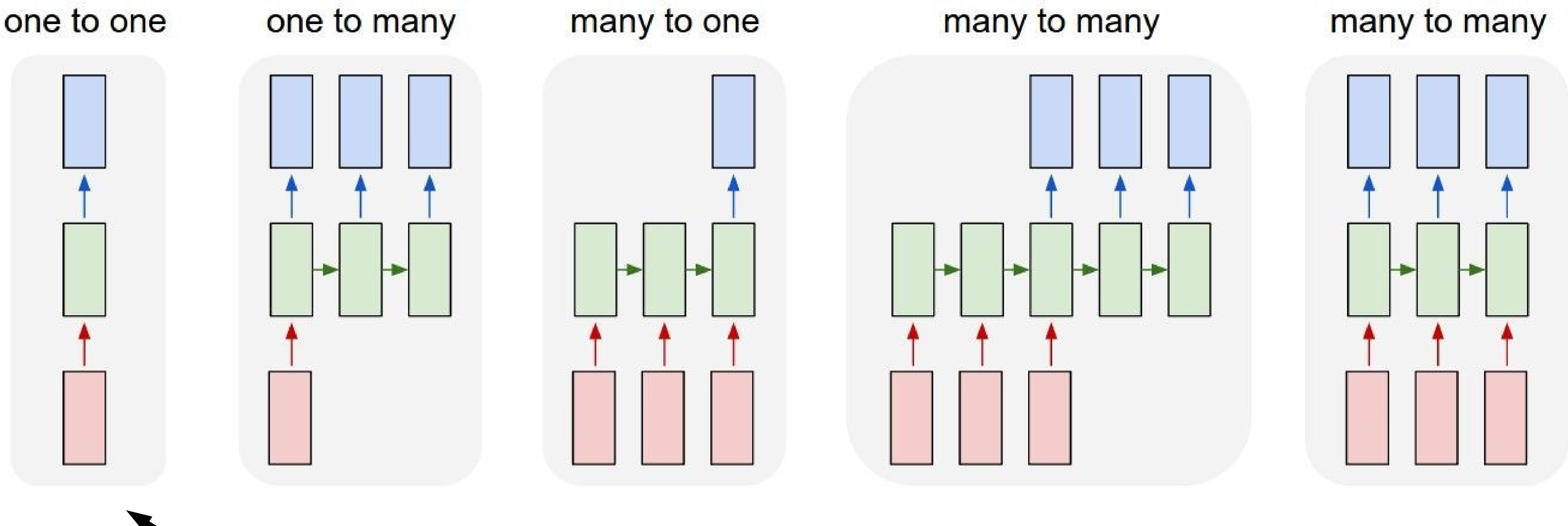
Recurrent Neural Networks
+
Multimodal Deep Learning (Vision+Language)

Jamie Ryan Kiros
University of Toronto

Recurrent neural networks

(images from Andrej Karpathy)



Feed-forward neural network

Recurrent neural networks

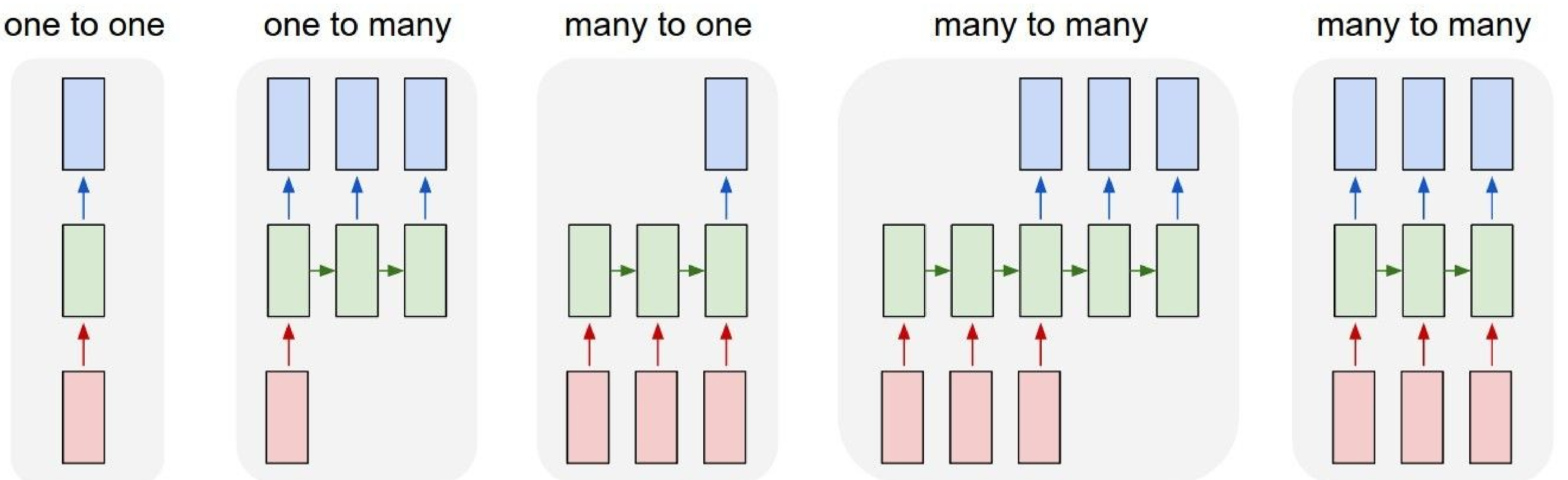
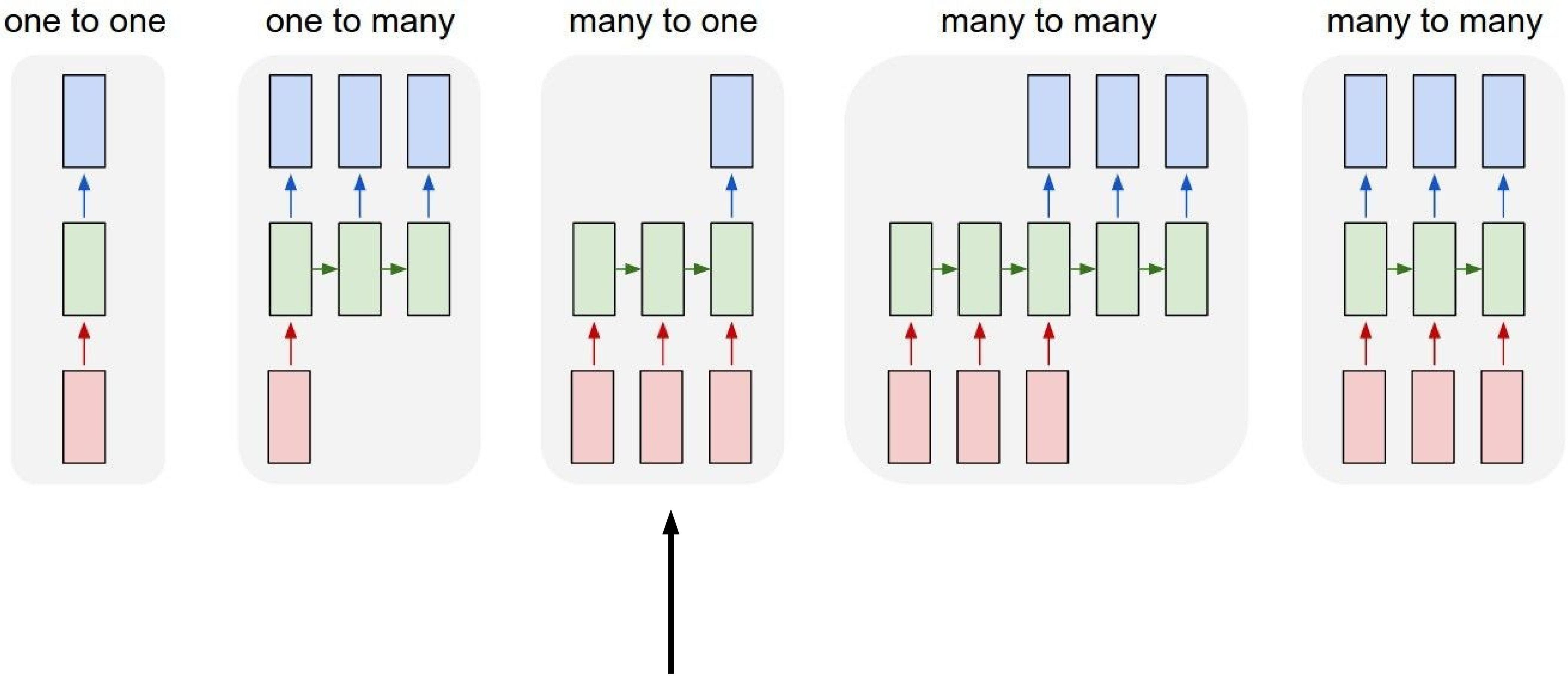


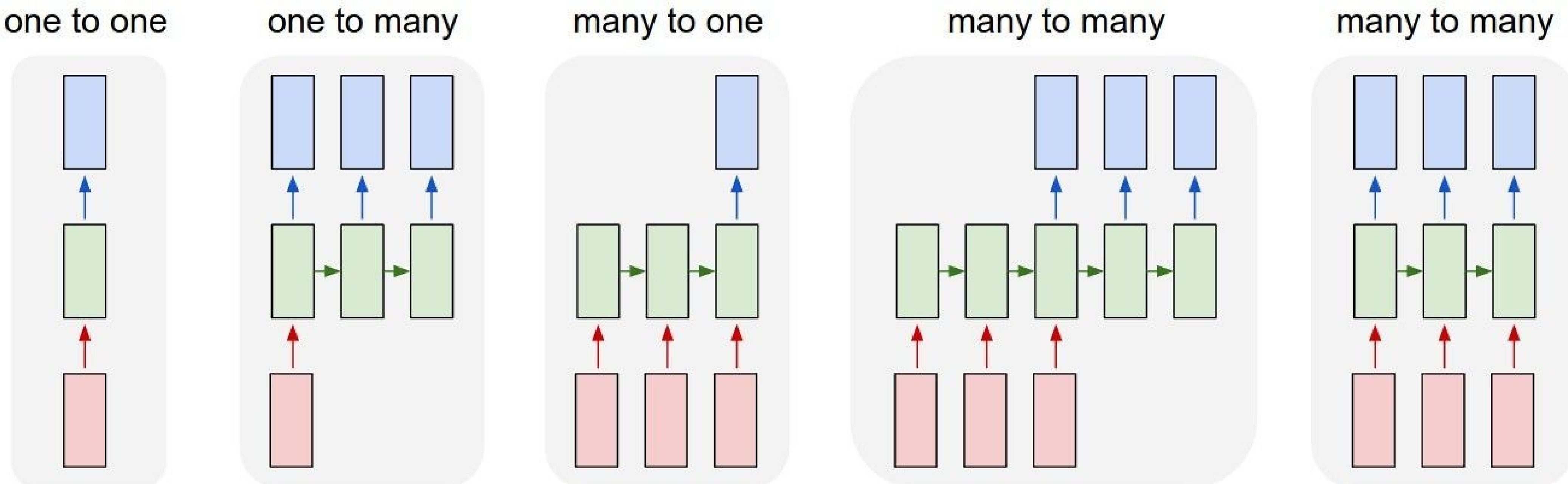
Image-captioning
(image -> sequence of words)

Recurrent neural networks



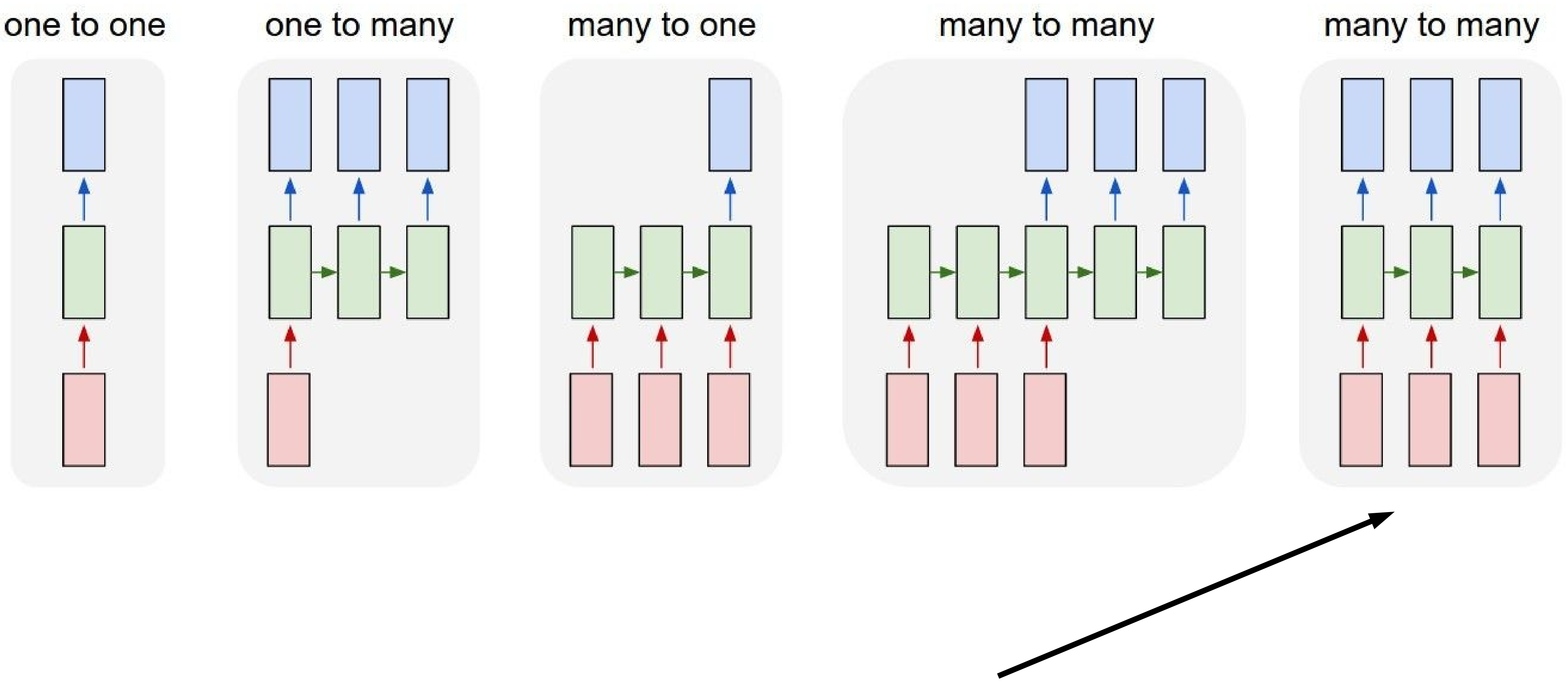
e.g. Sentiment classification
sequence of words ->sentiment

Recurrent neural networks



e.g. Machine Translation
sequence of words -> sequence of words

Recurrent neural networks



e.g. Language modelling
Video frame classification

Recurrent Neural Network

We can process a sequence of vectors **x** by applying a recurrence formula at every time step:

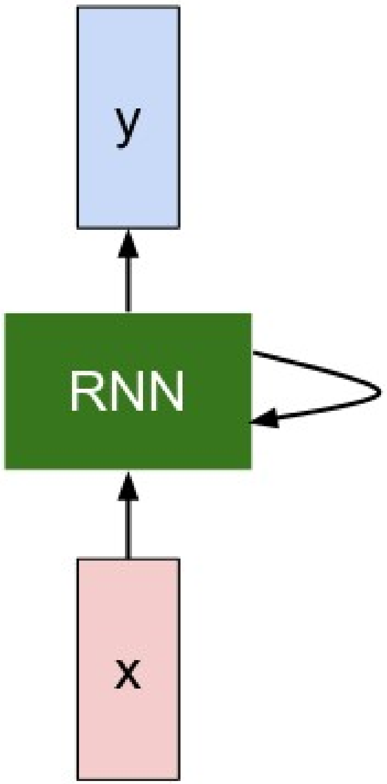
$$\boxed{h_t} = \boxed{f_W}(\boxed{h_{t-1}}, \boxed{x_t})$$

new state

old state

input vector at some time step

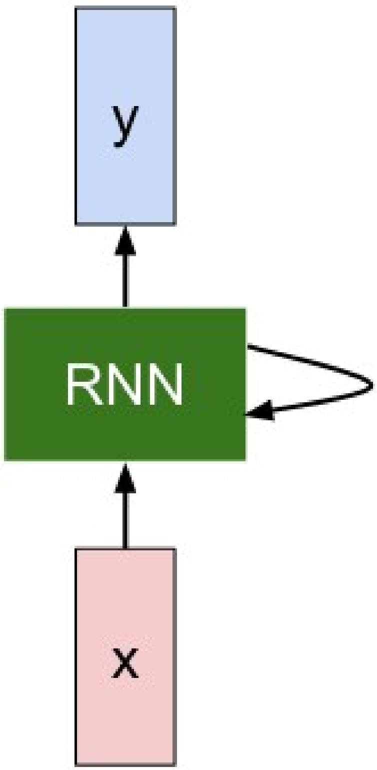
some function with parameters W



(images from Andrej Karpathy)

(Vanilla) Recurrent Neural Network

The state consists of a single *"hidden"* vector **h**:



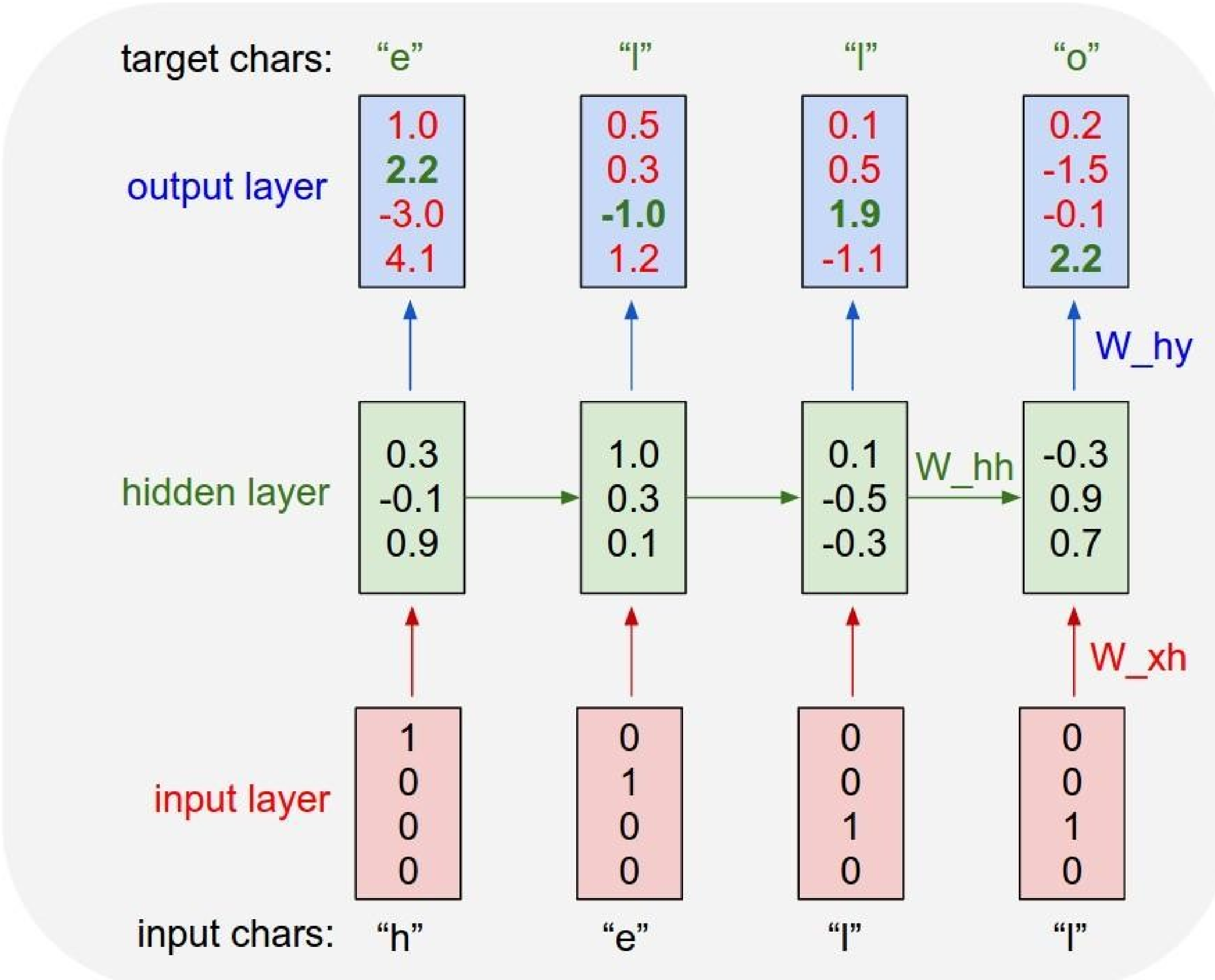
$$h_t = f_W(h_{t-1}, x_t)$$

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

$$y_t = W_{hy}h_t$$

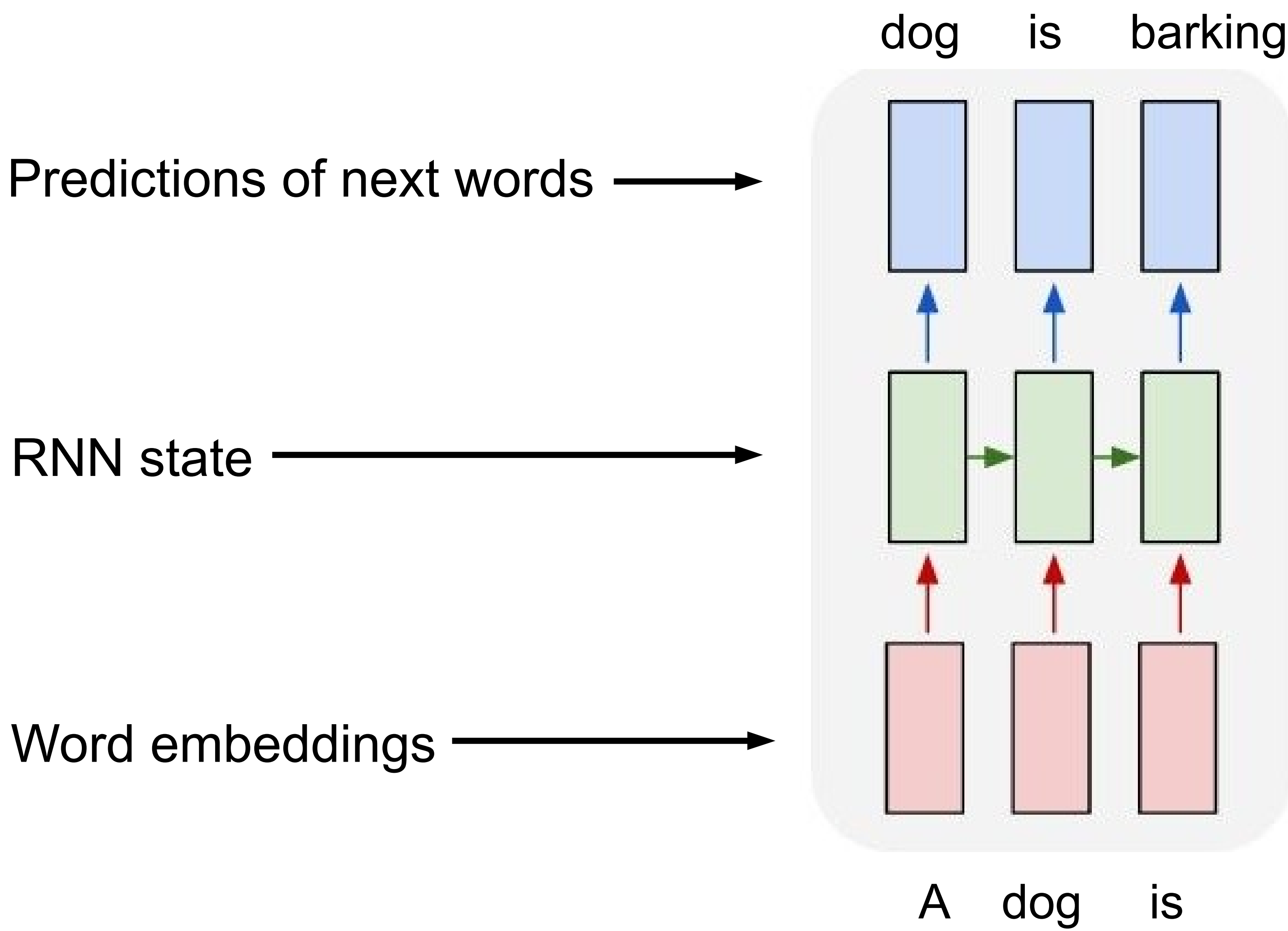
(images from Andrej Karpathy)

Example: character-level language models

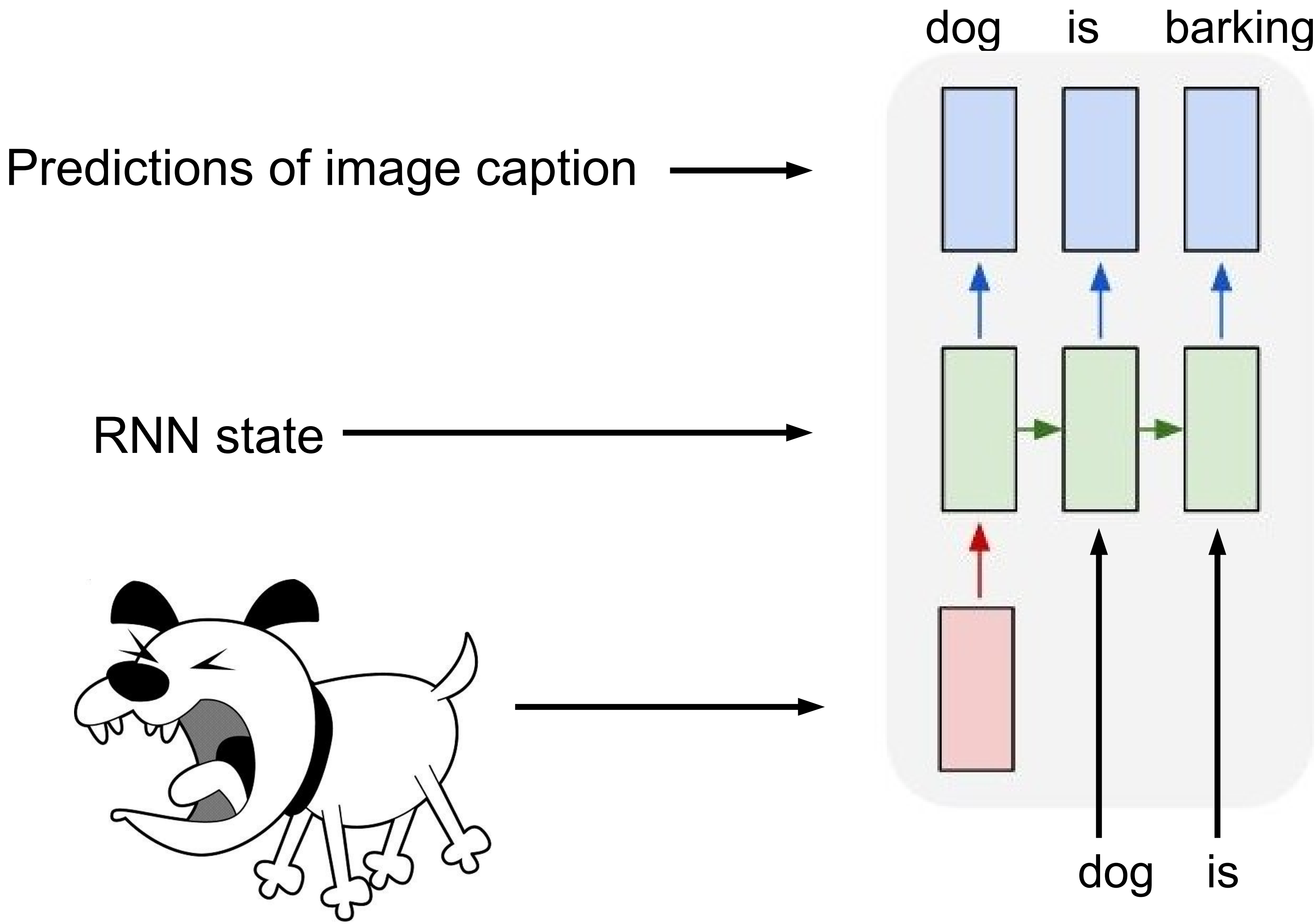


(images from Andrej Karpathy)

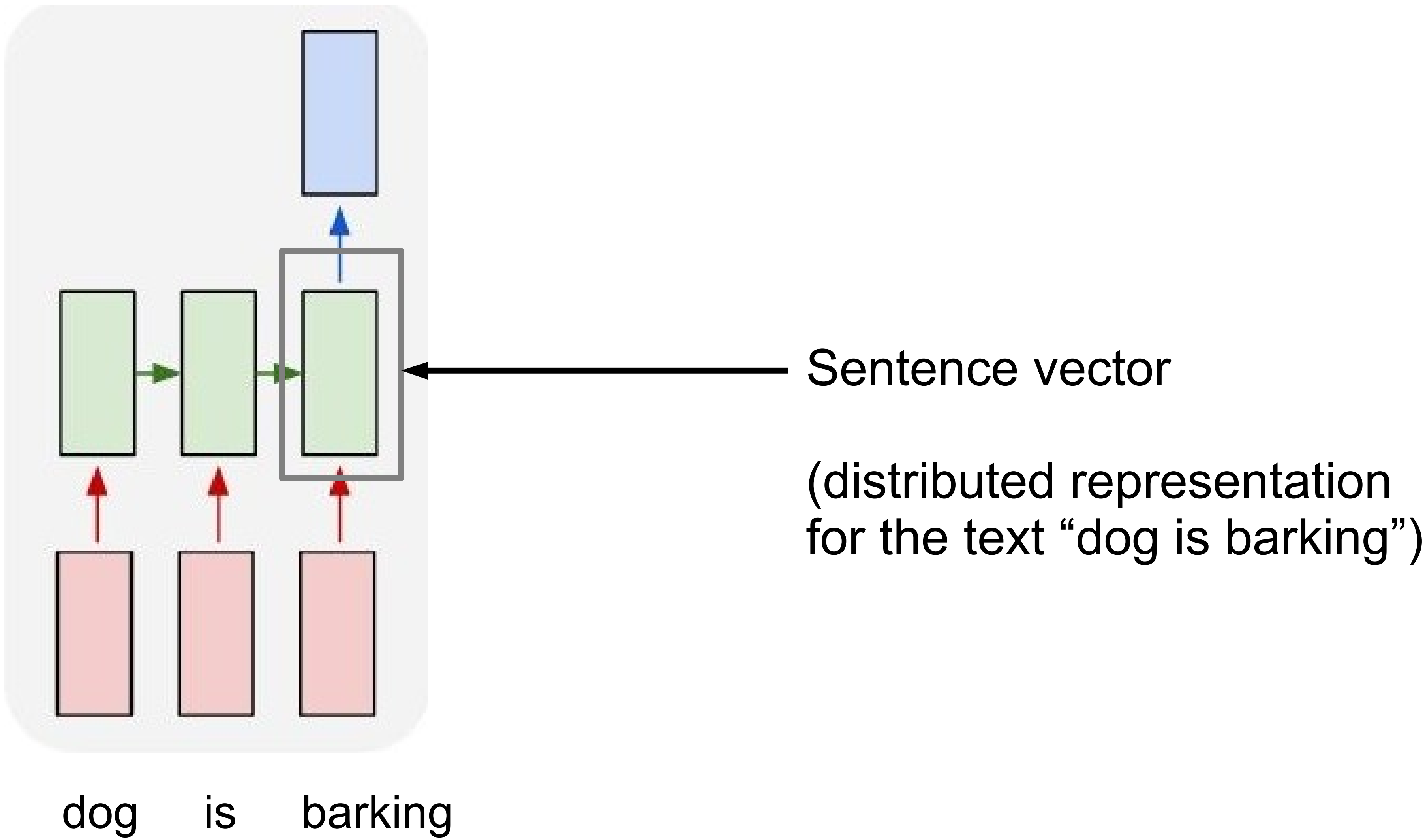
RNNs for language: language models



RNNs for language: decoders (conditional language models)

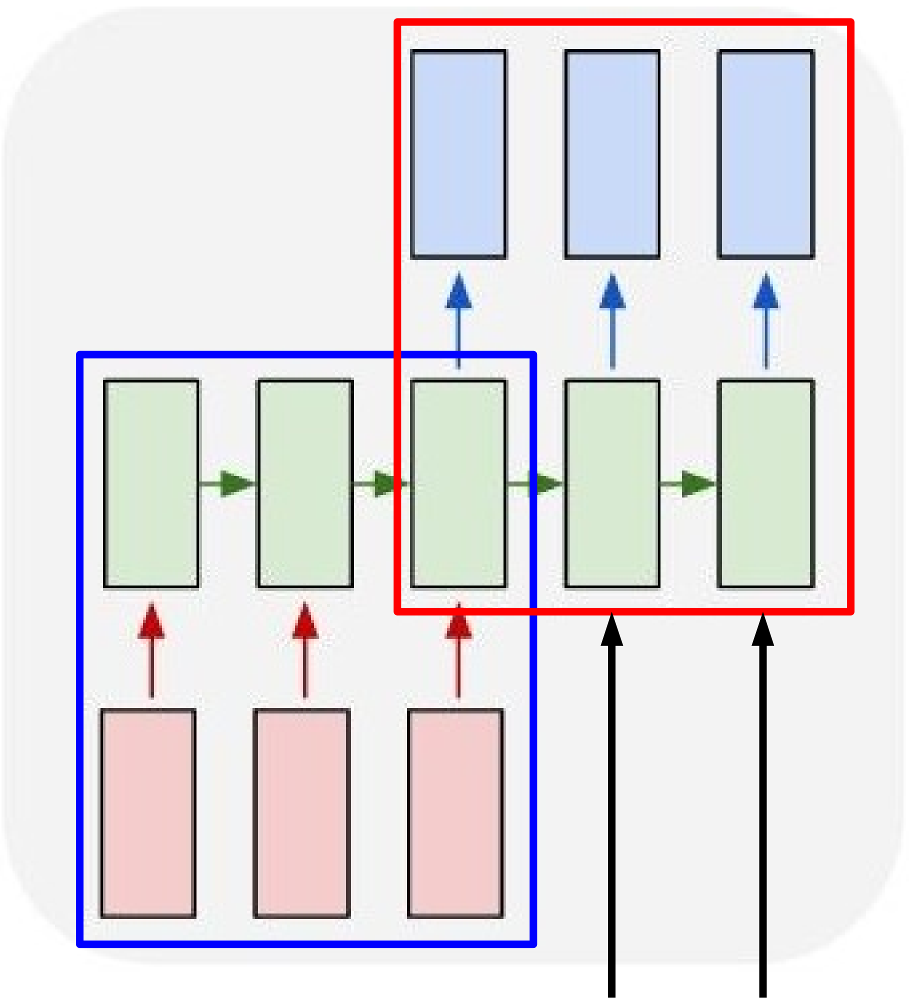


RNNs for language: encoders



RNNs for language: encoder-decoders

Decoder for french sentence



Encoder for english sentence

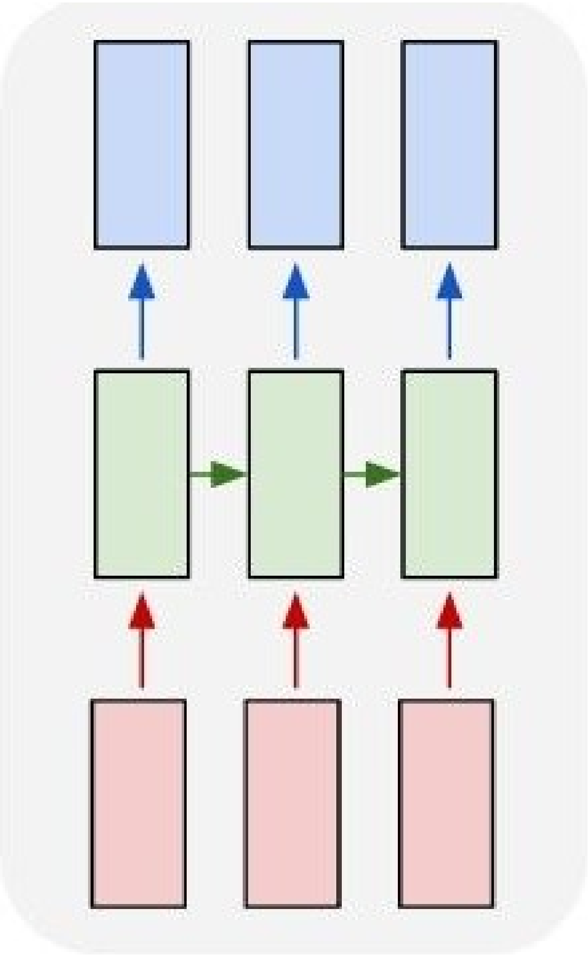
Problems with vanilla RNNs

Vanishing gradient problem

-> use LSTM

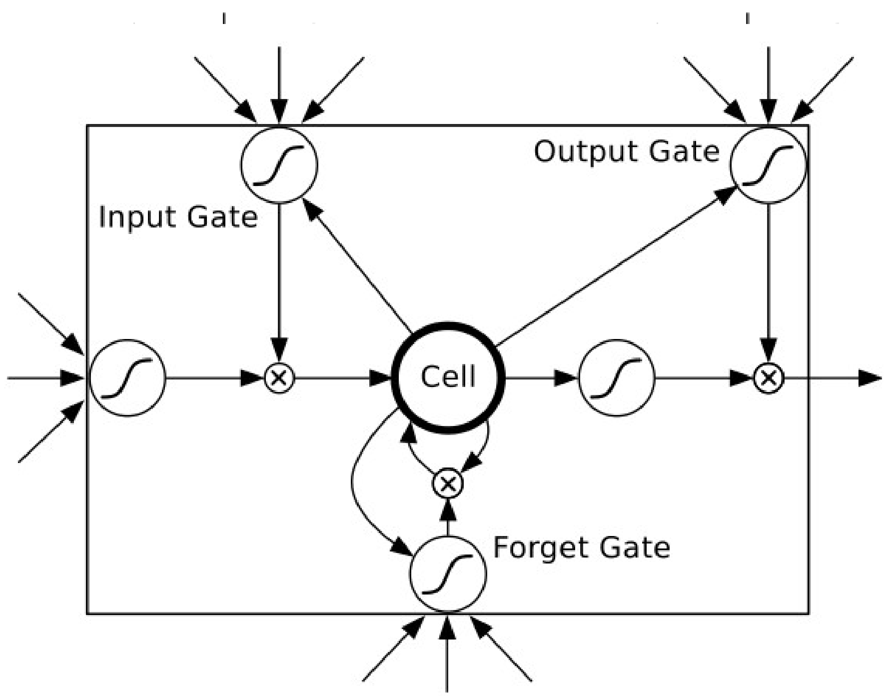
Exploding gradient problem

-> use gradient clipping



Long short-term memory (LSTM)
(slide from Alex Graves)

- LSTM is an RNN architecture designed to have a better memory. It uses linear memory cells surrounded by multiplicative gate units to store read, write and reset information



Input gate: scales input to cell (write)

Output gate: scales output from cell (read)

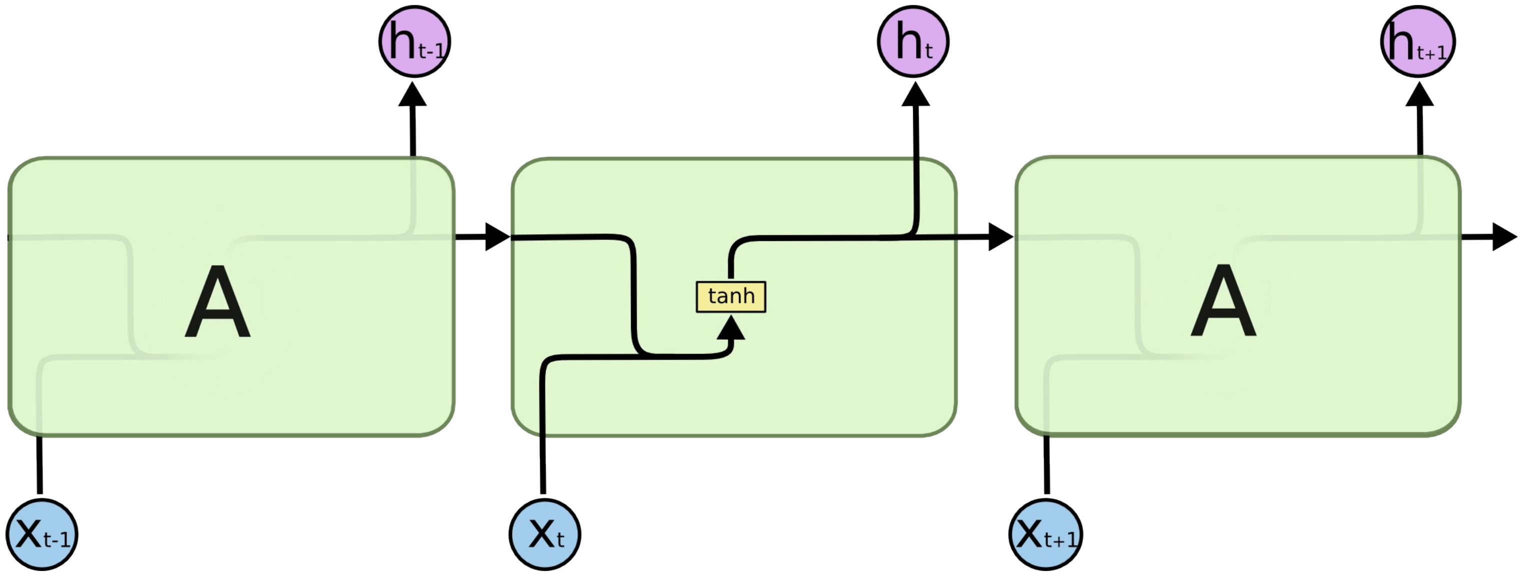
Forget gate: scales old cell value (reset)

- S. Hochreiter and J. Schmidhuber, "*Long Short-term Memory*" Neural Computation 1997

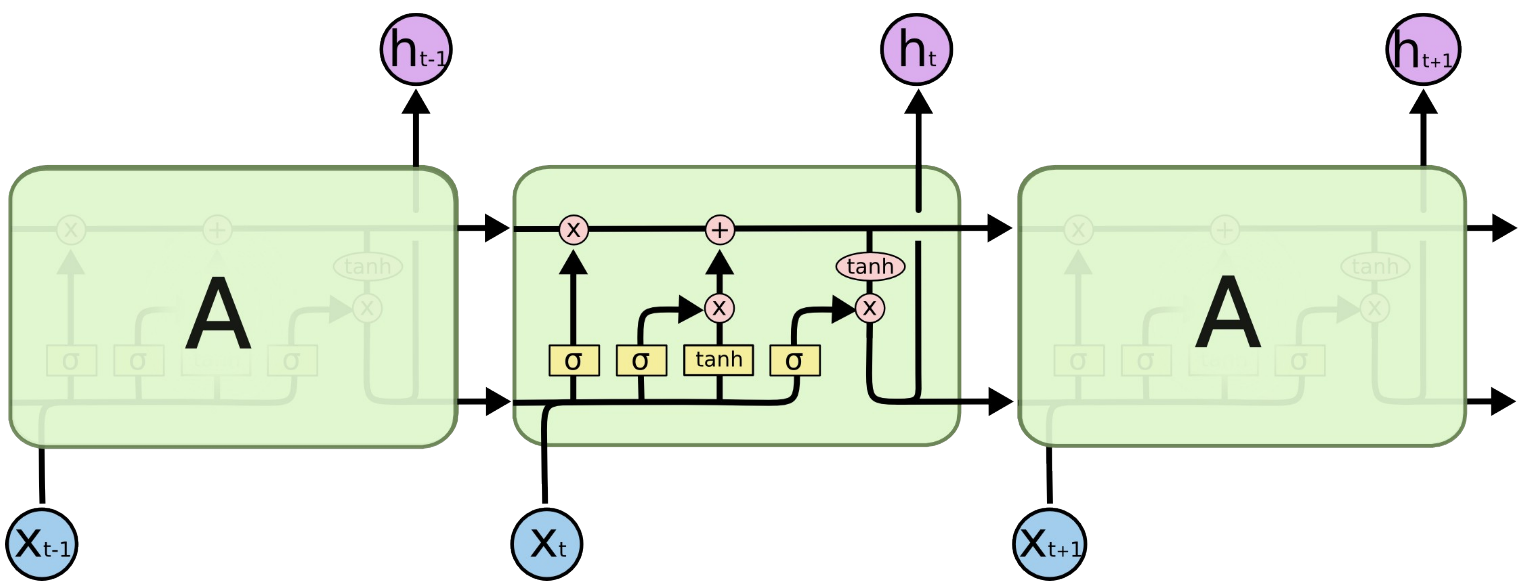
LSTM vs simple RNN

(images from Chris Olah)

Simple RNN



LSTM



Some successes of the LSTM from 2013-2014 (many more since then!) (list from Schmidhuber)

1. Text-to-speech synthesis (Fan et al., Microsoft, Interspeech 2014)
2. Language identification (Gonzalez-Dominguez et al., Google, Interspeech 2014)
3. Large vocabulary speech recognition (Sak et al., Google, Interspeech 2014)
4. Prosody contour prediction (Fernandez et al., IBM, Interspeech 2014)
5. Medium vocabulary speech recognition (Geiger et al., Interspeech 2014)
6. English to French translation (Sutskever et al., Google, NIPS 2014)
7. Audio onset detection (Marchi et al., ICASSP 2014)
8. Social signal classification (Brueckner & Schulter, ICASSP 2014)
9. Arabic handwriting recognition (Bluche et al., DAS 2014)
10. TIMIT phoneme recognition (Graves et al., ICASSP 2013)
11. Optical character recognition (Breuel et al., ICDAR 2013)
12. Image caption generation (Vinyals et al., Google, 2014)
13. Video to textual description (Donahue et al., 2014)
14. Syntactic parsing for Natural Language Processing (Vinyals et al., Google, 2014)
15. Photo-real talking heads (Soong and Wang, Microsoft, 2014).

Applications to Multimodal tasks (Language+Vision)

(#1): Multimodal image-sentence embeddings

(#2): Image caption generation

(#3): Skip-thought vectors

(#4): Aligning books and movies

(#5): Style analogies + Neural storyteller

Applications to Multimodal tasks (Language+Vision)

(#1): Multimodal image-sentence embeddings

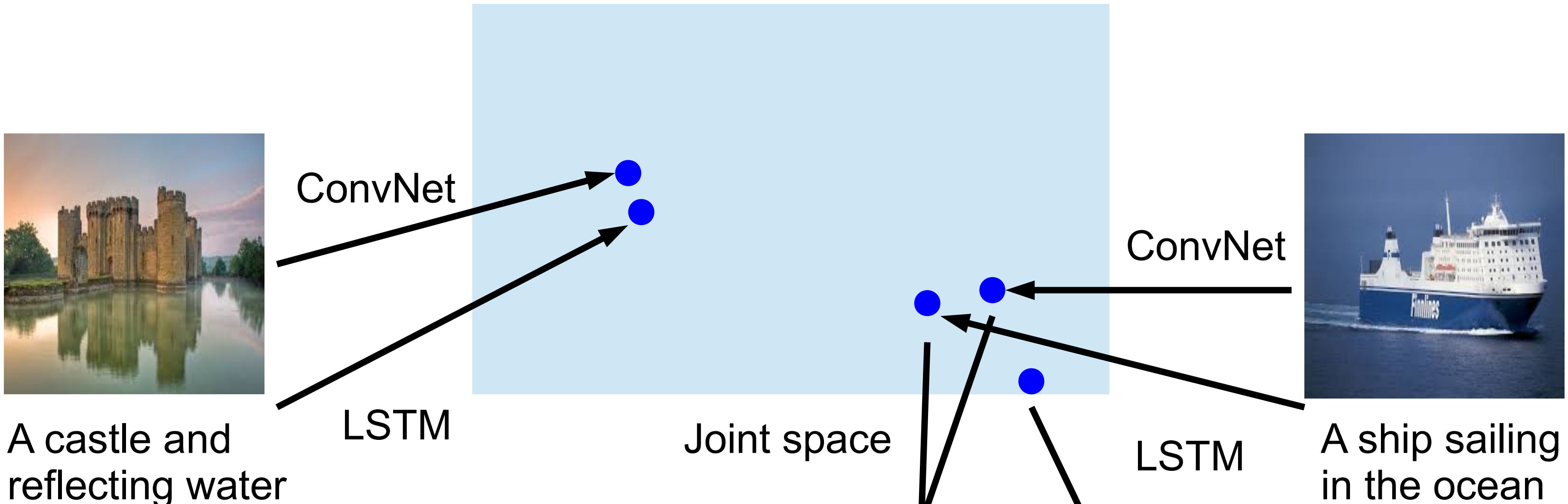
(#2): Image caption generation

(#3): Skip-thought vectors

(#4): Aligning books and movies

(#5): Style analogies + Neural storyteller

A joint image-text embedding (ConvNet - LSTM)



Minimize the following objective:

$$\sum_{\mathbf{x}} \sum_{k} \max\{0, \alpha - s(\mathbf{x}, \mathbf{v}) + s(\mathbf{x}, \mathbf{v}_k)\} +$$

images → $\mathbf{x}$

$$\sum_{\mathbf{v}} \sum_{k} \max\{0, \alpha - s(\mathbf{v}, \mathbf{x}) + s(\mathbf{v}, \mathbf{x}_k)\}$$

text → $\mathbf{v}$

Train globally, retrieve locally

tower, building, cathedral, dome, castle

bowl, cup, soup, cups, coffee

kitchen, stove, oven, refrigerator, microwave

ski, skiing, skiers, skiiers, snowmobile

beach

snow

Adjectives

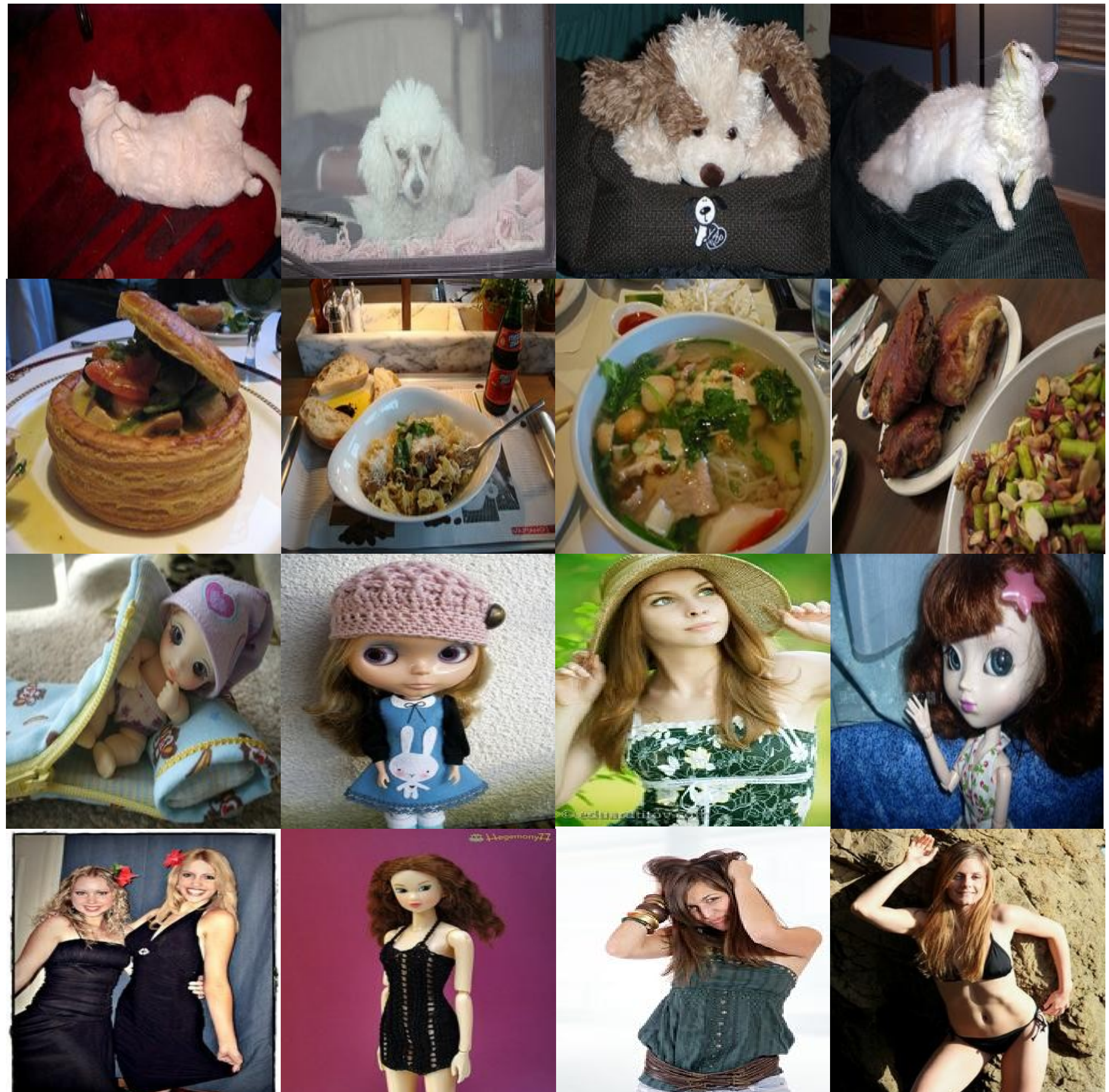
Nearest images

fluffy

delicious

adorable

sexy



Good retrieval results (sentences)

The dogs are in the snow in front of a fence .

Four men playing basketball , two from each team .

A boy skateboarding

Two men and a woman smile at the camera .

Women participate in a skit onstage .

A man is doing tricks on a bicycle on ramps in front of a crowd .

Not so good retrieval results

(these have ground truth ranked > 100)

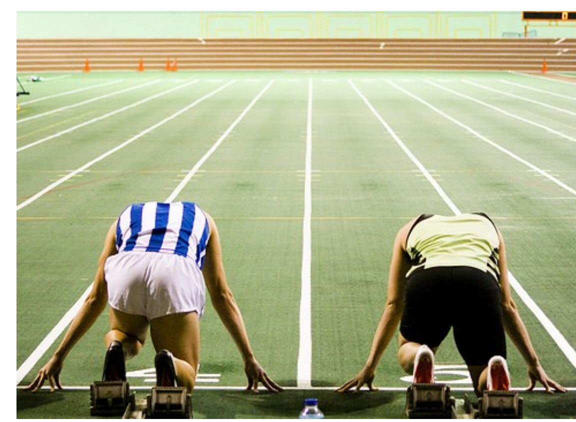
two people wearing white shirts and jeans each carrying a skateboard

a dog jumps over a bar with a ball in its mouth .

White medium sized dog is running through the ocean .

A lady holds a little boy While another little boy smiles at them .

A man and a woman walking down a street , carrying luggage .

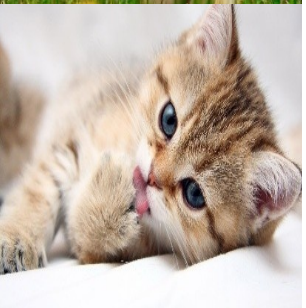


Woman in white dribbling basketball .

Multimodal linguistic regularities

Nearest images

- dog + cat =



- cat + dog =

- plane + bird =

- man + woman =

colours

Nearest images

- blue + red =

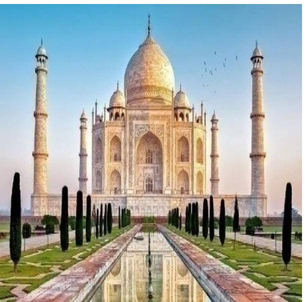
- blue + yellow =

- yellow + red =

- white + red =

Some interesting examples

Nearest images



- day + night =

- flying + sailing =

- bowl + box =

- box + bowl =

Sanity check

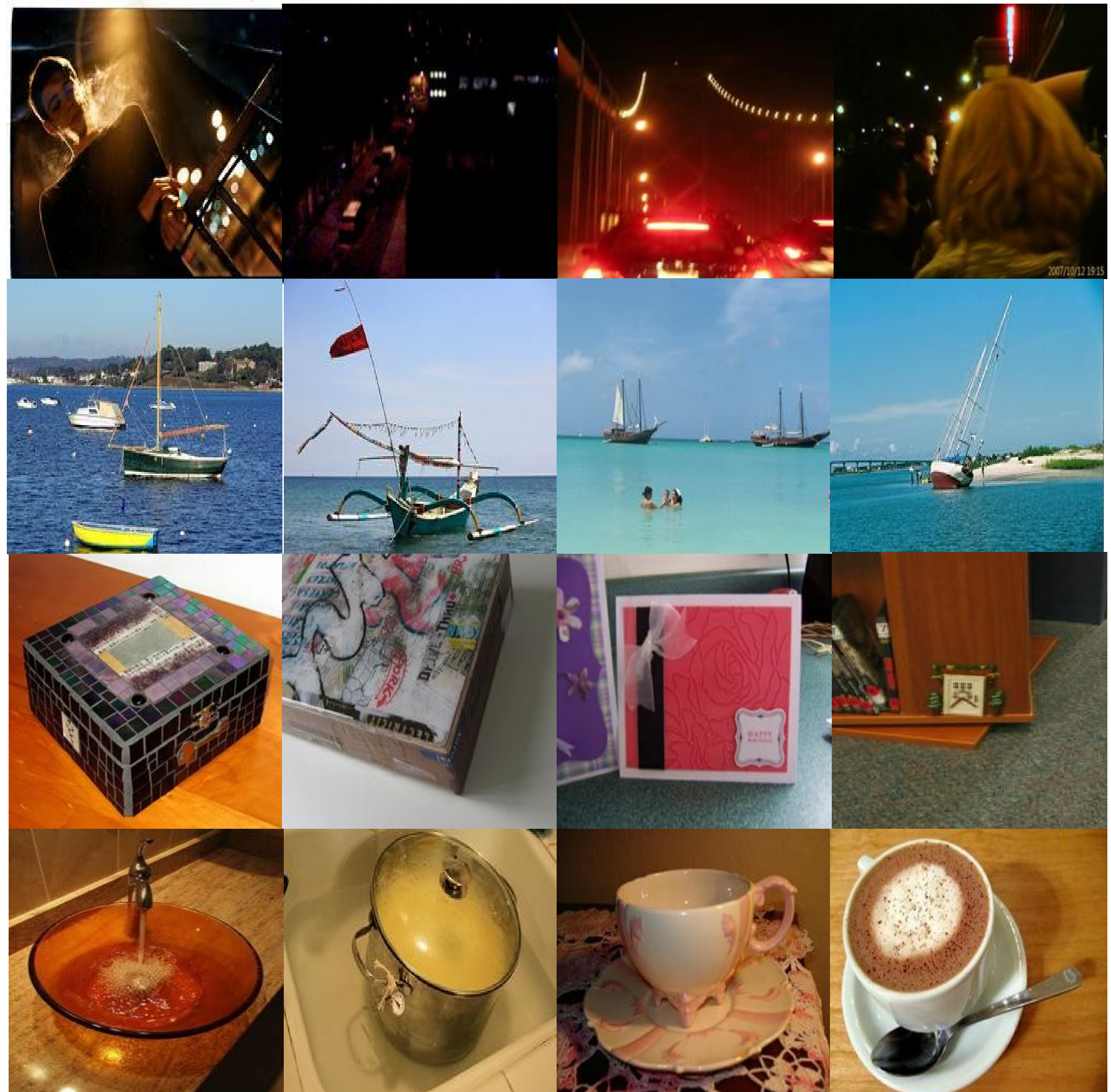
Nearest images

night

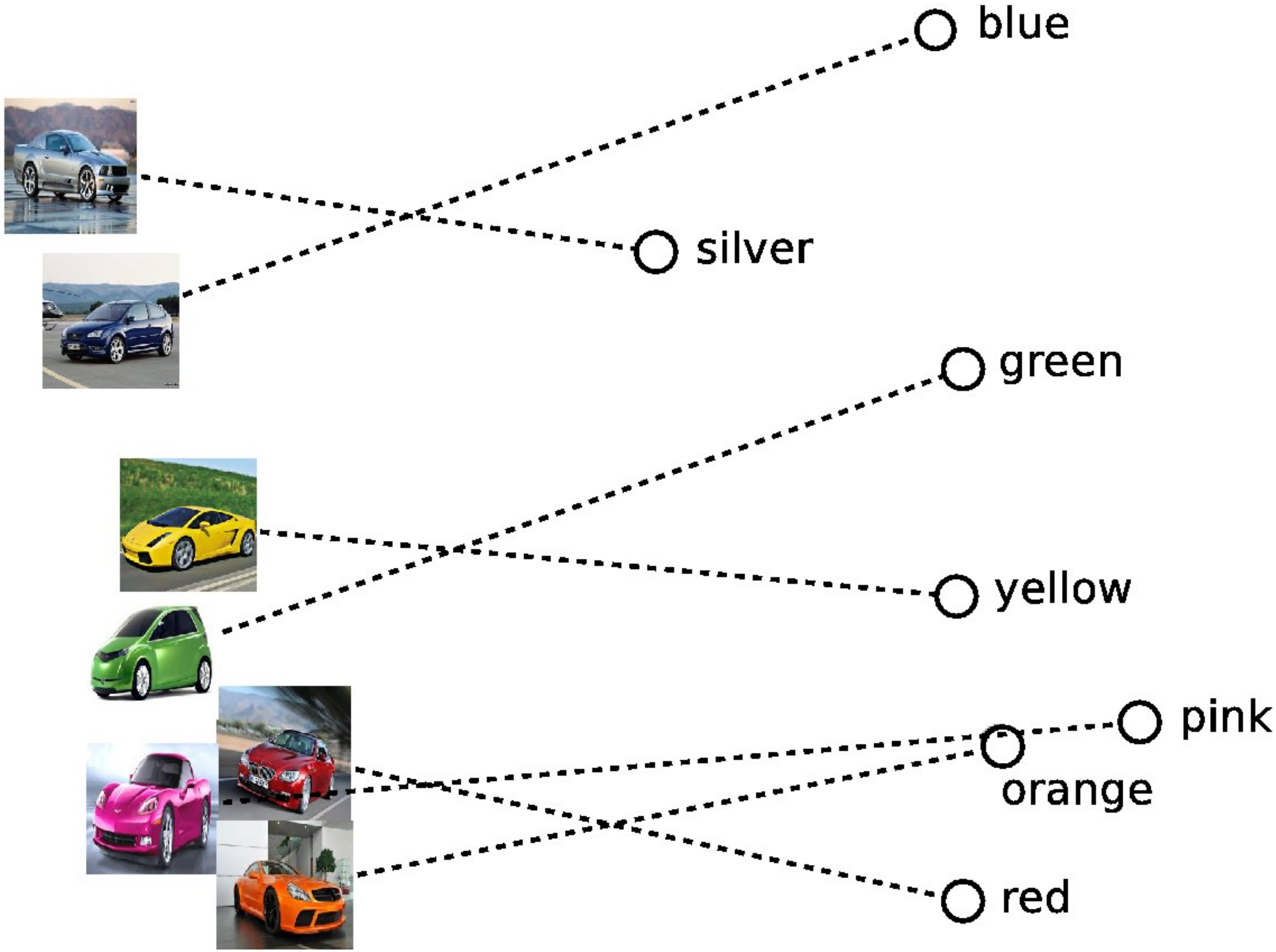
sailing

box

bowl



PCA embedding



Why does this work?

$\mathbf{v}_{car}$ $\mathbf{v}_{red}$ $\mathbf{v}_{blue}$: word vectors for 'car', 'red', 'blue'

$\mathbf{I}_{bcar}$ $\mathbf{I}_{rcar}$: embeddings of a blue car and a red car

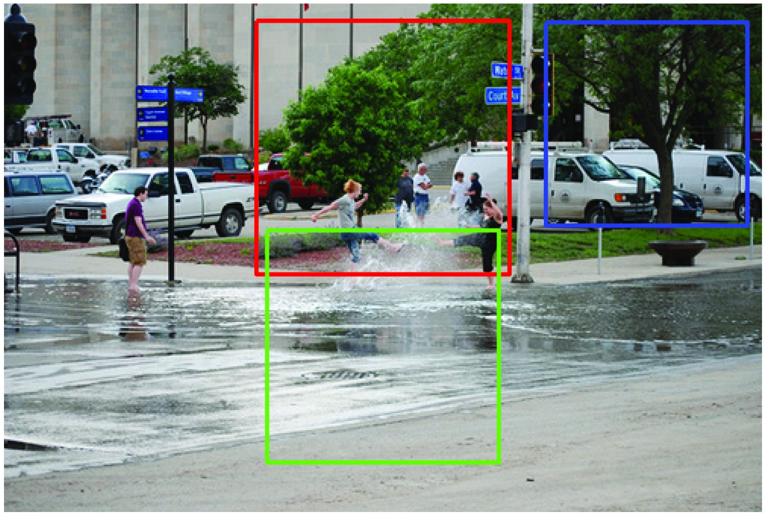
After training a linear encoder, the model has the property that:

$$\mathbf{v}_{blue} + \mathbf{v}_{car} \approx \mathbf{I}_{bcar} \text{ and } \mathbf{v}_{red} + \mathbf{v}_{car} \approx \mathbf{I}_{rcar}$$

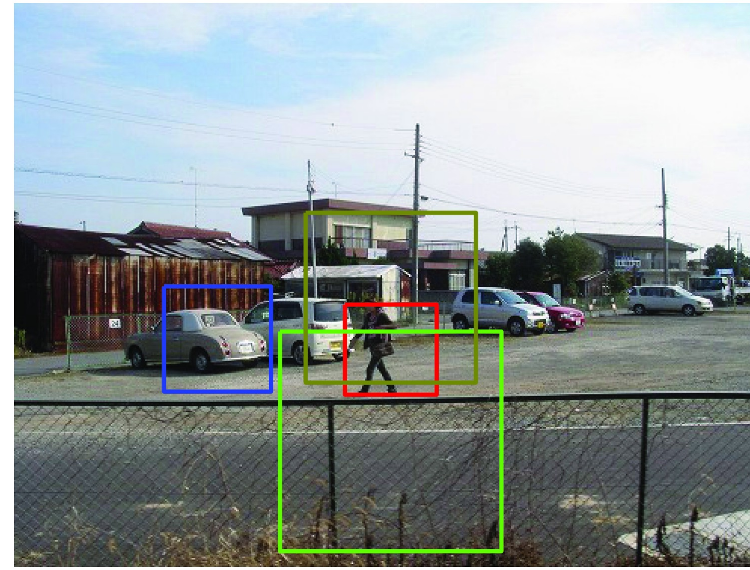
It follows that:

$$\begin{aligned}
\mathbf{v}_{car} &\approx \mathbf{I}_{bcar} - \mathbf{v}_{blue} && (1) \\
\mathbf{v}_{red} + \mathbf{v}_{car} &\approx \mathbf{I}_{bcar} - \mathbf{v}_{blue} + \mathbf{v}_{red} && (2) \\
\mathbf{I}_{rcar} &\approx \mathbf{I}_{bcar} - \mathbf{v}_{blue} + \mathbf{v}_{red} && (3)
\end{aligned}$$

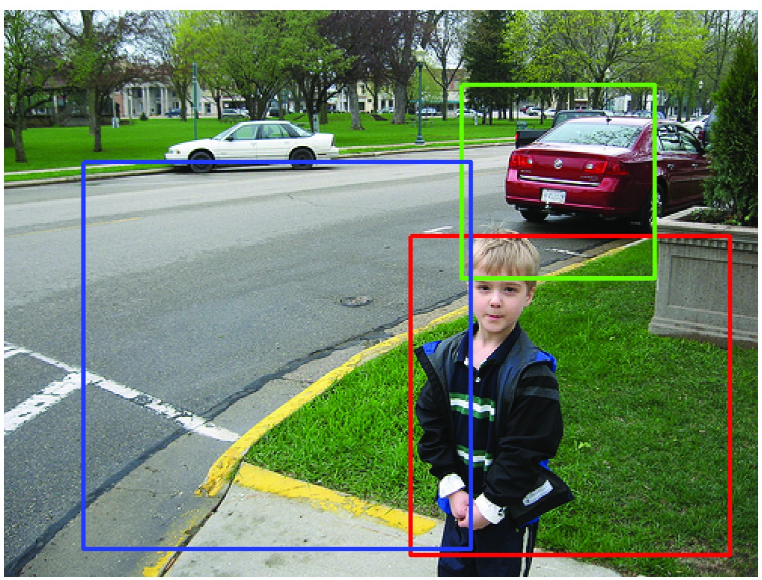
Image-text alignments from scratch



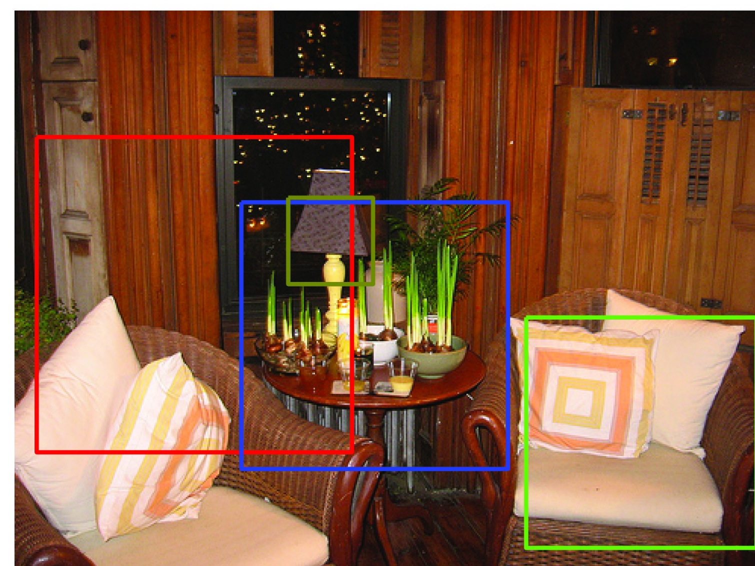
People, water, truck



Woman, fence, cars, building

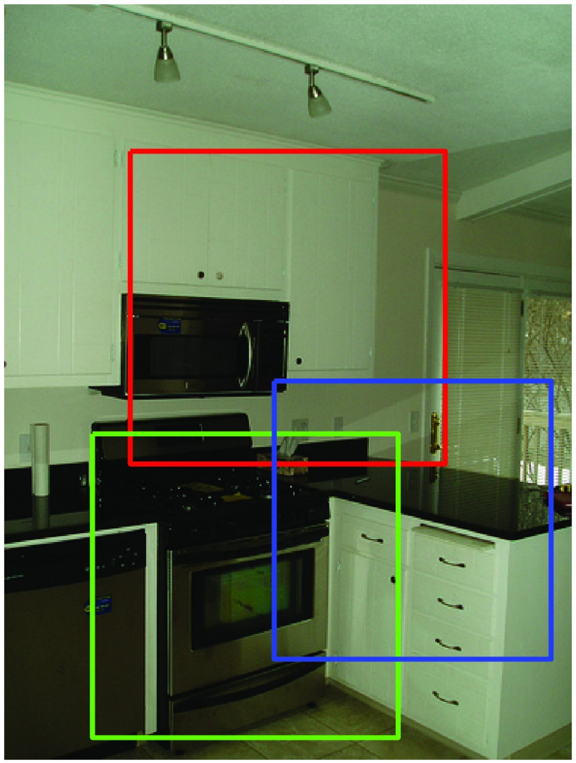


Boy, car, road

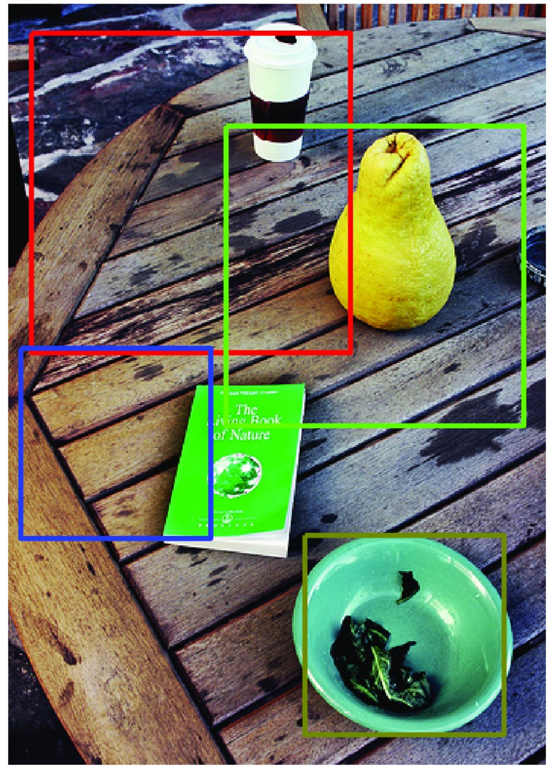


Chair, pillow, table, lamp

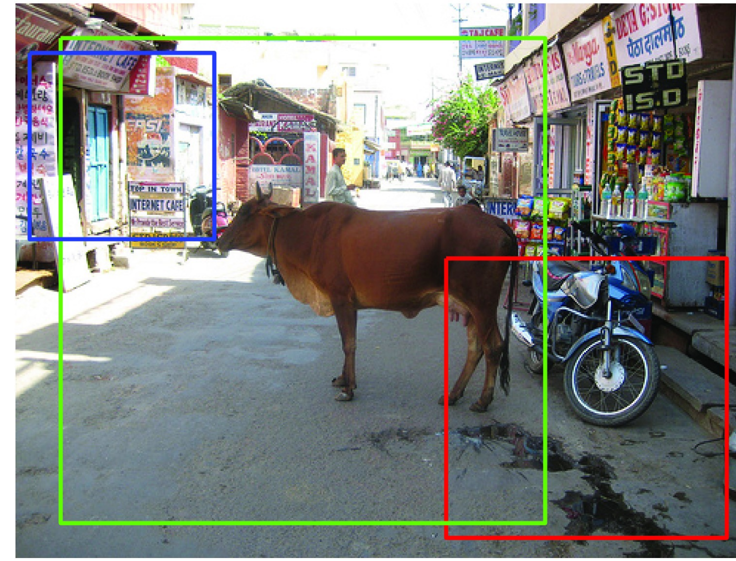
Image-text alignments



Oven, microwave, counter



Cup, pear, book, bowl

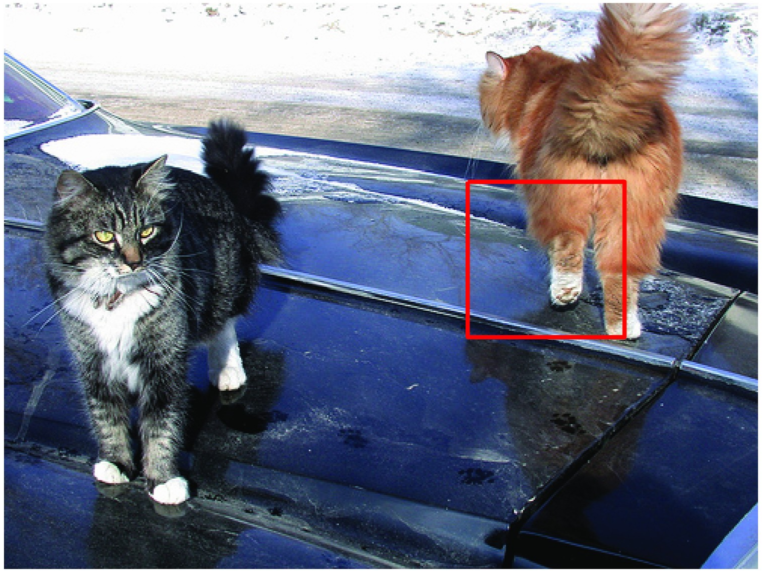


Motorcycle, cow, shop

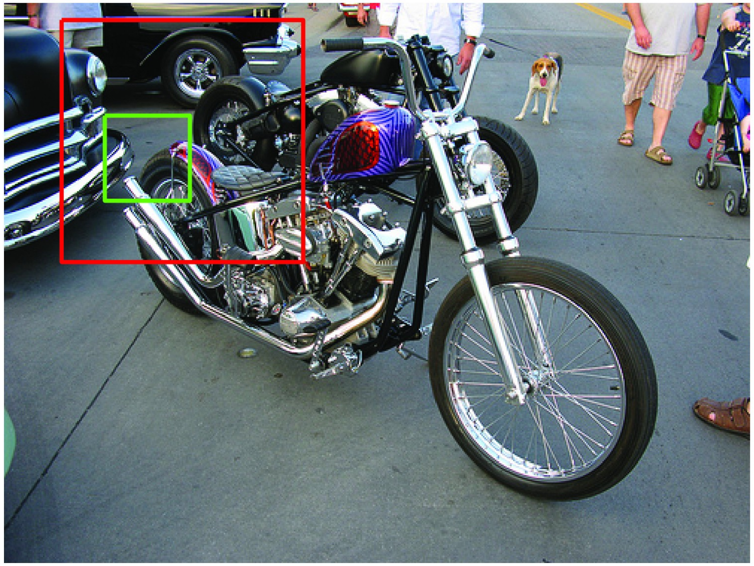
Screen, clock, window, shelf

adjectives

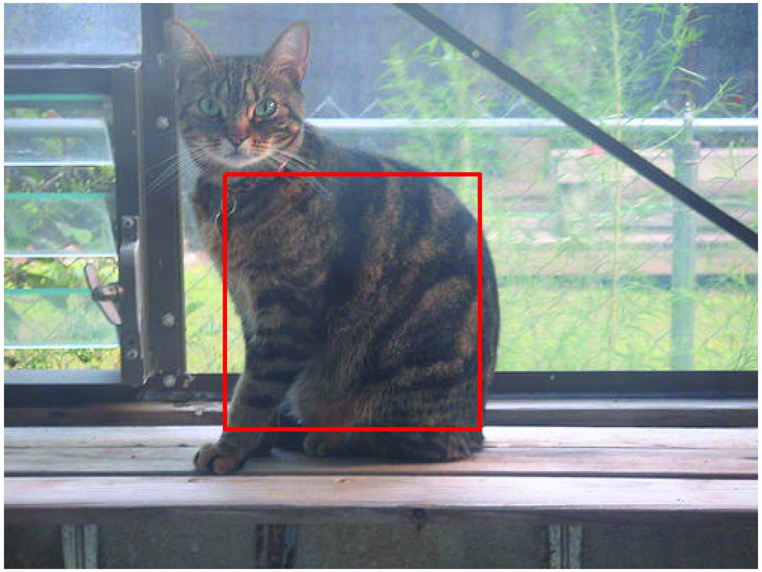
Delicious



fluffy

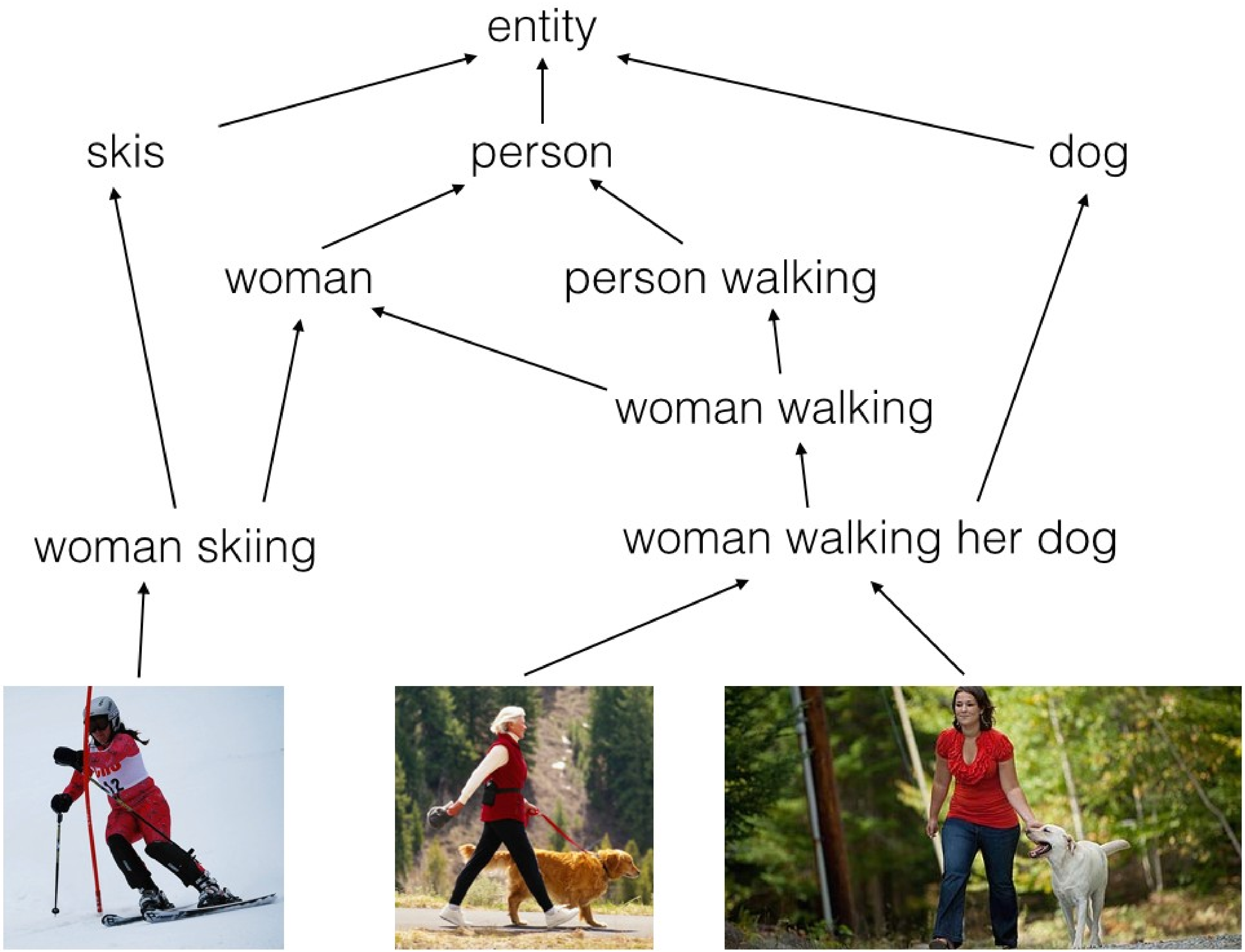


Shiny, round

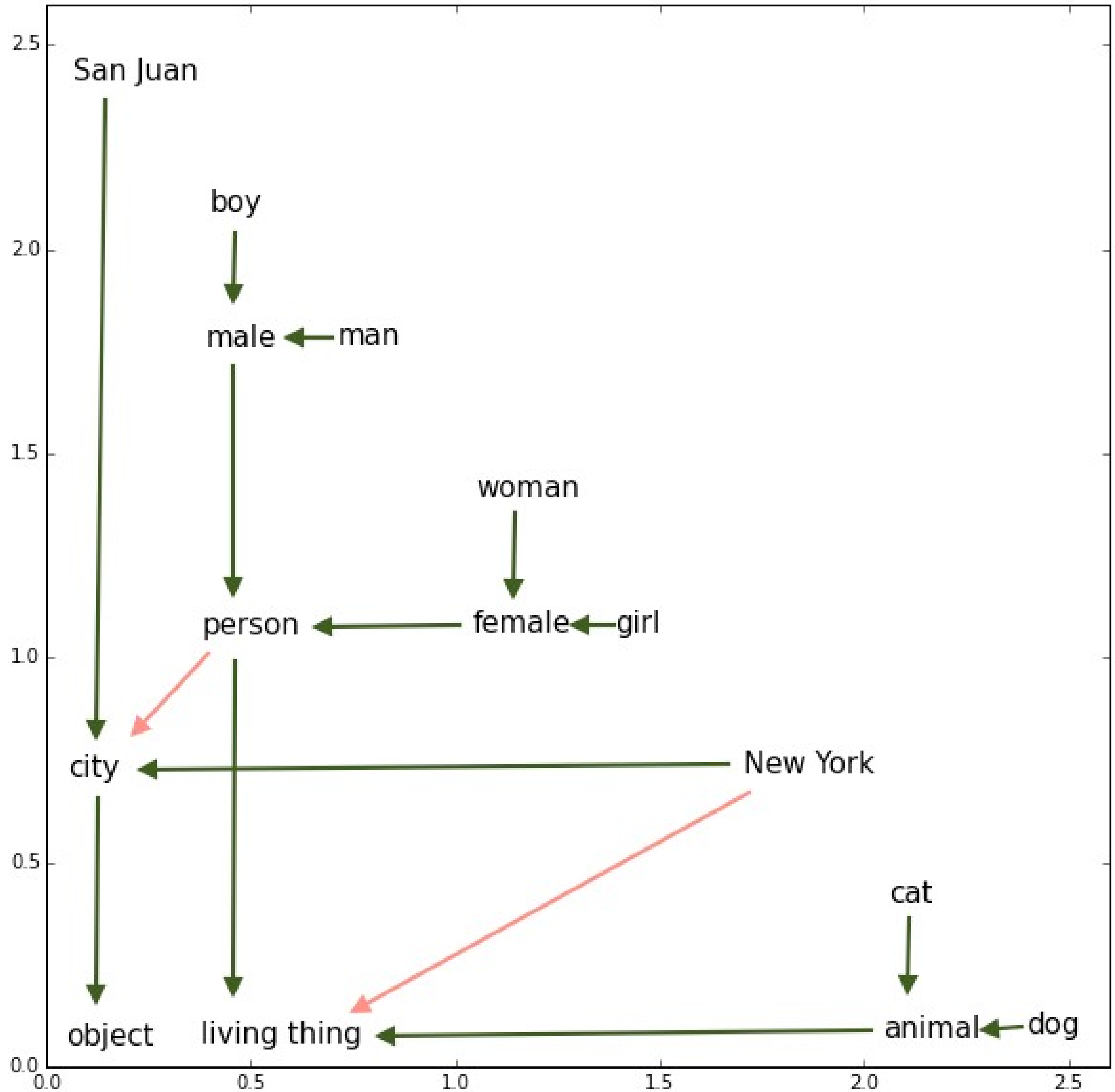


cute

Order-embeddings (Vendrov et al, 2016)



Order-embeddings (Vendrov et al, 2016)




Query
Nearest non-query images in COCO train
max("man", "cat")
max("black dog", "park")
max(,)
min(,)
min(, "dog")
max(, "man")
HORSEZONE

Applications to Multimodal tasks (Language+Vision)

(#1): Multimodal image-sentence embeddings

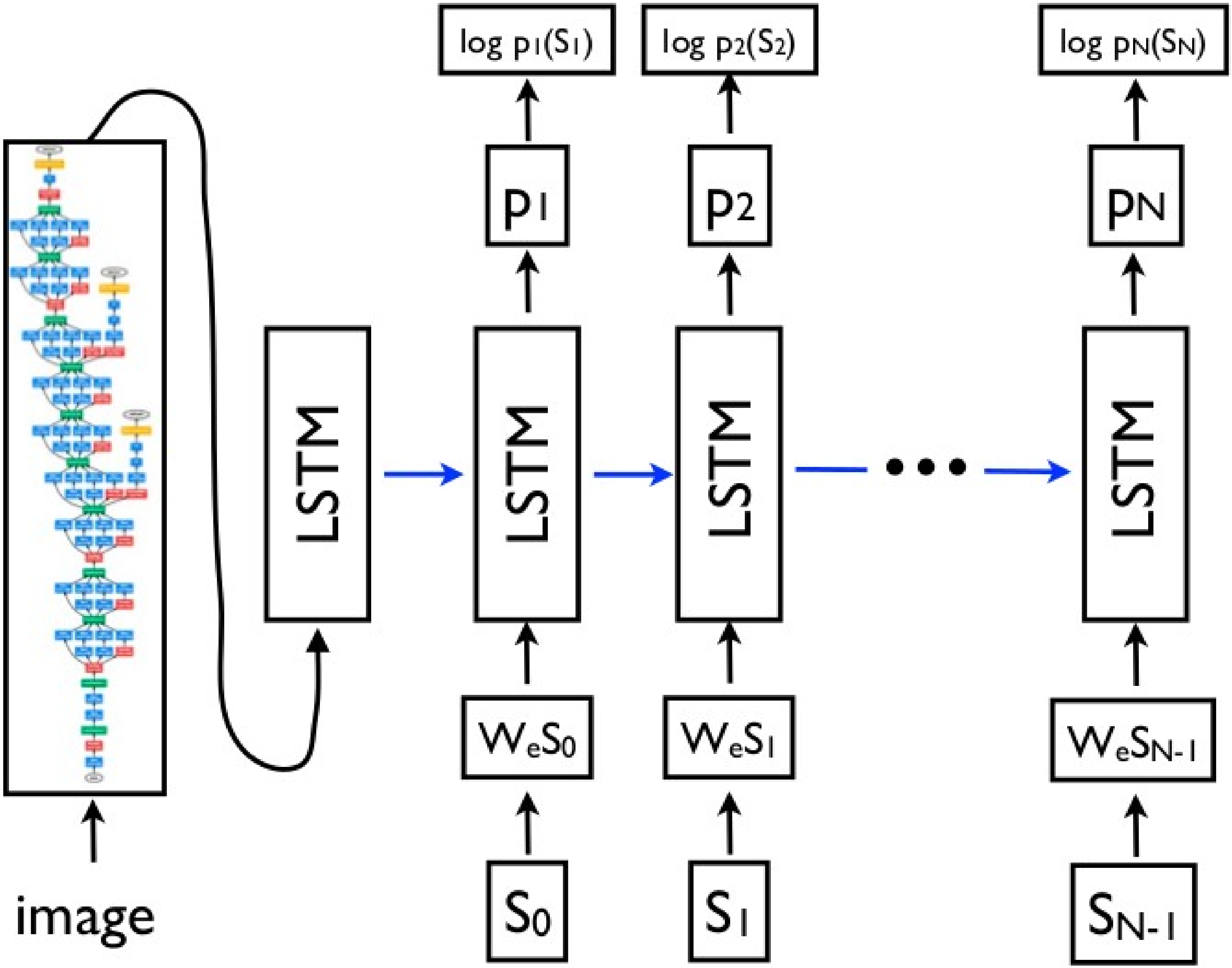
(#2): Image caption generation

(#3): Skip-thought vectors

(#4): Aligning books and movies

(#5): Style analogies + Neural storyteller

Google model: Multimodal LSTM (Vinyals et al, 2015)



Google model: Multimodal LSTM

Toronto model: Structure-content NLMs

$$P(w_n | w_{1:n-1}, t_{n:n+k}, \mathbf{x})$$

n-th word word context POS context

__________ (NN VBN IN DT NN)
DT

A __________ (VBN IN DT NN -)
NN

A bicycle __________ (IN DT NN - -)
VBN

A bicycle parked __________ (DT NN - - -)
IN

A bicycle parked on __________ (NN - - - -)
DT

A bicycle parked on the __________ (- - - - -)
NN

Some good results - generation

L.Z

a car is parked in
the middle of nowhere .

a wooden table and chairs
arranged in a room .

there is a cat sitting on a shelf .

a ferry boat on a marina
with a group of people .

a little boy with a bunch
of friends on the street .

Some failure types

the two birds are trying
to be seen in the water .
(can't count)

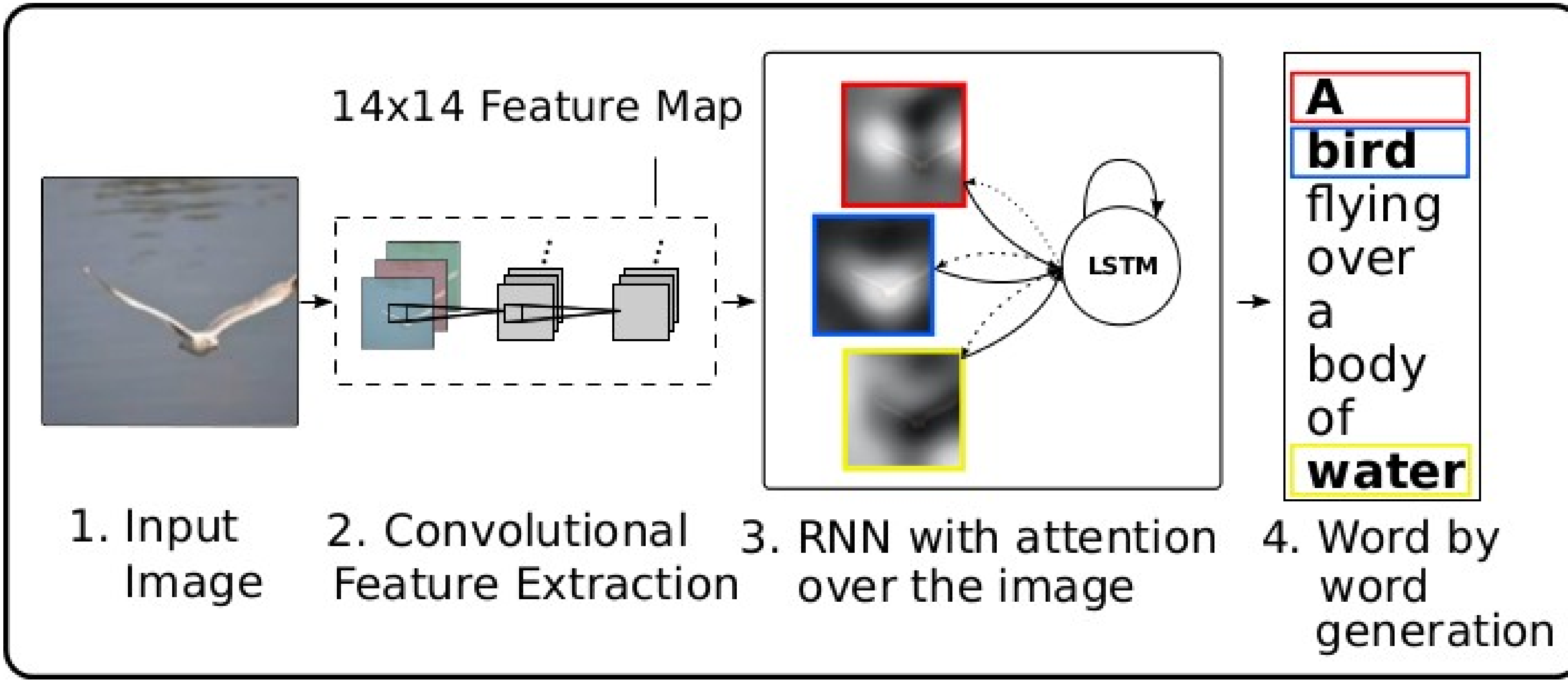
a giraffe is standing next
to a fence in a field .
(hallucination)

a parked car while
driving down the road .
(contradiction)

the handlebars are trying
to ride a bike rack .
(nonsensical)

a woman and a bottle of wine
in a garden . (gender)

Montreal+Toronto: LSTM with attention (Xu et al, 2015)



Montreal+Toronto: LSTM with attention (Xu et al, 2015)

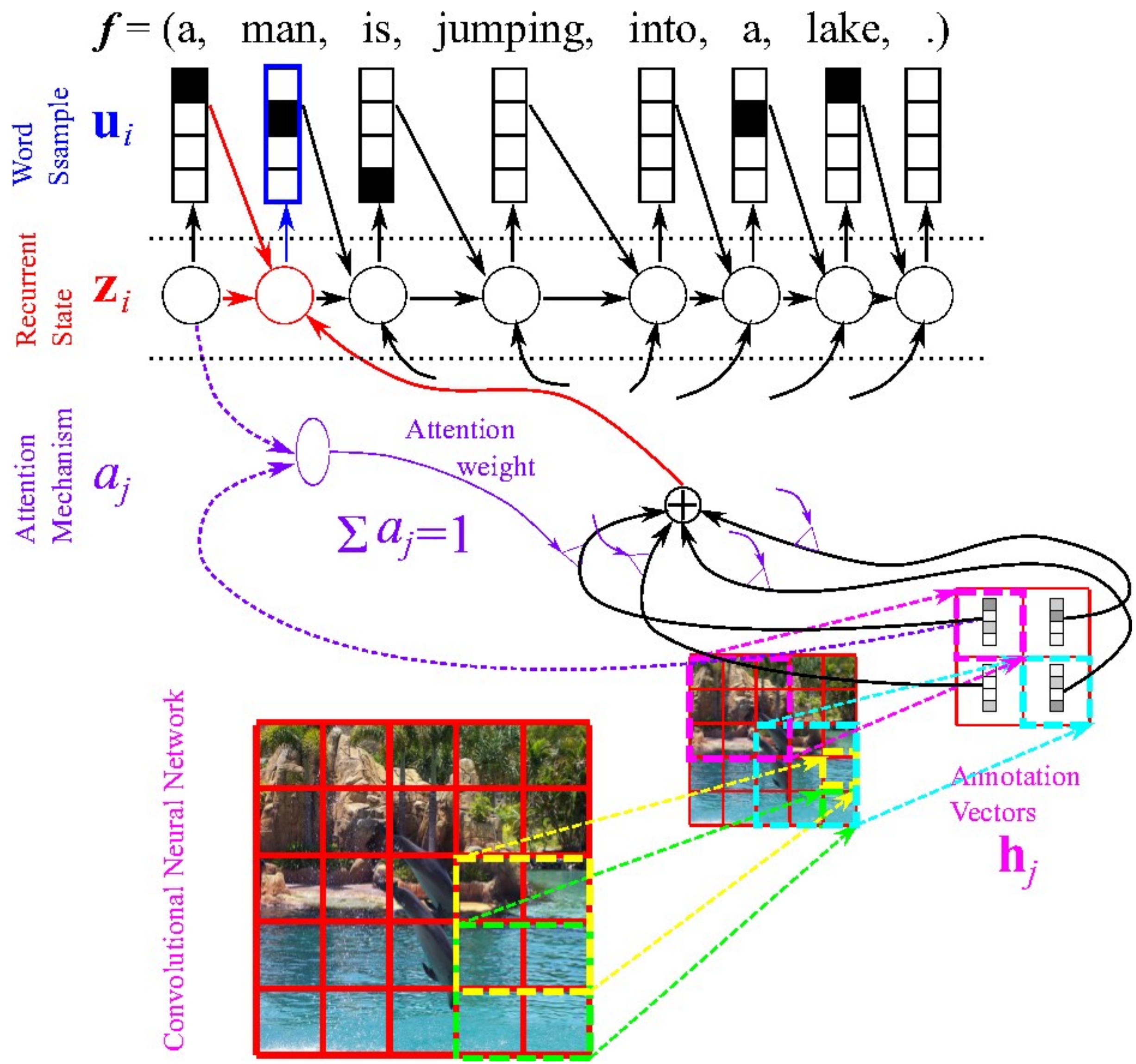


Image caption generation with attention

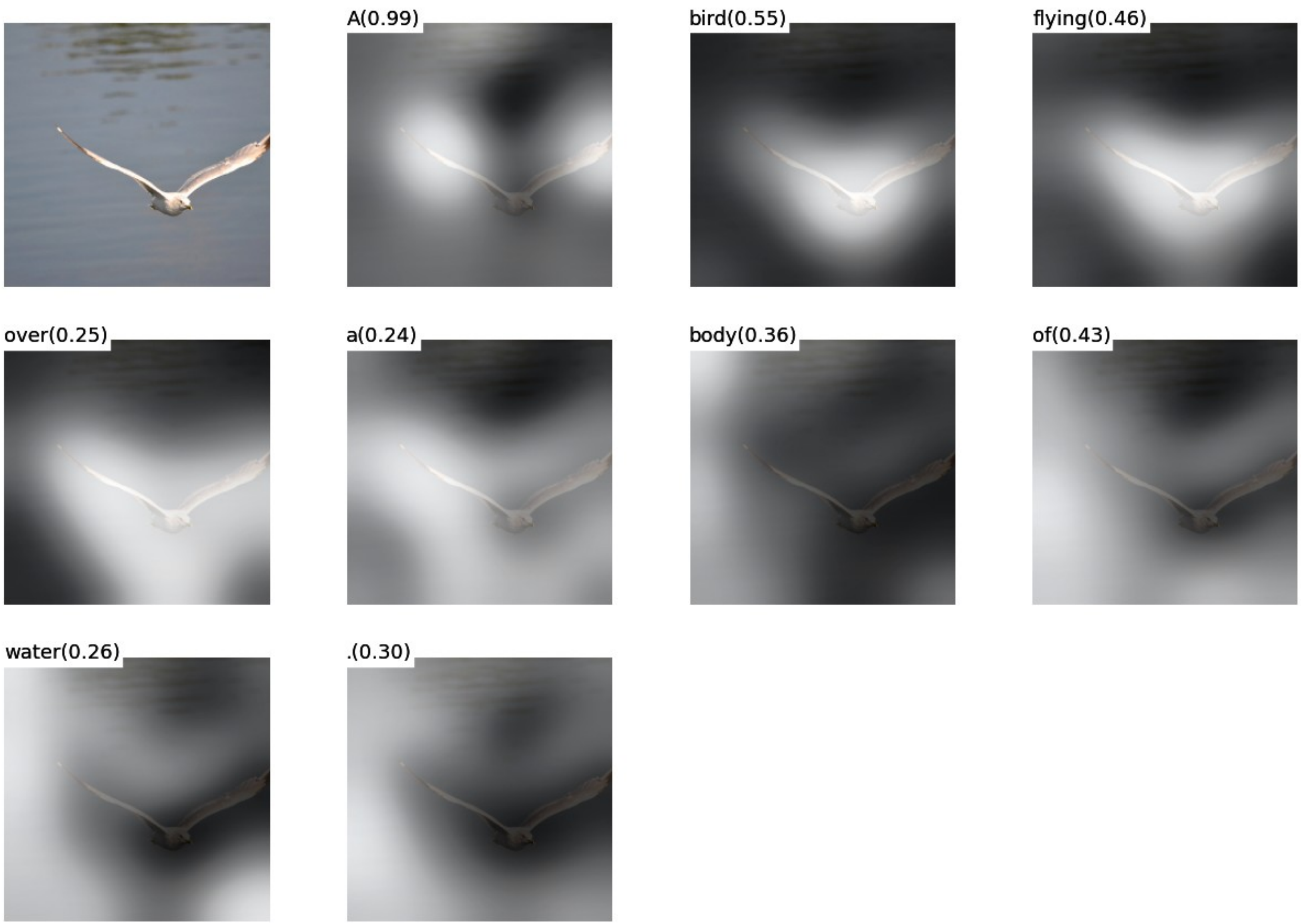


Image caption generation with attention

Image caption generation with attention

Many, many captioning papers...

https://github.com/kjw0612/awesome-rnn#image-captioning

See the above for a full list! (15 papers and counting)

Who is best? Microsoft COCO Competition

Team	M1	M2	Total	Ranking
Google	5	4	9	1st (tie)
MSR	4	5	9	1st (tie)
Montreal-Toronto	3	2	5	3rd (tie)
MSR Captivator	2	3	5	3rd (tie)
Berkeley	1	1	2	5th

M1: Percentage of captions that are evaluated as better or equal to human.
M2: Percentage of captions that pass the Turing Test.

Applications to Multimodal tasks (Language+Vision)

(#1): Multimodal image-sentence embeddings

(#2): Image caption generation

(#3): Skip-thought vectors

(#4): Aligning books and movies

(#5): Style analogies + Neural storyteller

Unsupervised Distributed Representations for words and sentences

- There is a **massive** amount of available text data

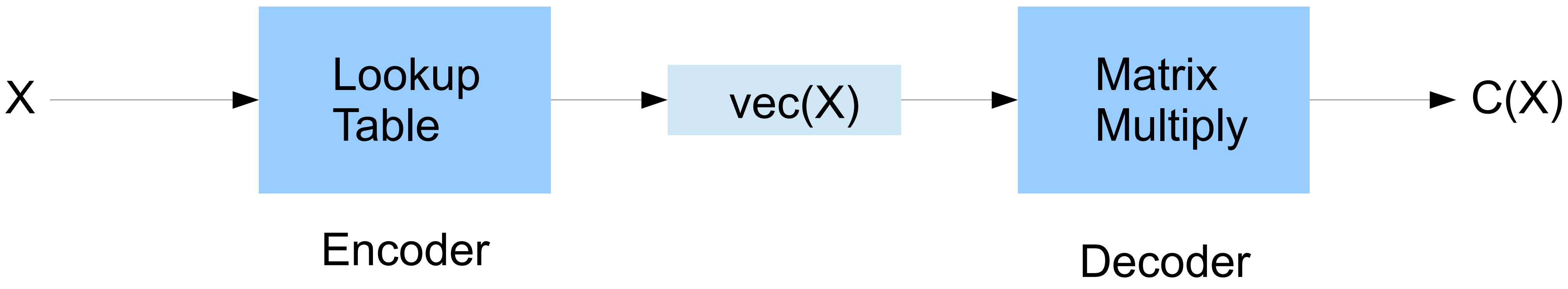
- Good word vectors utilize distributional hypothesis + tons of data
  Context as a learning signal (implicit or explicit)

- Sentence representations, on the other hand are usually task specific
  Backprop through the "composition function" using labelled data

Can we abstract how we learn word vectors to construct new objectives for learning sentence vectors?

Does the concept of a non-task specific sentence vector even make sense?

Revisiting skip-gram (Mikolov et al. 2013)

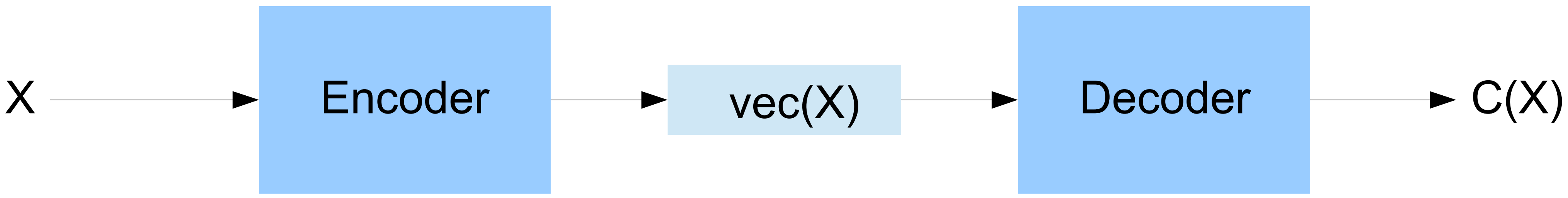


- Skip-gram is an encoder-decoder model:

  - Input X: A word
  - Encoder: Lookup table
  - Decoder: Matrix multiply
  - Context C(X): Predictions of surrounding words

- Minimize NLL of context predictions given X

Contextual Encoder-Decoders



- Skip-gram is just an instance of a more generic family of models!

- If we want to learn a vector for X, we just need to specify:

  - the encoder which maps X to vec(X)
  - the context C(X)
  - the decoder which maps vec(X) to predictions of C(X)

- Does X have to be a word? Why not a sentence? Paragraph? Etc...

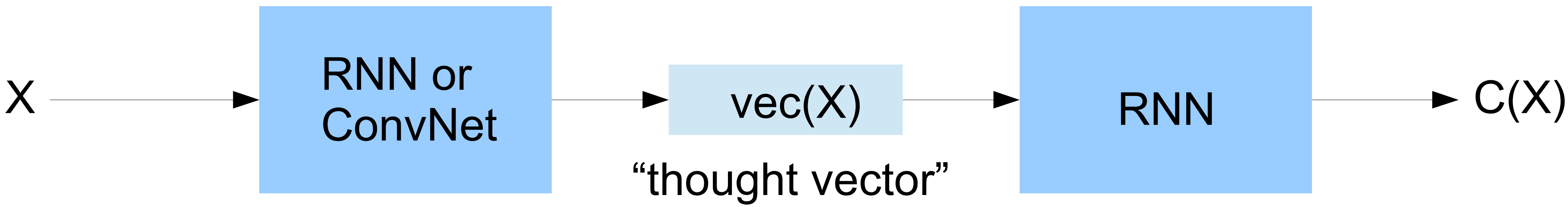
From words to sentences

Input	Word	Sentence
Encoder	Lookup table	RNN or ConvNet
Context	Surrounding words	Surrounding sentences
Decoder	Matrix multiply	RNN

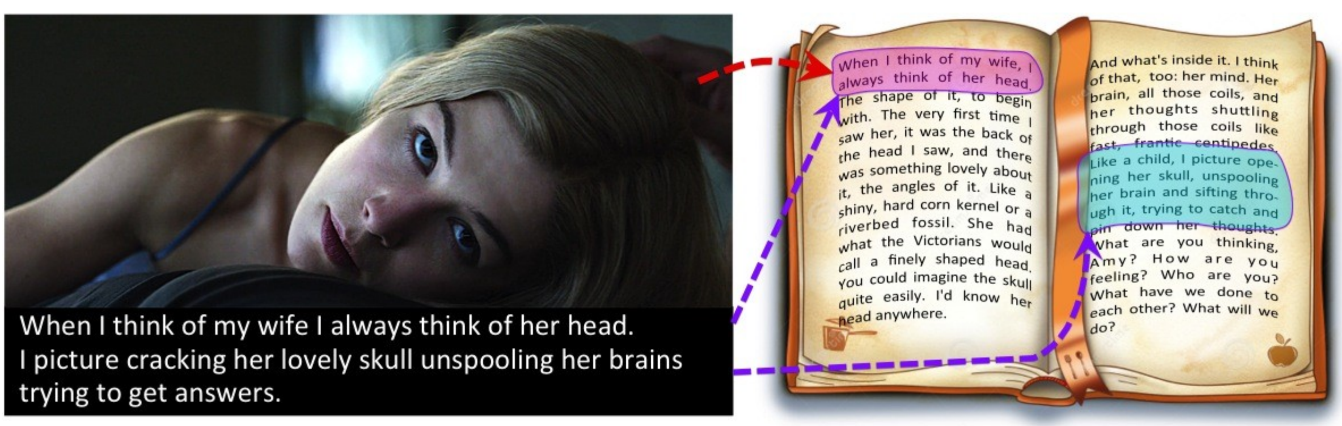
- What is a good context for sentences?

  - We use surrounding words for word context
  - Why not use surrounding sentences for sentence context?

Skip-thought vectors (Kiros et al. 2015)



- Given a sentence X, predict sentences before and after
- Note that we need contiguous text for training!
- BookCorpus dataset: 10K+ books, 70M+ sentences, ~1B words (Zhu+Kiros et al. 2015)



What does it learn? Nearest neighbours:

he ran his hand inside his coat , double-checking that the unopened letter was still there

he slipped his hand between his coat and his shirt , where the folded copies lay in a brown envelope .

im sure youll have a glamorous evening , she said , giving an exaggerated wink .

im really glad you came to the party tonight , he said , turning to her .

although she could tell he had n't been too invested in any of their other chitchat , he seemed genuinely curious about this .

although he had n't been following her career with a microscope , he 'd definitely taken notice of her appearances .

Related models and ideas

- Paragraph Vector (Le & Mikolov, 2014)

  - Encoder: Lookup table
  - Decode: Words in the sentence/paragraph

- Sequence Autoencoders (Dai & Le, 2015)

  - Encoder: LSTM
  - Decode: Words in the sentence

- C-PHRASE (Pham et al., 2015)

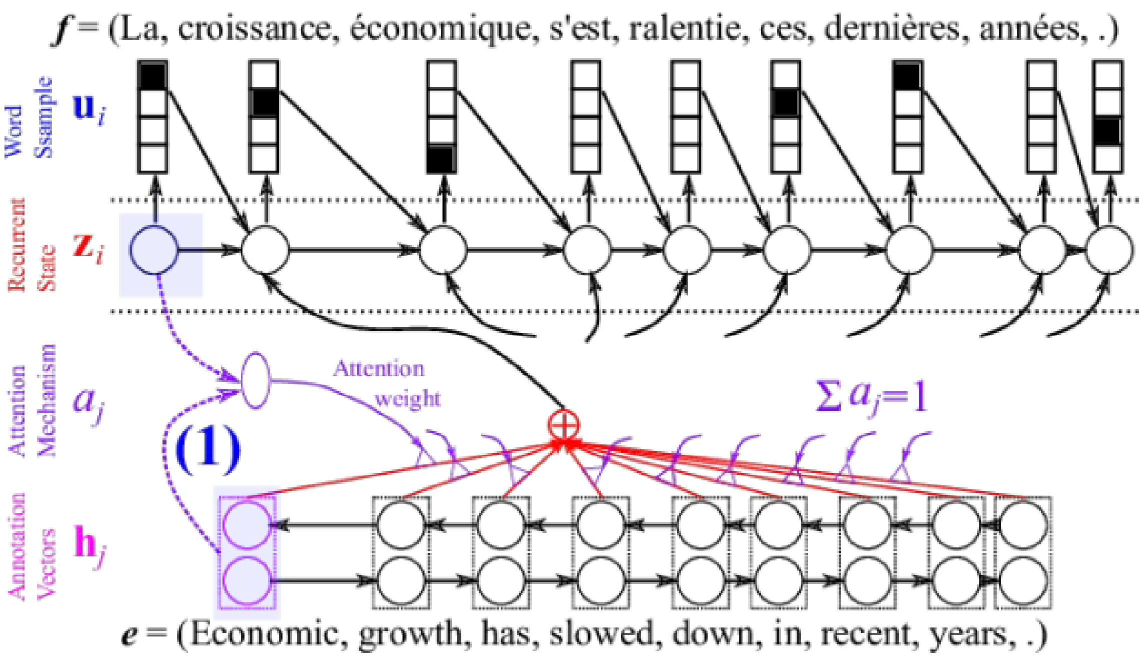
  - Encoder: Sum of word vectors
  - Decode: Syntactic context each each level of hierarchy

Main weakness: These models only look at the current sentence!
Ignores the context of which the sentence occurs

Cramming everything into a vector

- For some tasks (MT, QA, reading comprehension), this doesn't make a whole lot of sense

  - Instead, dynamically update the representation of a sentence
  - "Zone in" on the relevant parts at any given time
  - Attention mechanisms, memory networks, etc



(Bahdanau et al., 2014)

Can we still make use of unsupervised sentence vectors for these tasks?

How can we utilize skip-thought vectors for multimodal tasks?

Applications to Multimodal tasks (Language+Vision)

(#1): Multimodal image-sentence embeddings

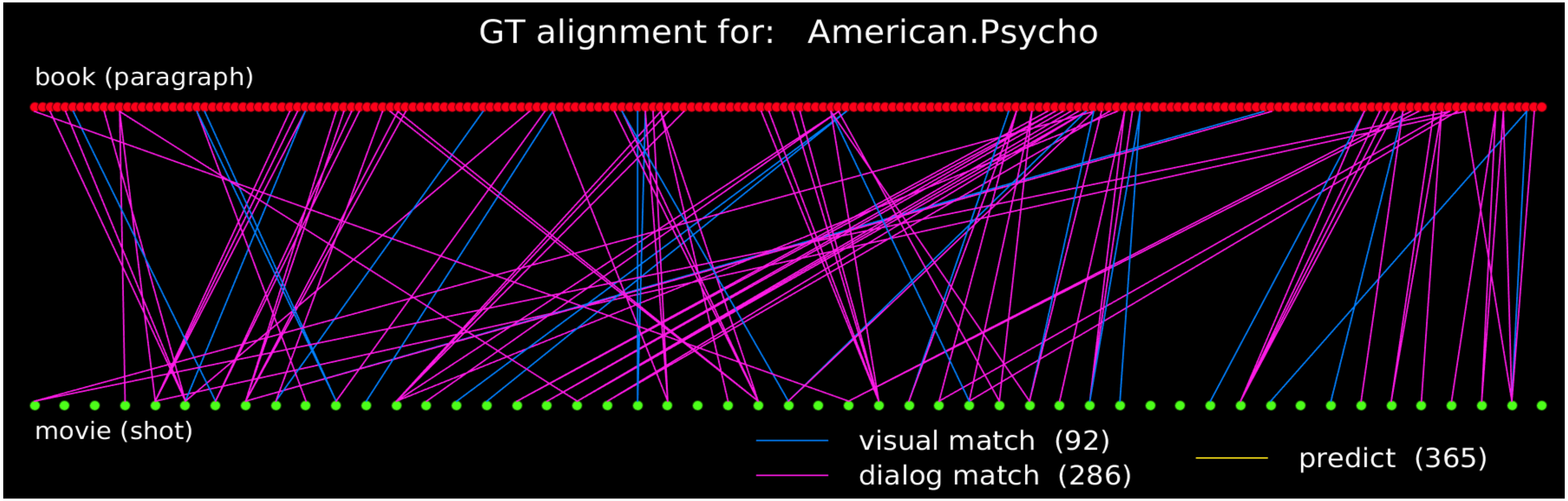
(#2): Image caption generation

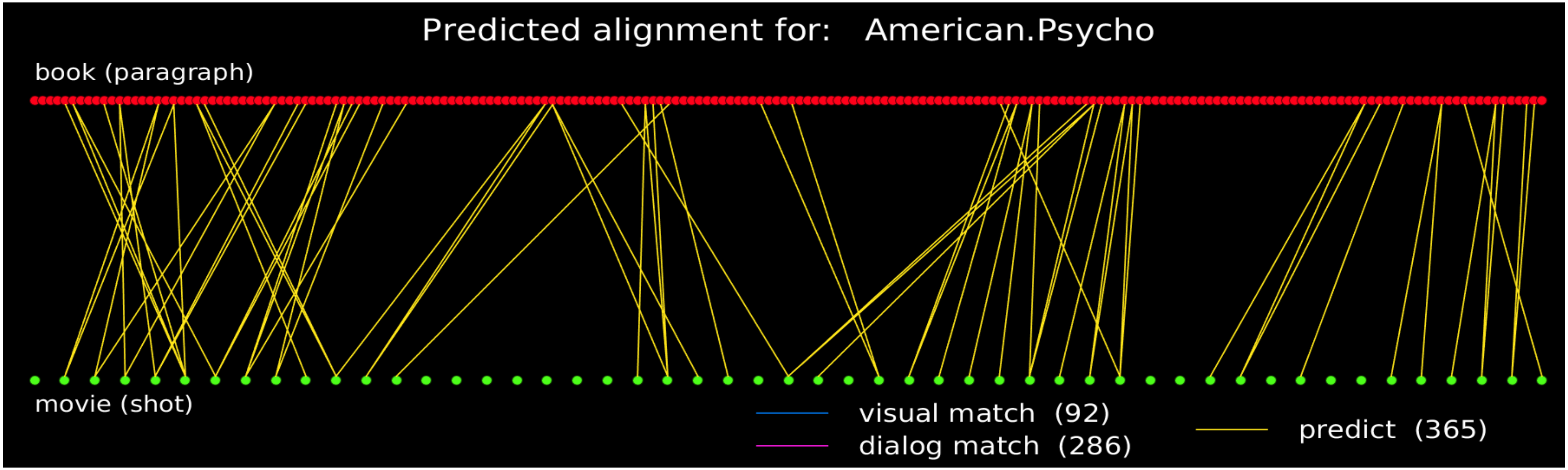
(#3): Skip-thought vectors

(#4): Aligning books and movies

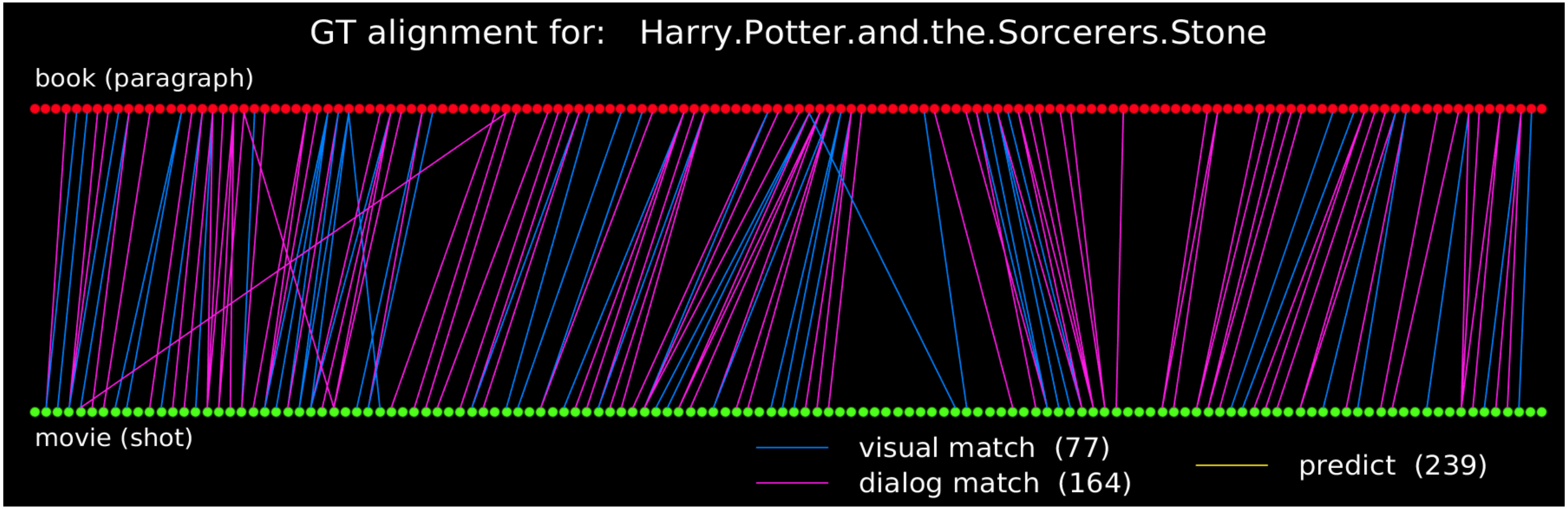
(#5): Style analogies + Neural storyteller

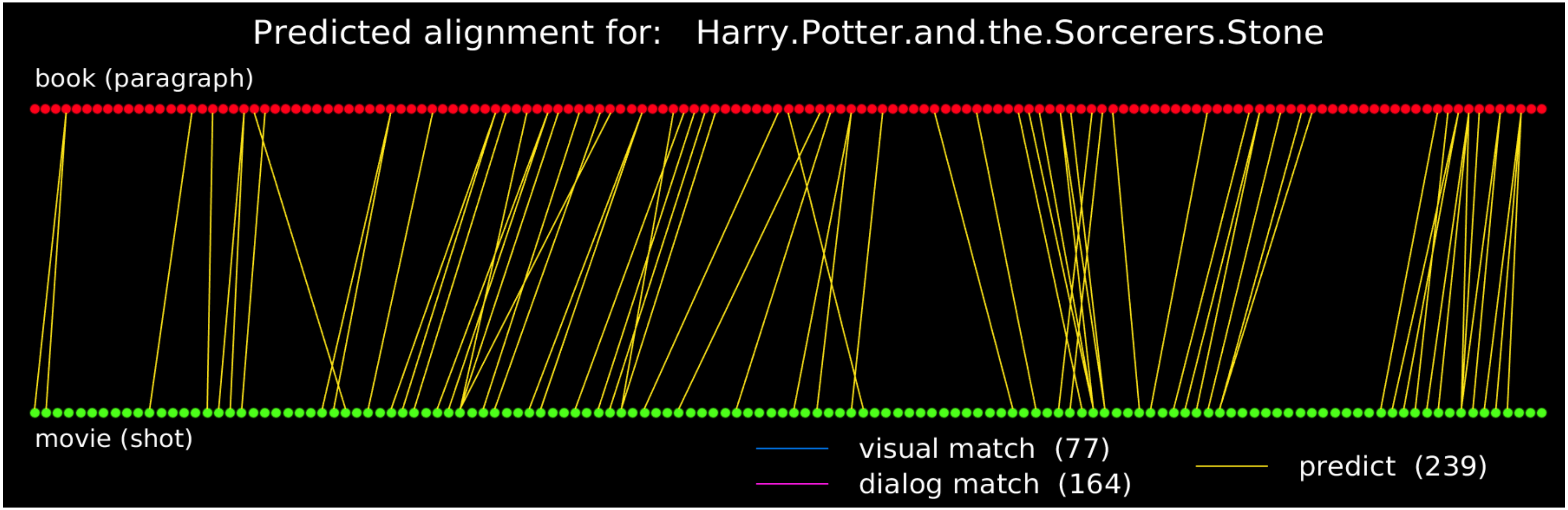
Aligning books and movies (Zhu+Kiros et al., 2015)





Aligning books and movies (Zhu+Kiros et al., 2015)

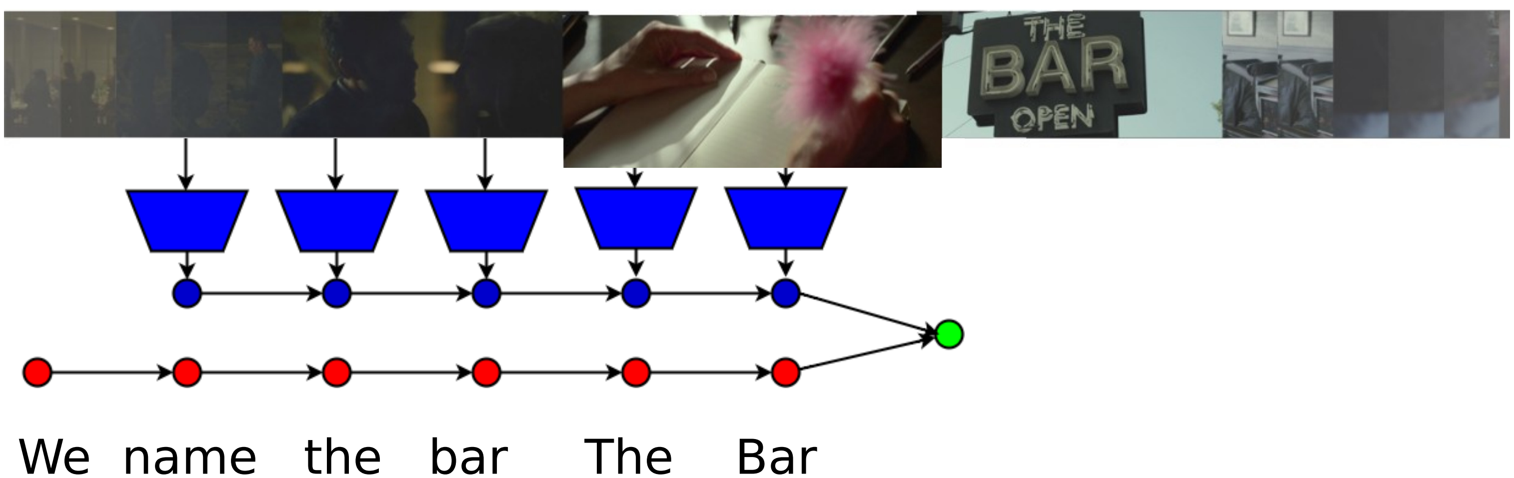




How it works

- Context-aware CNN that combines:

  - Text-to-text similarity: Skip-thoughts + tf-idf + BLEU(1-5)
  - Video-to-text similarity: Visual-semantic embedding of clips and DVS



- Chain CRF: Unaries are CNN outputs
  Pairwise terms for consistency between nearby alignments

Alignment results

[02:14:29:02:14:32] Good afternoon, Harry.

... He realized he must be in the hospital wing. He was lying in a bed with white linen sheets, and next to him was a table piled high with what looked like half the candy shop.

"Tokens from your friends and admirers," said Dumbledore, beaming. "What happened down in the dungeons between you and Professor Quirrell is a complete secret, so, naturally, the whole school knows. I believe your friends Misters Fred and George Weasley were responsible for trying to send you a toilet seat. No doubt they thought it would amuse you. Madam Pomfrey, however, felt it might not be very hygienic, and confiscated it."
...

Alignment results (X movie/book)

Batman.Begins

[01:38:41:01:38:44] I'm gonna give you a sedative. You'll wake up back at home.

A Captive s Submission

"I believe you will enjoy your time here. I am not a harsh master but I am strict. When we are with others, I expect you to present yourself properly. What we do here in your room and in the dungeon is between you and I. It is a testament to the trust and respect we have for each other and no one else needs to know about our arrangement. I'm sure the past few days have been overwhelming thus far but I have tried to give you as much information as possible. Do you have any questions?"

Retrieving stories for images

the club was a little emptier than i would have expected for the late afternoon , and the bartender , in red waistcoat and bowtie , was busy wiping down his counter , replacing peanuts and putting out new coasters .
a television with the latest la liga news was hung in an upper corner , and behind him , rows of bottles were reflected in a giant bar mirror .
above the stools , a pergola-type overhead structure held rows of wine glasses .
it was a classy place , with ferns in the corner , and not the kind of bar to which i was accustomed .
my places usually had a more ... relaxed feel .

he felt like an idiot for yelling at the child , but his frustration and trepidation was getting the better of him .
he glanced toward the shadowed hall and quickly nodded toward melissa before making his way forward .
he came across more children sitting upon a couch in the living room .
they watched him , but did n't move and did n't speak .
his skin started to feel like hundreds of tiny spiders were running up and down it and he hurried on .

Can we generate stories instead?

Applications to Multimodal tasks (Language+Vision)

(#1): Multimodal image-sentence embeddings

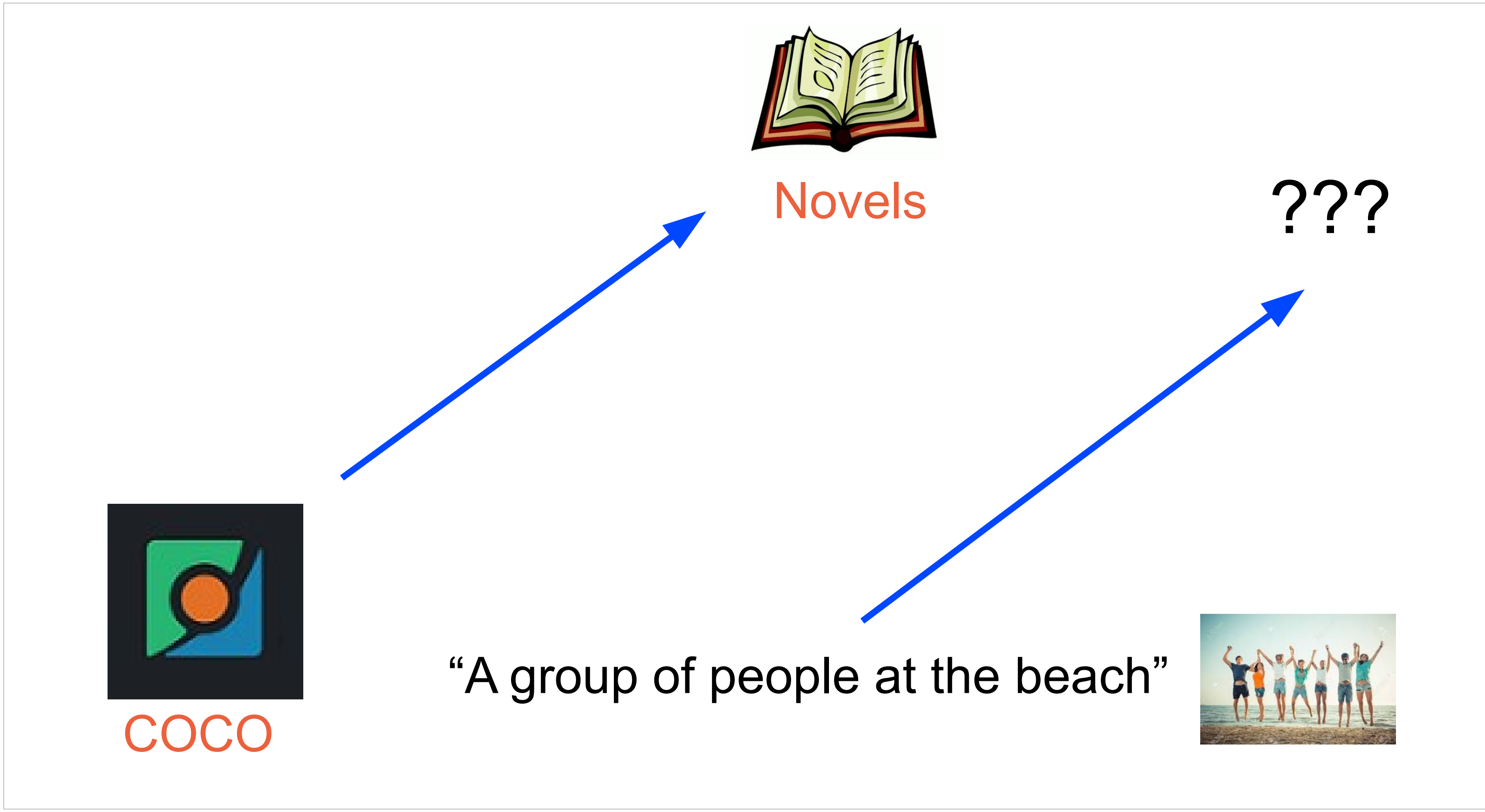
(#2): Image caption generation

(#3): Skip-thought vectors

(#4): Aligning books and movies

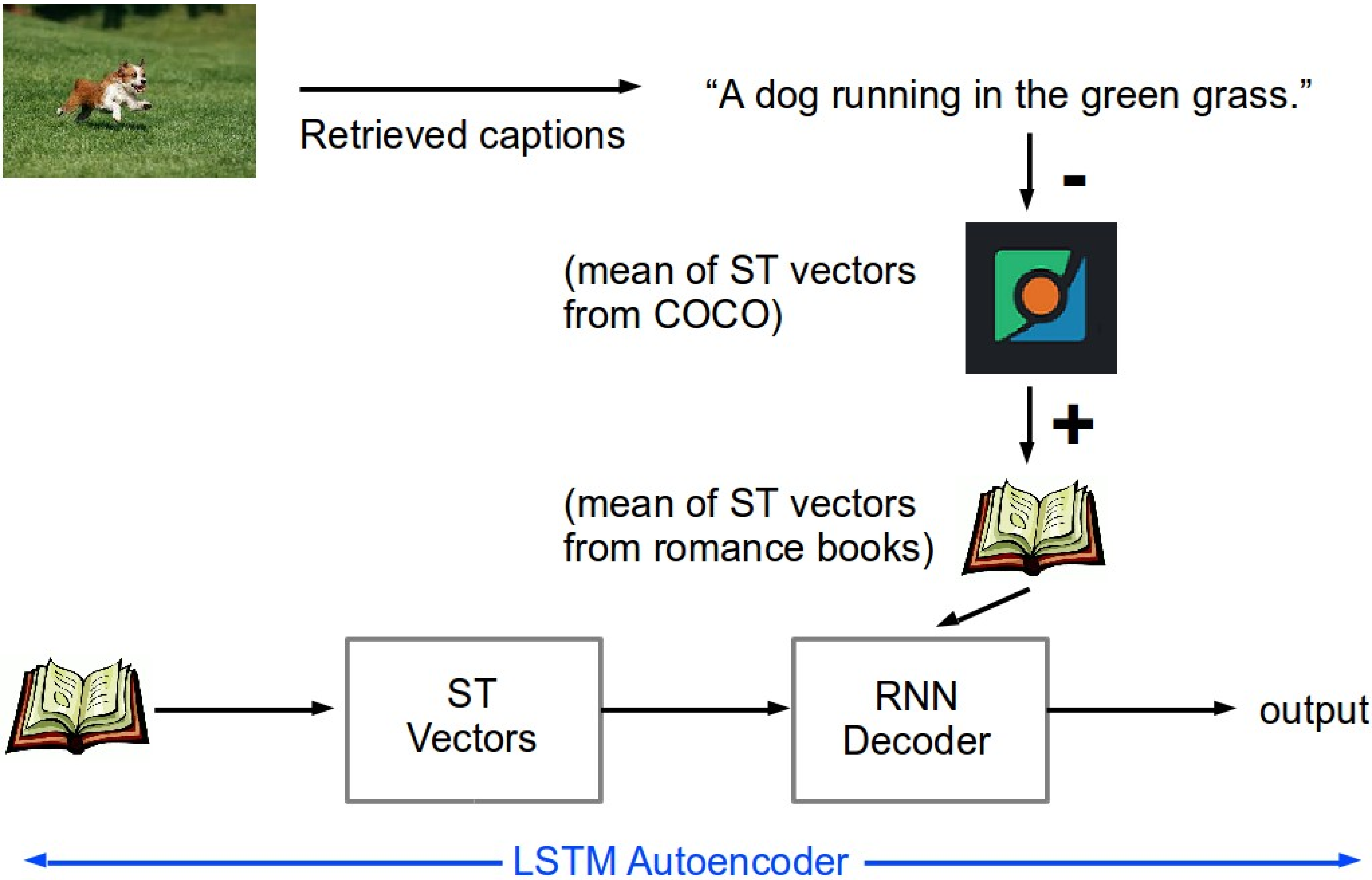
(#5): Style analogies + Neural storyteller

(#3) Style analogies with thought vectors

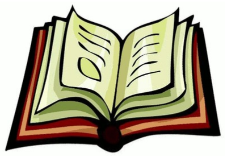


“A group of people at the beach” - + = ???

The Neural-Storyteller



Results



We were barely able to catch the breeze at the beach , and it felt as if someone stepped out of my mind . She was in love with him for the first time in months , so she had no intention of escaping . The sun had risen from the ocean , making her feel more alive than normal . She 's beautiful , but the truth is that I do n't know what to do . The sun was just starting to fade away , leaving people scattered around the Atlantic Ocean . I d seen the men in his life , who guided me at the beach once more .

Results (Samim Winiger*)

*"Generating stories about images"

Generated story about image

Model: Romantic Novels

"He was a shirtless man in the back of his mind, and I let out a curse as he leaned over to kiss me on the shoulder.

He wanted to strangle me, considering the be-atiful boy I'd become wearing his boxers."

Results (Samim Winiger*)

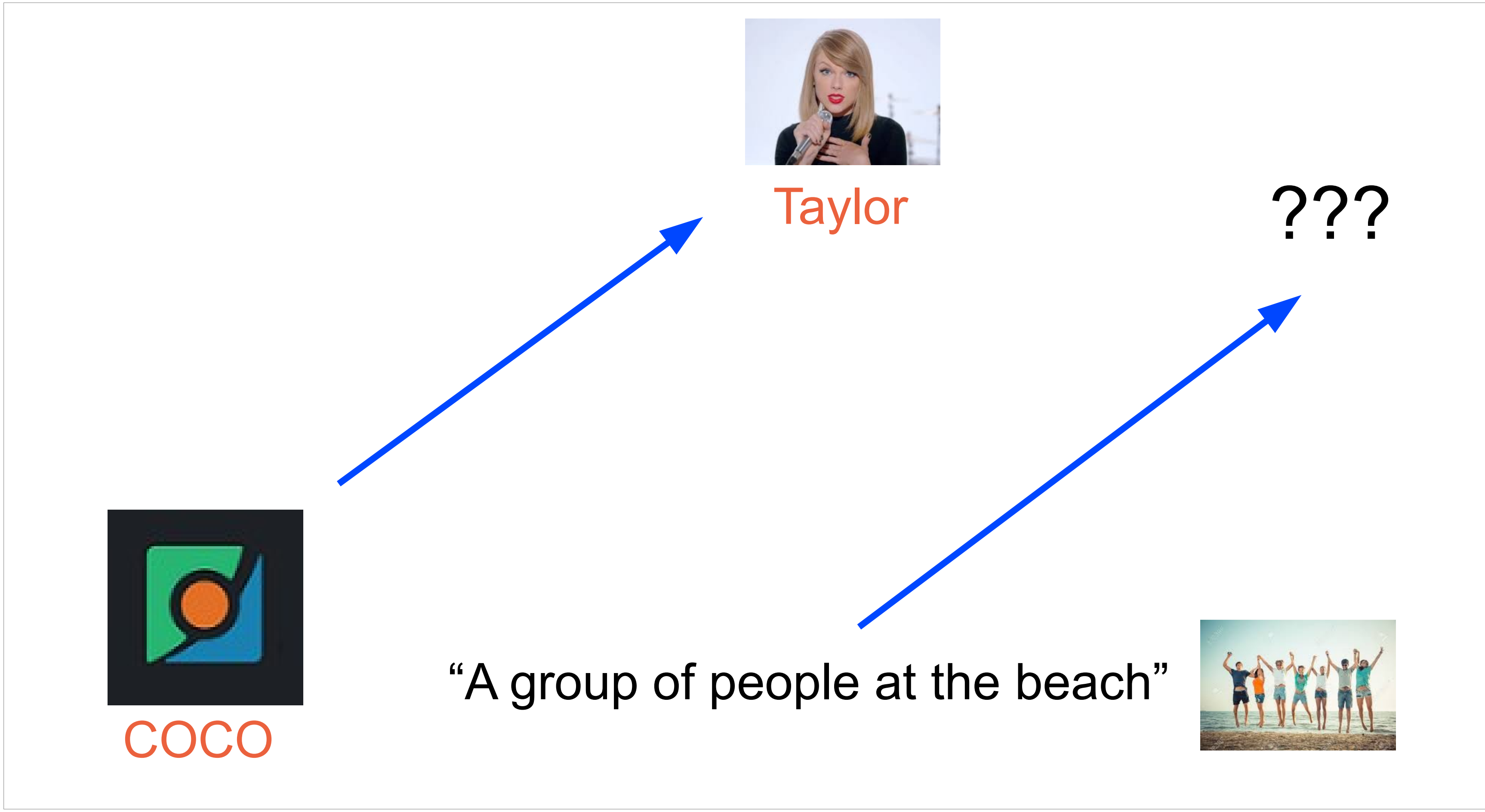
*"Generating stories about images"

Generated story about image
Model: Romantic Novels

"We men were in a tense position at the end of the meeting. And i looked up at my best friend.

Of course, i had no intention of letting him go. I don't know what else to say, but he is also the most beautiful man you ever meet."

What about Taylor Swift?



“A group of people at the beach” - +  = ???

Results

You re the only person on the beach right now
you know
I do n't think I will ever fall in love with you
and when the sea breeze hits me
I thought
Hey

(#3) Results (Samim Winiger*)

*"Generating stories about images"

Generated story about image
Model: Taylor Swift Lyrics

"I give you a man , I don't know what 's happening to me , and when I look back at the stage, I say, God , I love you more than I should."

(#3) Results (Samim Winiger*)

*"Generating stories about images"

Generated story about image
Model: Taylor Swift Lyrics

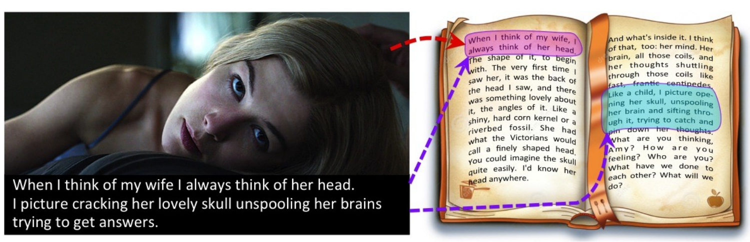
"Like I 'm standing right now , man, it s going to be a sidewalk in the street, I thought, Oh my God, I don't see you walking away."

Resources and Code

Resources and code

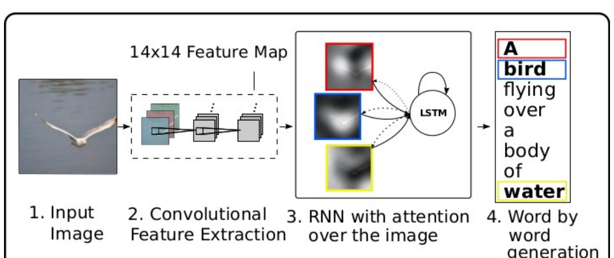
https://github.com/ryankiros

- Skip-thought vectors
- Neural-Storyteller
- Visual Semantic Embeddings



http://www.cs.toronto.edu/~mbweb/

- BookCorpus dataset
- Ground-truth Movie/Book alignments



https://github.com/kelvinxu/arctic-captions

- “Show, attend and Tell” code

Andrej Karpathy

https://github.com/karpathy
http://cs231n.stanford.edu/

- char-rnn
- neuraltalk
- neuraltalk2
- randomfun

http://cs231n.stanford.edu/slides/winter1516_lecture10.pdf

Learn more about RNNs

- https://github.com/kjw0612/awesome-rnn

Large collection of lectures, papers and code

My awesome collaborators (for multimodal learning)

Toronto

Richard Zemel
Ruslan Salakhutdinov
Raquel Urtasun
Sanja Fidler
Kevin Swersky
Jimmy Ba
Yukun Zhu
Ivan Vendrov
Mengye Ren
Shikhar Sharma

NYU

Kyunghyun Cho

Montreal

Yoshua Bengio
Aaron Courville
Kelvin Xu

Harvard

Ryan Adams
Jasper Snoek
Oren Rippel

MIT

Antonio Torralba