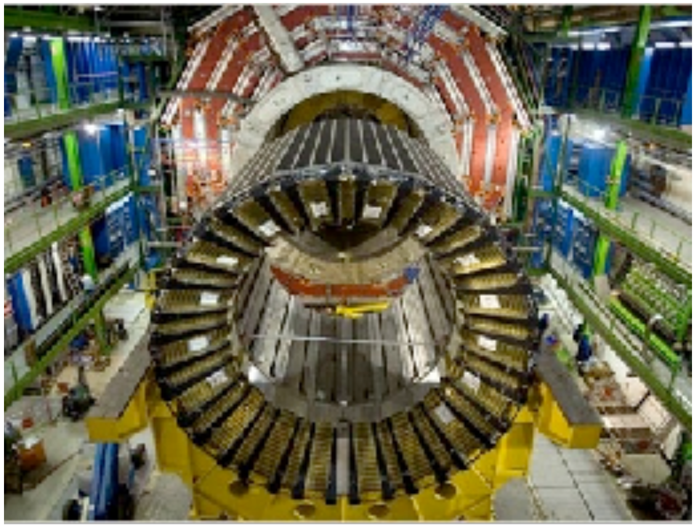
Lecture 20

Words and pictures


Crowdsourcing

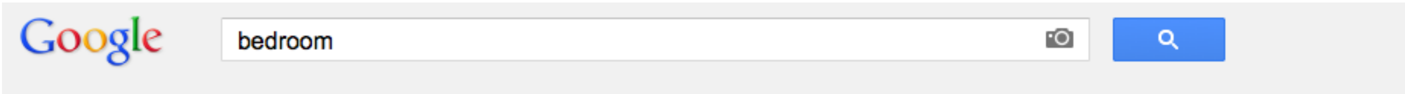
The value of data

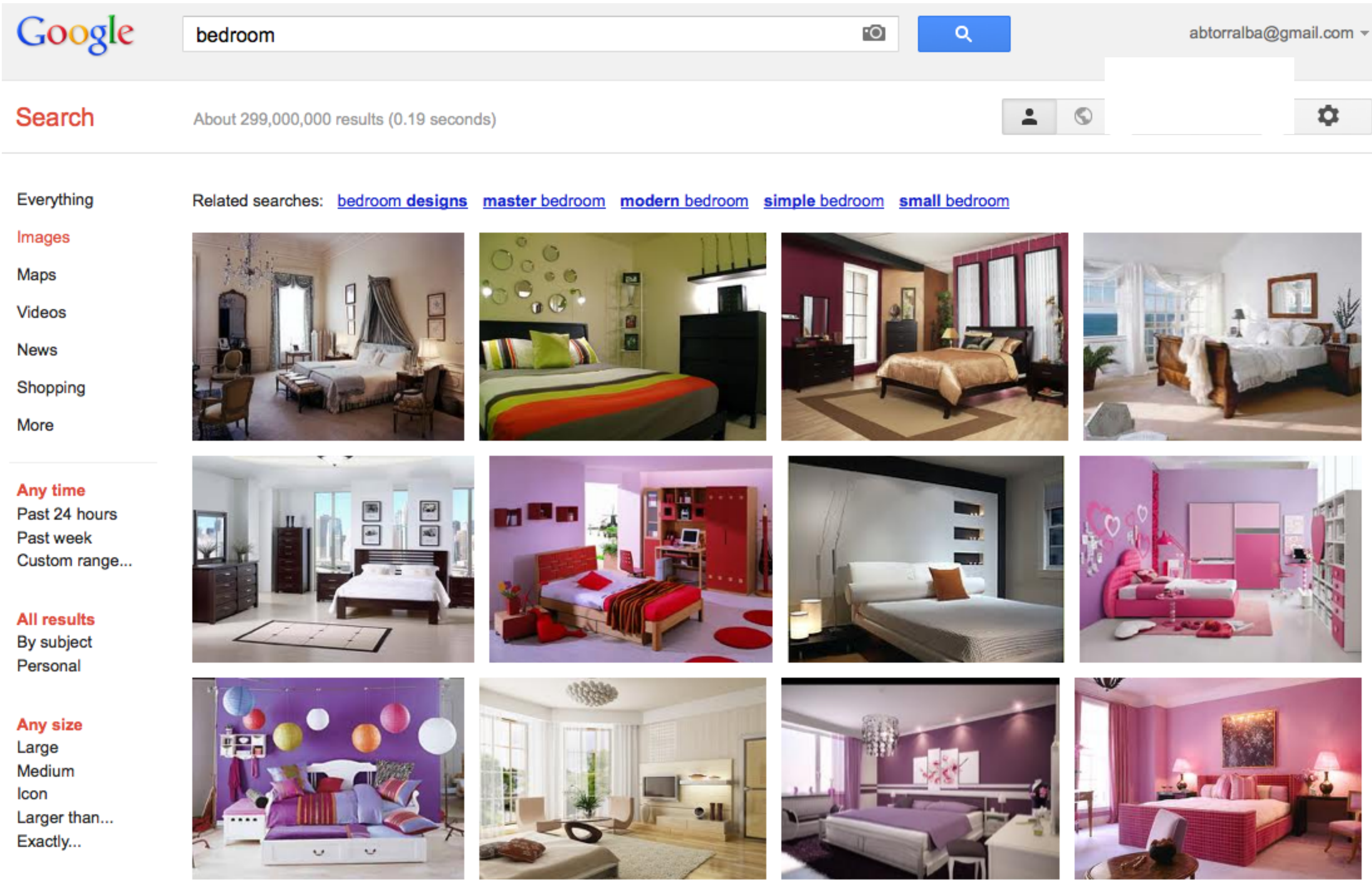


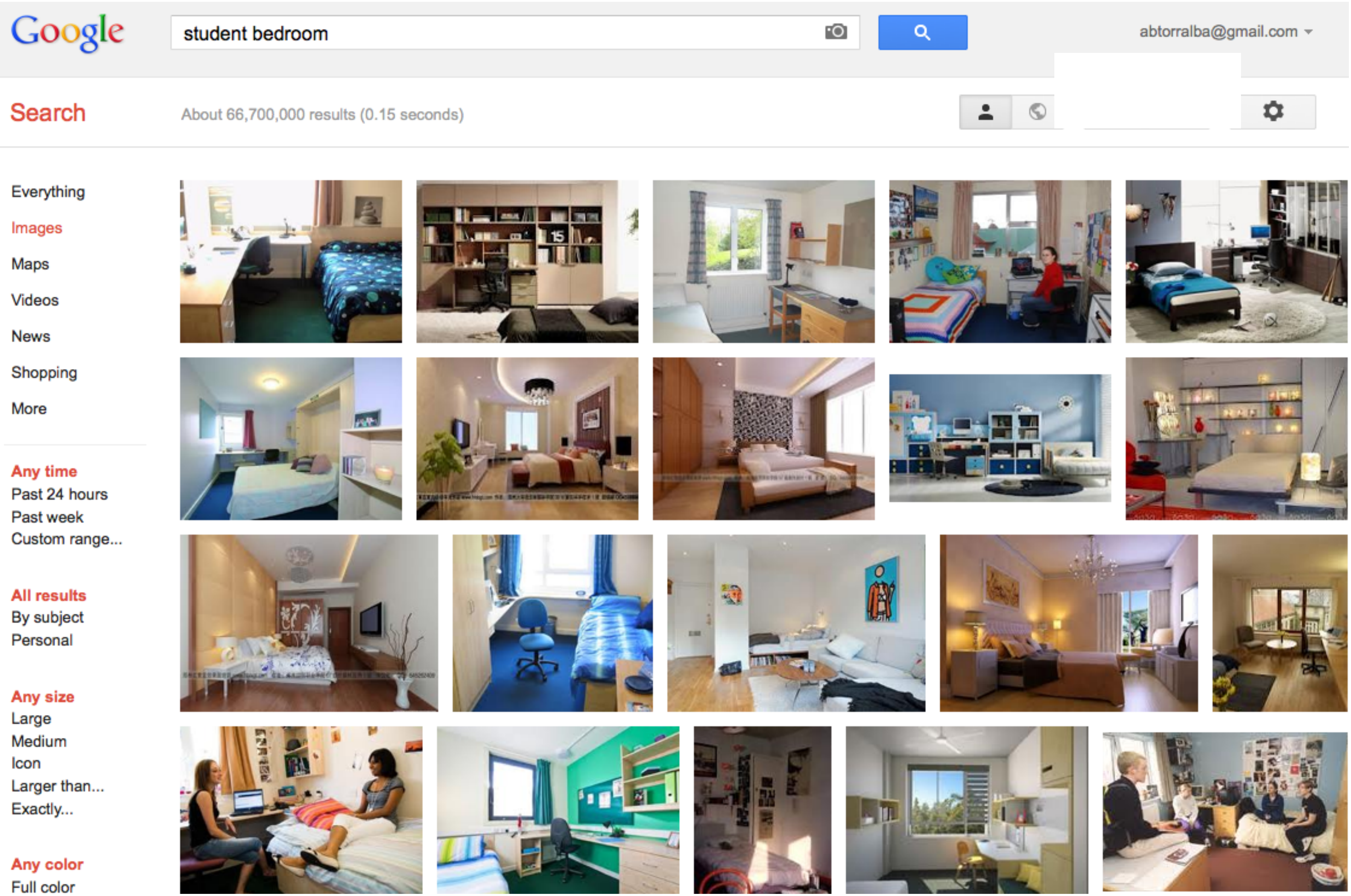
The Large Hadron Collider
$ 10^{10}

Amazon Mechanical Turk
$ 10^{2} - 10^{4}

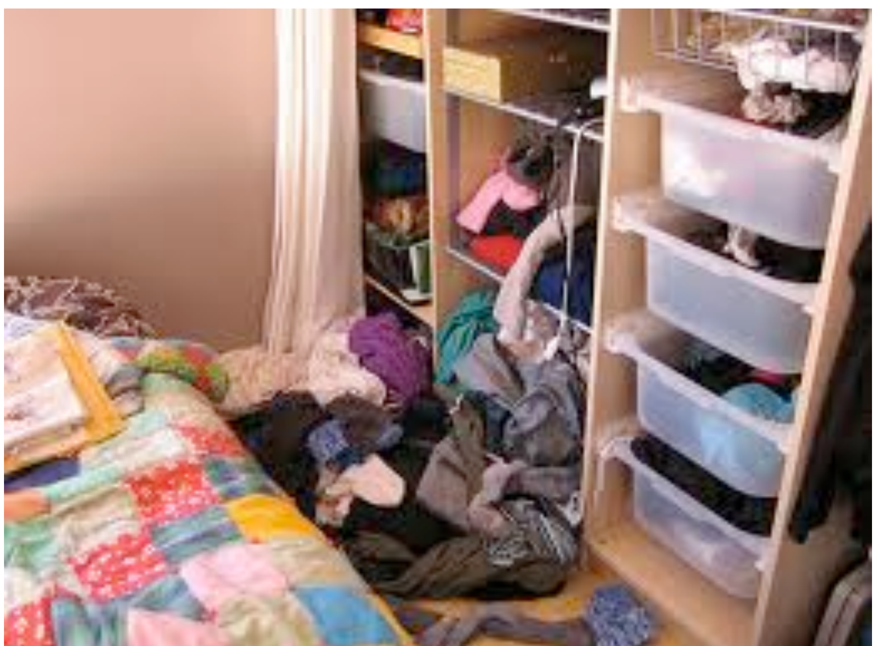
But can humans collect good data?


bedroom


Google
bedroom
abtorralba@gmail.com
Search
About 299,000,000 results (0.19 seconds)
Everything
Images
Maps
Videos
News
Shopping
More
Related searches: bedroom designs master bedroom modern bedroom simple bedroom small bedroom
Any time
Past 24 hours
Past week
Custom range...
All results
By subject
Personal
Any size
Large
Medium
Icon
Larger than...
Exactly...


Google
student bedroom
abtorralba@gmail.com
Search
About 66,700,000 results (0.15 seconds)
Everything
Images
Maps
Videos
News
Shopping
More
Any time
Past 24 hours
Past week
Custom range...
All results
By subject
Personal
Any size
Large
Medium
Icon
Larger than...
Exactly...
Any color
Full color


www.bigstock.com · 7067629




Google
mug
I OWN YOU
Mom

Getting more humans in the annotation loop

Labeling to get a Ph.D.

Labeling for fun

Luis Von Ahn and Laura Dabbish 2004

Labeling for money

(Sorokin, Forsyth, 2008)

amazonmechanical turk

Artificial Artificial Intelligence

Labeling because it gives you added value

Visipedia

(Belongie, Perona, et al)

Just for labeling

LabelMe

Sorokin, Forsyth, 2008

Carl Vondrick, Deva Ramanan, Don Patterson

N. Kumar, A. C. Berg,
P. N. Belhumeur, and S. K. Nayar, ICCV 2009

Farhadi Endres Hoiem Forsyth CVPR 2008

And many more…

Beware of the human in your loop

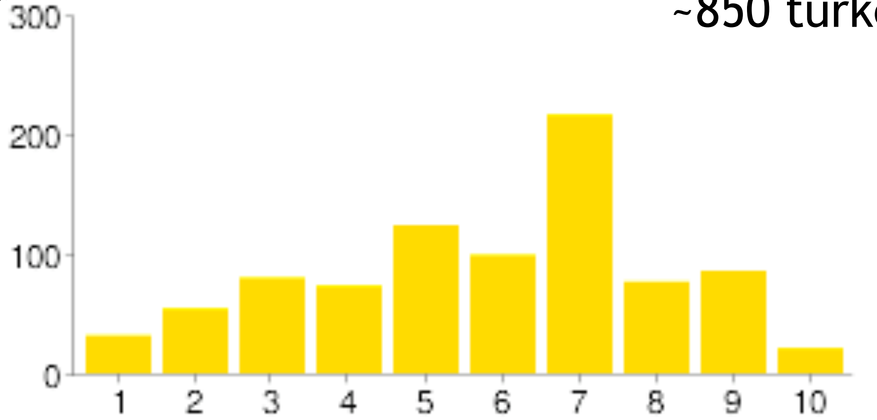
- What do you know about them?
- Will they do the work you pay for?

Let’s check a few simple experiments

People has biases…

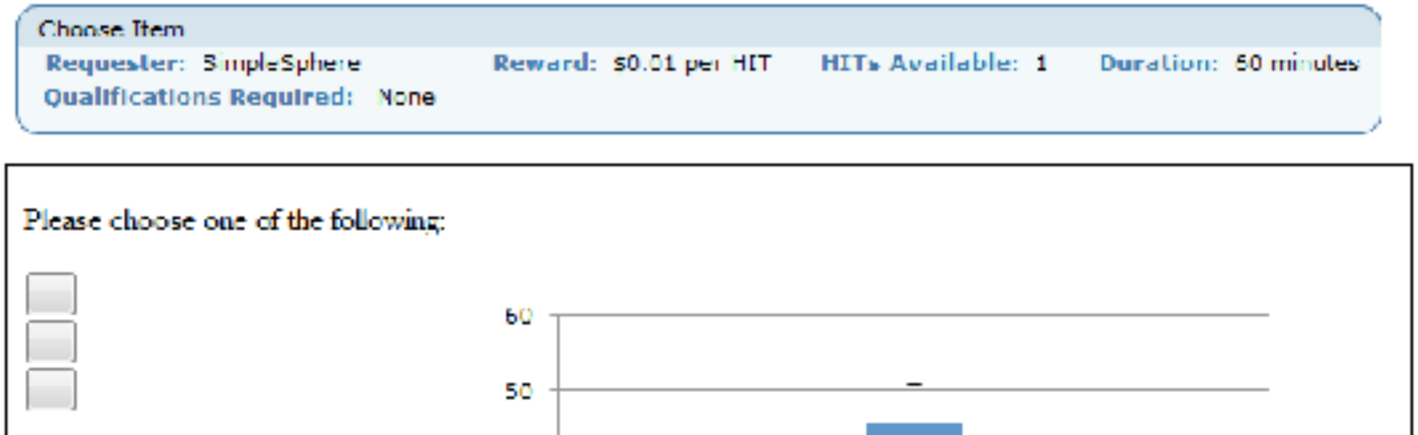
Turkers were offered 1 cent to pick a number from 1 to 10.

~850 turkers

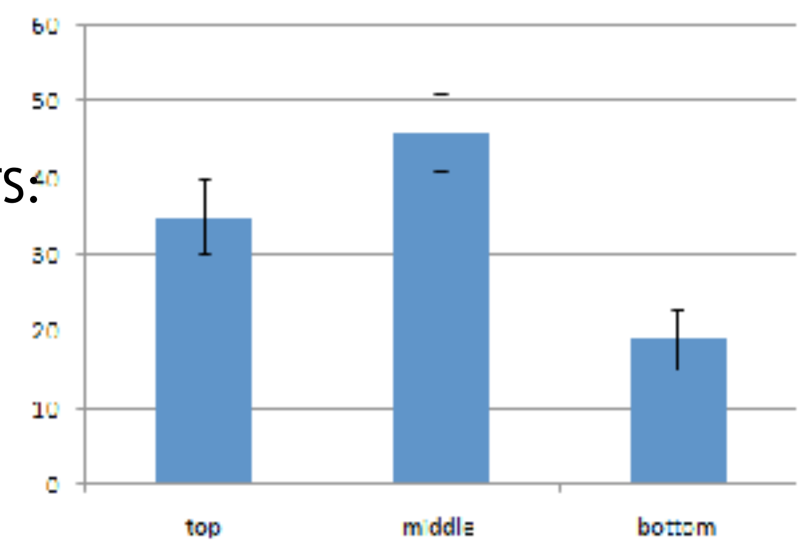


Experiment by Greg Little
From http://groups.csail.mit.edu/uid/deneme/

Do humans have consistent biases?

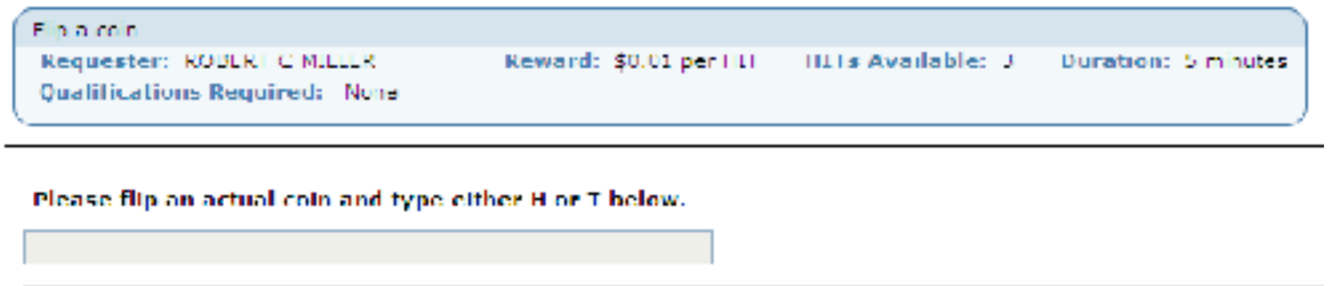


Results form 100 HITS:

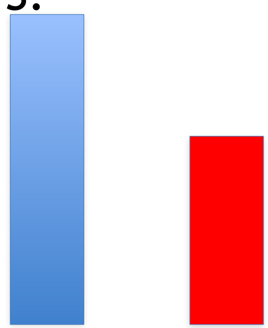


Experiment by Greg Little
From http://groups.csail.mit.edu/uid/deneme/

Do humans do what you ask for?

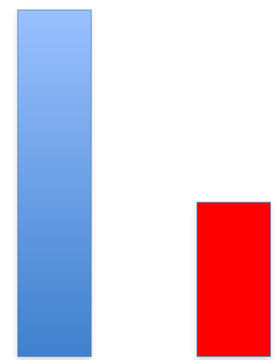


After 50 HITS:



31 heads, 19 tails

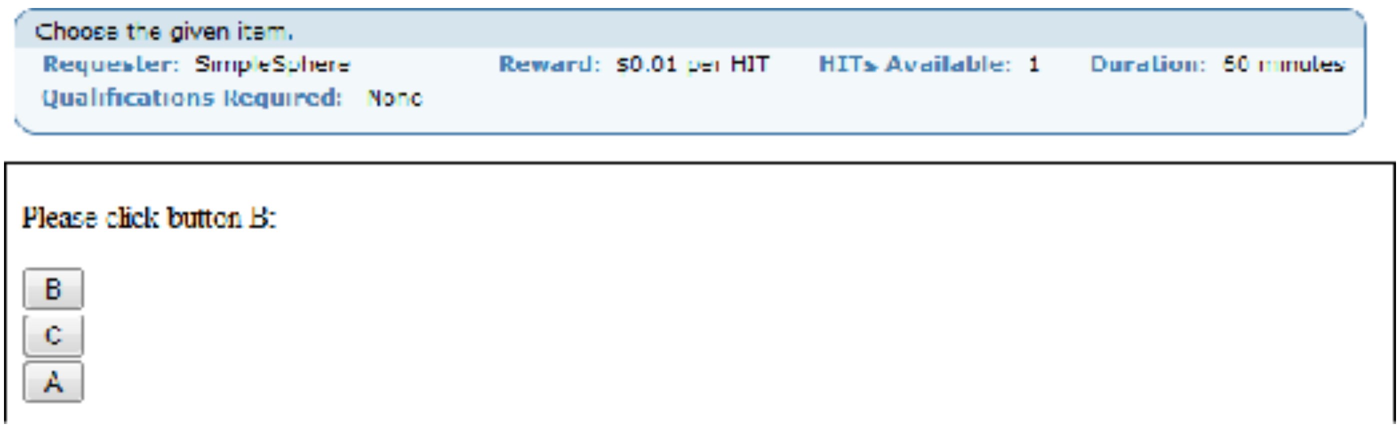
And 50 more:



34 heads, 16 tails

Experiment by Rob Miller
From http://groups.csail.mit.edu/uid/deneme/

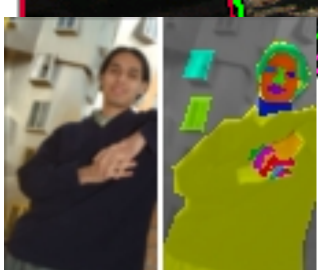
Are humans reliable even in simple tasks?



Results of 100 HITS

A: 2
B: 96
C: 2

Experiment by Greg Little
From http://groups.csail.mit.edu/uid/deneme/



Tool went online July 1st, 2005

Labelme.csail.mit.edu

B. Russell, A. Torralba, K. Murphy, W.T. Freeman. IJCV 2008

With Bryan Russell

1 cent

Task: Label one object in this image

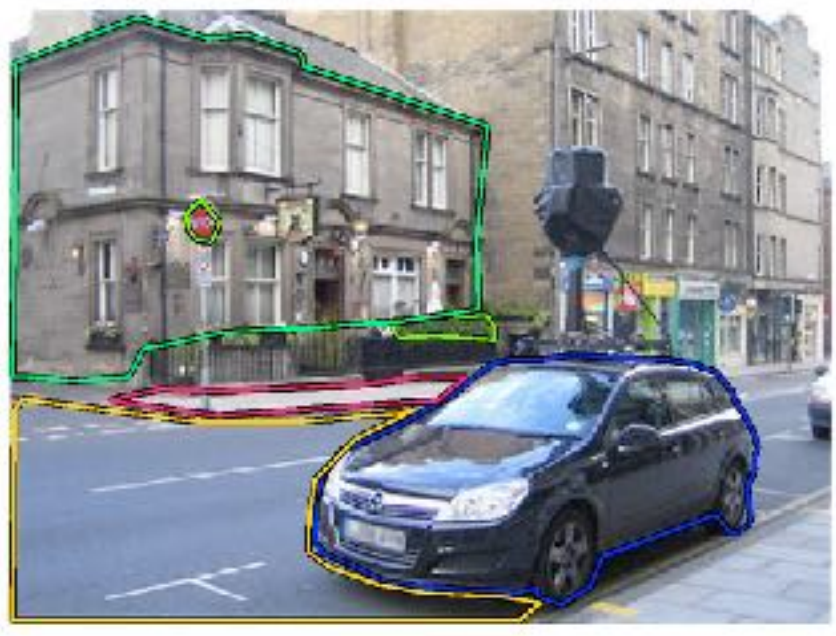

car
building
building
lampost
planter box
This is a window.
This is the street.
This is a balcony.
door
entrance
Traffic sign
SKY
cloud
arch
street light

LabelMe iterations

1) Label as many objects as you can
2) Delete any wrong polygon
3) Go to 1

1

Label some objects

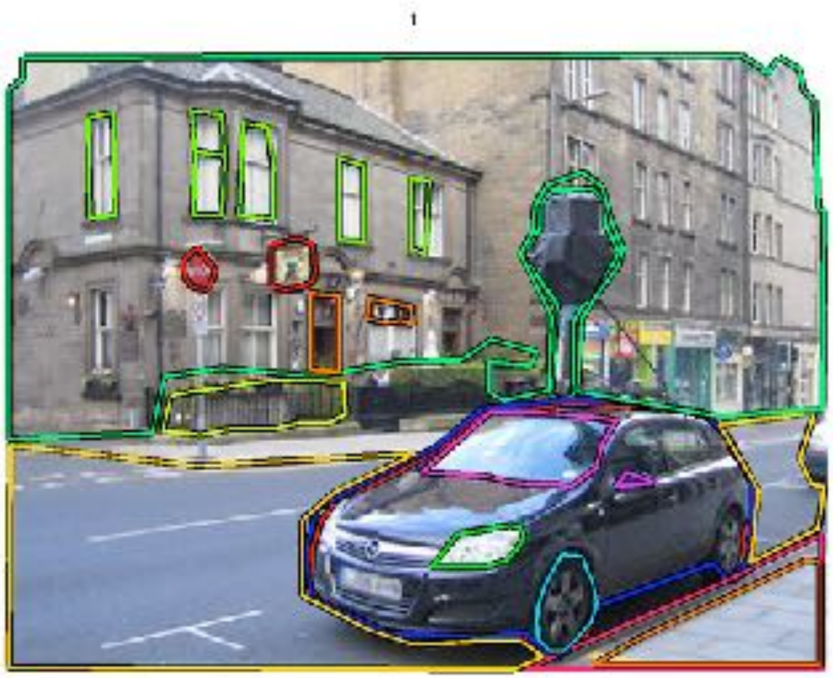


Delete any wrong polygons

Label some objects

Delete any wrong polygons

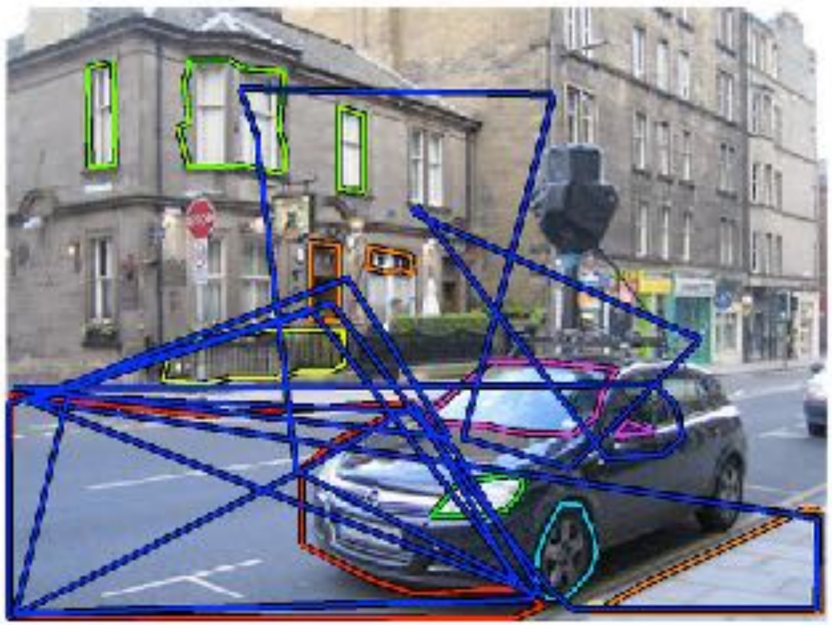
Label some objects



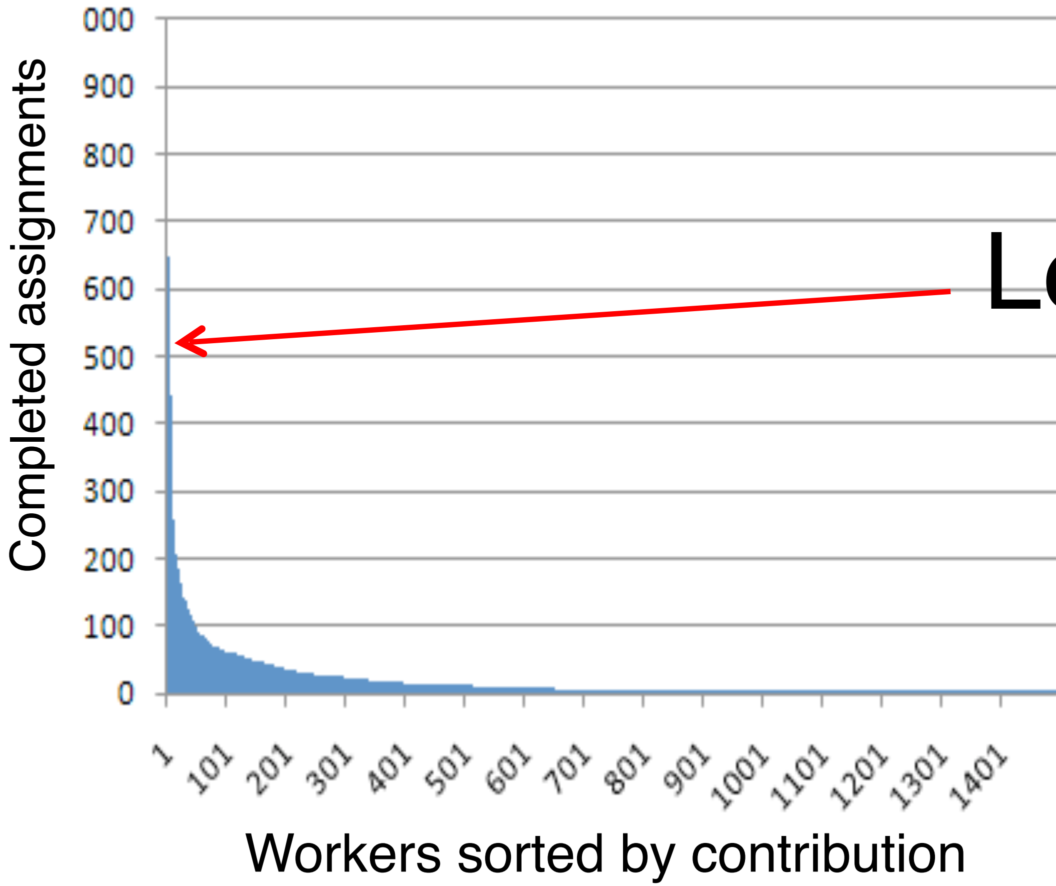
Delete any wrong polygons

Label some objects

1

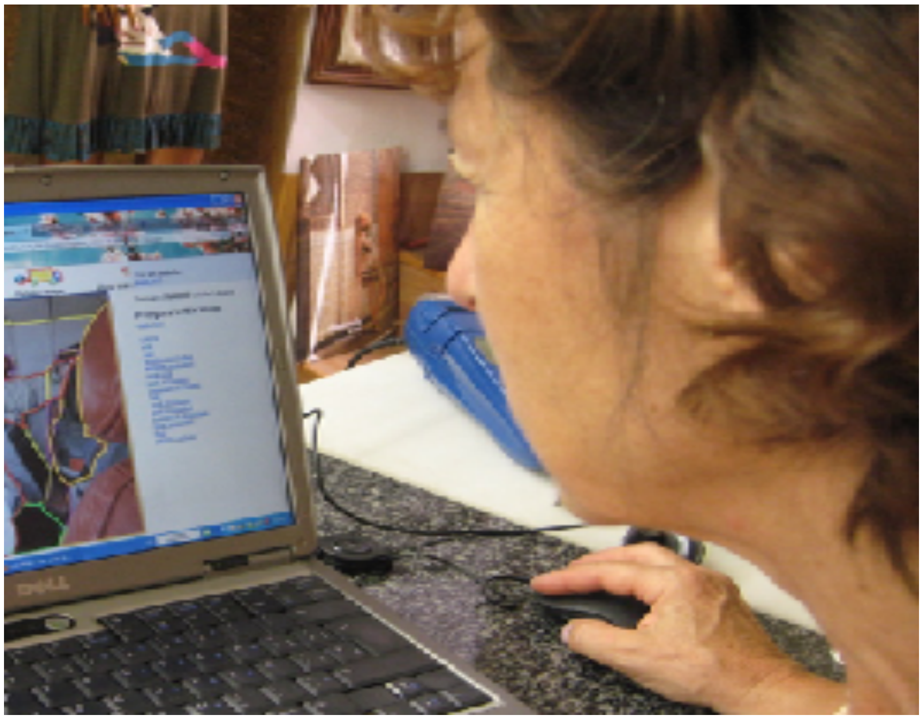


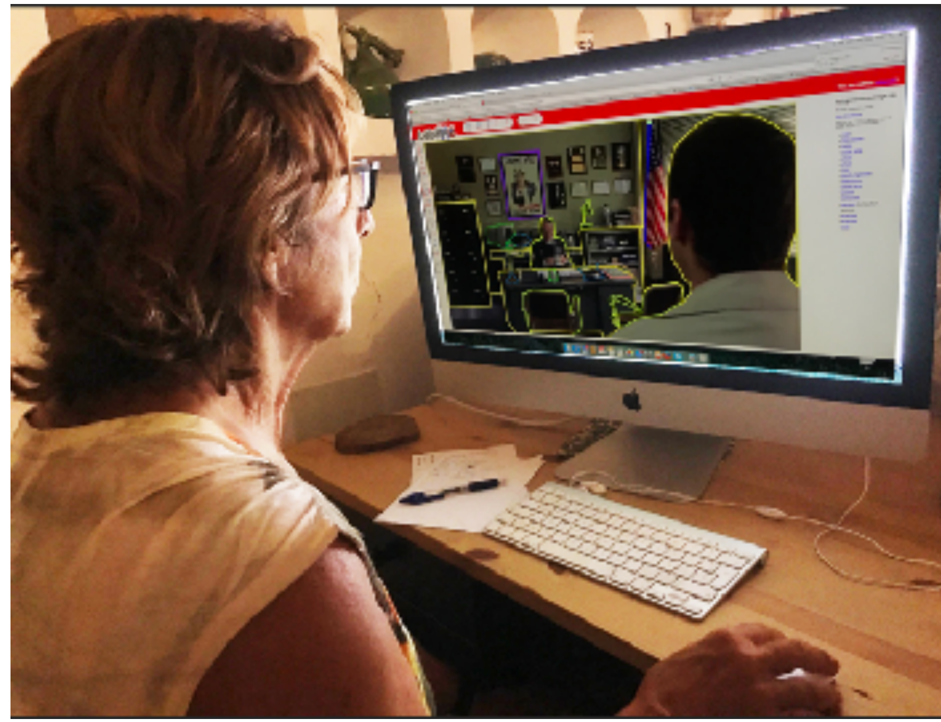
Who does the work?

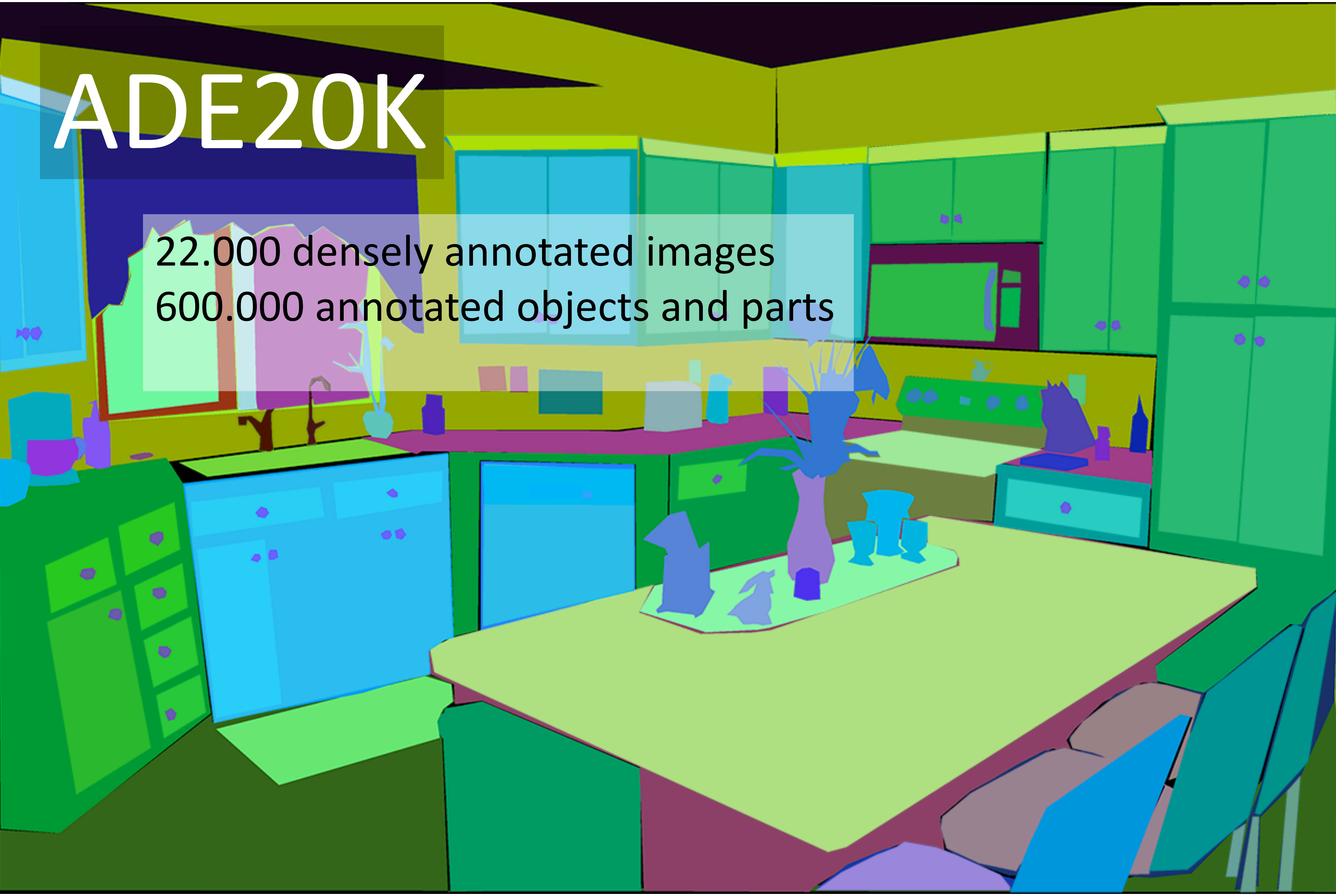


Let's hire that

From http://groups.csail.mit.edu/uid/deneme/






ADE20K
22.000 densely annotated images
600.000 annotated objects and parts

COCO ADE20K

	Images	Obj. inst.	Obj. classes	Part inst.	Part classes	Obj. classes per image
COCO	123,287	886,284	91	0	0	3.5
ImageNet*	476,688	534,309	200	0	0	1.7
NYU Depth V2	1,449	34,064	894	0	0	14.1
Cityscapes	25,000	N/A	30	0	0	N/A
SUN	16,873	313,884	4,479	0	0	9.8
OpenSurfaces	22,214	71,460	160	0	0	N/A
PascalContext	10,103	~104,398**	540	181,770	40	5.1
ADE20K	22,000	415,099	2,944	171,148	354	10.5

* has only bounding boxes (no pixel-level segmentation). Sparse annotations.

** PascalContext dataset does not have instance segmentation. In order to estimate the number of instances, we find connected components (having at least 150pixels) for each class label.


https://www.youtube.com/watch?v=AIEsakeXvMM
THE
SNUGGL
EAL

https://www.youtube.com/watch?v=AIEeakeXv

Cross modal learning
text and images

Two man sitting behind a long table.

Inspired from COCO captio

Q: Is everyone of these two holding a wine glass? A: No

Q: How many people are there? A: 2

Q: How many are awake? A: 1

Inspired from http://visualqa.org/index.ht

Story-like Description

Pietro had a long day of talks at the workshop. At the end of the session, Pietro was invited to participate in a panel. As he is a mature and confident professor, he decided to take a short nap during the discussion. The chair was comfortable. Nobody dared to wake him up as there were other less confident professors at the panel that could answer the questions. …

“Pictures and words”

- Barnard, Duygulu, de Freitas, Forsyth, Blei, Jordan, Matching words and pictures, JMLR, 2003
- Duygulu, Barnard, de Freitas, Forsyth, Object Recognition as Machine Translation: Learning a lexicon for a fixed image vocabulary , ECCV, 2003
- Blei & Jordan, Modeling annotated data, ACM SIGIR, 2003
- Chang, Goh, Sychay, & Wu, *Soft* annotation using Bayes point machines, IEEE Transactions on Circuits and Systems for Video Technology, 2003
- Goh, Chang, & Cheng, Ensemble of SVM-based classifiers for annotation, 2003
- ….

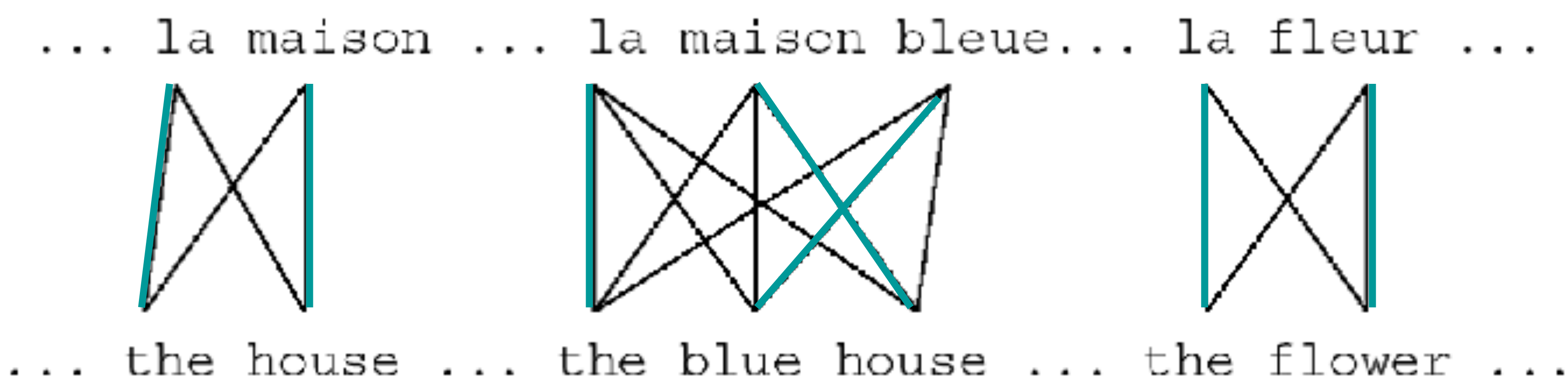
Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary

P. Duygulu[1] , K. Barnard[1] , J.F.G. de Freitas[2] and D.A. Forsyth[1]

Computer Science Division, U.C. Berkeley, Berkeley, CA 94720
Department of Computer Science, University of British Columbia,Vancouver
{duygulu, kobus, daf}@cs.berkeley.edu, nando@cs.ubc.ca

Statistical Machine Translation

- Statistically link words in one language to words in another
- Requires aligned bitext
  - eg. Hansard for Canadian parliament

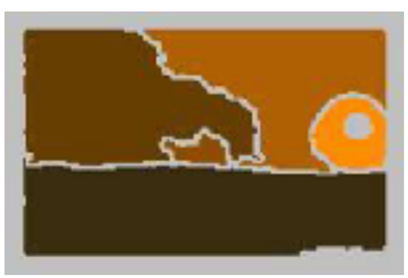


Slide: H Dunlop

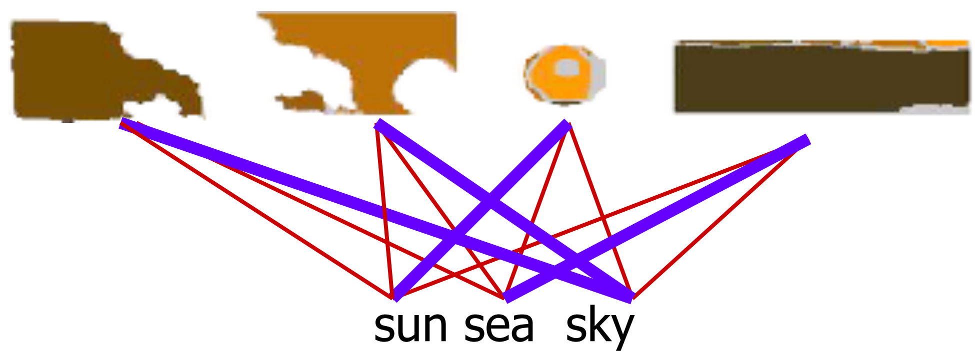
Multimedia Translation

- Data:

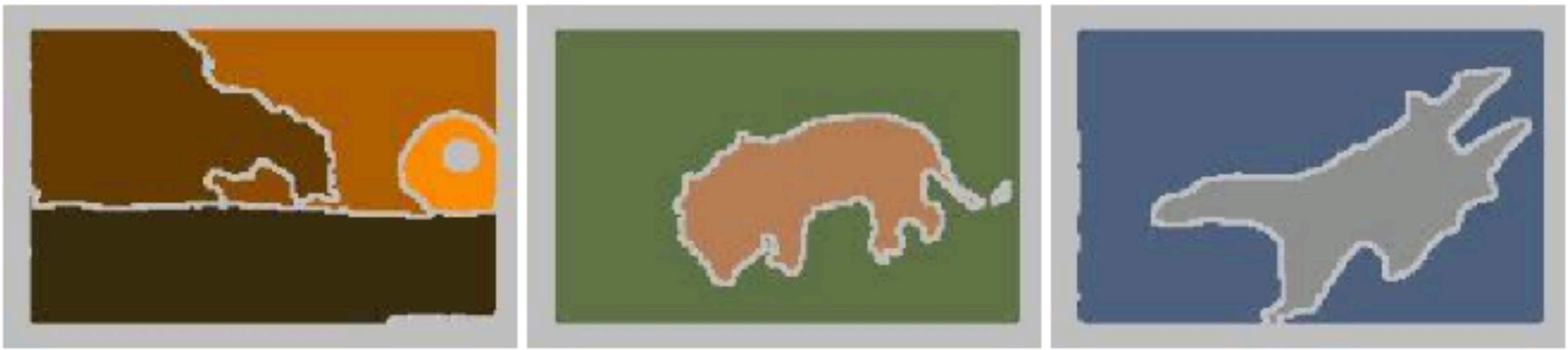
  - Words are associated with images, but correspondences are unknown



sun sea sky



Slide: H Dunlop



sea sky sun waves   cat forest grass tiger   jet plane sky

Fig. 1. *Examples from the Corel data set. We have associated keywords and segments for each image, but we don't know which word corresponds to which segment. The number of words and segments can be different; even when they are same, we may have more than one segment for a single word, or more than one word for a single blob. We try to align the words and segments, so that for example an orange stripy blob will correspond to the word tiger.*

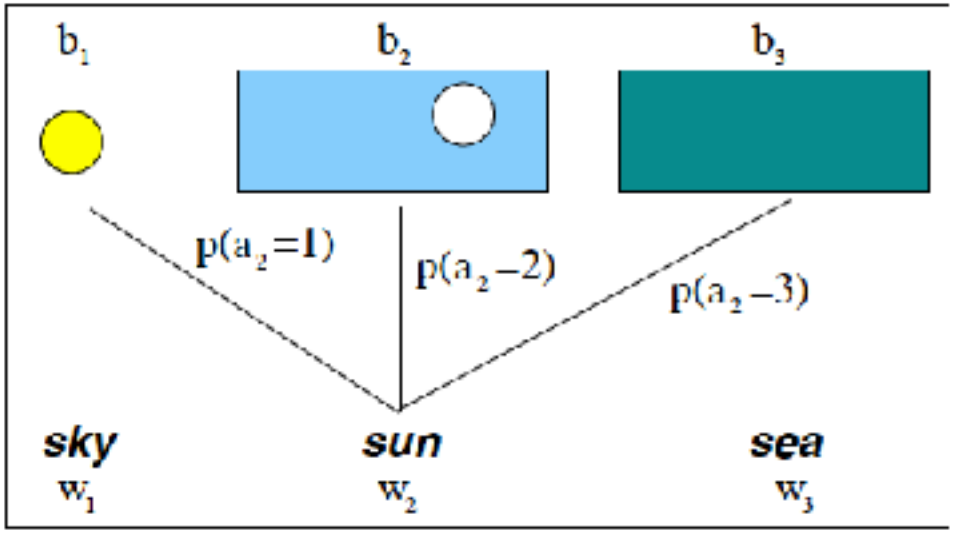


Fig. 3. *Example : Each word is predicted with some probability by each blob, meaning that we have a mixture model for each word. The association probabilities provide the correspondences (assignments) between each word and the various image segments. Assume that these assignments are known; then computing the mixture model is a matter of counting. Similarly, assume that the association probabilities are known; then the correspondences can be predicted. This means that EM is an appropriate estimation algorithm.*

Fig. 8. *Some examples of the labelling results. The words overlaid on the images are the words predicted with top probability for corresponding blob. We are very successful in predicting words like* `sky`, `tree` *and* `grass` *which have high recall. Sometimes, the words are correct but not in the right place like* `tree` *and* `buildings` *in the center image.*

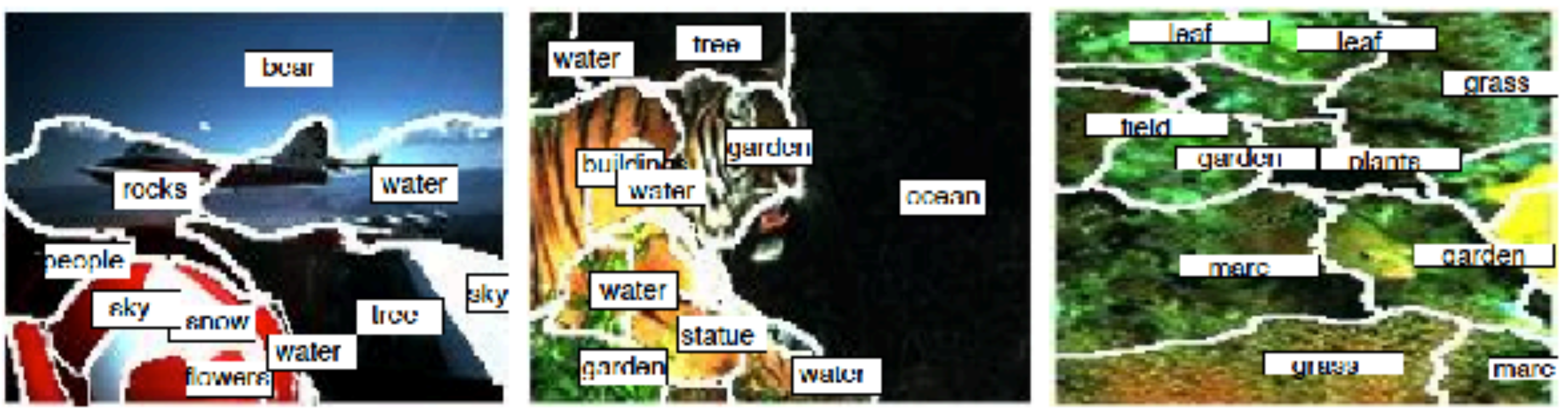
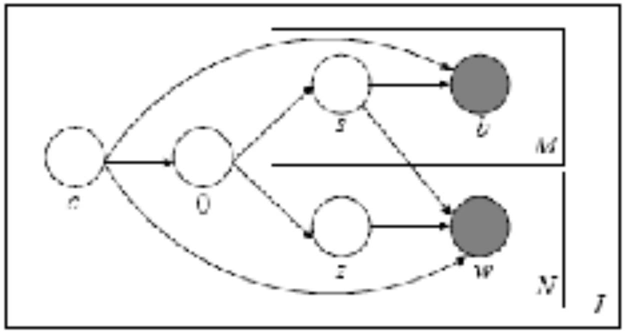
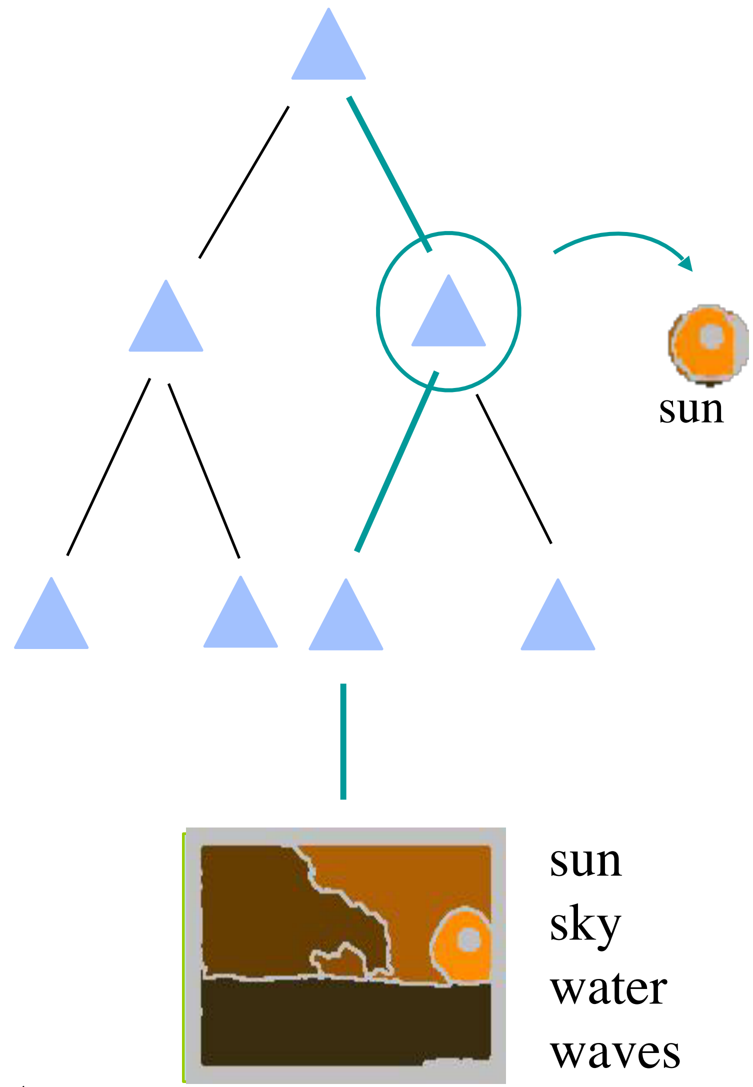


Fig. 9. *Some test results which are not satisfactory. Words that are wrongly predicted are the ones with very low recall values. The problem mostly seen in the third image is since green blobs coocur mostly with* `grass`, `plants` *or* `leaf` *rather than the under water plants.*



- A generative model for assembling image data sets from multimodal clusters
  - Chose an image cluster by p(c)
  - Chose multimodal concept clusters using p(s|c)
  - From each multimodal cluster, sample a Gaussian for blob features, p(b|s), and a multinomial for words, p(w|s)
  - (Skip with some probability to account for mismatched numbers of words and blobs)
  - For a given correspondence*

$$p(\{w \Leftrightarrow b\}) = \sum_{c} p(c) \prod_{\{w \Leftrightarrow b\}} \left(\sum_{l} p(w \mid l) p(b \mid l) p(l \mid c) \right)$$

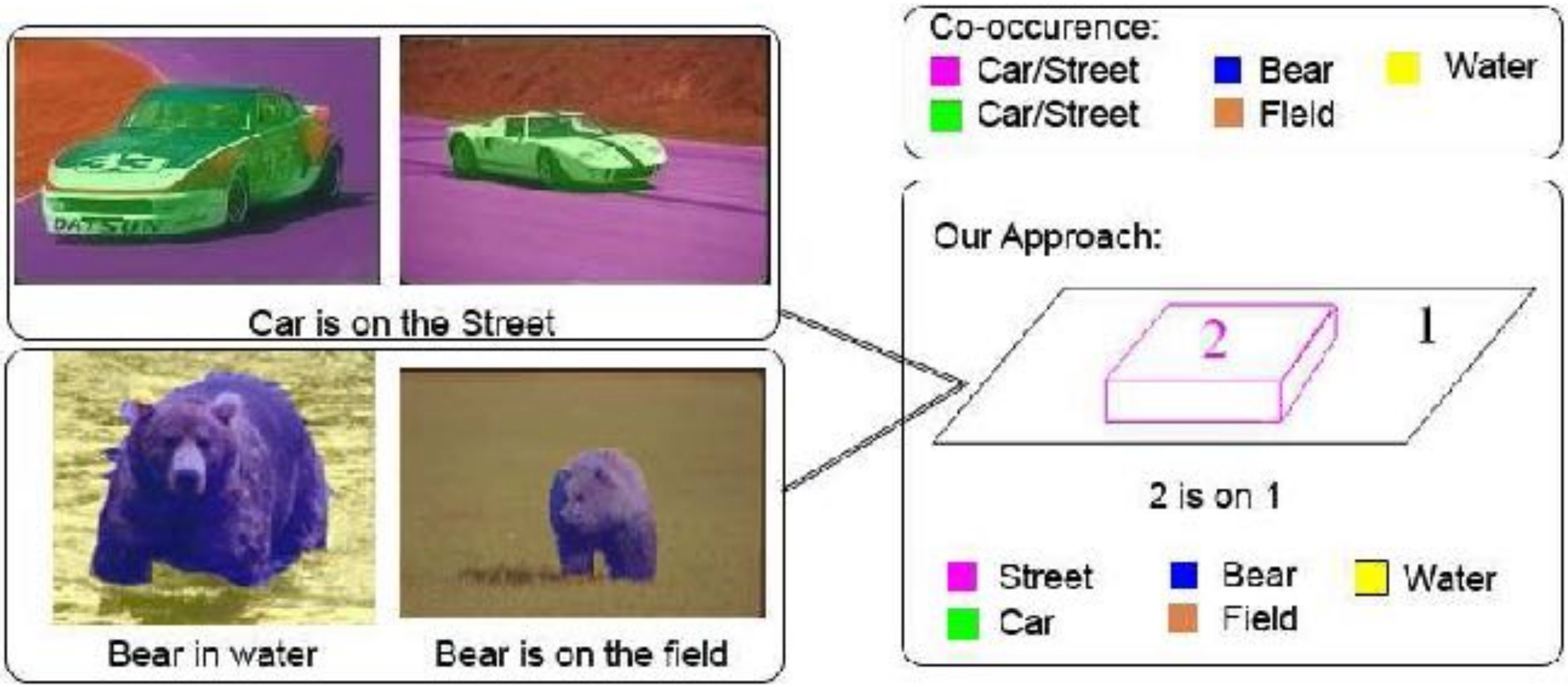


Slide courtesy of Kobus Barnard

Barnard et al. JMLR, 2005

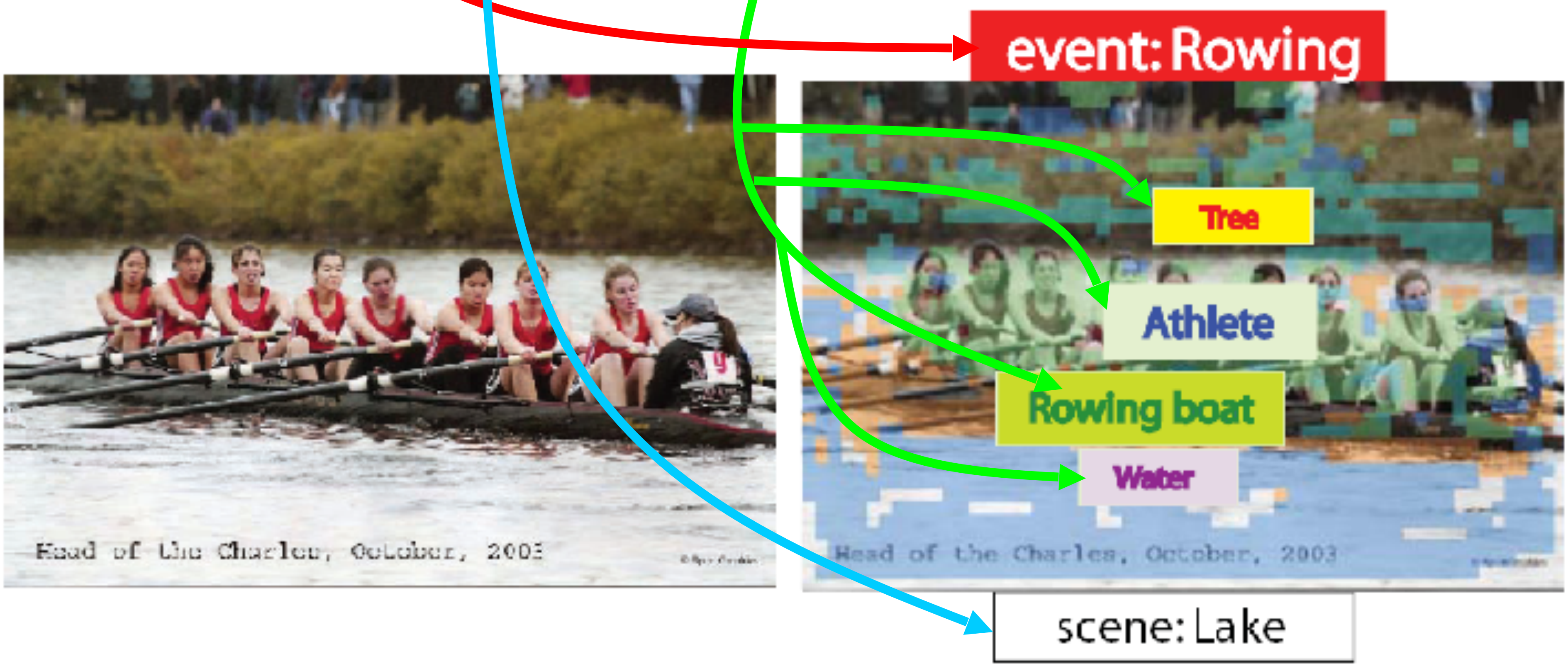
Barnard et al. JMLR, 2005

"Beyond nouns"



Gupta & Davis, EECV, 2008

What, where and who? Classifying events by scene and object recognition

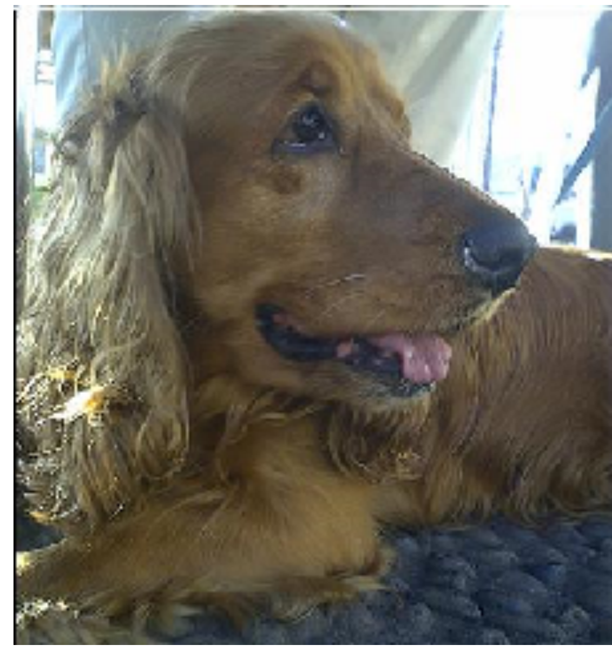


L-J Li & L. Fei-Fei, ICCV 2007

Attribute Examples

Shape:
Part: Head, Ear, Nose, Mouth, Hair, Face, Torso, Hand, Arm
Material: Skin, Cloth

Shape:
Part: Head, Ear, Snout, Eye
Material: Furry



Shape:
Part: Head, Ear, Snout, Eye, Torso, Leg
Material: Furry

Slide: Derek Hoiem

"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."

"boy is doing backflip on wakeboard."

"girl in pink dress is jumping in air."

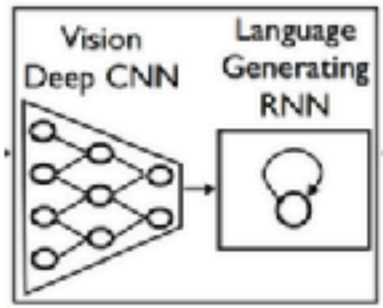
"black and white dog jumps over bar."

"young girl in pink shirt is swinging on swing."

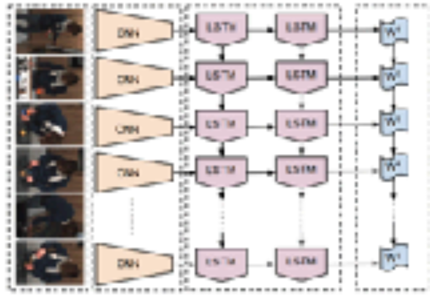
"man in blue wetsuit is surfing on wave."

Slides Credits: Andrej Karpathy, FeiFei Li

Image captioning is receiving a lot of attention



Vinyals et al., 2015

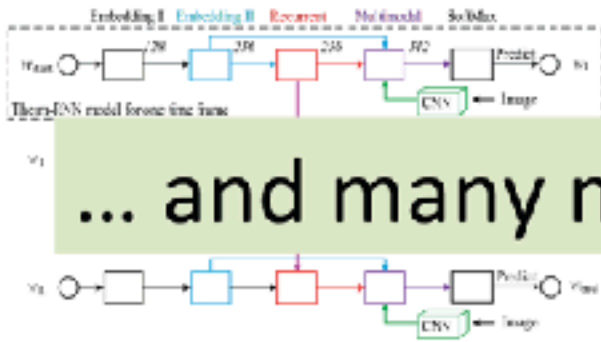


Donahue et al., 2015

Karpathy and Fei-Fei, 2015

Hodosh et al., 2013

Fang et al., 2015

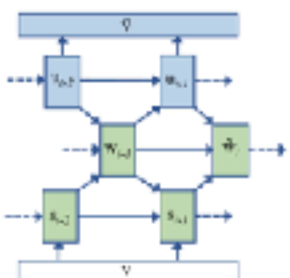


Mao et al., 2015

... and many more

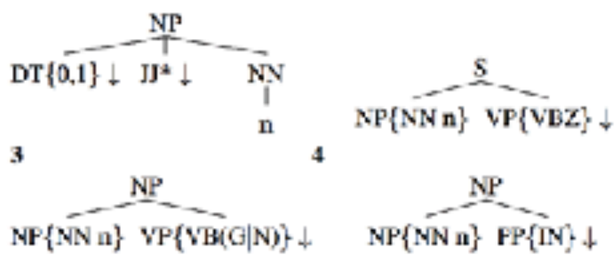
Ordonez et al., 2011

Kulkarni et al., 2011

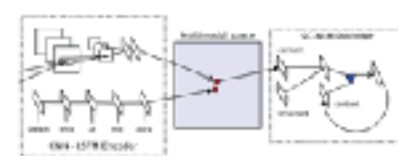


Chen and Zitnick, 2015

Farhadi et al., 2010



Mitchell et al., 2012



Kiros et al., 2015

Neural Image Caption (NIC) (CVPR 2015)

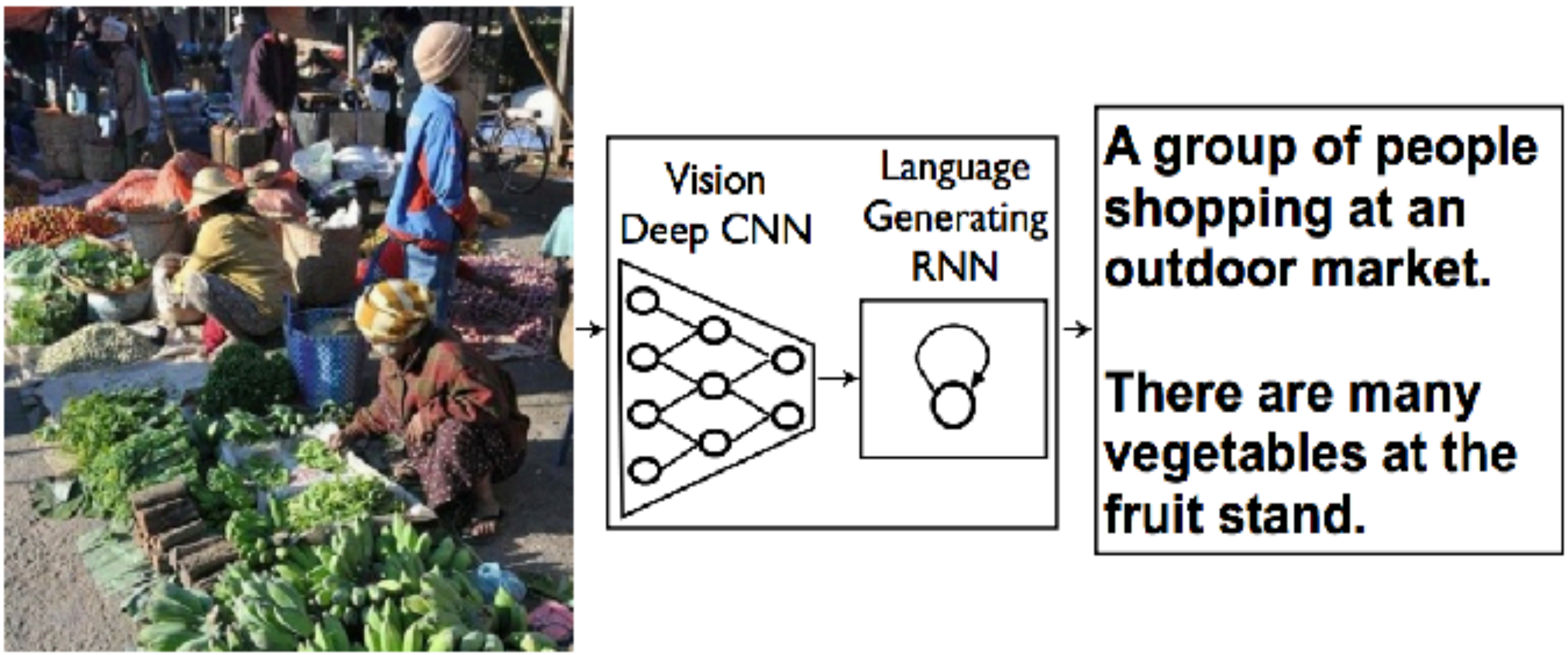
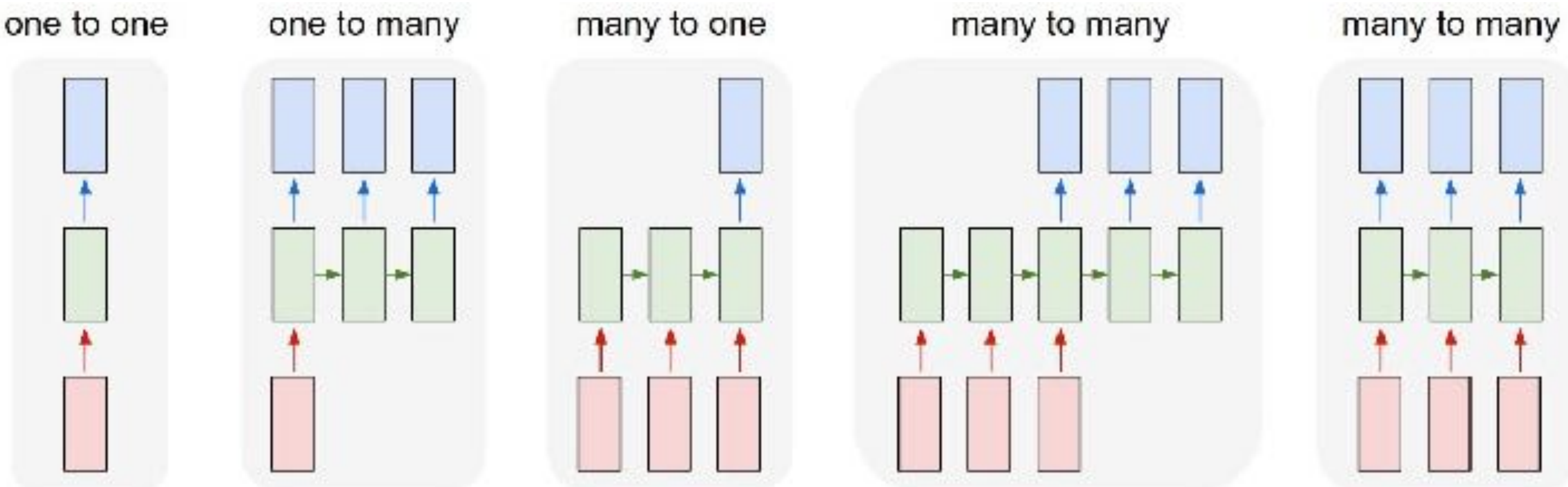


Figure Credits: Show and Tell: A Neural Image Caption Generator

How do we model sequences?



one to one

Input: No sequence

Output: No sequence

Example: "standard" classification / regression problems

one to many

Input: No sequence

Output: Sequence

Example: Im2Caption

many to one

Input: Sequence

Output: No sequence

Example: sentence classification, multiple-choice question answering

many to many

Input: Sequence

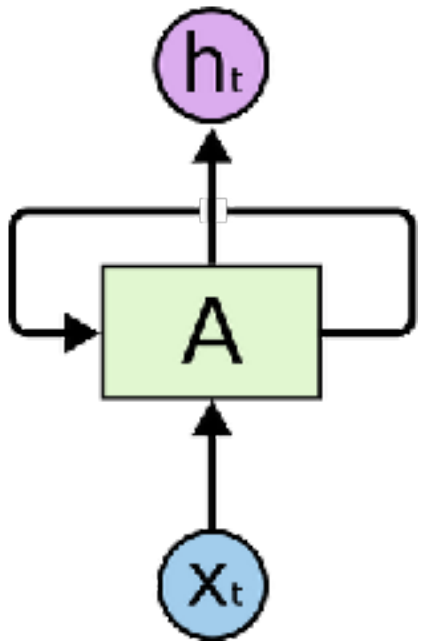
Output: Sequence

Example: machine translation, video captioning, open-ended question answering, video question answering

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Credit: Dhruv Batra, Andrej Karpathy

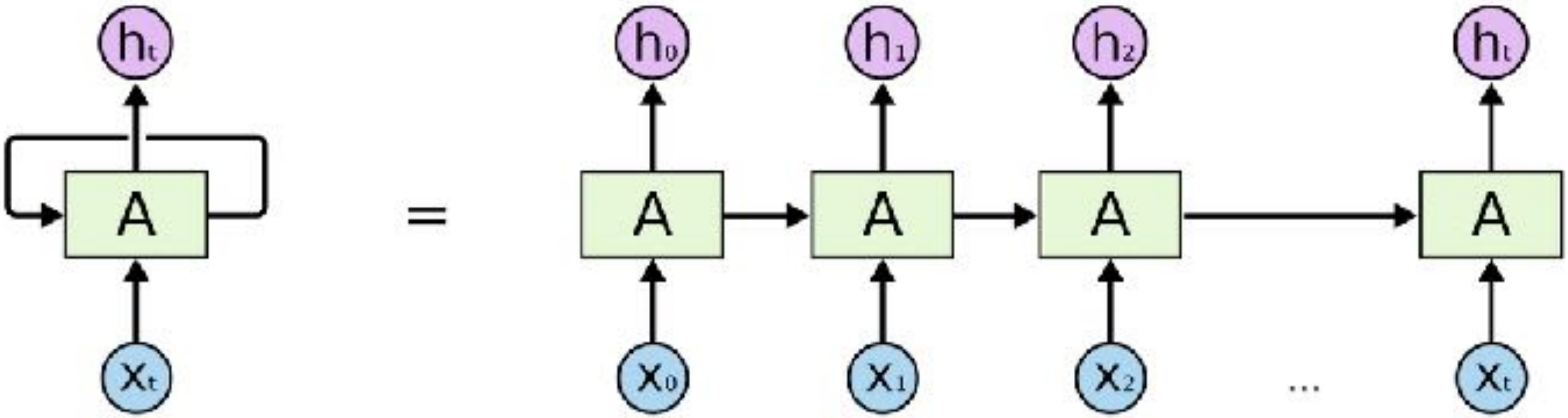
Recurrent Neural Networks (RNNs)



In the above diagram, a chunk of neural network, A, looks at some input x_i and outputs a value h_i. A loop allows information to be passed from one step of the network to the next.

Credit: Christopher Olah

Recurrent Neural Networks (RNNs)

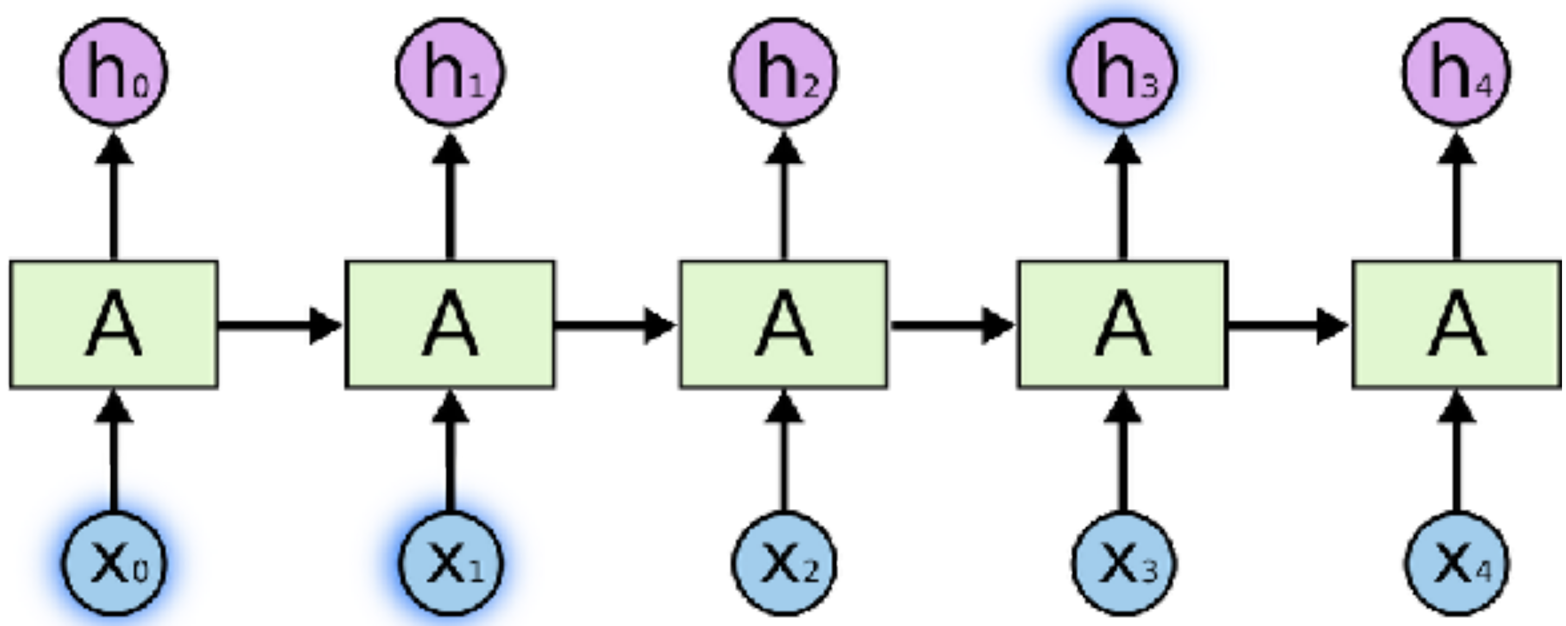


An unrolled recurrent neural network.

A recurrent neural network can be thought of as multiple copies of the same network, each passing a message to a successor

Credit: Christopher Olah

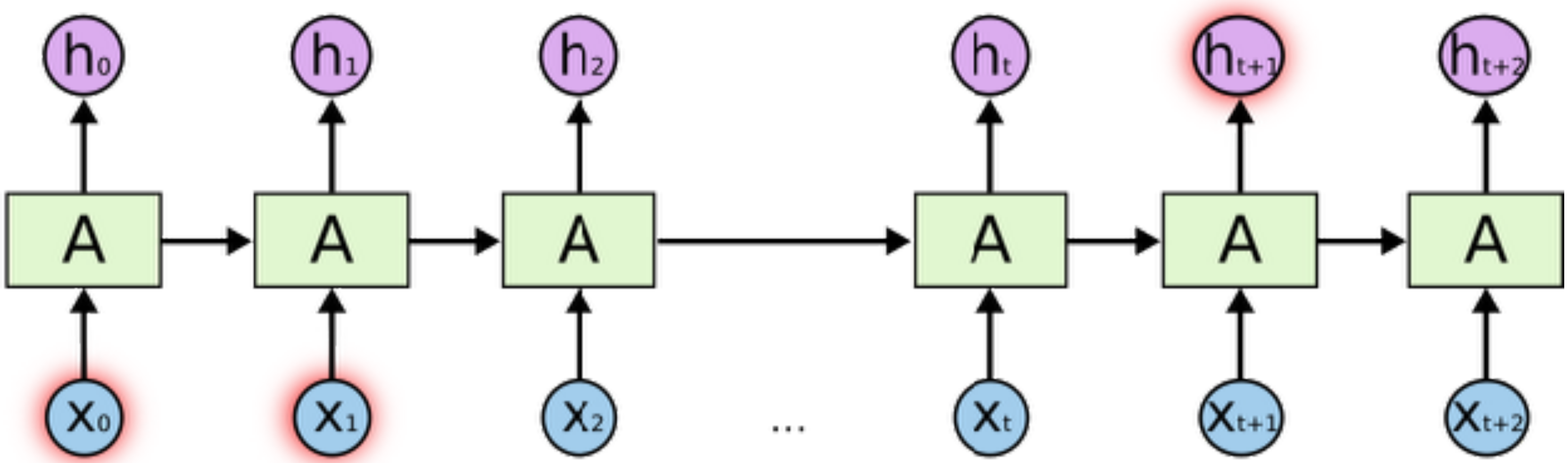
Recurrent Neural Networks (RNNs)



When the gap between the relevant information and the place that it's needed is small, RNNs can learn to use the past information

Credit: Christopher Olah

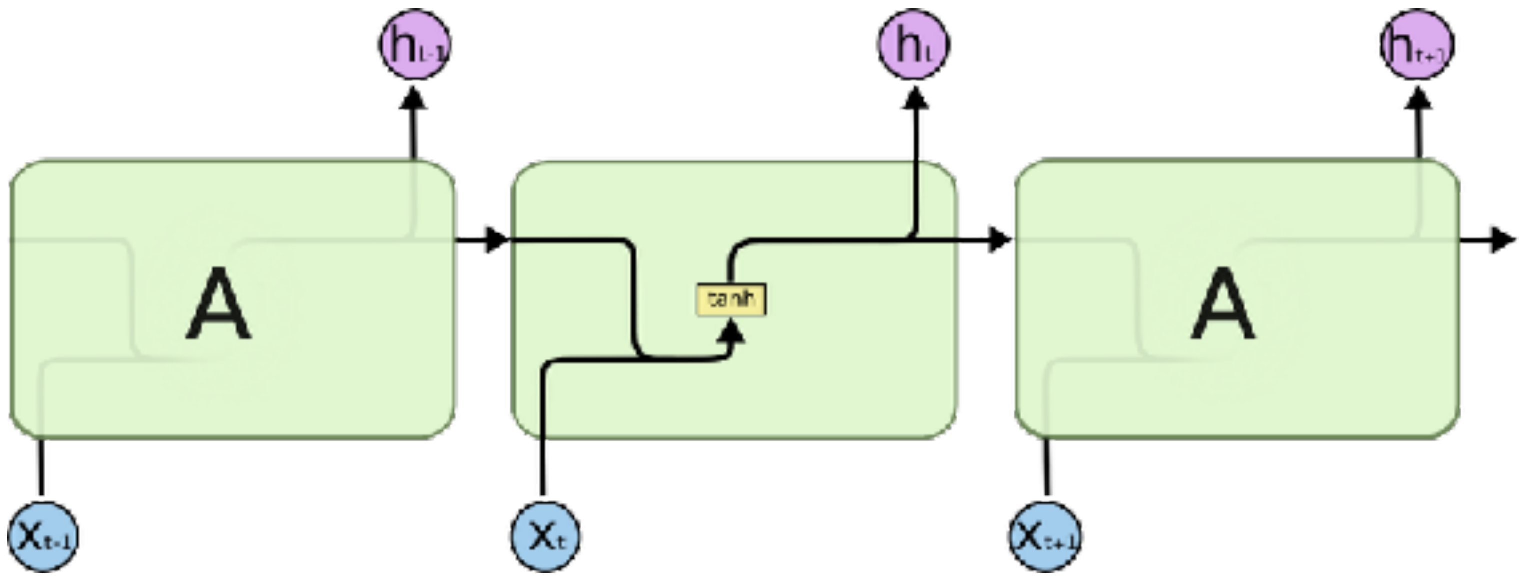
Long-term dependencies - hard to model!



But there are also cases where we need more context.

Credit: Christopher Olah

From plain RNNs to LSTMs

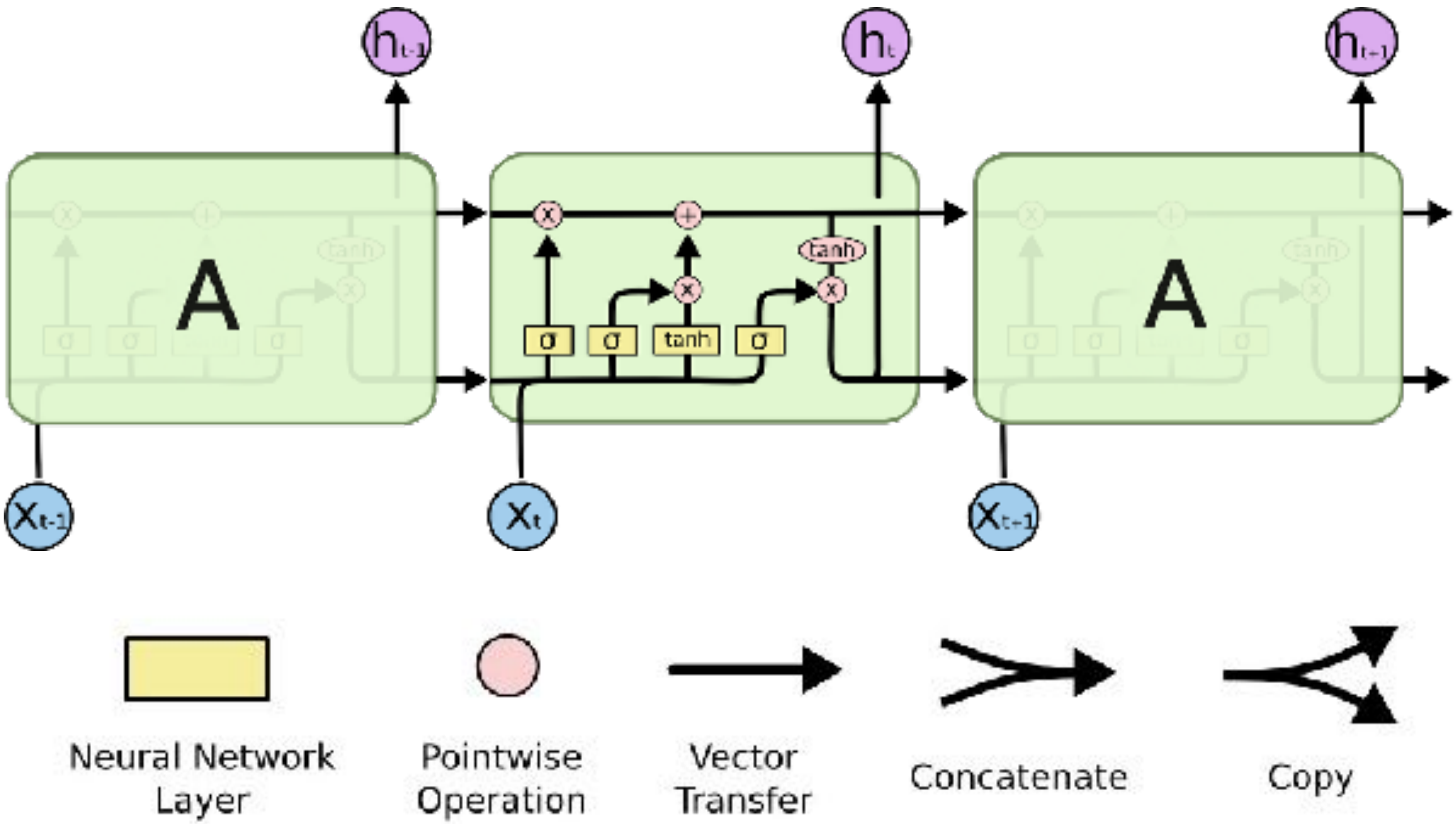


(LSTM: Long Short Term Memory Networks)

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Credit: Christopher Olah

From plain RNNs to LSTMs

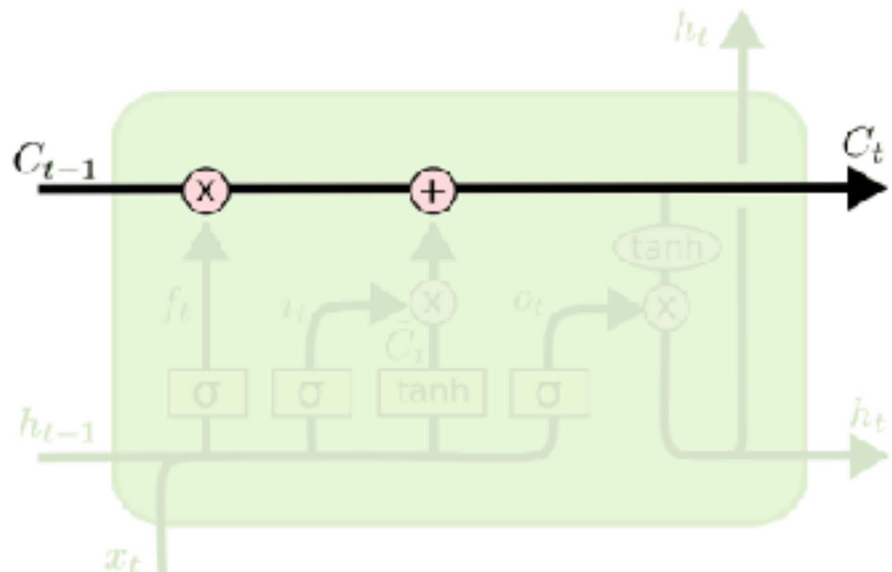


(LSTM: Long Short Term Memory Networks)

Credit: Christopher Olah

LSTMs Step by Step: Memory

Cell State / Memory

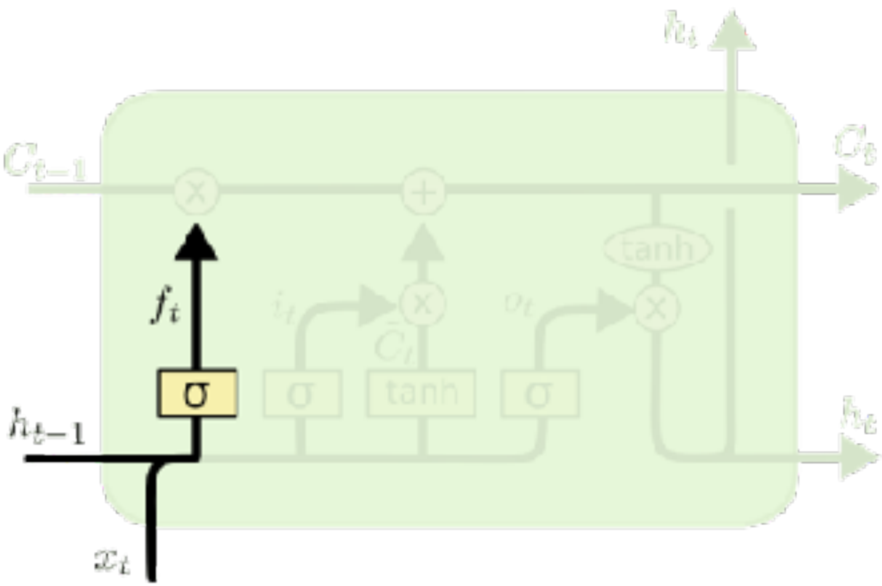


The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates

Credit: Christopher Olah

LSTMs Step by Step: Forget Gate

Should we continue to remember this "bit" of information or not?



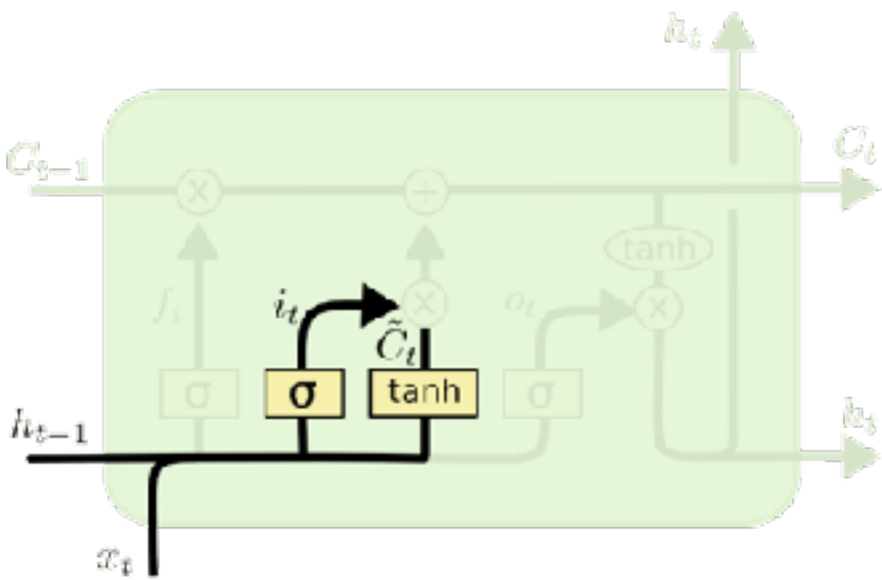
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

The first step in our LSTM is to decide what information we're going to throw away from the cell state. This decision is made by a sigmoid layer called the "forget gate layer."

Credit: Christopher Olah

LSTMs Step by Step: Input Gate

Should we update this “bit” of information or not? If so, with what?



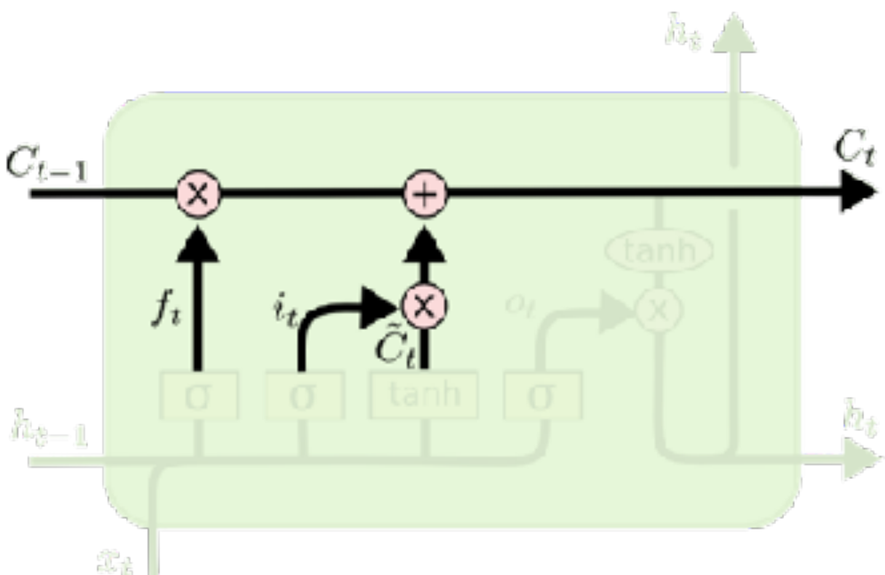
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

The next step is to decide what new information we’re going to store in the cell state. This has two parts. First, a sigmoid layer called the “input gate layer” decides which values we’ll update. Next, a tanh layer creates a vector of new candidate values, $\tilde{C}_t$, that could be added to the state.

Credit: Christopher Olah

LSTMs Step by Step: Memory Update

Decide what will be kept in the cell state/memory



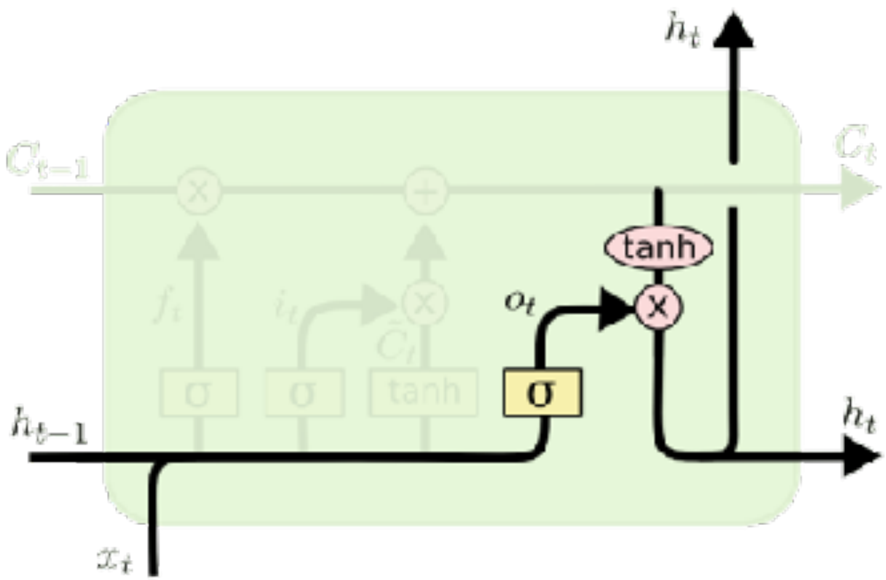
Forget that **Memorize this**

$$C_t = \underbrace{f_t * C_{t-1}}_{\text{Forget that}} + \underbrace{i_t * \tilde{C}_t}_{\text{Memorize this}}$$

Credit: Christopher Olah

LSTMs Step by Step: Output Gate

Should we output this “bit” of information?

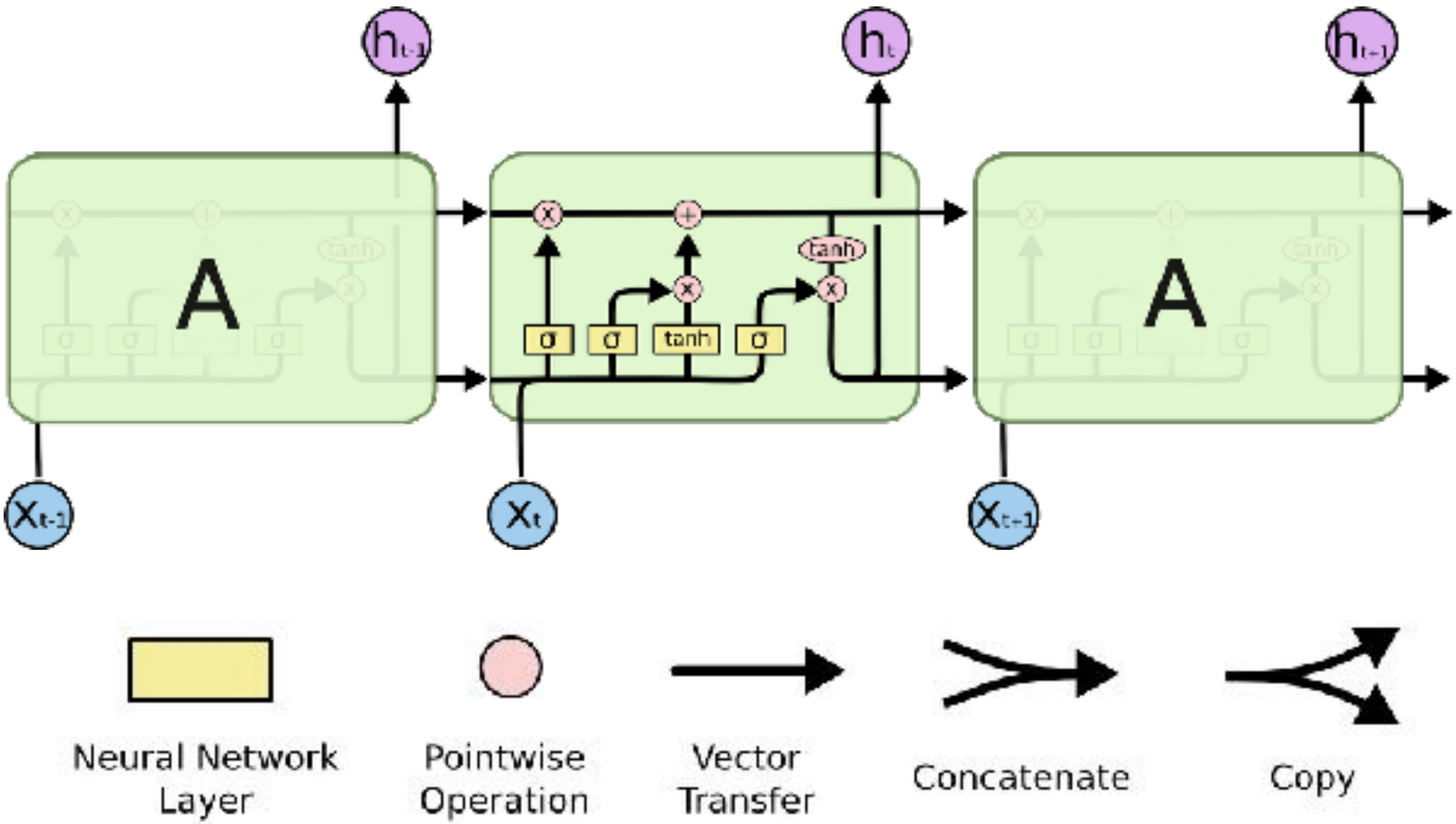


$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

This output will be based on our cell state, but will be a filtered version. First, we run a sigmoid layer which decides what parts of the cell state we’re going to output. Then, we put the cell state through tanh (to push the values to be between −1 and 1) and multiply it by the output of the sigmoid gate, so that we only output the parts we decided to.

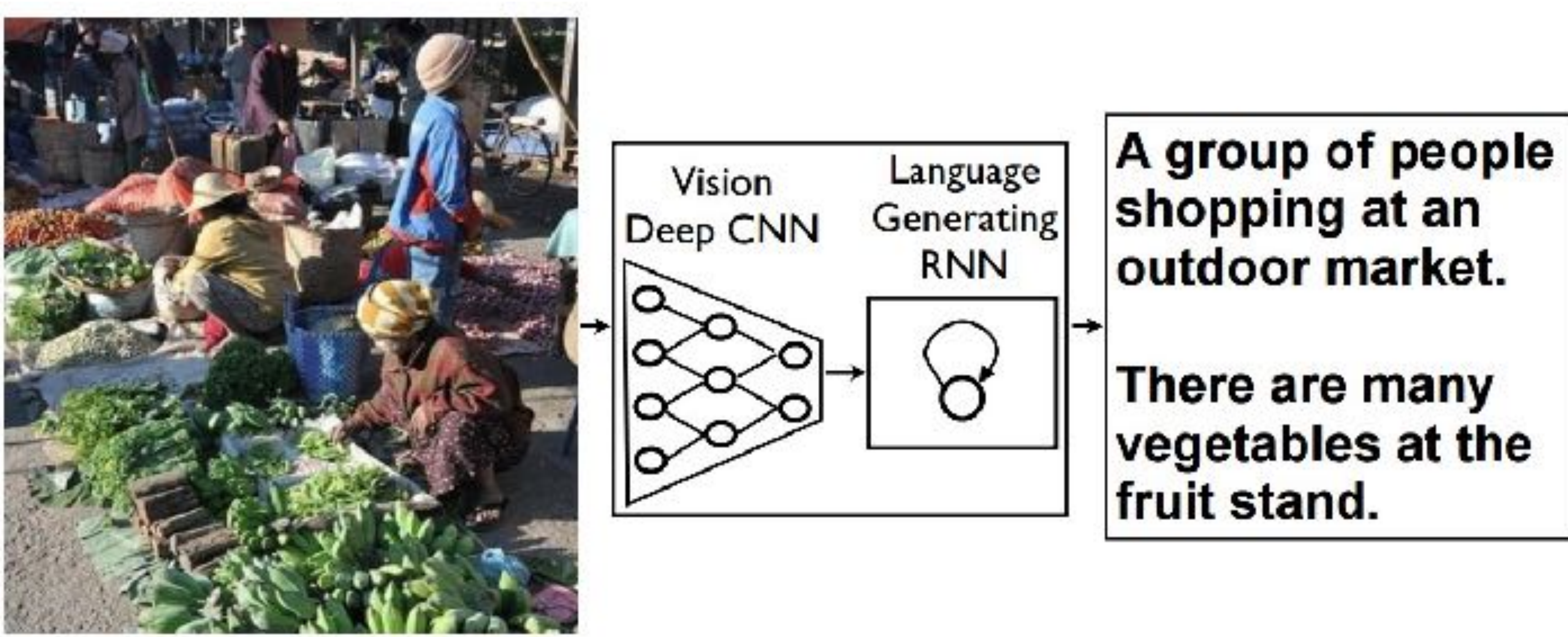
Credit: Christopher Olah

Complete LSTM - A pretty sophisticated cell



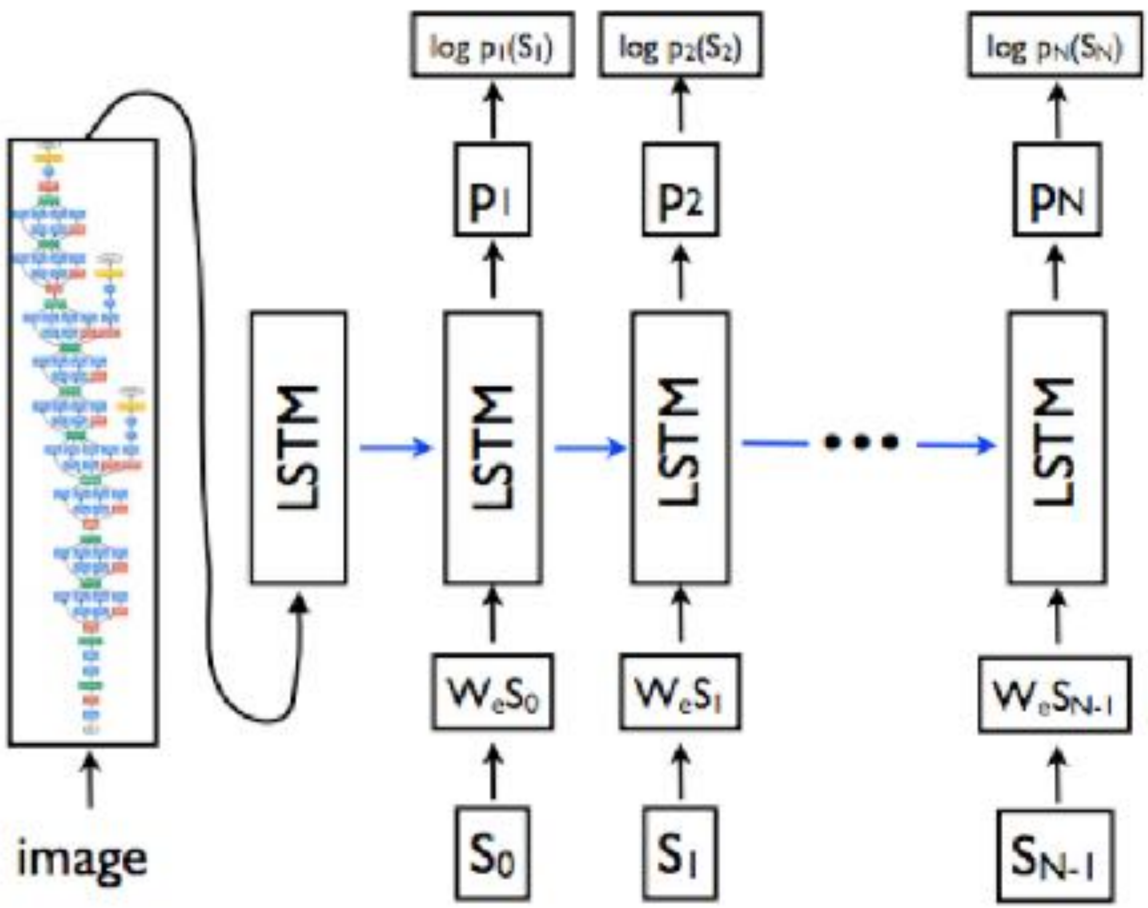
Credit: Christopher Olah

Show and Tell: A Neural Image Caption Generator



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

Show and Tell: A Neural Image Caption Generator



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

Image Caption Generator Results

https://pdollar.wordpress.com/2015/01/21/image-captioning/

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

Cross-modal learning

Description (eg, Wikipedia article)

Snares penguin

From Wikipedia, the free encyclopedia

The **Snares penguin** (*Eudyptes robustus*), also known as the **Snares crested penguin** and the **Snares Islands penguin**, is a penguin from New Zealand. The species breeds on The Snares, a group of islands off the southern coast of the South Island. This is a medium-small, yellow-crested penguin, at a size of 50–70 cm (19.5–27.5 in) and a weight of 2.5–4 kg (5.5–8.8 lb). It has dark blue-black upperparts and white underparts. It has a bright yellow eyebrow-stripe which extends over the eye to form a drooping, bushy crest. It has bare pink skin at the base of its large red-brown bill.

- Lots of descriptions/entries in Wikipedia available

Images





Zero-shot Learning

Description (eg, Wikipedia article)

Cardinal (bird)

From Wikipedia, the free encyclopedia

This article is about the bird family. For other uses, see Cardinal.

Cardinals, in the family **Cardinalidae**, are passerine birds found in North and South America. They are also known as cardinal-grosbeaks and cardinal-buntings. The South American cardinals in the genus *Paroaria* are placed in another family, the Thraupidae (previously placed in Emberizidae).

Can we predict an image classifier from a description alone?

Assume:

- In training we have access to wiki articles and labeled images
- For test classes we only have wiki articles
- We want to classify a new image (it can belong to any class)

Zero-shot Learning

- Goal: learn to predict an image classifier from a description
- Linear binary 1-vs-all classifier:

$$y_c = w_c^T x$$

- x … image feature vector
- w_c ... classifier weight vector for class c

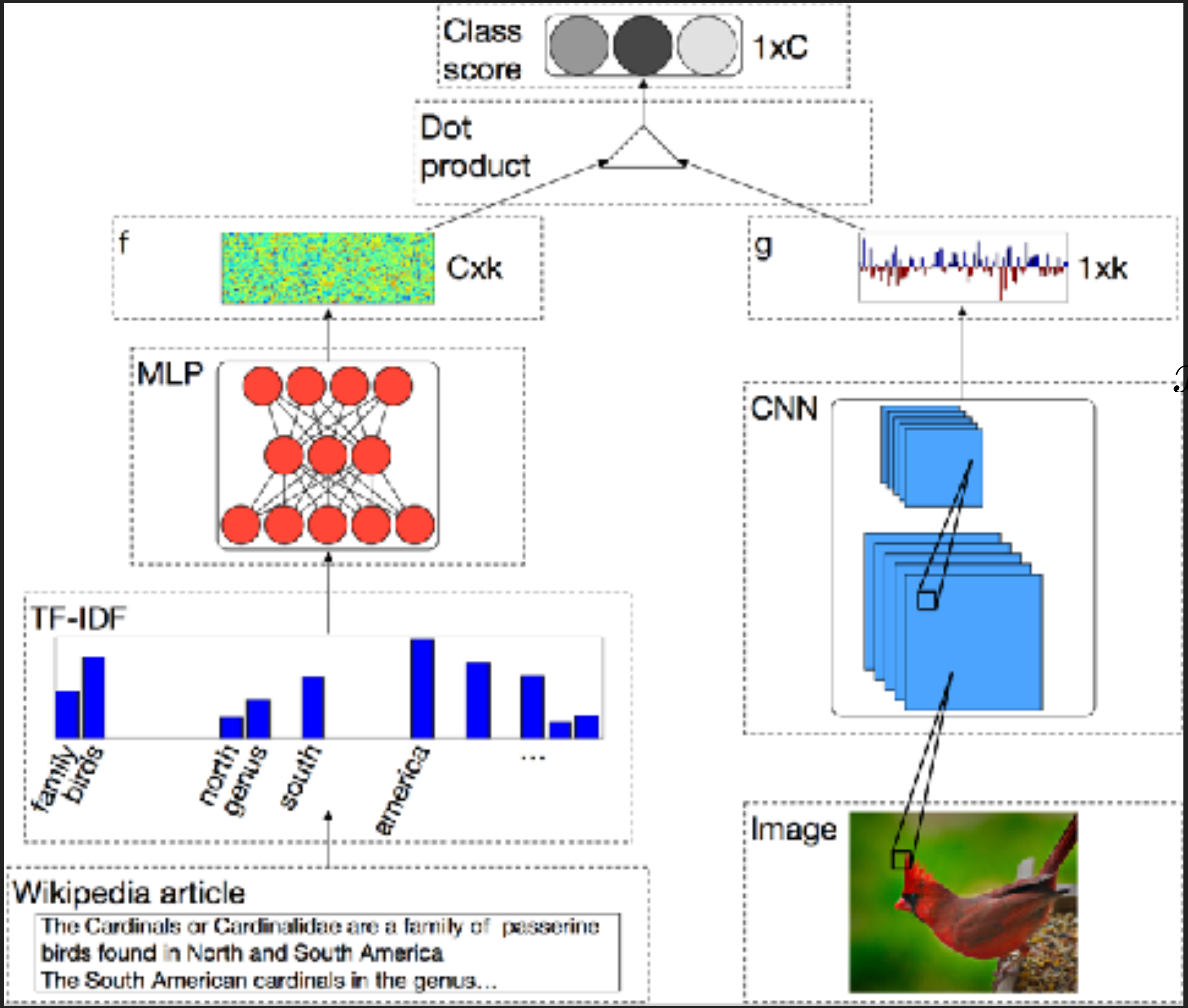
- We are also given t_c, a vector representing a textual description about class c
- We want:

$$w_c = f_t(t_c)$$

- f_c … a mapping $\mathbb{R}^p \to \mathbb{R}^d$ that transforms text features to the visual image feature space

Zero-shot Learning

- f_t can be a neural network



g used to compress x to a k<<d dim

x

Red faced Cormorant

The Red-faced Cormorant, Red-faced Shag or Violet Shag, Phalacrocorax urile, is a species of cormorant that is found in the far north of the Pacific Ocean and Bering Sea, from the eastern tip of Hokkaidō in Japan, via the Kuril Islands, the southern tip of the Kamchatka Peninsula and the Aleutian Islands to the Alaska Peninsula and Gulf of Alaska. The Red-faced Cormorant is closely related to the Pelagic Cormorant P. pelagicus, which has a similar range, and like the Pelagic Cormorant is placed by some authors (e.g. Johnsgaard) in a genus Leucocarbo. Where it nests alongside the Pelagic Cormorant, the Red-faced Cormorant generally breeds the more successfully of the two species, and it is currently increasing in numbers, at least in the easterly parts of its range. It is however listed as being of conservation concern{Verify source|date=September 2009}, partly because relatively little is so far known about it.

The adult bird has glossy plumage that is a deep greenish blue in colour, becoming purplish or bronze on the back and sides. In breeding condition it has a double crest,

......

Red faced Cormorant

The Red-faced Cormorant, Red-faced Shag or Violet Shag, Phalacrocorax urile, is a species of cormorant that is found in the far north of the Pacific Ocean and Bering Sea, from the eastern tip of Hokkaidō in Japan, via the Kuril Islands, the southern tip of the Kamchatka Peninsula and the Aleutian Islands to the Alaska Peninsula and Gulf of Alaska. The Red-faced Cormorant is closely related to the Pelagic Cormorant P. pelagicus, which has a similar range, and like the Pelagic Cormorant is placed by some authors (e.g. Johnsgaard) in a genus Leucocarbo. Where it nests alongside the Pelagic Cormorant, the Red-faced Cormorant generally breeds the more successfully of the two species, and it is currently increasing in numbers, at least in the easterly parts of its range. It is however listed as being of conservation concern{Verify source|date=September 2009}, partly because relatively little is so far known about it.

The adult bird has glossy plumage that is a deep greenish blue in colour, becoming purplish or bronze on the back and sides. In breeding condition it has a double crest,

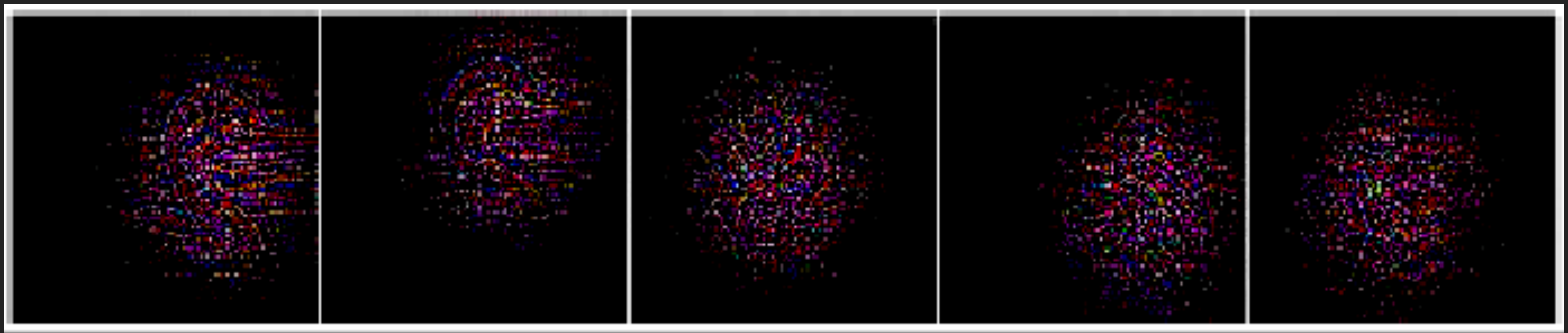
......

Red faced Cormorant

The Red-faced Cormorant, Red-faced Shag or Violet Shag, Phalacrocorax urile, is a species of cormorant that is found in the far north of the Pacific Ocean and Bering Sea, from the eastern tip of Hokkaidō in Japan, via the Kuril Islands, the southern tip of the Kamchatka Peninsula and the Aleutian Islands to the Alaska Peninsula and Gulf of Alaska. The Red-faced Cormorant is closely related to the Pelagic Cormorant P. pelagicus, which has a similar range, and like the Pelagic Cormorant is placed by some authors (e.g. Johnsgaard) in a genus Leucocarbo. Where it nests alongside the Pelagic Cormorant, the Red-faced Cormorant generally breeds the more successfully of the two species, and it is currently increasing in numbers, at least in the easterly parts of its range. It is however listed as being of conservation concern{Verify source|date=September 2009}, partly because relatively little is so far known about it.

The adult bird has glossy plumage that is a deep greenish blue in colour, becoming purplish or bronze on the back and sides. In breeding condition it has a double crest,

......

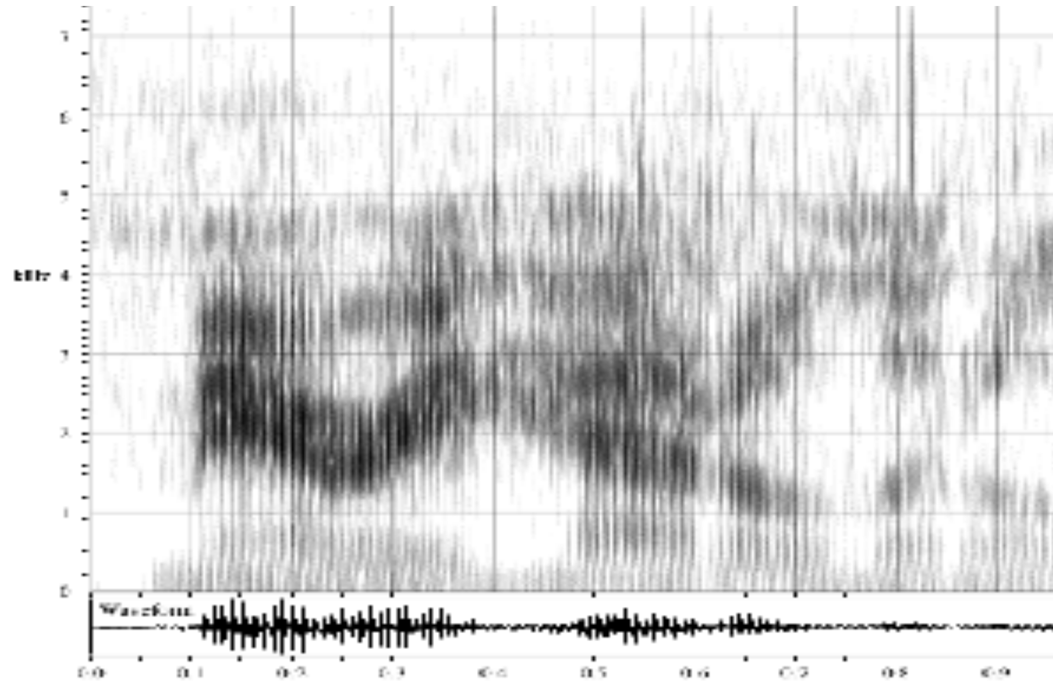


visualization by Zeiler & Fergus, ECCV'14.


https://www.youtube.com/watch?v=AIEsakeXvMM
THE

Raw pixels

Raw audio speech



- [Unsupervised Learning of Spoken Language with Visual Context.](#) David Harwath, Antonio Torralba, James Glass. Advances in Neural Information Processing Systems (NIPS), 2016.

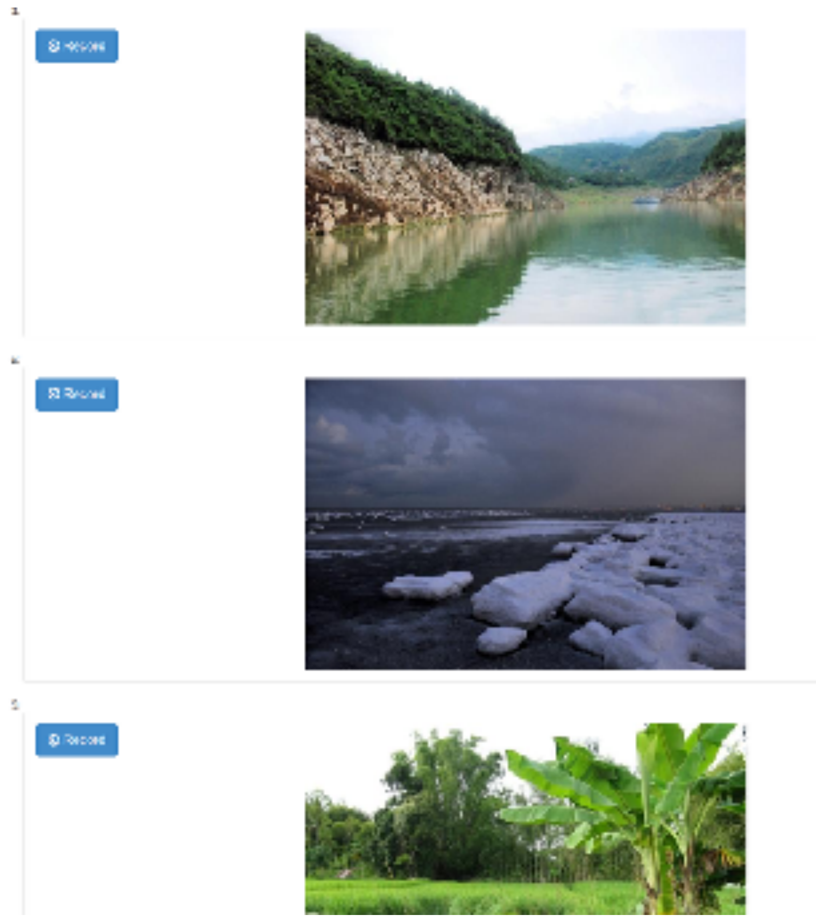
Unsupervised Learning of Spoken Language with Visual Context.

David Harwath

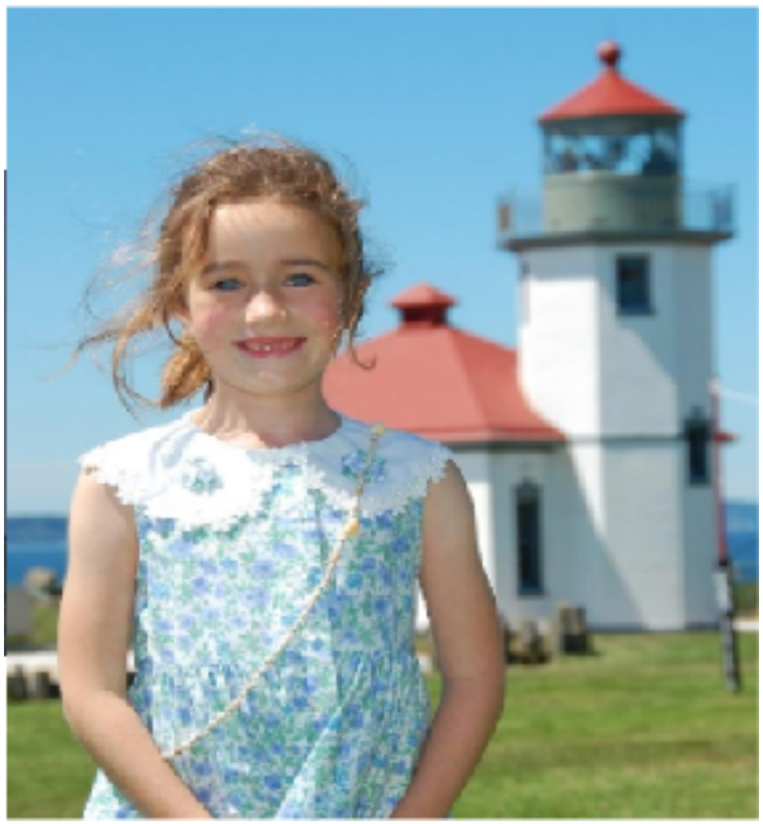
Jim Glass

[Unsupervised Learning of Spoken Language with Visual Context.](#) David Harwath, Antonio Torralba, James Glass. Advances in Neural Information Processing Systems (NIPS), 2016.

Crowdsourcing Audio-Visual Data



Crowdsourcing Audio-Visual Data

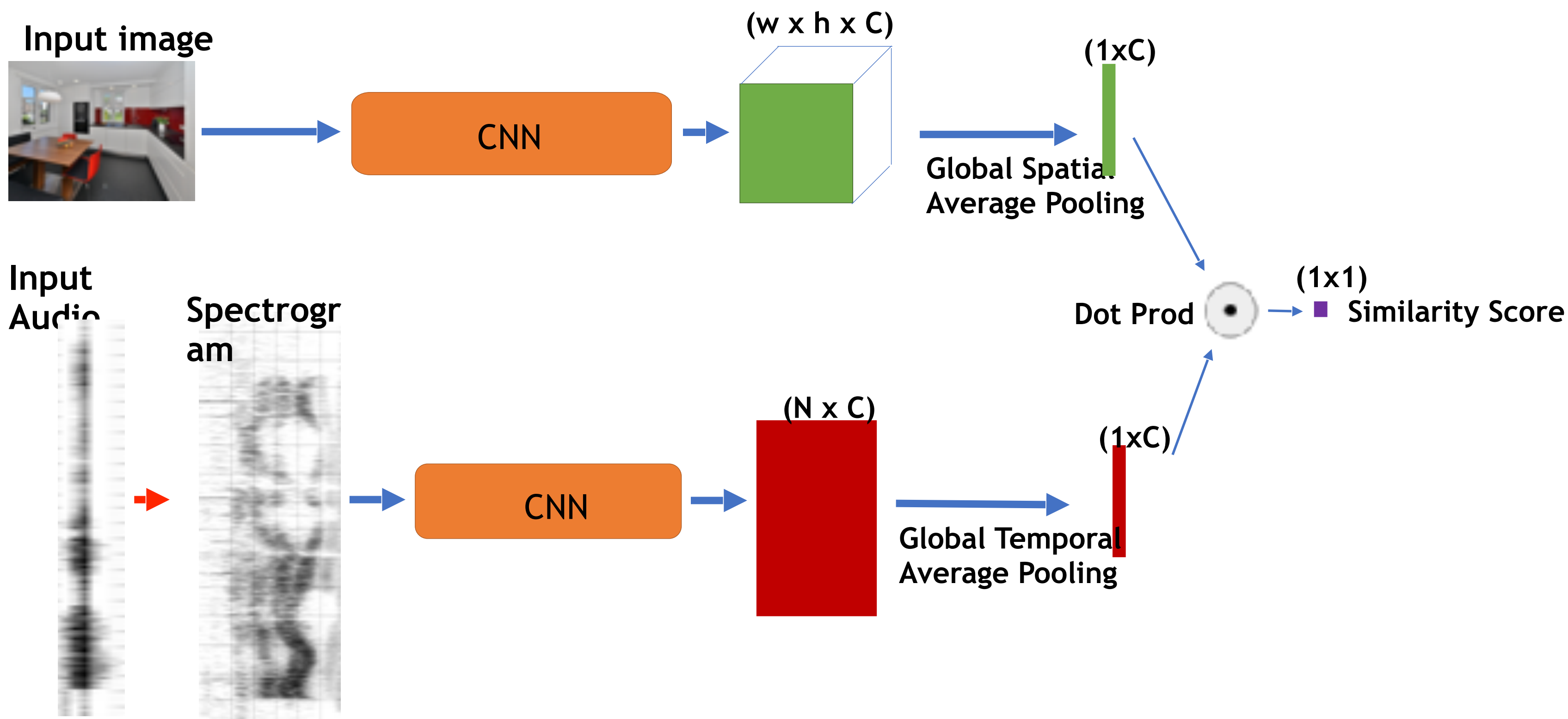


382.060 Speech descriptions on Images from Places dataset.

Crowdsourcing Audio-Visual Data

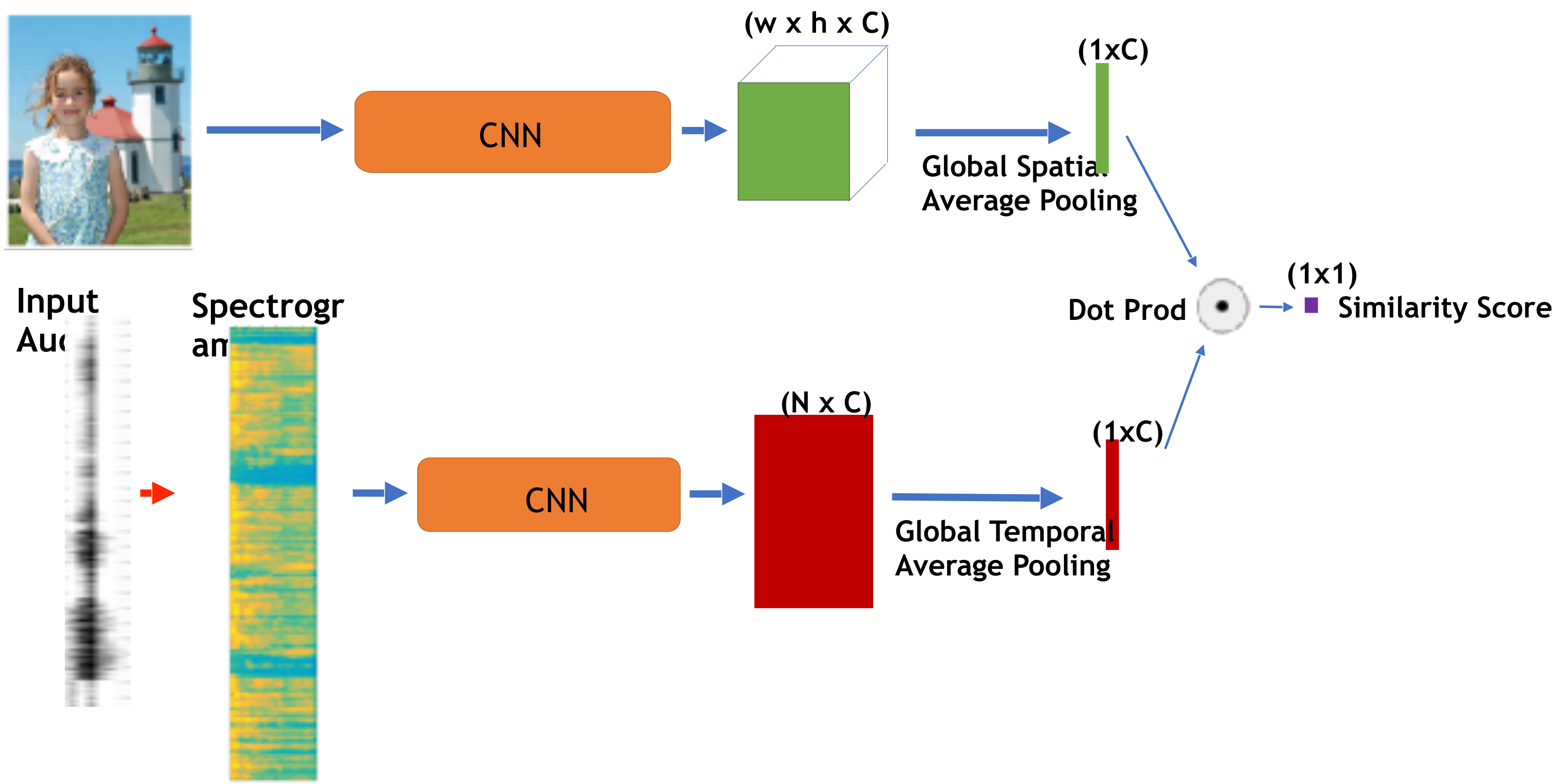
382.060 Speech descriptions on
nages from Places dataset.

Joint Audio-Visual Architecture



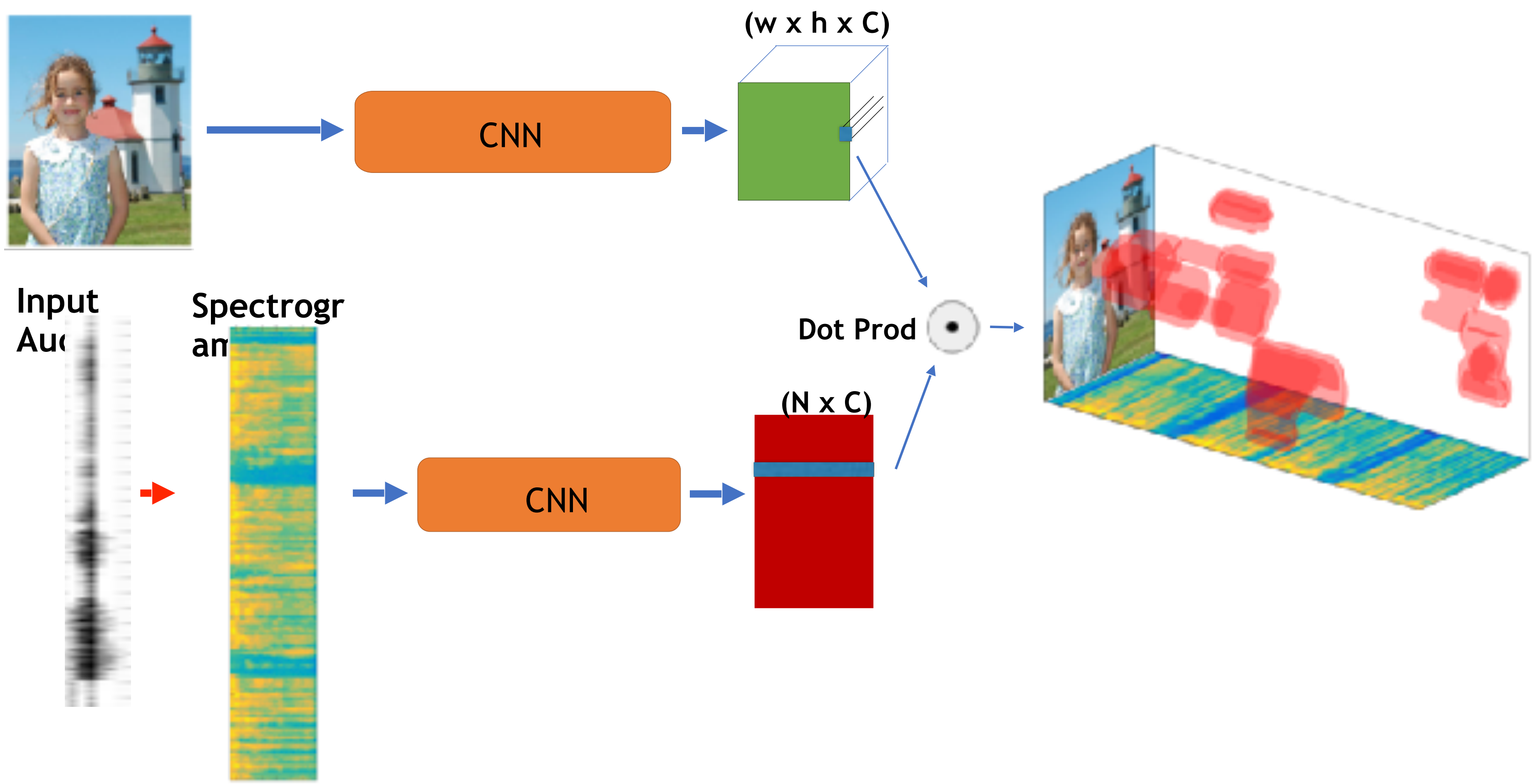
Harwath et al., NIPS 2016

Joint Audio-Visual Architecture



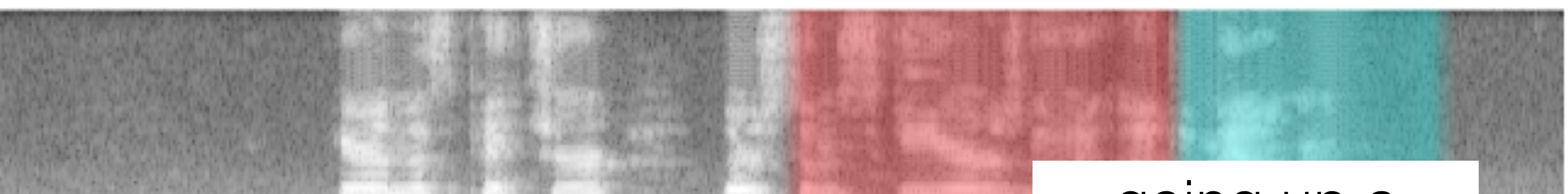
Harwath et al., NIPS 2016

Joint Audio-Visual Architecture



Harwath et al., NIPS 2016

Co-segmentation of speech and image



Here in this pic there some skier going up a mountain

Words

- Need ways to compare words

Next to the 'sofa' is a desk, and a 'person' is sitting behind it.

'armchair'	'man'
'bench'	'woman'
'chair'	'child'
'deck chair'	'teenager'
'ottoman'	'girl'
'seat'	'boy'
'stool'	'baby'
'swivel chair’	'daughter’
‘loveseat’	‘son’
…	…

Encoding words into vectors

- Need ways to compare words

So that if two words I and j are similar then wi and wj are close

T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

Encoding words into vectors

- Need ways to compare words

One-hot representations

'sofa' → $[1, 0, 0, 0, \ldots 0]_V$

'person' → $[0, 1, 0, 0, \ldots 0]_V$

'car' → $[0, 0, 1, 0, \ldots 0]_V$

'tree' → $[0, 0, 0, 1, \ldots 0]_V$

1-of-V coding, where V is size of the vocabulary

word2vec

- Find better vector encodings

word → encoder → w

'sofa' → w1

'person' → w2

'car' → w3

'tree' → w4

So that if two words I and j are similar then wi and wj are close

T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

But we do not have word similarities…

How do we learn the vectors?

We will use a different task, and hope that similarity will emerge…

We will train a classifier to predict the words surrounding each word.

word2vec

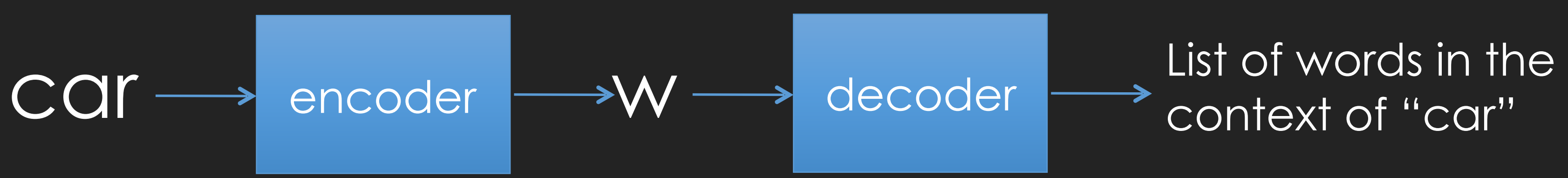
I parked the car in a nearby street. It is a red car with two doors, ...

I parked the vehicle in a nearby street...

T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

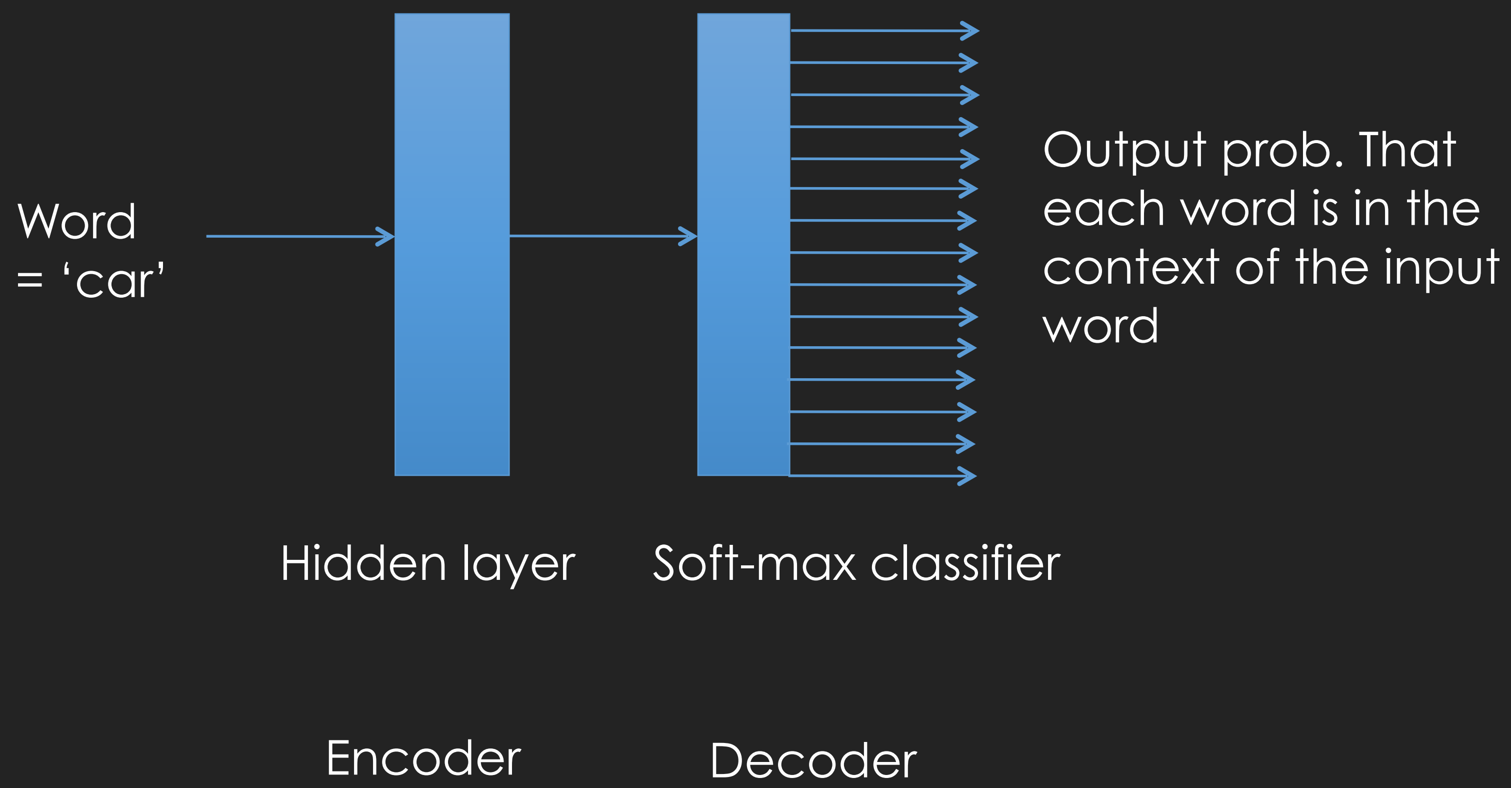
word2vec

I parked the car in a nearby street. It is a red car with two doors, …



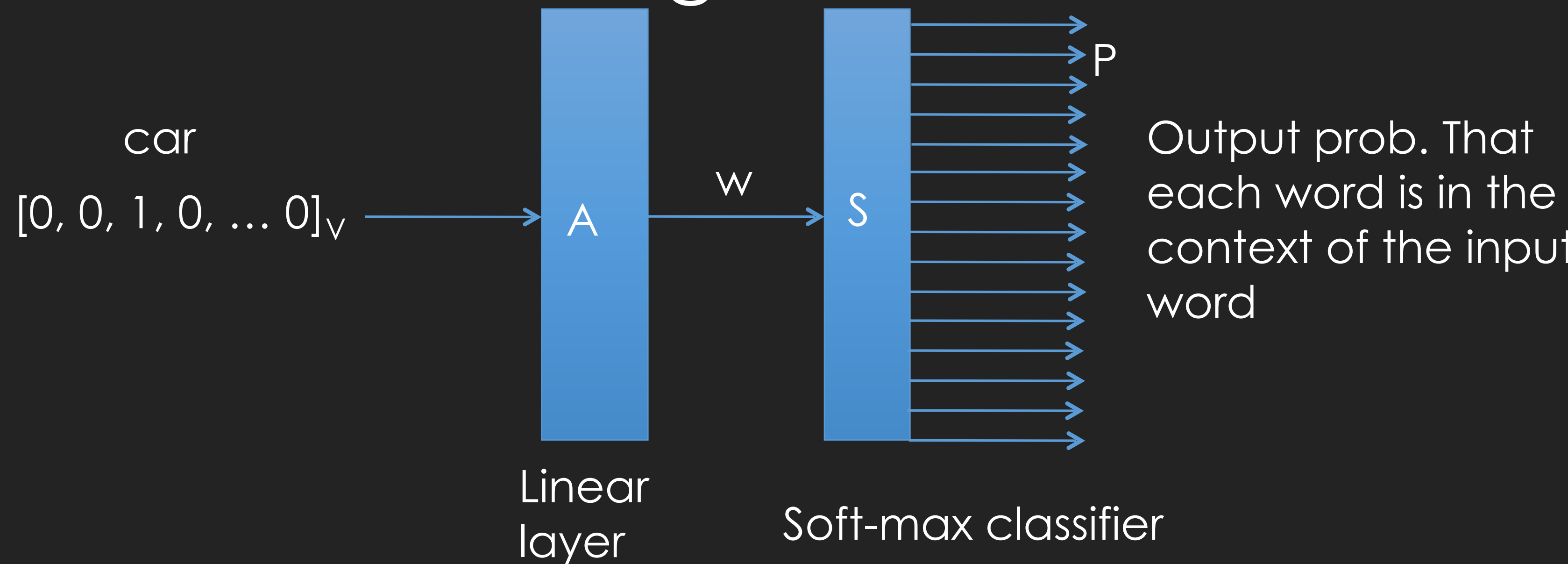
T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

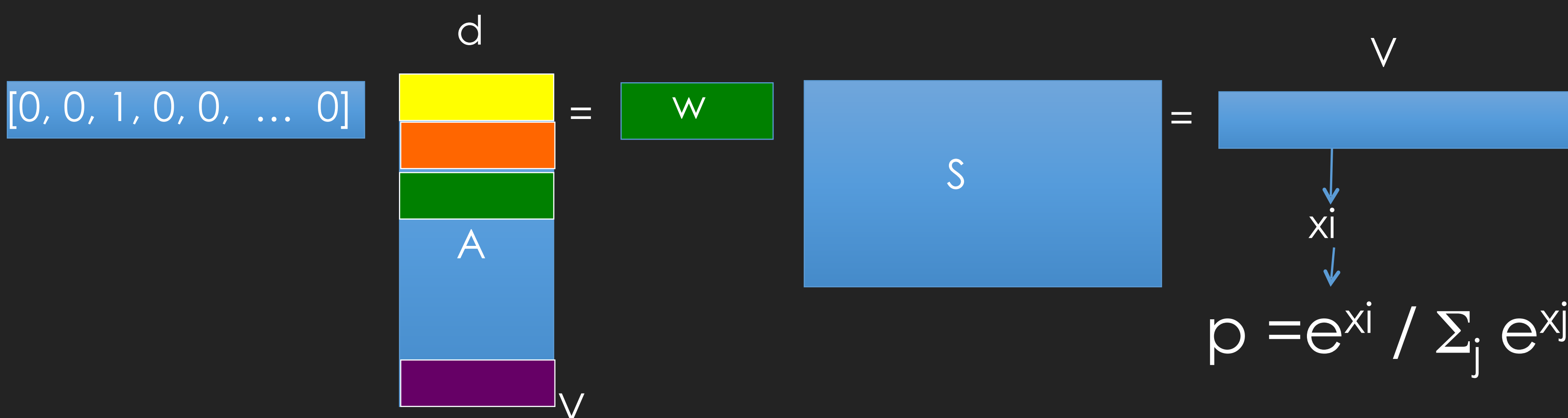
word2vec



T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec, training





T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec, training

- In training maximize log-likelihood over the training set:

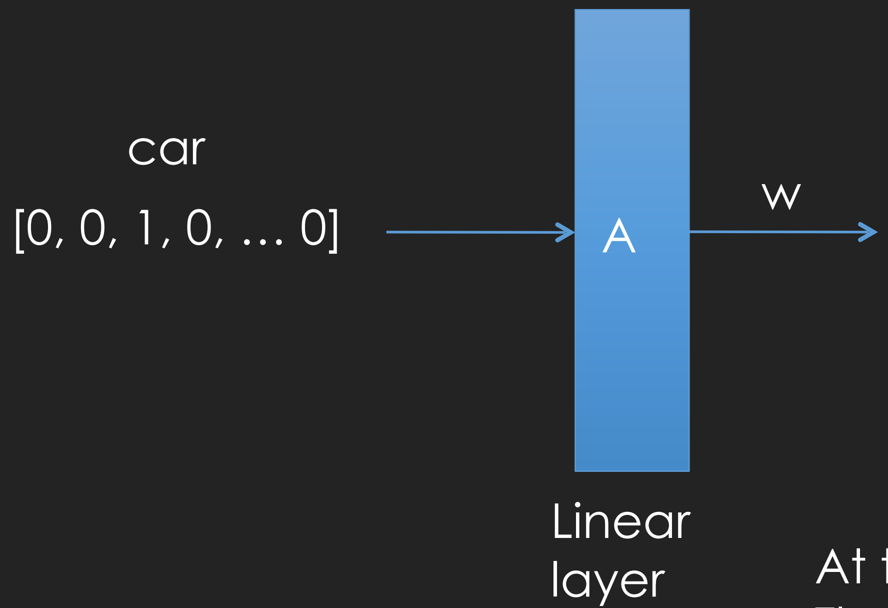
$$\sum_{t=1}^{T} \sum_{i=-c}^{c} \log p(w_{t+i} | w_t)$$

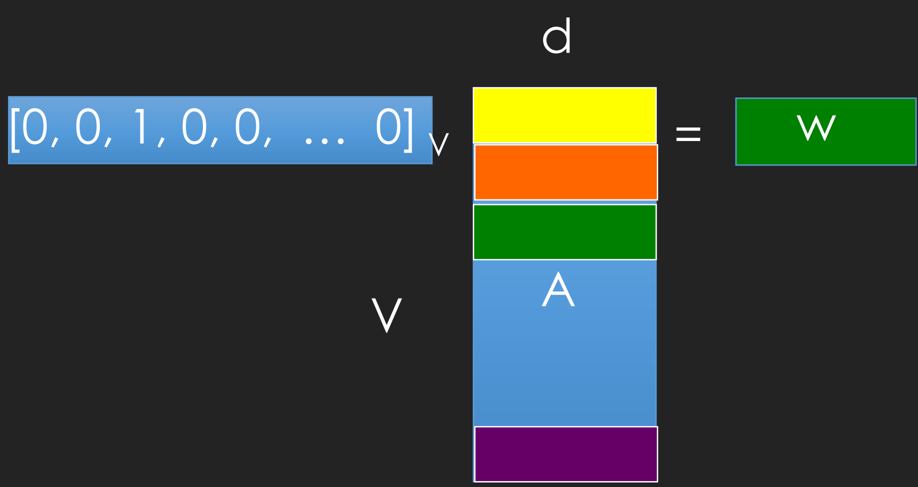
T … training set size
c … context window size

$$p = e^{xi} / \Sigma_j e^{xj}$$

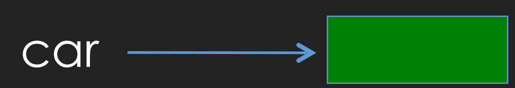
T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec, test time





At test time, w is our word embedding.
The encoding is just a look up table.



T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

Algebraic operations with the vector representation of words

X = Vector(“Paris”) – vector(“France”) + vector(“Italy”)

Closest nearest neighbor to X is vector(“Rome”)

Remember: Subtitle – Sentence Similarity

01:00:58 --> 01:01:03
I'm telling you, it's spooky. She knows more about you than you do.

01:01:03 --> 01:01:05
Who doesn't?

Subtitle – Sentence Similarity

She and Uncle Vernon went the lumpy bed next door, and Harry was left to find the softest bit of floor he could and to curl up under the thinnest, most ragged blanket.

Sentences

Skip-Thought Vectors

training corpus: 11K books

....

They called from outside.

She smiled and tried to open the door.

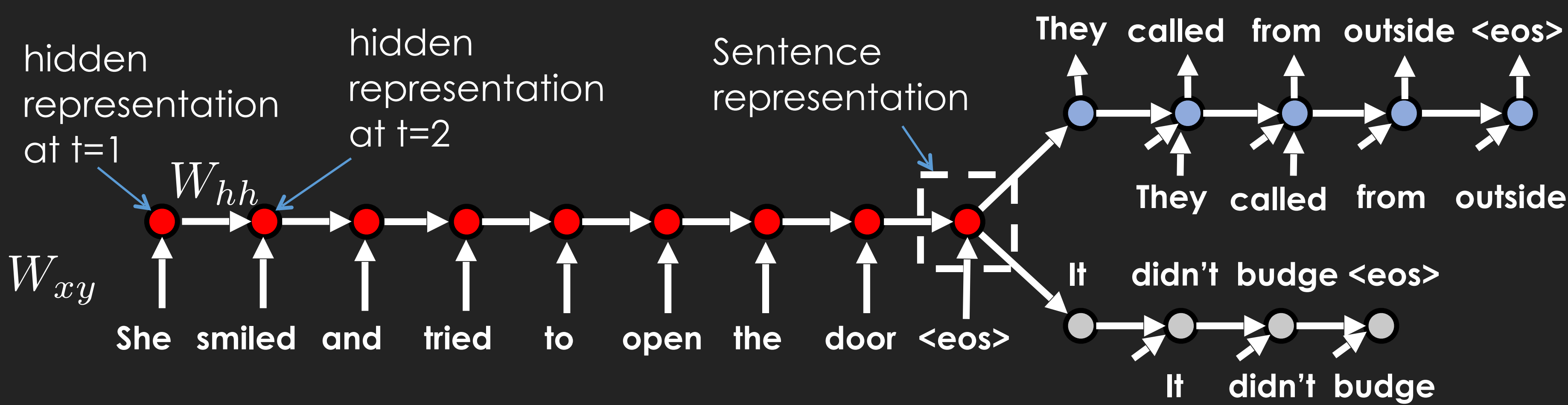
It didn't budge.

......

R. Kiros, Y. Zhu, R. Salakhutdinov, R. Zemel, A. Torralba, R. Urtasun, S. Fidler.
Skip-Thought Vectors. NIPS'2015

Sentences

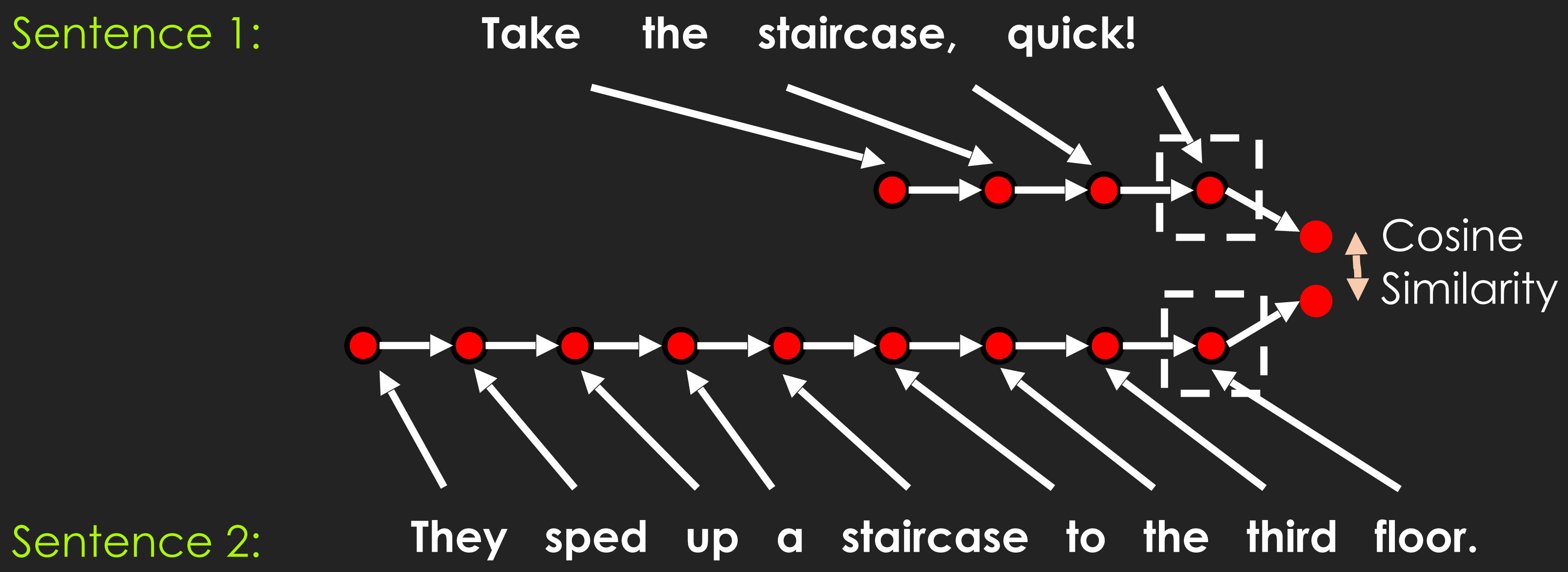
Skip-Thought Vectors



R. Kiros, Y. Zhu, R. Salakhutdinov, R. Zemel, A. Torralba, R. Urtasun, S. Fidler. Skip-Thought Vectors. NIPS'2015

Sentences

On test time:

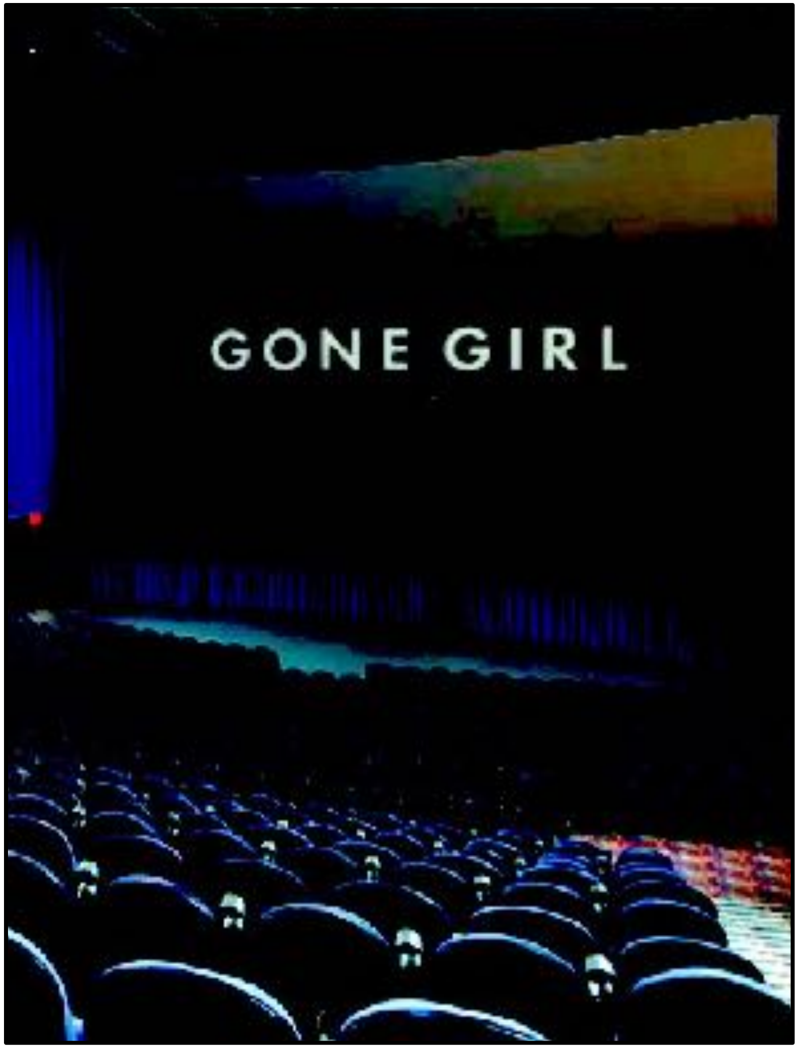


Books Contain Rich Descriptions

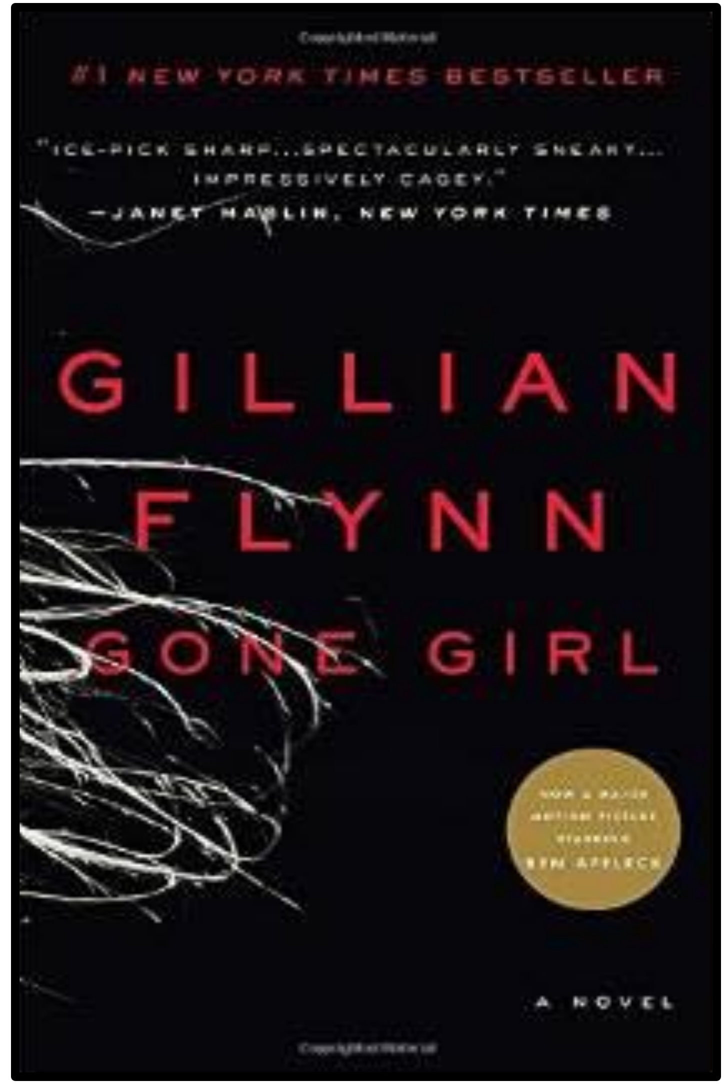
But lack rich visual content

Movies contain rich visual content

Lots of paired books and movies



movie

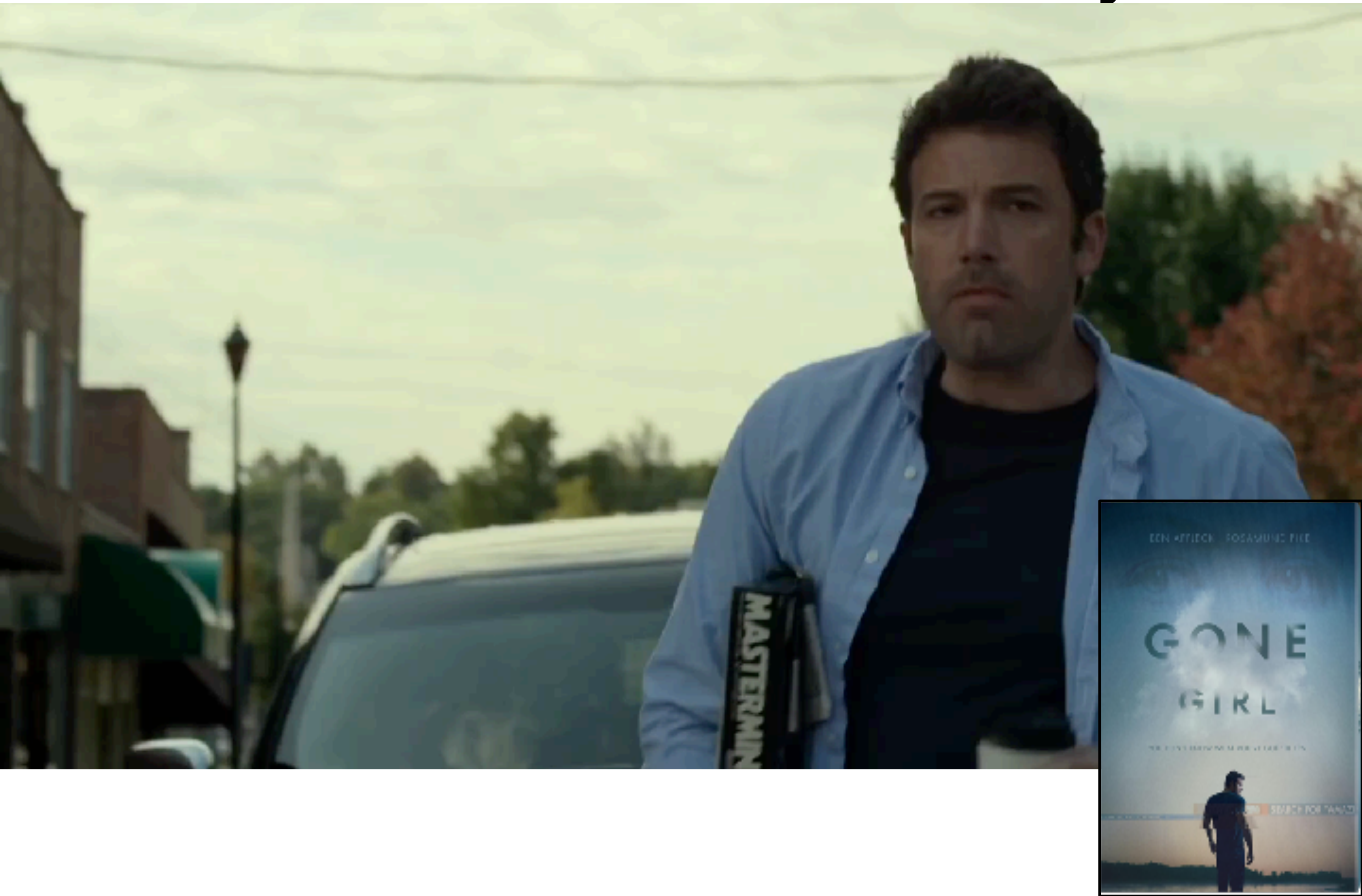


book

Book: Tells A Story

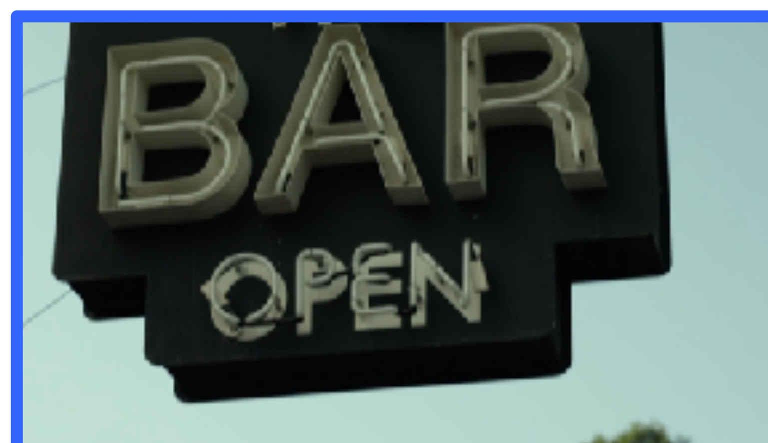
As I walked toward the bar across the concrete parking lot, I looked straight down the road and saw the river. Moving apace with the river was a long single line of men, eyes aimed at their feet, walking steadfastly nowhere. I felt an immediate need to get inside.

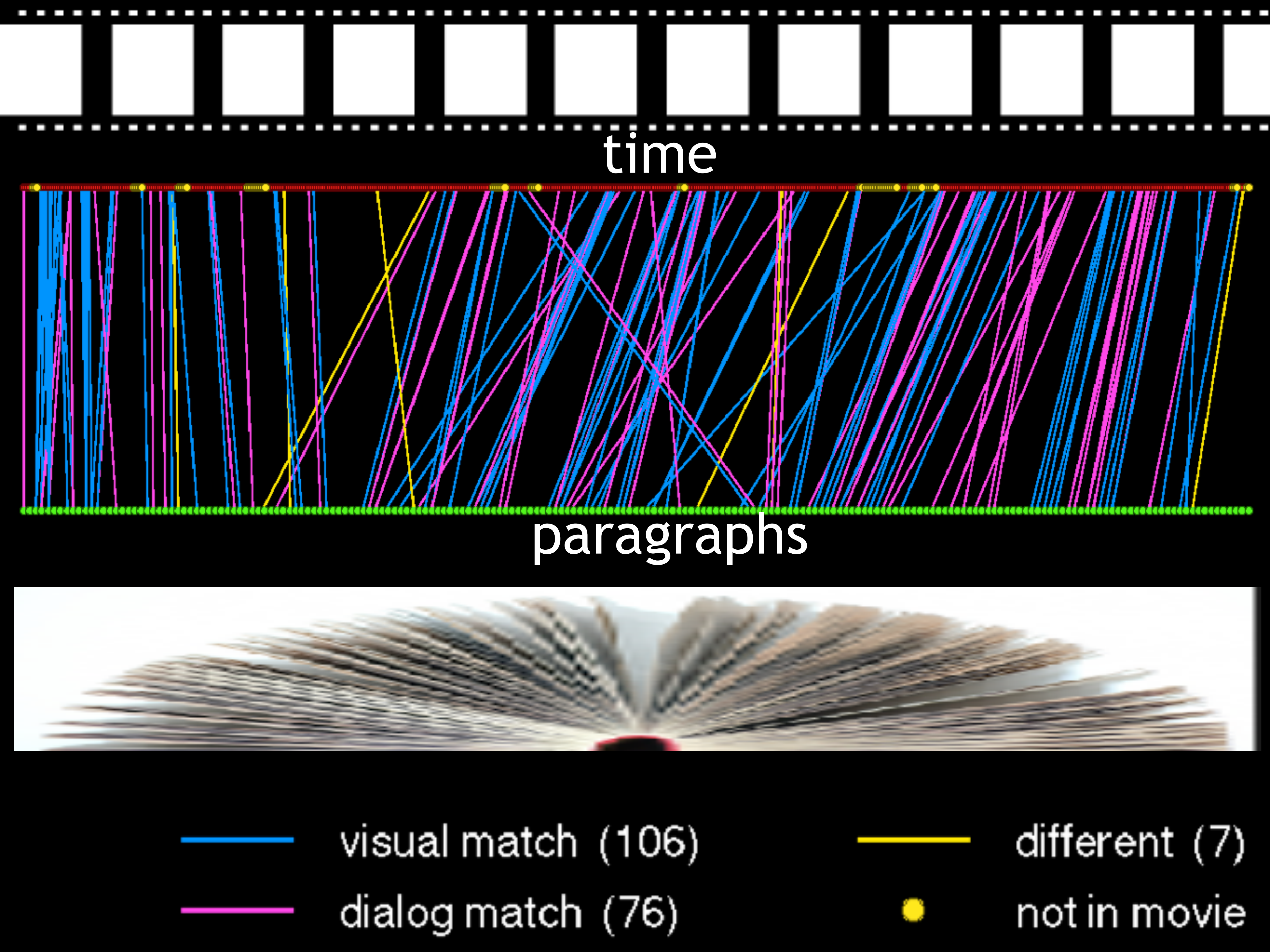
Movie: Visualizes A Story



As I walked toward the bar across
the concrete parking lot, I looked straight down the road and saw the river. Moving apace with the river was a long single line of men, eyes aimed at their feet, walking steadfastly nowhere. I felt an immediate need to get inside.

As I walked toward the bar across the concrete parking lot, I looked straight down the road and saw the river. Moving apace with the river was a long single line of men, eyes aimed at their feet, walking steadfastly nowhere. I felt an immediate need to get inside.




time
paragraphs
visual match (106)
dialog match (76)
different (7)
not in movie

Sentence Query Example

Query:

- He drove down the street off into the distance.

Sentence Query Example

Query:
- He drove down the street off into the distance.

Top Retrieved Sentences using Skip-Thoughts:

- He started the car, left the parking lot and merged onto the highway a few miles down the road.
- She watched the lights flicker through the trees as the men drove toward the road.

Skip-Thought Vectors

Near state-of-the-art results on standard NLP tasks:

- Semantic relatedness
- Paraphrase detection
- Image-sentence ranking
- Movie review sentiment prediction
- Question type classification

R. Kiros, Y. Zhu, R. Salakhutdinov, R. Zemel, A. Torralba, R. Urtasun, S. Fidler.
Skip-Thought Vectors. NIPS'2015

Cross-modal learning

Description (eg, Wikipedia article)

Snares penguin

From Wikipedia, the free encyclopedia

The **Snares penguin** (*Eudyptes robustus*), also known as the **Snares crested penguin** and the **Snares Islands penguin**, is a penguin from New Zealand. The species breeds on The Snares, a group of islands off the southern coast of the South Island. This is a medium-small, yellow-crested penguin, at a size of 50–70 cm (19.5–27.5 in) and a weight of 2.5–4 kg (5.5–8.8 lb). It has dark blue-black upperparts and white underparts. It has a bright yellow eyebrow-stripe which extends over the eye to form a drooping, bushy crest. It has bare pink skin at the base of its large red-brown bill.

- Lots of descriptions/entries in Wikipedia available

Images





Aligning Movies with Books

01:00:58 --> 01:01:03
I'm telling you, it's spooky. She knows more about you than you do.

01:01:03 --> 01:01:05
Who doesn't?

01:01:06 --> 01:01:08
What's happening?

01:01:08 --> 01:01:13
The staircases change, remember?

... As night fell, the promised storm blew up around them. Spray from the high waves splattered the walls of the hut and a fierce wind rattled the filthy windows.

Aunt Petunia found a few moldy blankets in the second room and made up a bed for Dudley on the moth-eaten sofa.

She and Uncle Vernon went the lumpy bed next door, and Harry was left to find the softest bit of floor he could and to curl up under the thinnest, most ragged blanket.

The storm raged more and more ferociously as the night went on, and Harry couldn't sleep.

Aligning Movies with Books

01:00:58 --> 01:01:03
I'm telling you, it's spooky. She knows more about you than you do.

01:01:03 --> 01:01:05
Who doesn't?

01:01:06 --> 01:01:08
What's happening?

01:01:08 --> 01:01:13
The staircases change, remember?

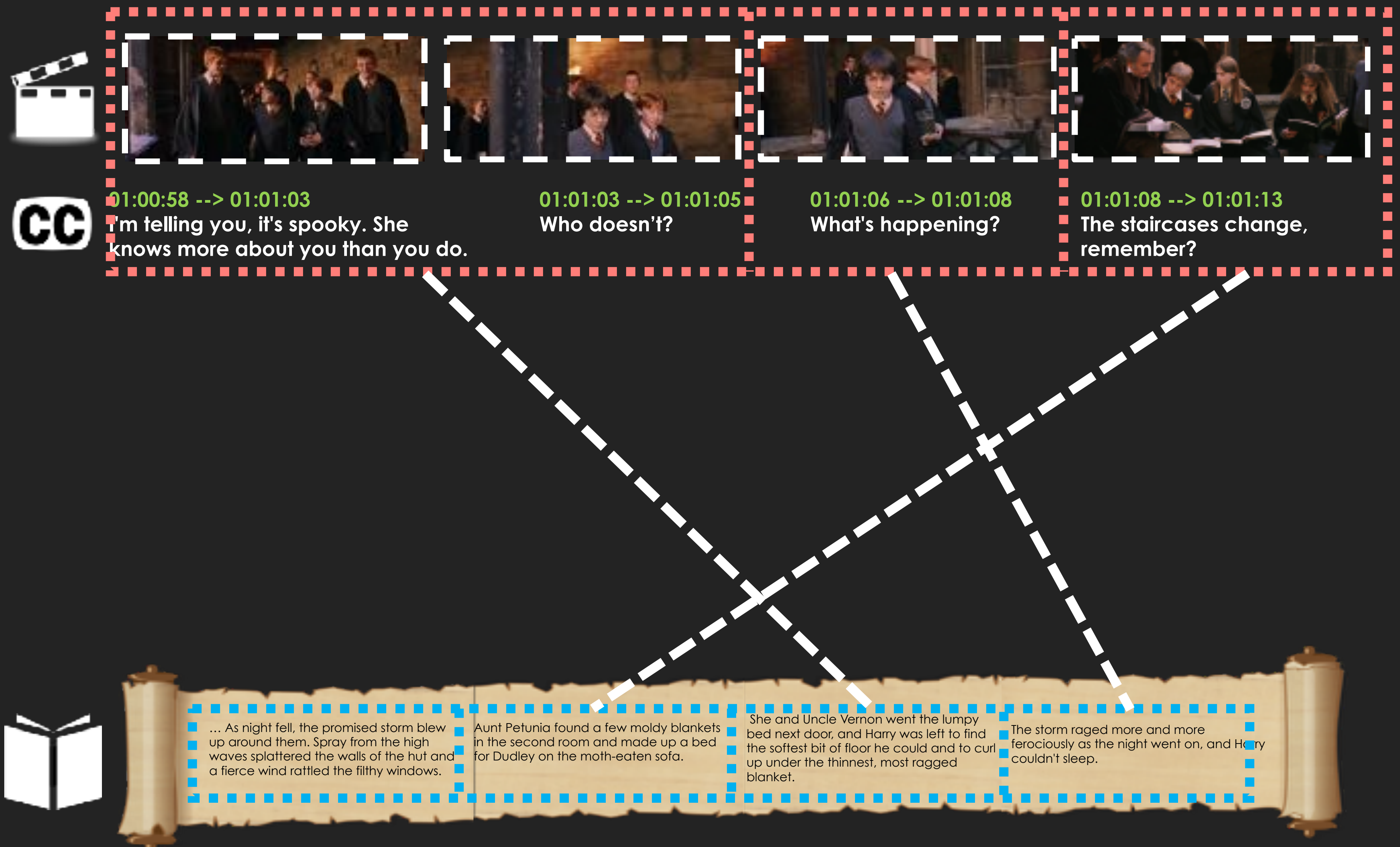
… As night fell, the promised storm blew up around them. Spray from the high waves splattered the walls of the hut and a fierce wind rattled the filthy windows.

Aunt Petunia found a few moldy blankets in the second room and made up a bed for Dudley on the moth-eaten sofa.

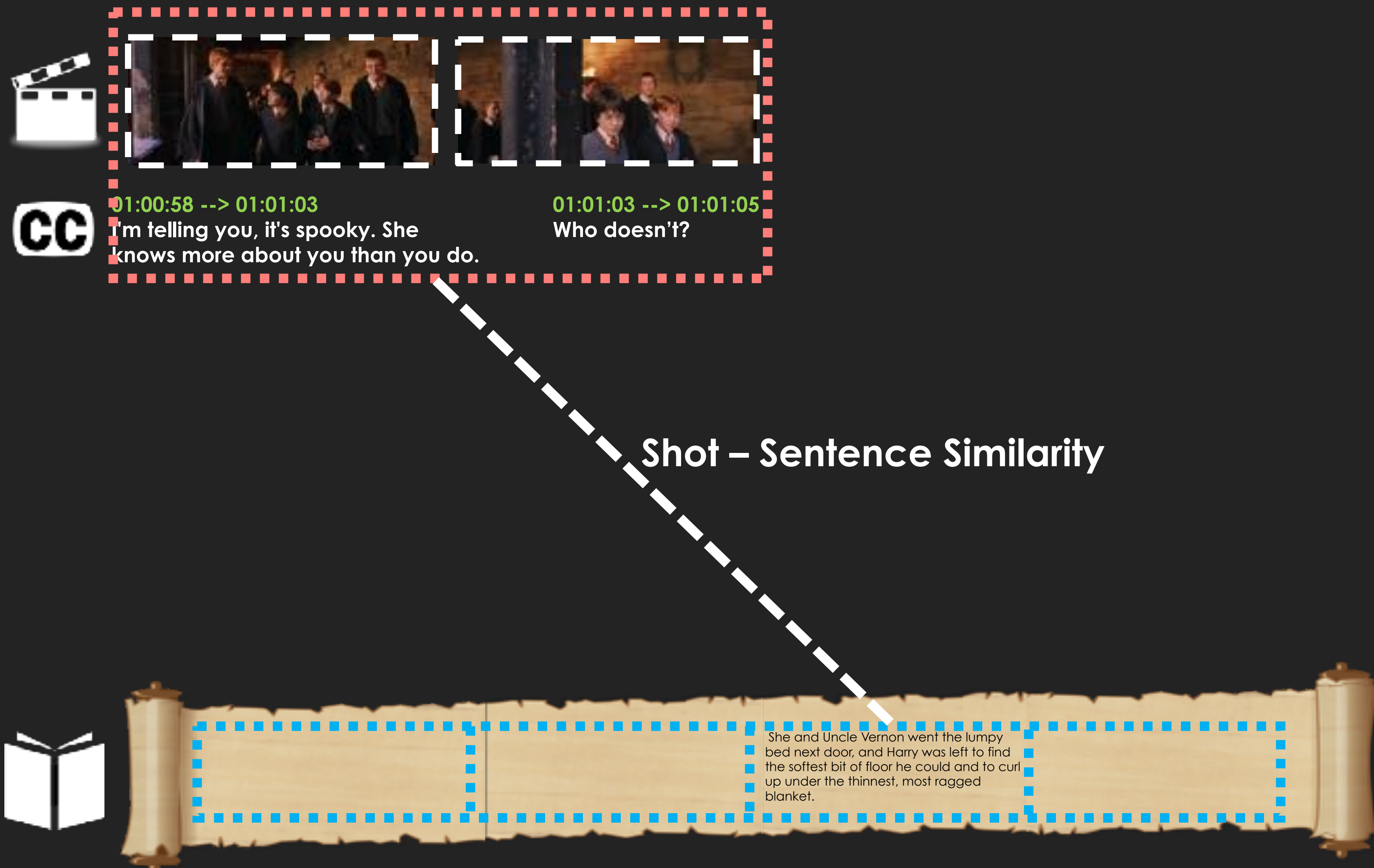
She and Uncle Vernon went the lumpy bed next door, and Harry was left to find the softest bit of floor he could and to curl up under the thinnest, most ragged blanket.

The storm raged more and more ferociously as the night went on, and Harry couldn't sleep.

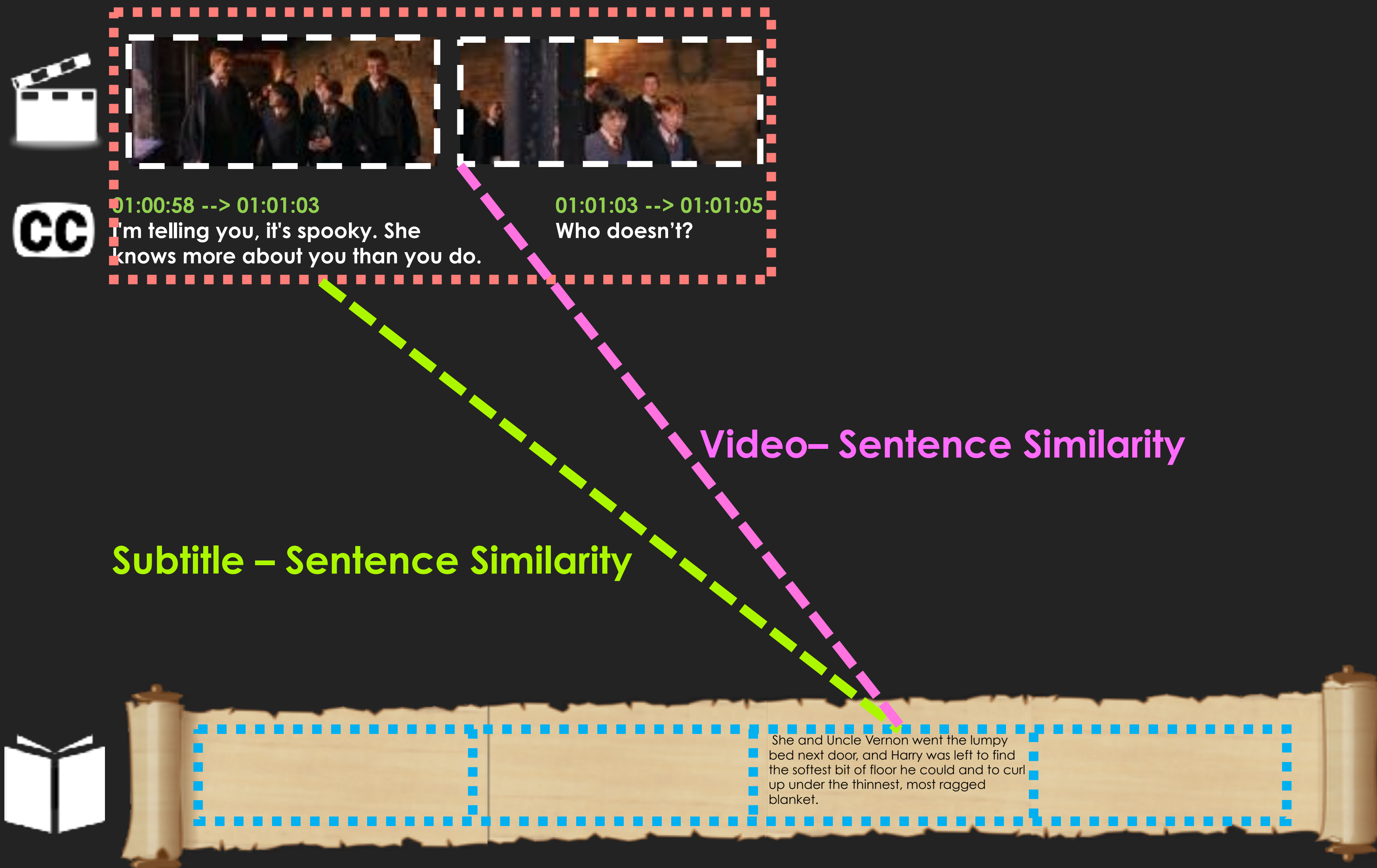
Aligning Movies with Books



Aligning Movies with Books



Aligning Movies with Books



Our Method

01:00:58 --> 01:01:03
I'm telling you, it's spooky. She knows more about you than you do.

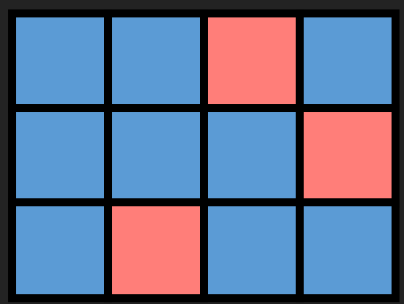
01:01:03 --> 01:01:05
Who doesn't?

01:01:06 --> 01:01:08
What's happening?

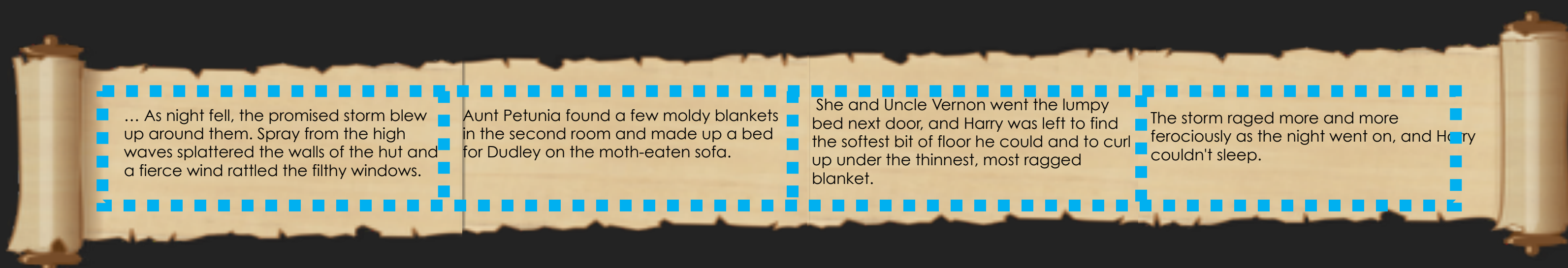
01:01:08 --> 01:01:13
The staircases change, remember?

Shot – Sentence Similarity Matrix

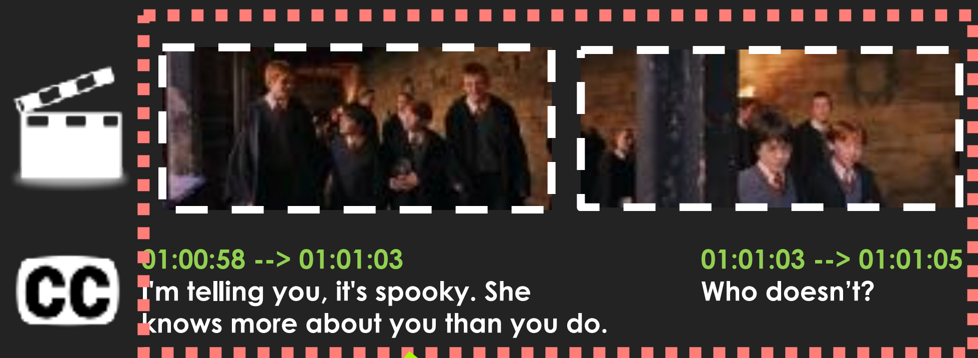
shots in movie



sentences in book

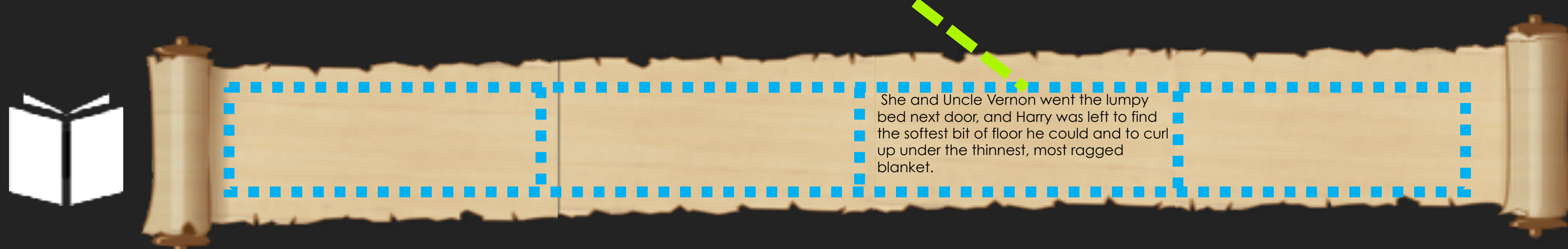


Subtitle – Sentence Similarity



- BLEU scores
- Longest subsequence matching + TF-IDF
- Sentence embedding (Skip-Thoughts)

Subtitle – Sentence Similarity



Subtitle – Sentence Similarity

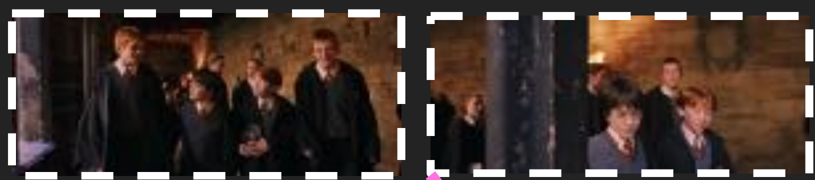
01:00:58 --> 01:01:03
I'm telling you, it's spooky. She knows more about you than you do.

01:01:03 --> 01:01:05
Who doesn't?

Subtitle – Sentence Similarity

She and Uncle Vernon went the lumpy bed next door, and Harry was left to find the softest bit of floor he could and to curl up under the thinnest, most ragged blanket.

Video – Sentence Similarity

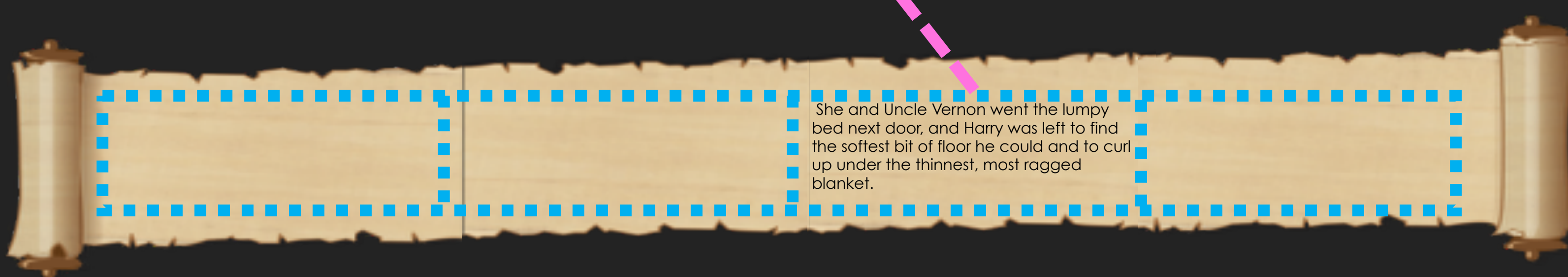


01:00:58 --> 01:01:03
I'm telling you, it's spooky. She knows more about you than you do.

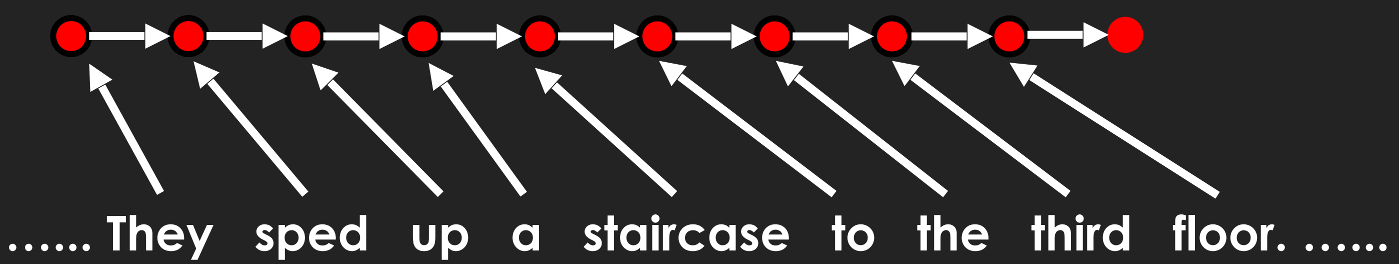
01:01:03 --> 01:01:05
Who doesn't?

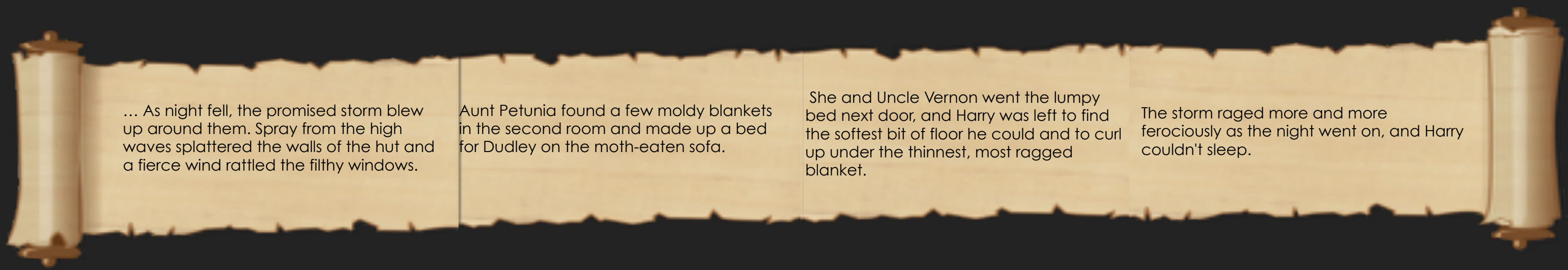
- Visual Semantic Embedding

Video– Sentence Similarity



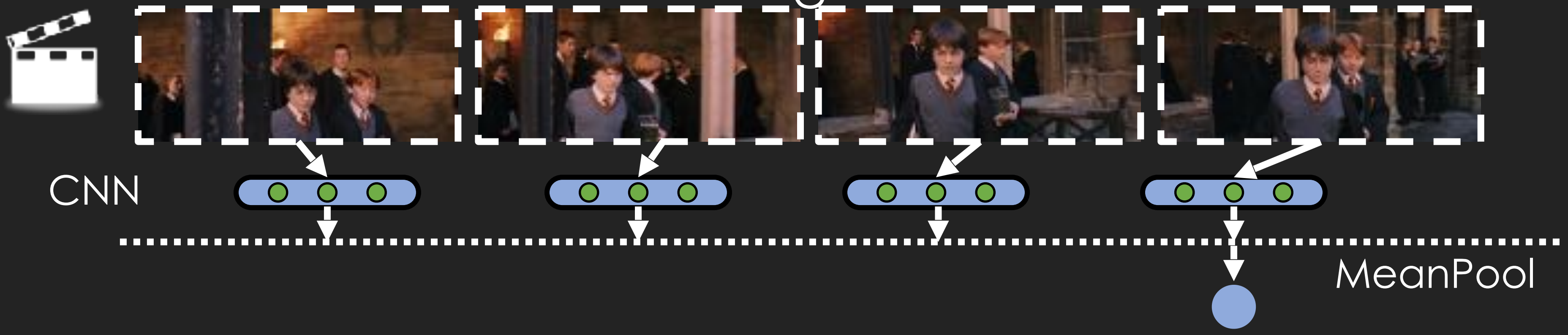
Video – Sentence Similarity





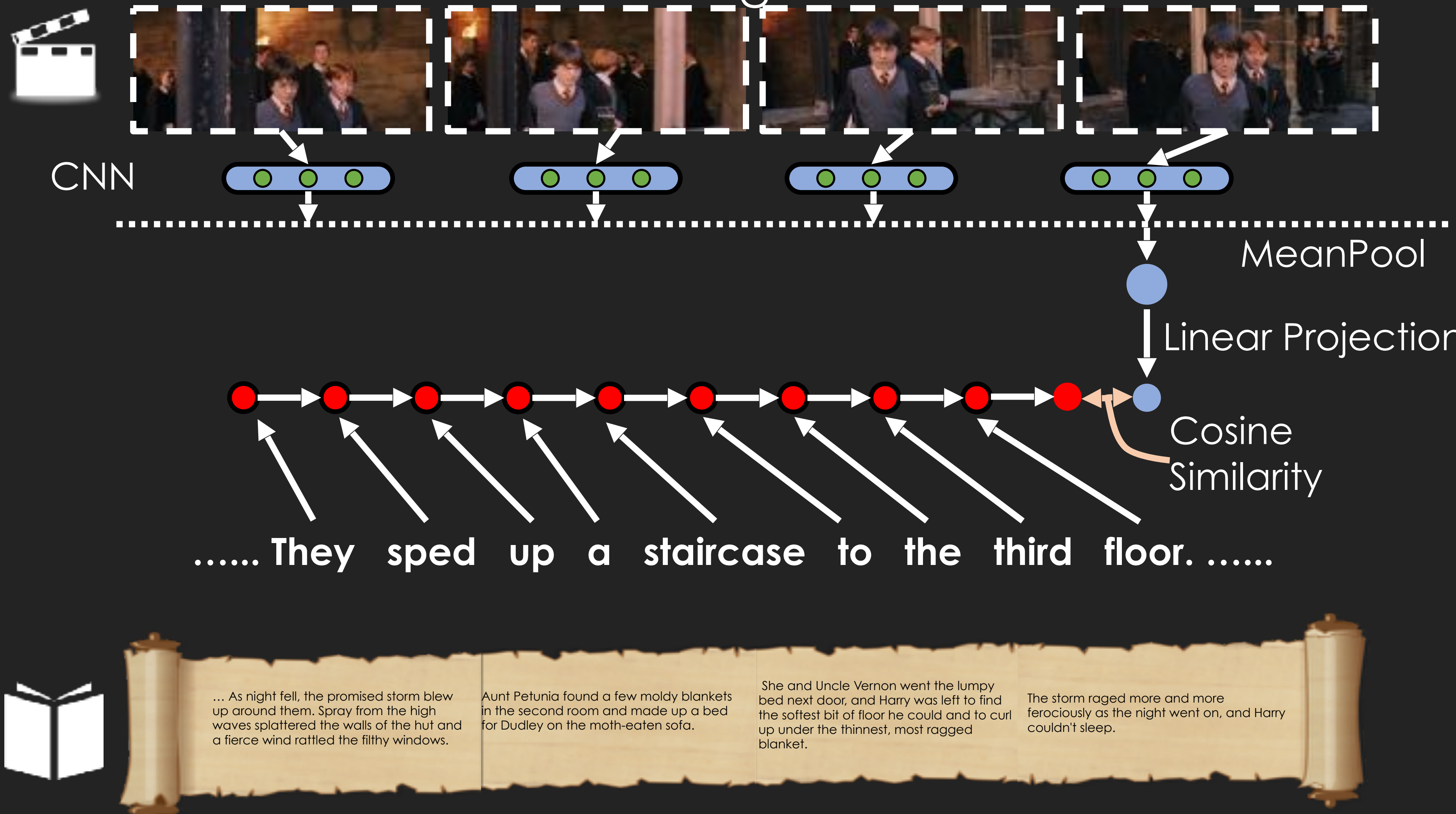
Video – Sentence Similarity

Visual Semantic Embedding



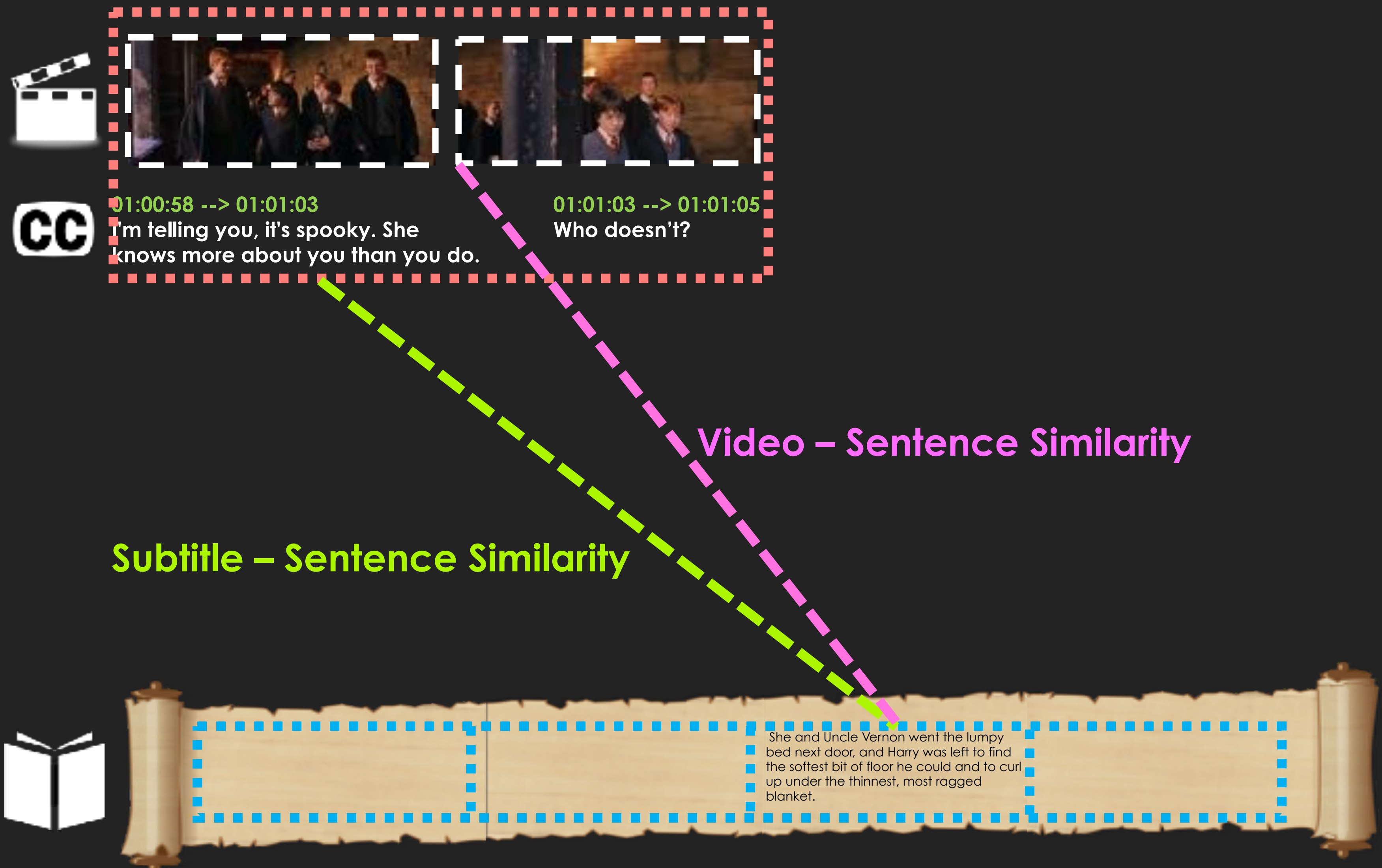
Video – Sentence Similarity

Visual Semantic Embedding



A. Rohrbachet al. A Dataset for Movie Description. In CVPR'15

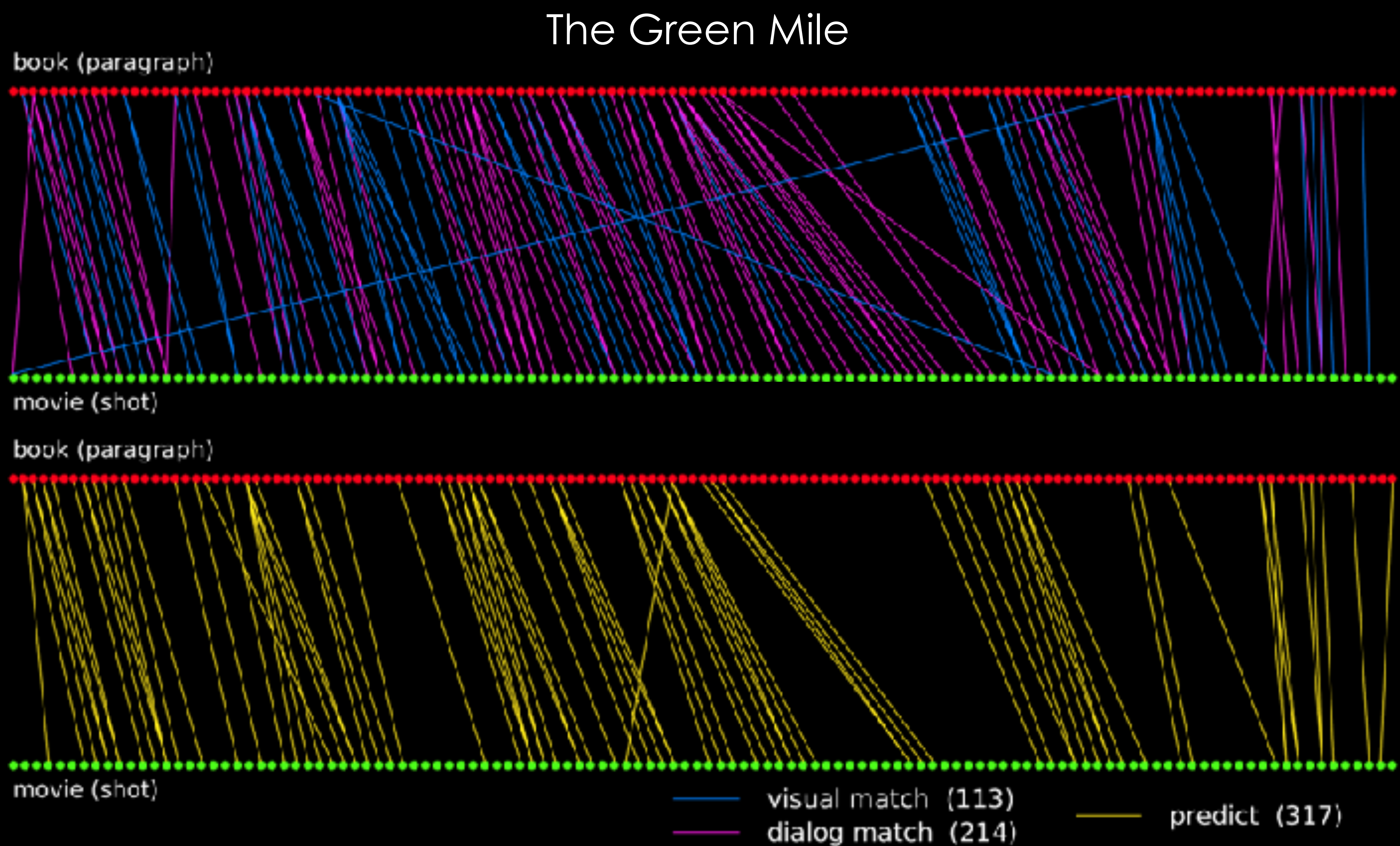
Aligning Movies with Books



Dataset

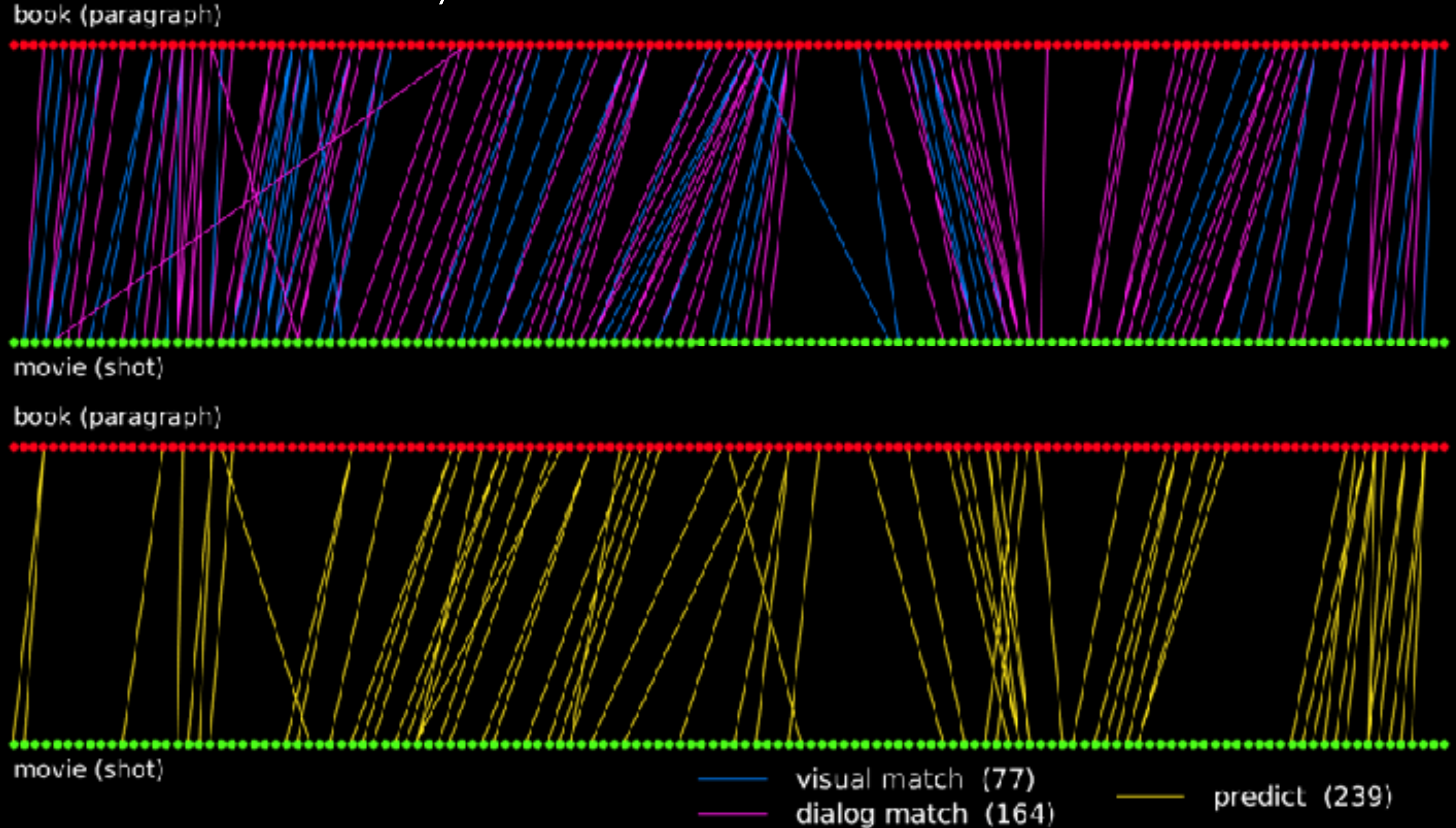
11 movie-book pairs annotated with 2,070 correspondences

Qualitative Results



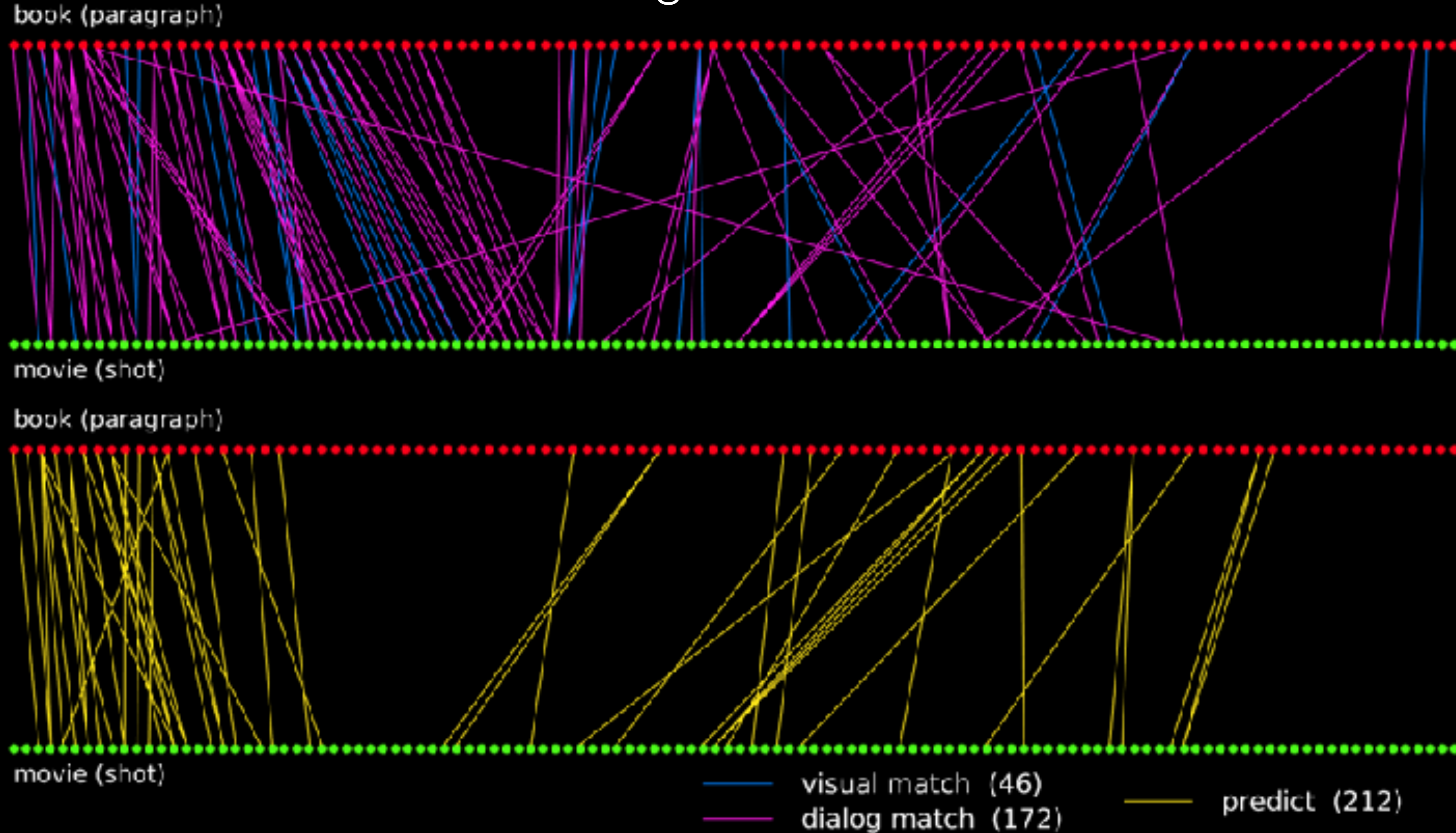
Qualitative Results

Harry Potter and the Sorcerers Stone



Qualitative Results

Fight Club



Qualitative Results on Alignment

We narrate the matched paragraph from the Harry Potter book.

Qualitative Results on Alignment