Image captioning
&
Visual question answering

Hakan Bilen

Machine Learning Practical - MLP Lecture 17
27 Feb 2019

How would you describe this image?

- A man taking picture of the landscape
- A man facing the mountains to take a picture
- There is snow on the mountains
- …
- “the birthday guy, part one”

Image source

Today's goal

- Tasks beyond single modality
  - image and text
- Tasks beyond "what" and "where"
  - relations in natural language
- Mimicking human intelligence
- How to design a learning machine for the task of interest
  - Customise network architecture
  - Integrate multiple modalities

"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."

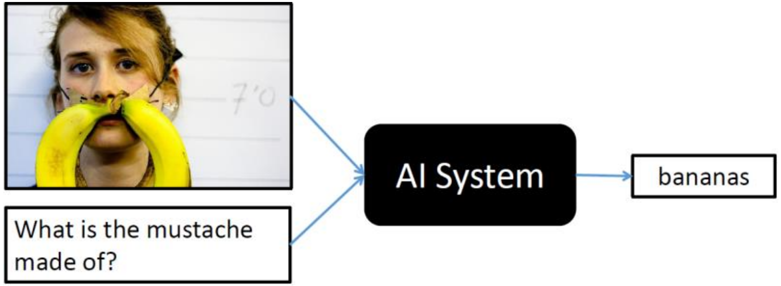


Image captioning

Photo credit: Hodosh et al

- A man is doing tricks on a bicycle on ramps in front of a crowd.
- A man on a bike executes a jump as part of a competition while the crowd watches.
- A man rides a yellow bike over a ramp while others watch.
- Bike rider jumping obstacles.
- Bmx biker jumps off of ramp.

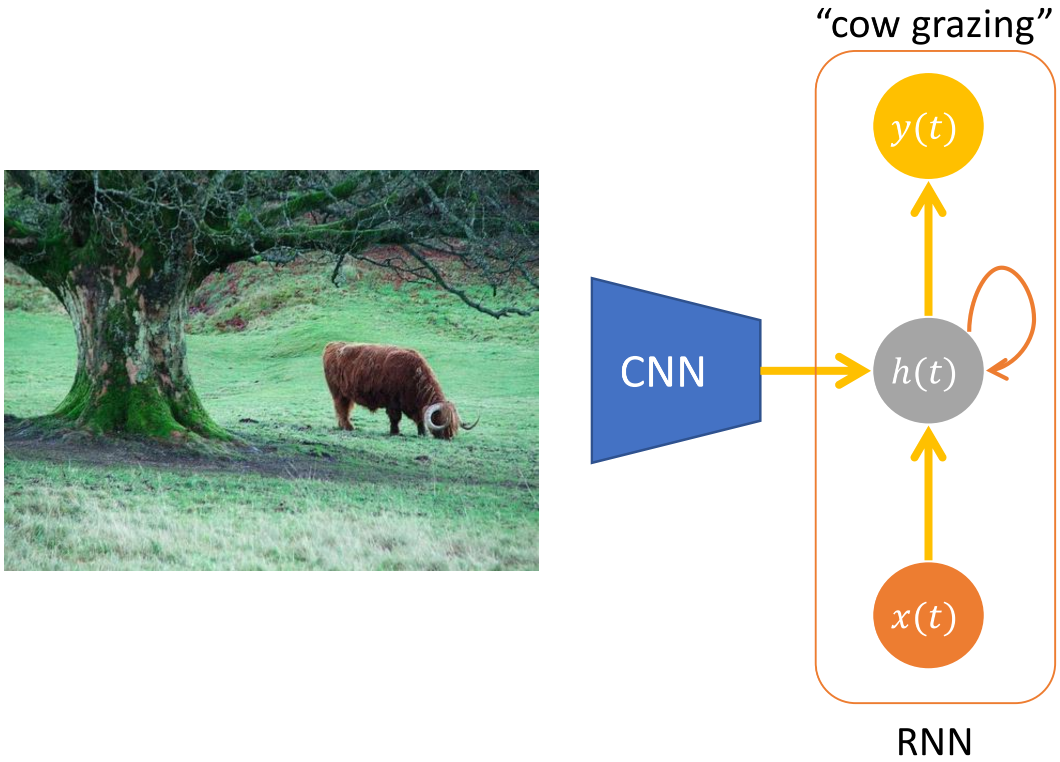
Image captioning

Photo credit: Hodosh et al

- **Goal**: Automatically generate an accurate caption for a given image using proper (English) language
  - Naming objects in the image
  - Relations between objects, their attributes and activities
- **Challenges**
  - Object categories are not pre-specified (no closed set)
  - Do not describe unimportant details (depending on visual salience)
  - Human descriptions vary

Slide credit: Lazebnik CS 598 LAZ

CNN + RNN

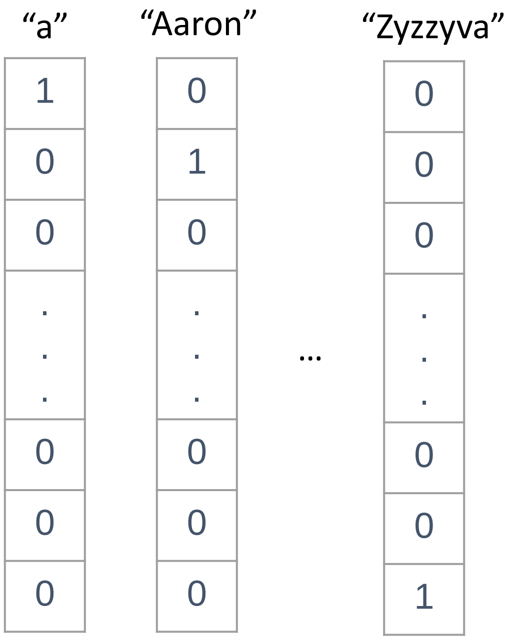


[Mao et al. (2015) "Explain Images with Multimodal Recurrent Neural Networks", ICLR]
[Karpathy and Fei-Fei (2015), Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR]
[Vinyals et al (2015), Show and Tell: A Neural Image Caption Generator, ICCV]

Word representation

One hot representation

- Each word is represented by a sparse vector $x^o \in R^{|V|}$

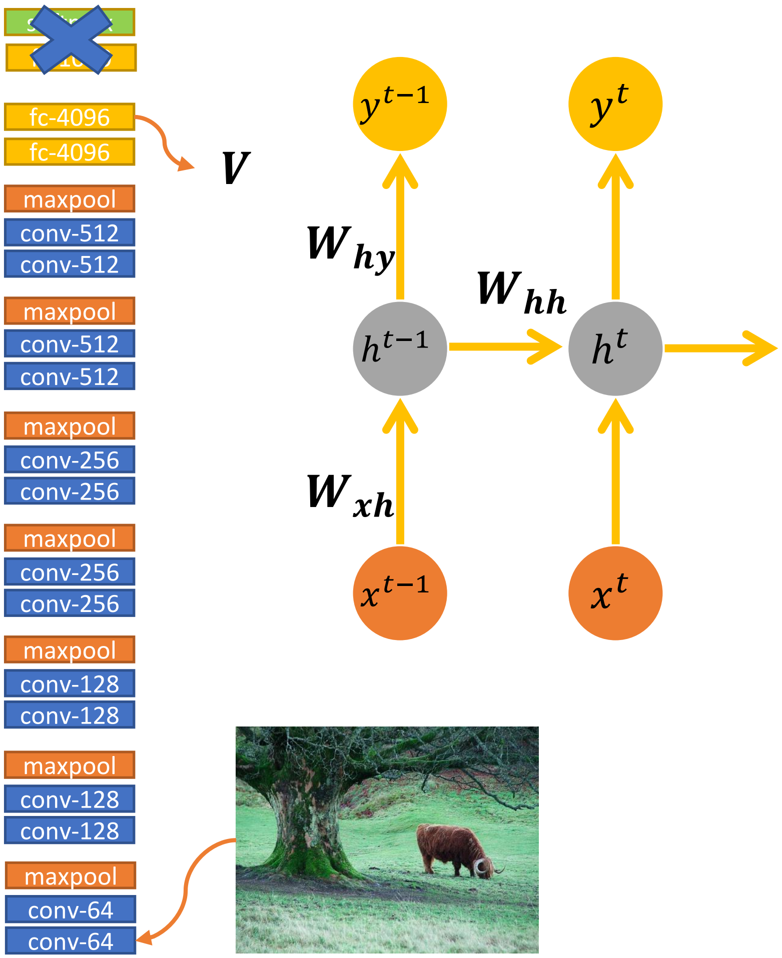


Word vectors

- Each word is represented by a dense vector $x \in R^D$ where $\mathrm{D} \ll |V|$
- Semantically close words are also close in the vector space
- Semantic relations are preserved
- "king" + "woman" – "man" = "queen"
- A word vector can be written as

$$x^W = W x^o$$

[Mikolov et al (2014), Distributed Representations of Words and Phrases and their Compositionality, NIPS]



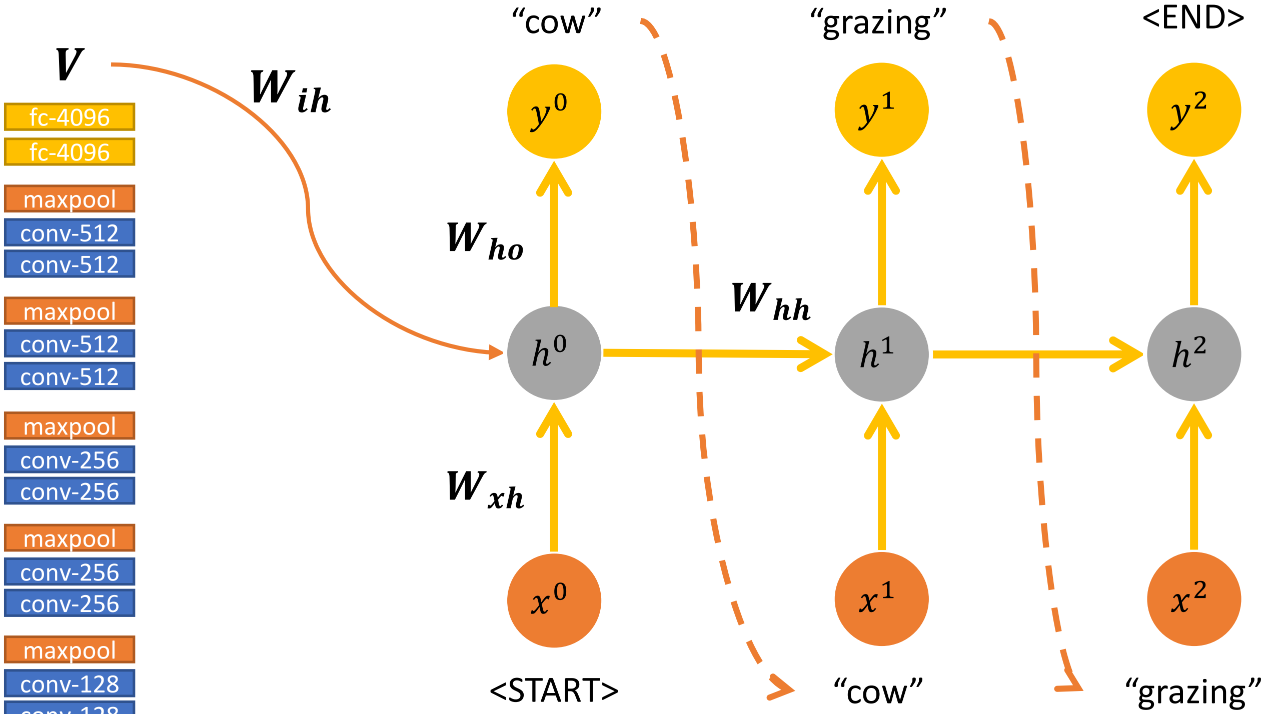
Vanilla RNN

- $h(t) = tanh(W_{xh}x(t) + W_{hh}h(t-1))$
- $y(t) = softmax(W_{ho}h(t))$

CNN+RNN

- $h(t) = tanh(W_{xh}x(t) + W_{hh}h(t-1) + W_{ih}V)$
- $y(t) = softmax(W_{ho}h(t))$

Karpathy and Fei-Fei (2015), Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR

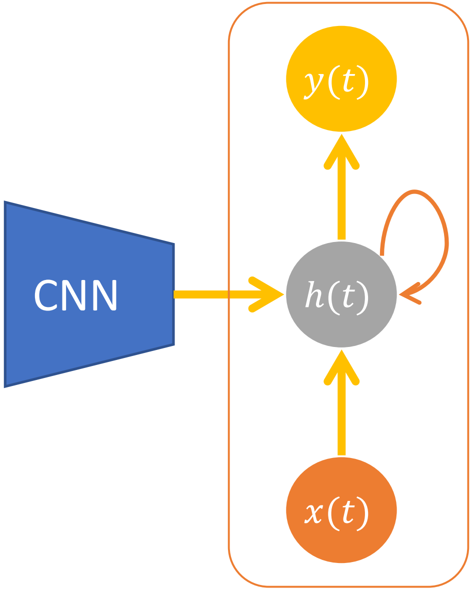


- CNN features are only passed at the first iteration (t=0)!
- Input vectors are word2vec embeddings

CNN+RNN

- $h(t) = tanh(W_{xh}x(t) + W_{hh}h(t-1) + W_{ih}V)$
- $y(t) = softmax(W_{ho}h(t))$

[Karpathy and Fei-Fei (2015), Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR]

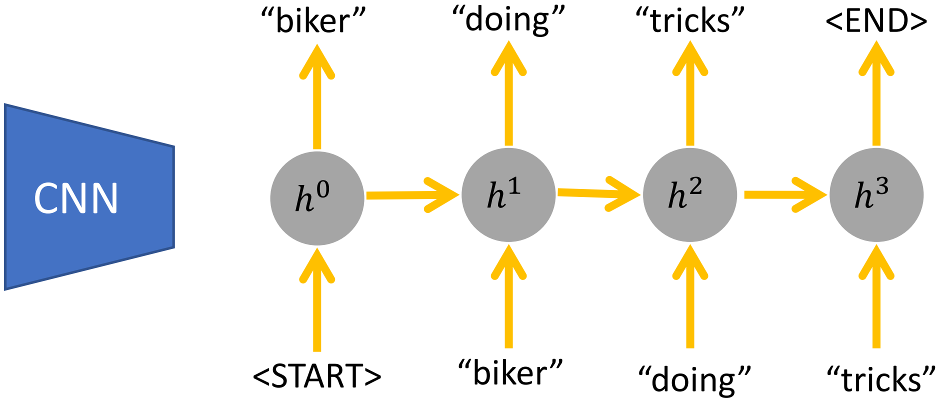


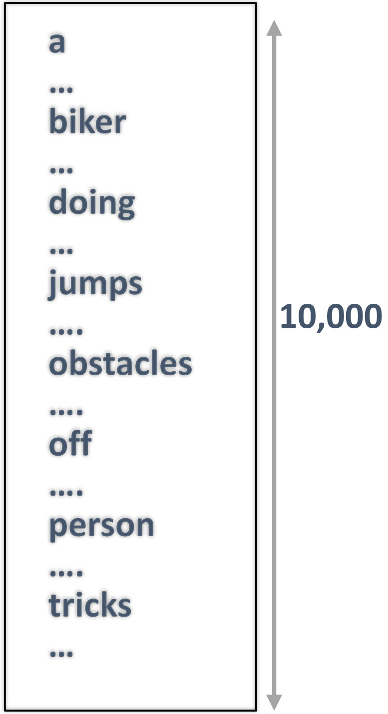
RNN

Question

How can we get multiple captions for an image using a model?

Beam search algorithm





Problem 1: Only one possible output

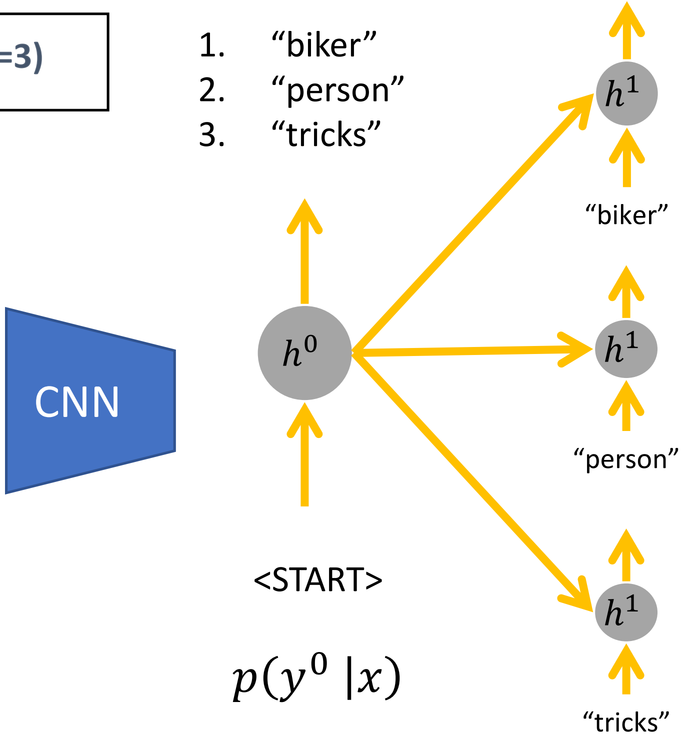
Problem 2: Output may not be the highest probability one for the given model

$$p(y^1, y^0|x) = p(y^0\ |x) \cdot p(y^1\ |x, y^0)$$

Karpathy and Fei-Fei (2015), Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR

Beam search algorithm

Beam width (e.g. B=3)



$$p(y^0 \,|x)$$

$$p(y^0 \,|x) \cdot p(y^1 \,|x, y^0)$$

- Each prediction y^t for "biker", "person" and "tricks" contains 10,000 probabilities
- Keep the best B of them
- Move to the iteration *t+1*

[Karpathy and Fei-Fei (2015), Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR]

Image captioning examples

A cat sitting on a suitcase on the floor

A cat is sitting on a tree branch

A dog is running in the grass with a frisbee

A white teddy bear sitting in the grass

Two people walking on the beach with surfboards

A tennis player in action on the court

Two giraffes standing in a grassy field

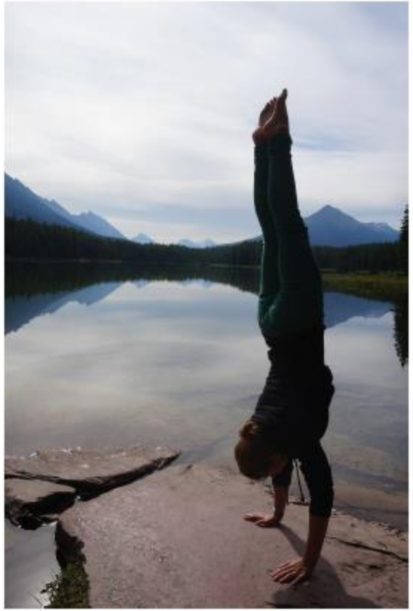
A man riding a dirt bike on a dirt track

Slide credit: Fei-Fei Li & J. Johnson & S. Young

Image captioning failure cases

A woman is holding a cat in her hand

A person holding a computer mouse on a desk



A woman standing on a beach holding a surfboard

A bird is perched on a tree branch

A man in a baseball uniform throwing a ball

Slide credit: Fei-Fei Li & J. Johnson & S. Young

Evaluation

Photo credit: Hodosh et al

References
- A man is doing tricks on a bicycle on ramps in front of a crowd.
- A man on a bike executes a jump as part of a competition while the crowd watches.
- A man rides a yellow bike over a ramp while others watch.
- Bike rider jumping obstacles.
- Bmx biker jumps off of ramp.

Prediction

A man spins around his bike.

Human evaluation

[Vinyals et al (2015), Show and Tell: A Neural Image Caption Generator, ICCV]()

Evaluation

BLEU (**Bi**Lingual **E**valuation **U**nderstudy)

- Substitutes expensive human judgement with automatic evaluation
- Measures overlap of *n*-grams between candidate and reference sentences

<u>Example:</u> The cat is on the mat.
Uni-gram: “the”, “cat”, “is”, “on”, “the”, “mat”
Bi-gram: “the cat”, “cat is”, “is on”, “on the”, “the mat”

<u>Example:</u> BLEU score on unigrams
Reference 1: <u>The</u> cat is on <u>the</u> mat.
Reference 2: There is a cat on <u>the</u> mat.
Candidate: the the the the the the the.

Precision: 7/7 Modified precision: 2/7 Count_{clip}(“the”) / Count(“the”)

[Papieni et al (2002), BLEU: a Method for Automatic Evaluation of Machine Translation, ACL]

Evaluation

Example: BLEU score on bigrams

Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.
Candidate: The cat the cat on the mat.

p_n = BLEU score on n-grams only
Combined score

$$BP \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4} p_n\right)$$

BP penalizes short sentences than reference sentence

	Count	$\text{Count}_{\text{clip}}$
the cat	2	1
cat the	1	0
cat on	1	1
on the	1	1
the mat	1	1

Modified precision: 4/6

Papieni et al (2002), BLEU: a Method for Automatic Evaluation of Machine Translation, ACL

What is wrong with BLEU score?

Problems with BLEU

- Lack of recall

(a) Ref: The cat is on the mat.
(b) Generated: The cat.

- N-gram overlap is insufficient to measure the similarity between meanings

(a) A young girl standing on top of a tennis court.
(b) A giraffe standing on top of a green field.

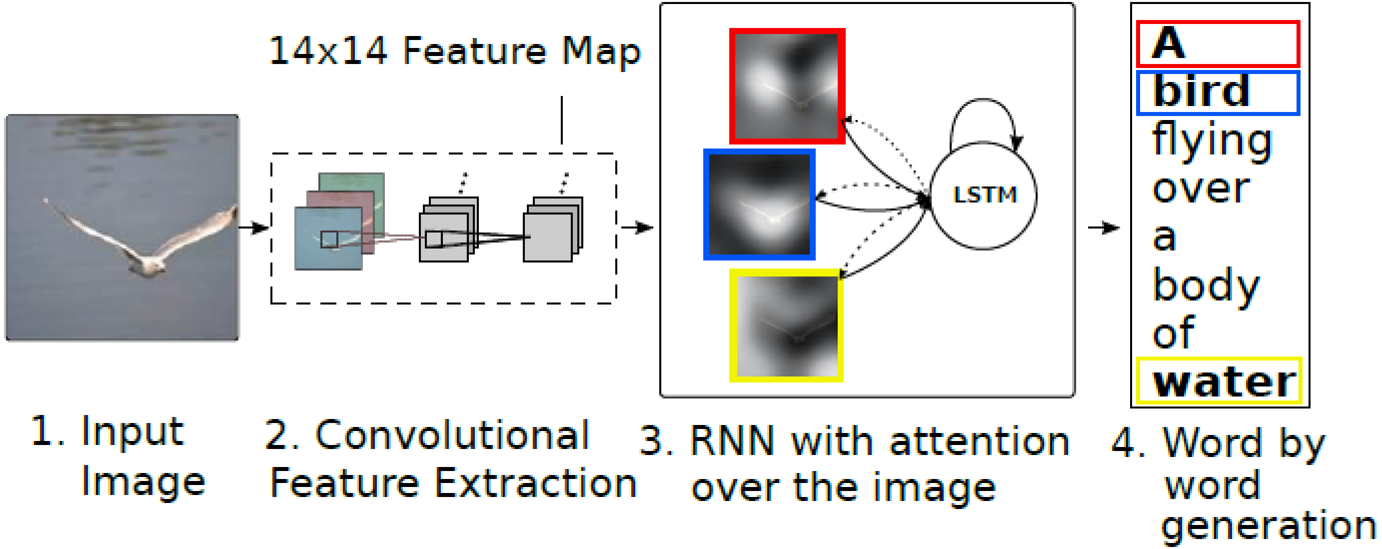
(c) A shiny metal pot filled with some diced veggies.
(d) The pan on the stove has chopped vegetables in it

- Other measures: CIDEr, METEOR, ROUGE-L, SPICE

Anderson et al (2016), SPICE: Semantic Propositional Image Caption, ECCV

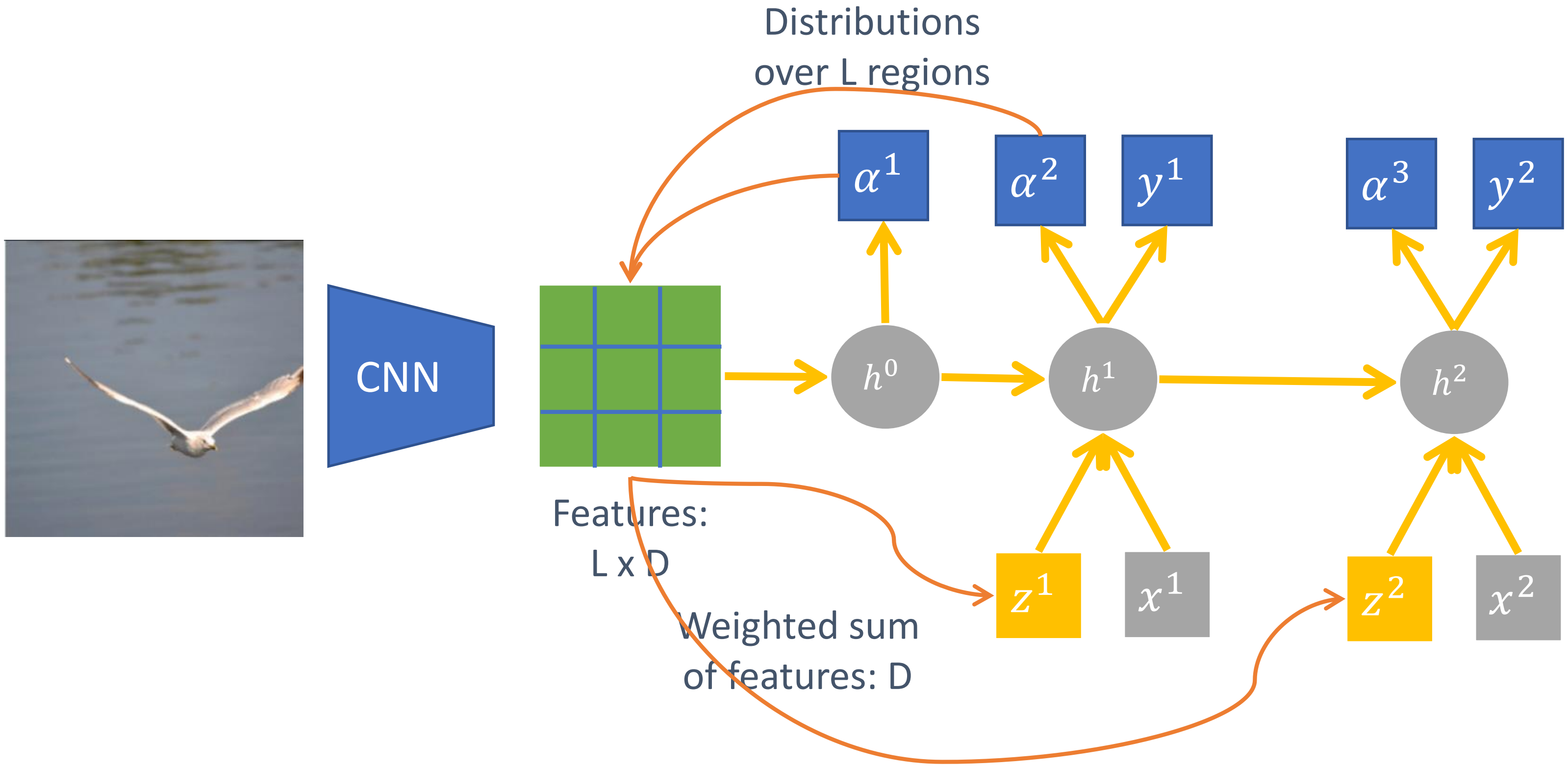
Image captioning with attention

- Human visual system is dynamic, attends to salient objects
- RNN looks at different parts of the image when generating each word



[Xu et al (2015), Show, Attend and Tell: Neural Image Caption, ICML](https://arxiv.org/abs/1502.03044)

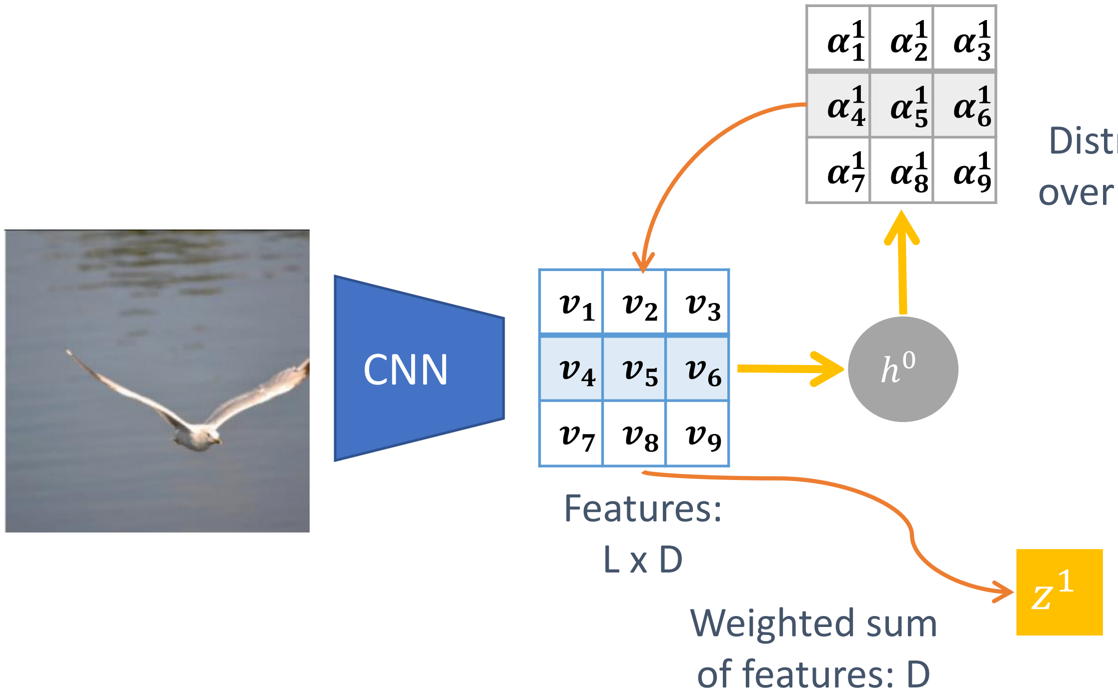
Image captioning with attention



Slide credit: Fei-Fei Li & J. Johnson & S. Young

Xu et al (2015), Show, Attend and Tell: Neural Image Caption, ICML

Image captioning with attention



Soft attention:

$$\boldsymbol{z^t = \sum_{l=1}^{L} \alpha_l^t v_l}$$

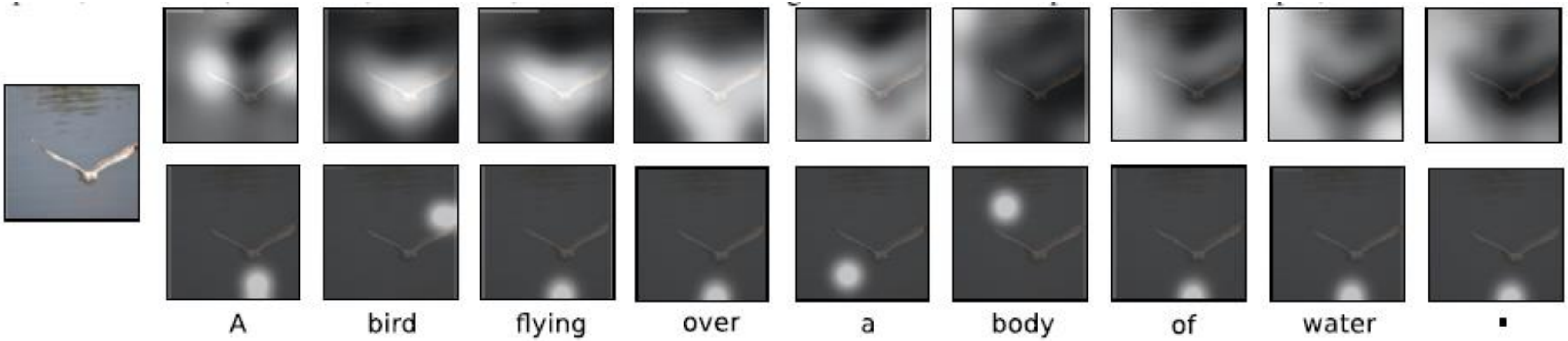
Hard attention:

$$\boldsymbol{z^t = v_{l^*}}$$

$$l^* = argmax_l \, \alpha_l$$

Xu et al (2015), Show, Attend and Tell: Neural Image Caption, ICML

Image captioning with attention



[Xu et al (2015), Show, Attend and Tell: Neural Image Caption, ICML]

Image captioning with attention

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.

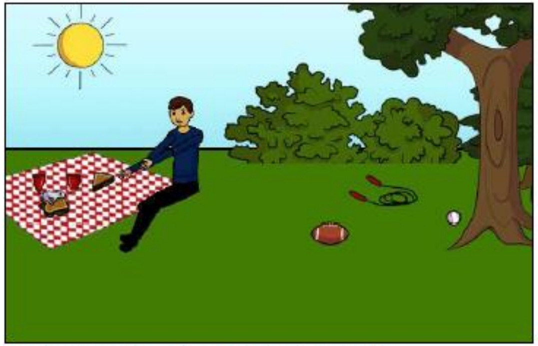
A giraffe standing in a forest with trees in the background.

Xu et al (2015), Show, Attend and Tell: Neural Image Caption, ICML

Visual question answering (VQA)

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?

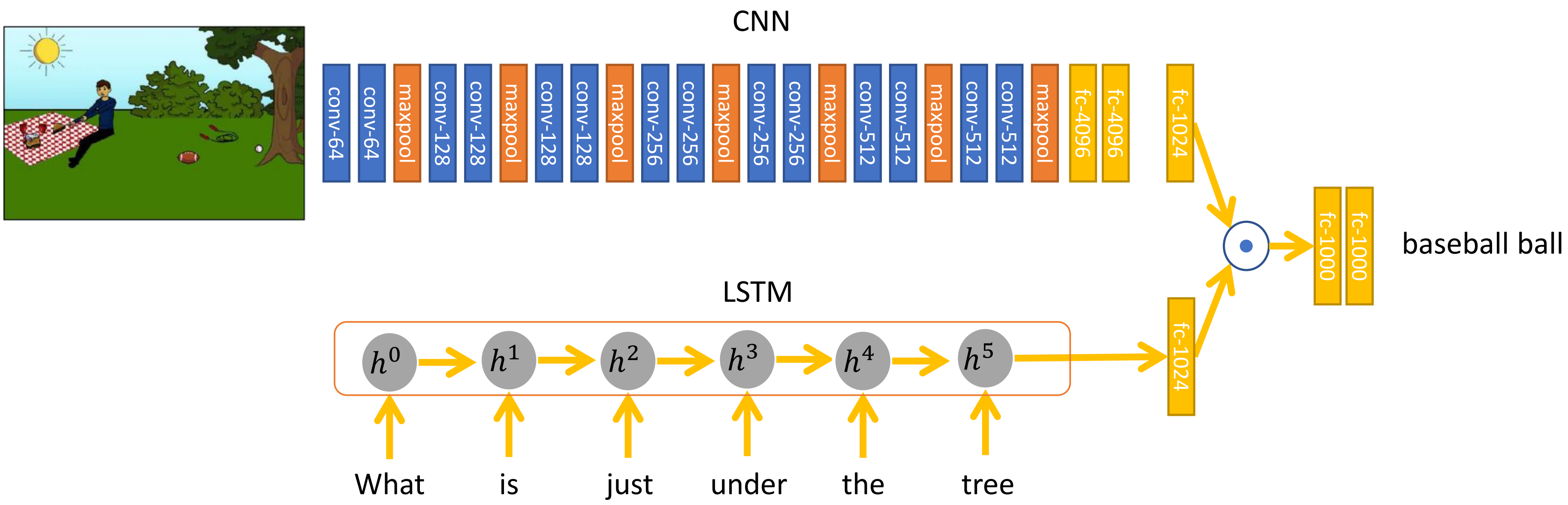
Does it appear to be rainy?
Does this person have 20/20 vision?

- Requires computer vision, natural language processing and knowledge representation & reasoning
- Open-ended answers and multiple choice answers
- Real and synthetic images
- Exact string matching

Accuracy = $\min\left(\frac{\#\,humans\ provided\ that\ answer}{3}, 1\right)$

Antol et al (2015), VQA: Visual Question Answering, ICCV

VQA



[Antol et al (2015), VQA: Visual Question Answering, ICCV](#)

Results

	Open-Answer				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
Question	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
Image	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
Q+I	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTM Q	48.76	78.20	35.68	26.59	54.75	78.22	36.82	38.78
LSTM Q+I	53.74	78.94	35.24	36.42	57.17	78.95	35.80	43.41
Caption	26.70	65.50	02.03	03.86	28.29	69.79	02.06	03.82
Q+C	54.70	75.82	40.12	42.56	59.85	75.89	41.16	52.53

Question Type	Open-Answer K = 1000 Q	Q + I	Q + C	Human Q	Q + I	Human Age To Be Able To Answer
what is (13.84)	23.57	34.28	43.88	16.86	73.68	09.07
what color (08.98)	33.37	43.53	48.61	28.71	86.06	06.60
what kind (02.49)	27.78	42.72	43.88	19.10	70.11	10.55
what are (02.32)	25.47	39.10	47.27	17.72	69.49	09.03
what type (01.78)	27.68	42.62	44.32	19.53	70.65	11.04
is the (10.16)	70.76	69.87	70.50	65.24	95.67	08.51
is this (08.26)	70.34	70.79	71.54	63.35	95.43	10.13
how many (10.28)	43.78	40.33	47.52	30.45	86.32	07.67
are (07.57)	73.96	73.58	72.43	67.10	95.24	08.65
does (02.75)	76.81	75.81	75.88	69.96	95.70	09.29
where (02.90)	16.21	23.49	29.47	11.09	43.56	09.54
is there (03.60)	86.50	86.37	85.88	72.48	96.43	08.25
why (01.20)	16.24	13.94	14.54	11.80	21.50	11.18
which (01.21)	29.50	34.83	40.84	25.64	67.44	09.27
do (01.15)	77.73	79.31	74.63	71.33	95.44	09.23
what does (01.12)	19.58	20.00	23.19	11.12	75.88	10.02
what time (00.67)	8.35	14.00	18.28	07.64	58.98	09.81
who (00.77)	19.75	20.43	27.28	14.69	56.93	09.49
what sport (00.81)	37.96	81.12	93.87	17.86	95.59	08.07
what animal (00.53)	23.12	59.70	71.02	17.67	92.51	06.75
what brand (00.36)	40.13	36.84	32.19	25.34	80.95	12.50

Summary

Image captioning & visual question answering

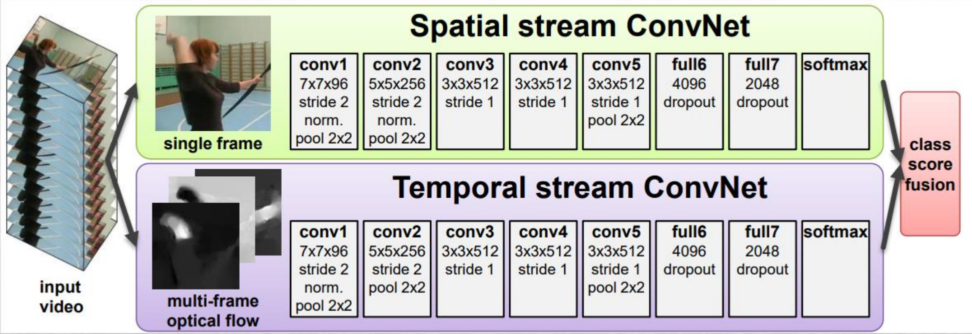
- Combination of computer vision, natural language processing and knowledge representation & reasoning
- Evaluation metrics

Recommended reading

- Vinyals et al (2015), Show and Tell: A Neural Image Caption Generator, ICCV

Additional reading

- Antol et al (2015), VQA: Visual Question Answering, ICCV



Next lecture