contributions
  - Incorporating medical domain knowledge
    - Image encoder – applied self-supervised pretraining using Radiology Objects in COntext (ROCO) dataset
    - Question encoder – used BioBERT, pretrained on same tasks as BERT, but using the PubMed corpus
  - Evidence verification
    - Gradient Weighted Class Activation Map (Grad-CAM) for evidence verification



Source: Canepa et al. (2023). Visual Question Answering in the Medical Domain. DICTA. 2023.

Results

- Test accuracy achieved by each model variation

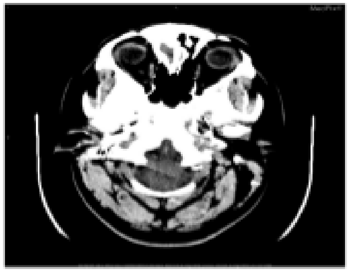
Model Variation	Test Accuracy
VGG-16 + LSTM + Concatenation	0.56
ResNet-50 + LSTM + Concatenation	0.54
ResNet-152 + LSTM + Concatenation	0.53
VGG-16 + BERT + Concatenation	**0.60**
VGG-16 + BERT + SAN	0.58
VGG-16 + BioBERT + Concatenation	**0.60**
Pretrained VGG-16 + BERT + Concatenation	**0.60**

- Accuracy of the baseline vs. BERT model per category type

Model Variation	Basceline	+BERT
Modality	0.64	0.76
Plane	0.78	0.77
Organ	0.74	0.74
Abnormality	0.06	0.08
Overall	0.56	**0.60**

Results

➢ Effect of using BERT vs. LSTM



Q: What imaging modality was used to take this image?
GT: CT with IV contrast
Baseline: Skull fracture from cell phone
+BERT: CT with IV contrast

(a) Category misclassification

Q: What was this image taken with?
GT: MR – PDW proton density
Baseline: Yes
+BERT: MR – PDW proton density

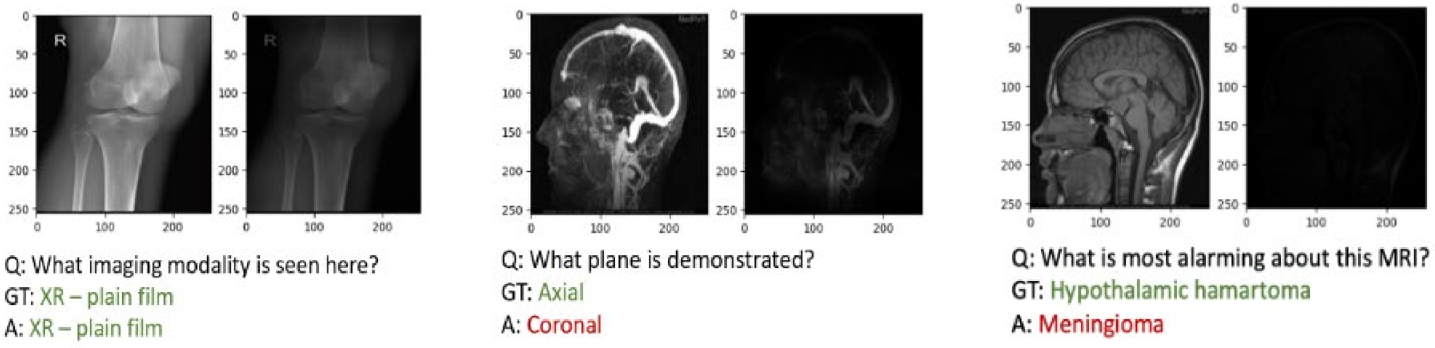
(b) Wrong answer type

Q: Is this a contrast or noncontrast MRI?
GT: Noncontrast
Baseline: MR – flair
+BERT: Noncontrast

(c) Not choosing from options

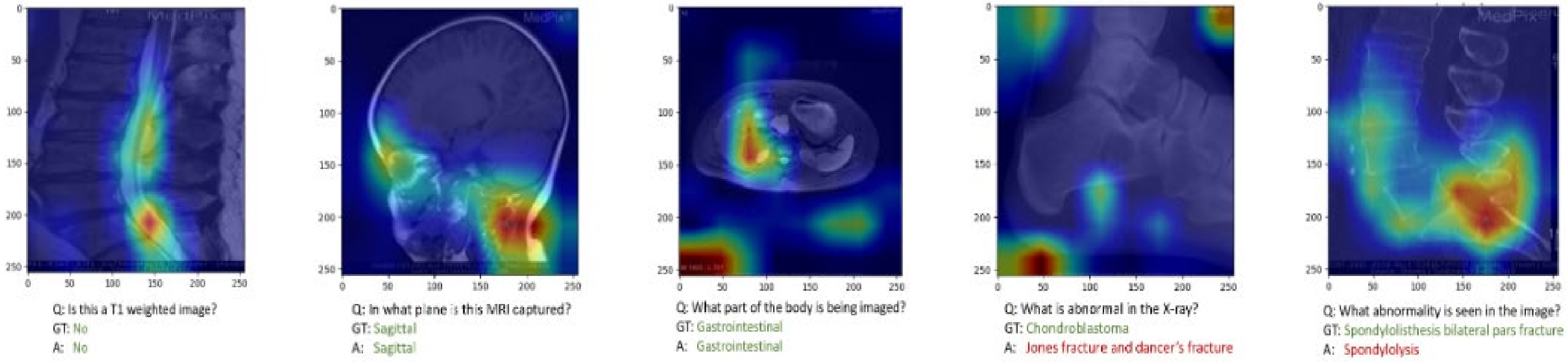
Evidence verification

- Attention distribution output by Stacked Attention Network (SAN) fusion method



(a) (b) (c)

- Grad-CAM output



(a) (b) (c) (d) (e)

Source: Canepa et al. (2023). Visual Question Answering in the Medical Domain. DICTA. 2023.

Let's see in action

- https://visualqa.org/visualize/
- https://demo.allennlp.org/visual-question-answering

Visual Dialog

- Ability to hold a meaningful dialog with humans in natural language about visual content
- **Given an image** I, a **history of dialog consisting** of sequence of **question-answer pairs**, and a natural language **follow-up question**, the task for the machine is to **answer the question** in free-from natural language.

Image Captioning vs. VQA vs. Visual Dialog

Captioning

Two people are in a wheelchair and one is holding a racket.

VQA

Q: How many people on wheelchairs ?

A: Two

Q: How many wheelchairs ?

A: One

Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a racket ?

A: The woman

Visual Dialog

Q: What is the gender of the one in the white shirt ?

A: She is a woman

Q: What is she doing ?

A: Playing a Wii game

Q: Is that a man to her right

A: No, it's a woman

Session Variables

Visual Dialog – Late Fusion Encoder

- Entire history H is concatenated and encoded by LSTM

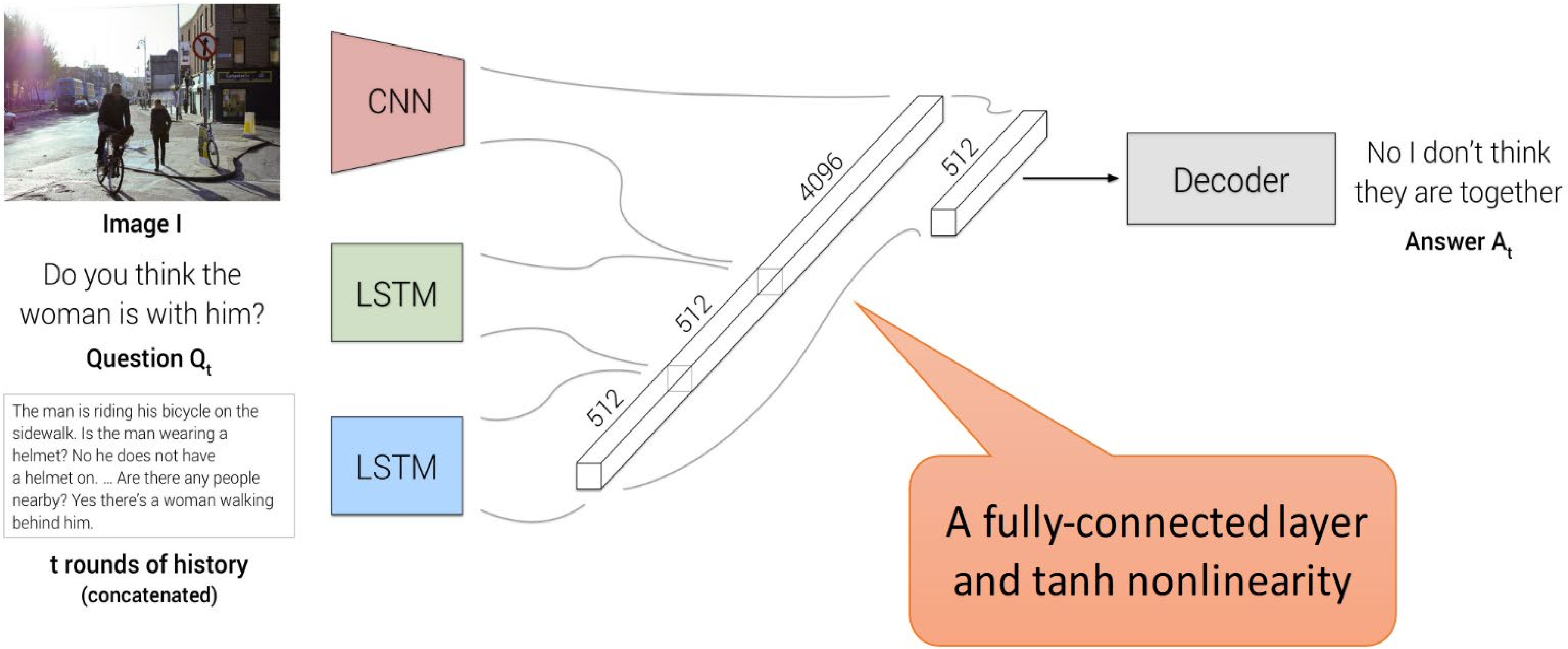
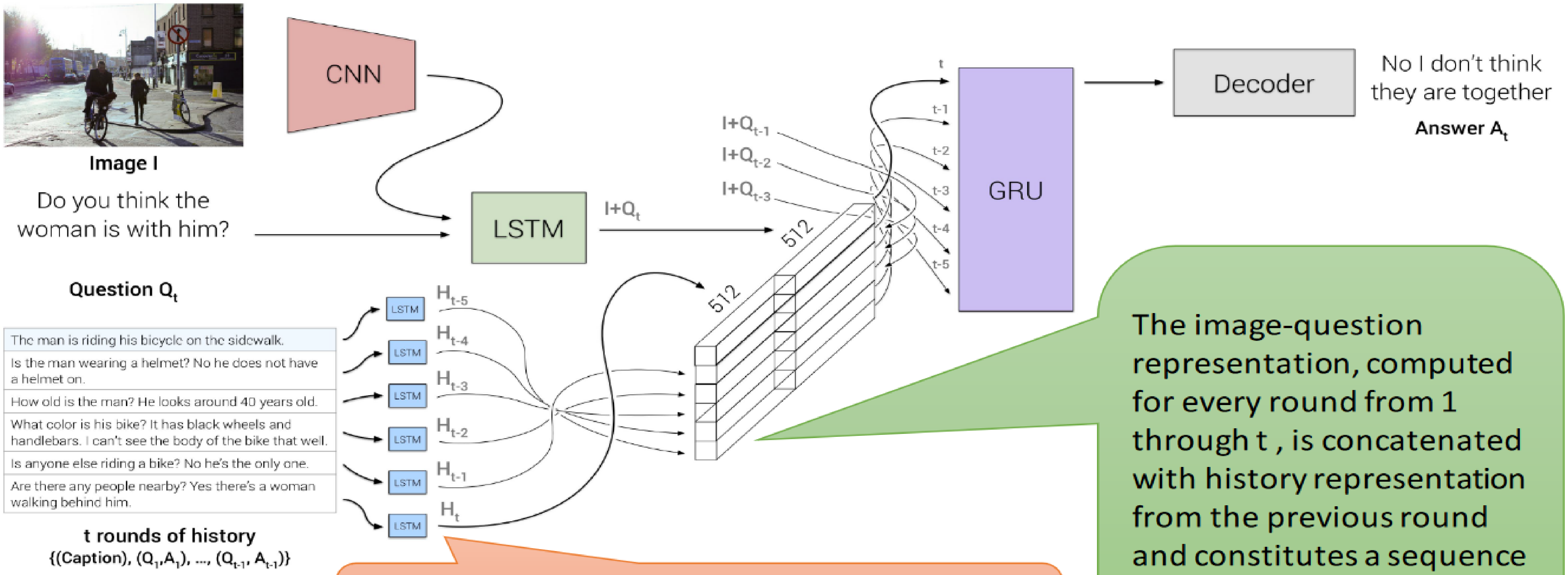


Image Credit: Das et al., (2017) Visual Dialog. https://visualdialog.org/

Visual Dialog – Hierarchical Recurrent Encoder

- Each QA-pairs in history H separately encoded by the LSTM

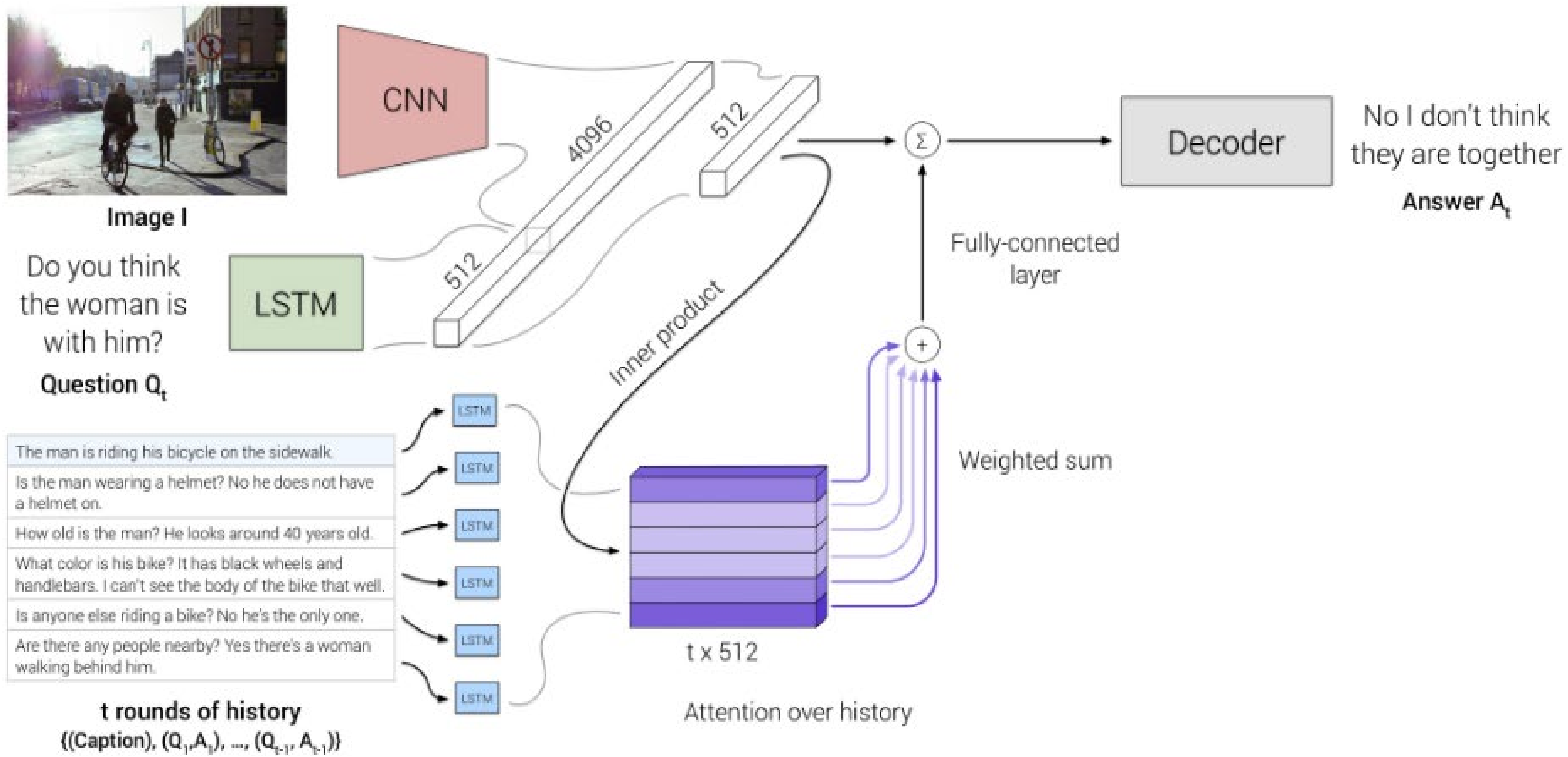


The image-question representation, computed for every round from 1 through t , is concatenated with history representation from the previous round and constitutes a sequence of question-history vectors

Each QA-pair in dialog history is independently encoded by another LSTM with shared weights

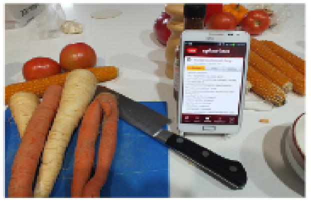
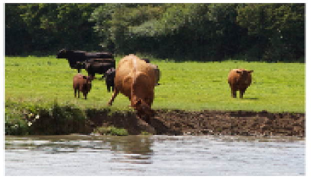
Visual Dialog – Hierarchical Recurrent Encoder

- Compute inner product of question vector with each history vector to get scores over previous rounds, which are fed to a SoftMax to get attention-over-history probabilities.



Improving Visual Dialog (and few examples)

➢ Generates more diverse, image relevant and fluent dialog.

Image + Caption	GT dialog (Das et al., 2017a)		Q-BOT-A-BOT dialog		Diverse-Q-BOT-A-BOT dialog	
	what color is the cutting board?	blue	is the photo in color ?	no	is the photo in color ?	1 picture is in color
	how big is it	about the size of a regular laptop screen	how many people do you see ?	no people	is the photo close up ?	black and white
	how many carrots are there ?	4	is this a close up ?	no	is the photo in color ?	no it 's black and white
	are they the orange kind ?	2 are orange , 2 are white	is this a close up ?	zebra is	is it a close up ?	no it 's a close up of the cake
	what kind of cell phone is it ?	a samsung	is this a close up ?	no	can you tell what brand it is ?	no i ca n't see the brand
	can you see any text on the screen ?	a recipe for UNK UNK soup	is this a close up ?	no	can you tell if it 's a male or female ?	no i ca n't tell
	do you see the room ?	just the counter	is this a close up ?	no	can you tell if it 's a male or female ?	no i ca n't tell
	are there people ?	not visible	is this a close up ?	no	is this in a kitchen ?	nope it 's in a home
	what color is the counter ?	white	is this a close up ?	no	are there any utensils ?	no
a cutting board with carrots next to a cell phone.			is this a close up ?	no	are there any utensils ?	nope
	is the field large ?	yes	what color are the cows ?	they are brown and white	do you see any people ?	1
	is the grass green or brown ?	green	how many cows are there ?	4	do you see a fence ?	a small 1
	can you see a barn ?	no	what color are the cows ?	they are brown and white	how many cows are there ?	5
	can you see a fence ?	no	can you see the sky ?	it is not visible	what color are the cows ?	black and white
	can you see a water supply ?	yes	do you see any trees ?	i do not see any trees	can you see the sky ?	some
	can you see any people ?	no	can you see the sky ?	green sky	can you see the sky ?	nope
	can you see trees ?	yes	are there any clouds ?	the sky is not visible	are there any other animals ?	not that i can see
	can you see mountains in the background ?	no	how many cows are there ?	2	can you see the sky ?	not at all
	is the grass tall or short ?	short	what color are the cows ?	the cows are brown and white	are there any trees ?	in the background
many brown and black cows standing in the grass			what color are the cows ?	they are brown	can you see the sky ?	not at all
	what type of building ?	may be part of a zoo	how many elephants are there ?	ca n't tell	how many elephants are there ?	1
	how many elephants ?	about 4	what color are the elephants ?	ca n't tell	any people ?	ca n't see any
	is it sunny ?	clear day , yes	how many elephants are there ?	ca n't tell	male or female ?	not sure
	any trees ?	yes	what color is the fence ?	i ca n't tell	what is the elephant doing ?	just standing there
	what color is the building ?	dark grey	can you see the sky ?	i ca n't see the sky	male or female ?	there are 2 males and 1 female
	are there other animals ?	no	are there any trees ?	just a few	can you see the sky ?	1 of it
	are there any people ?	yes	does this look like a zoo ?	1 of them is	what color is the building ?	just a black and white picture
	what are they wearing ?	shorts and shirts	are there any people ?	nothing in the picture	do you see any trees ?	there are trees in the background
	how many ?	3	how many elephants are there ?	just 1	are there any other animals ?	red and white
many elephants standing in an enclosure near a building			can you see the sky ?	just a little	do you see any trees ?	i see some in the background
	is the photo in color ?	yes	is the photo in color ?	5 of them are	is the photo in color ?	no
	is there humans in the photo ?	no	are there any people ?	it 's a close up so i ca n't see any people	is there any people ?	5
	is there any animals in the photo ?	no	what color is the sink ?	about the same color as the toilet	what color walls ?	i ca n't see the walls
	what color are the walls ?	tan	do you see a mirror ?	no	is there a mirror ?	yes
	what color is the sink ?	white	what color is the sink ?	yes it is white	is there a mirror ?	yes
	is there any towels ?	no	what color is the sink ?	yes it is white	is there a mirror ?	yes
	is there any pictures on the wall ?	no	what shape is the mirror ?	yes , it 's a rectangle	are the walls seen ?	no
	what color is the tub ?	white	are there any towels ?	he 's on a towel rack	what color are the walls ?	no walls
	can you see the floor ?	no	are there any towels ?	i do n't see any towels	what is the floor made of ?	yes
a reflection of the tub is in the bathroom mirror along with the sink , you can see all the room			are there any towels ?	i do n't see any towels	what color is the floor ?	i ca n't see the floor

Source: Murahari et al., (2019) Improving Generative Visual Dialog by Answering Diverse Questions.

Many more tasks at CV+NLP

➢ Vision-Language Navigation (VLN)

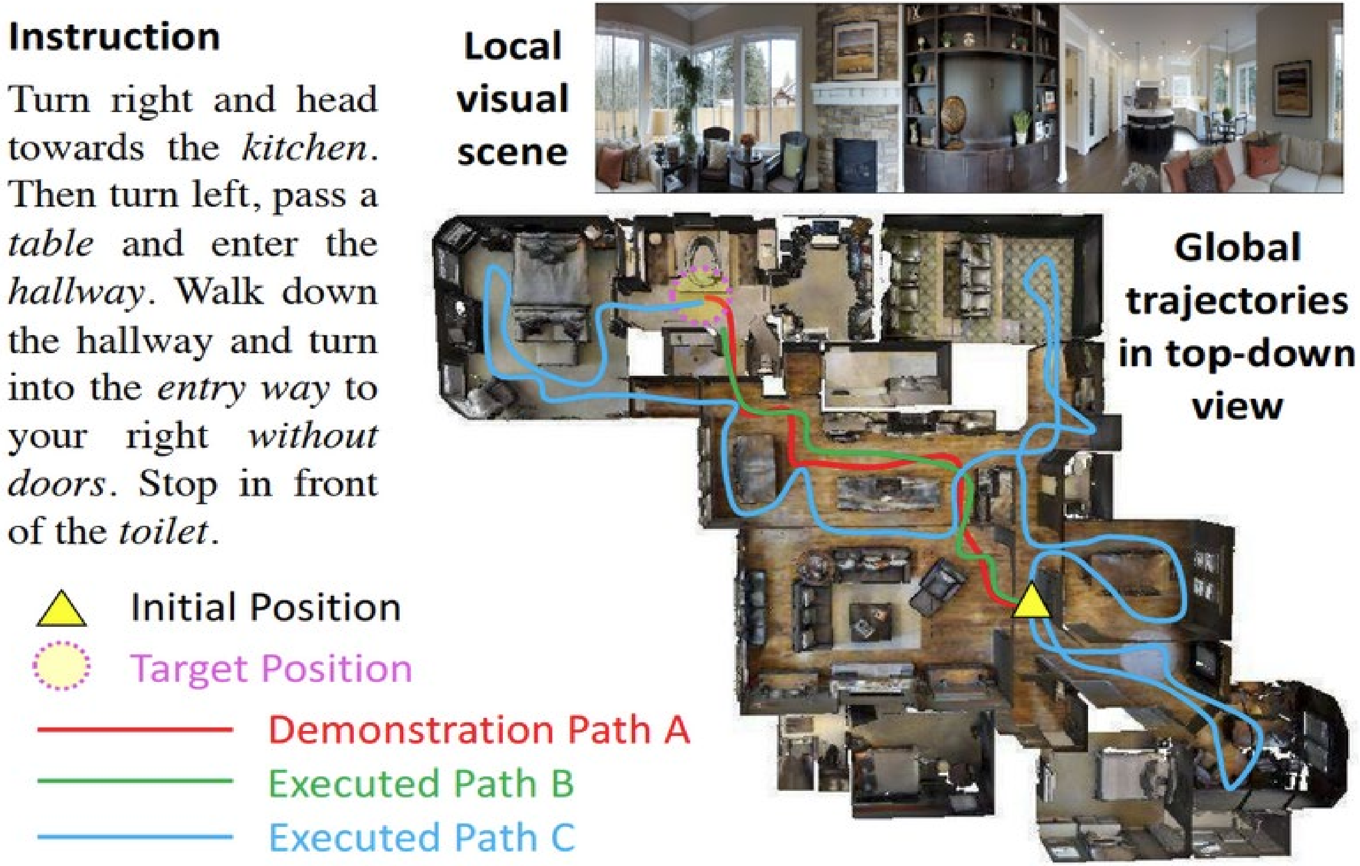


Figure 1: Demonstration of the VLN task. The instruction, the local visual scene, and the global trajectories in a top-down view. The agent does not have access to the top-down view. Path A is the demonstration path following the instruction. Path B and C represent two different paths executed by the agent. Figure credit: Wang et al. (2019).

Many more tasks at CV+NLP

- Visual grounding for referring expressions
  - Grounding an utterance to refer to something or someone in image

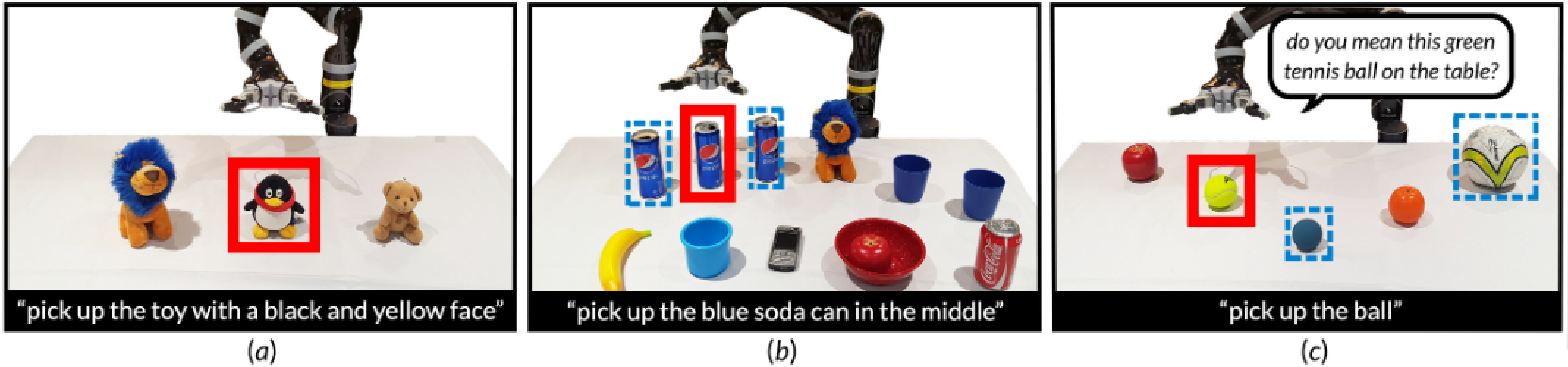


Fig. 1: Interactive visual grounding of referring expressions. (*a*) Ground self-referential expressions. (*b*) Ground relational expressions. (*c*) Ask questions to resolve ambiguity. Red boxes indicate referred objects. Blue dashed boxes indicate candidate objects. See also the accompanying video at http://bit.ly/INGRESSvid.

- Let's see video in action: http://bit.ly/INGRESSvid

Source: Sridhar et al. (2018). Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction

Many more tasks at CV+NLP

- Fun stuff: Comic books

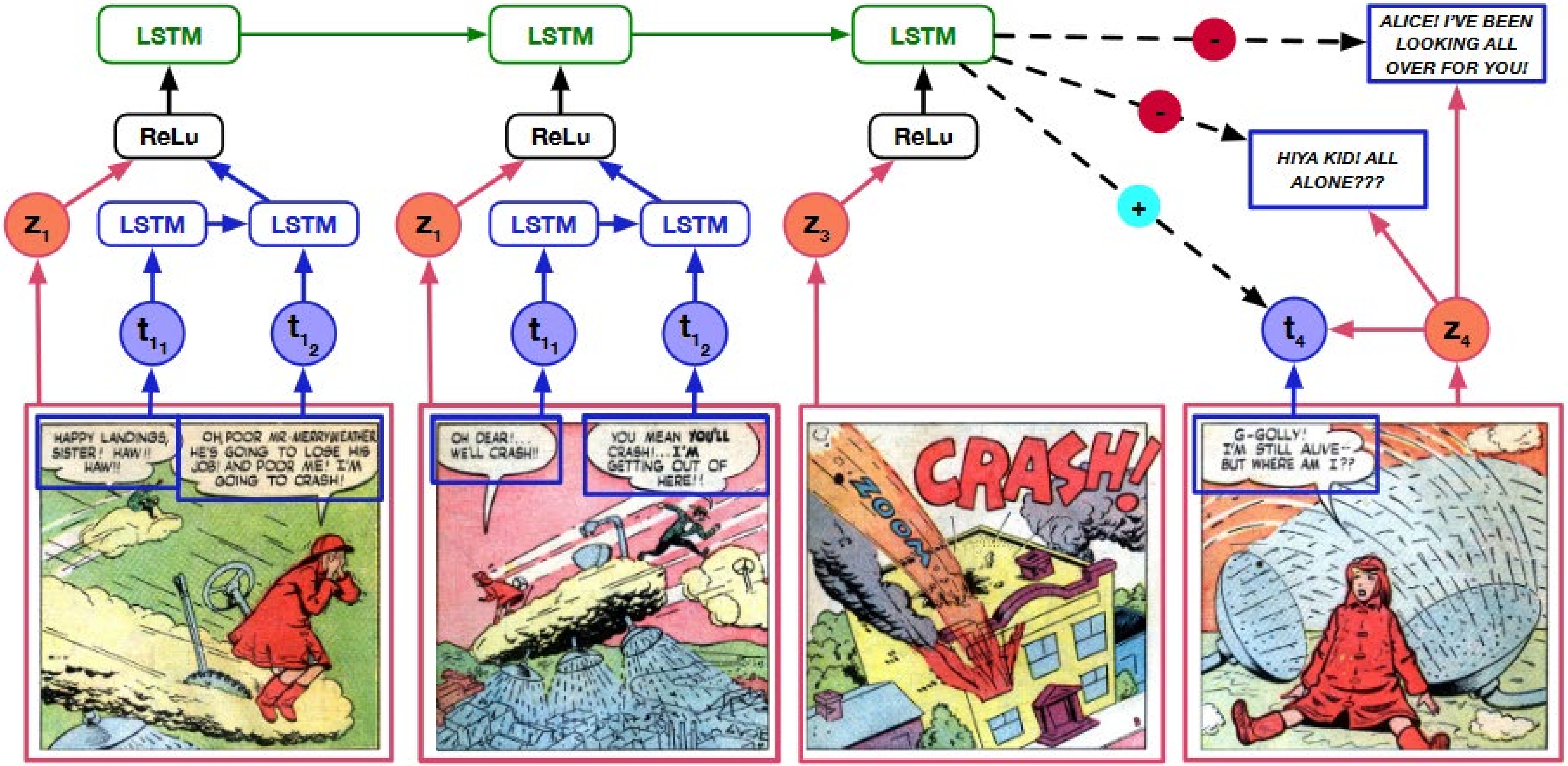
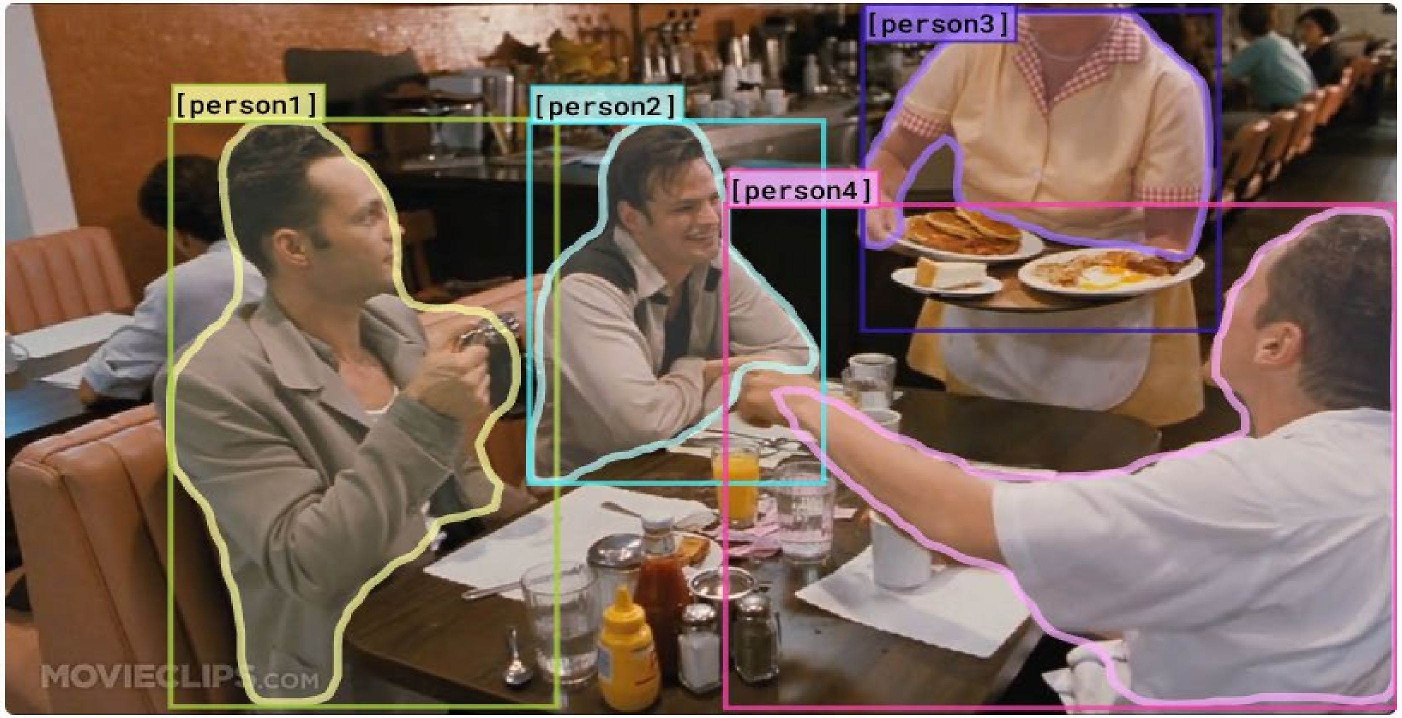


Figure 5. The **image-text** architecture applied to an instance of the *text cloze* task. Pretrained image features are combined with learned text features in a hierarchical LSTM architecture to form a context representation, which is then used to score text candidates.

Many more tasks at CV+NLP

➢ From Recognition to Cognition: Visual Commonsense Reasoning



hide all | show all | [person1] | [person2] | [person3] | [person4]

more objects »

Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

Rationale: I think so because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.

Source: Zellers et al. (2019) From Recognition to Cognition: Visual Commonsense Reasoning

Summary

- Vision + Language: Help us to understand human brain functioning better
- Visual Turing Test for modern AI systems
- A step towards Artificial General Intelligence (AGI)
- Variety of applications:
  - Helping visually-impaired people
  - Early child education
  - Personal assistants
  - Robot navigation
  - Video surveillance systems
  - Search engines

Questions?

If you feel fascinated and want to get involved in research, feel fee to reach out at sonit.singh@unsw.edu.au