DEEP VISUAL-SEMANTIC ALIGNMENTS FOR GENERATING IMAGE DESCRIPTIONS

Authors:

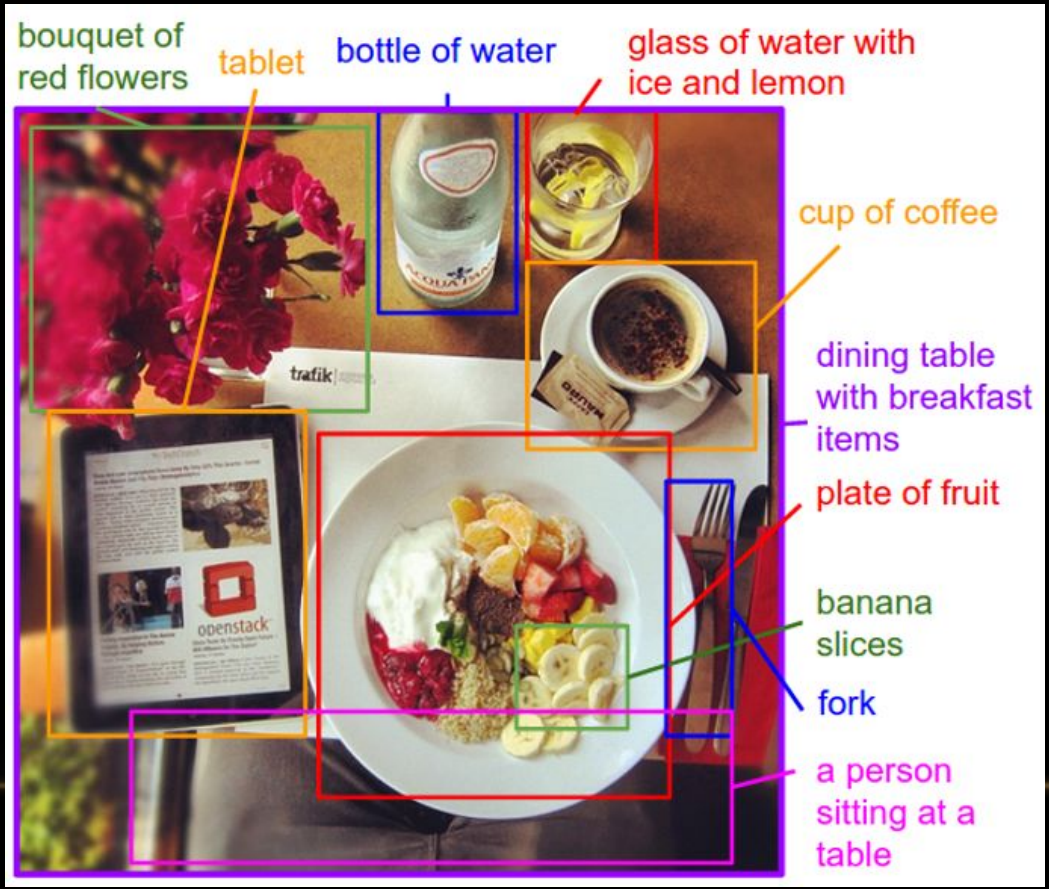
Andrej Karpathy, Li Fei-Fei

Presented by:

Jonathan Hohrath, Justin Lee

TASK / OBJECTIVE

- To generate natural language descriptions of images and their regions

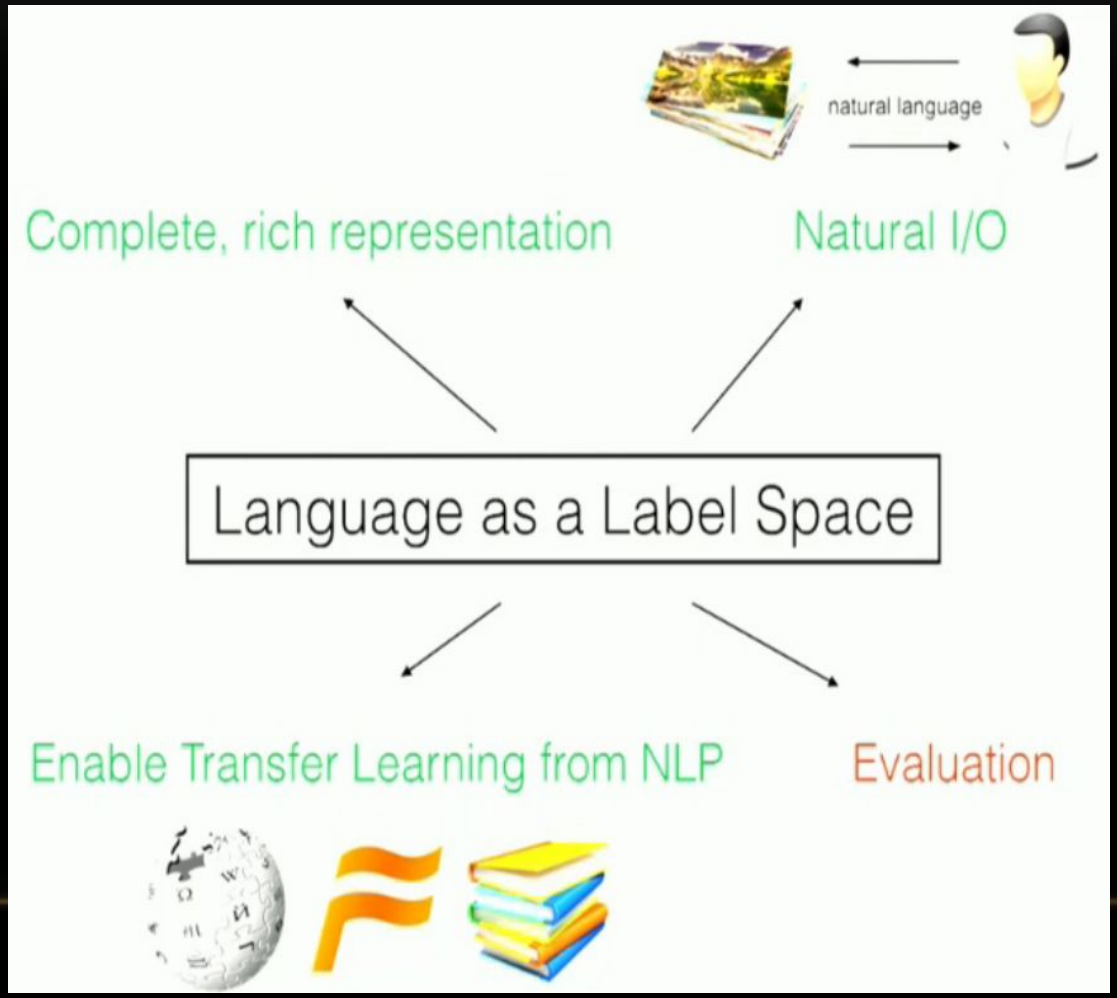


https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

CHALLENGE

- Model has to reason about image contents AND their representation in natural language
- Captioned images datasets are available, but they generally do not do not include entity locations in image

MOTIVATION: WHY LANGUAGE LABELS INSTEAD OF FIXED-CATEGORY LABELS?



http://techtalks.tv/talks/deep-visual-semantic-alignments-for-generating-image-descriptions/61593/

RELATED WORK: DENSE IMAGE ANNOTATION

K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei,and M.I.Jordan. Matchingwordsandpictures. JMLR, 2003.

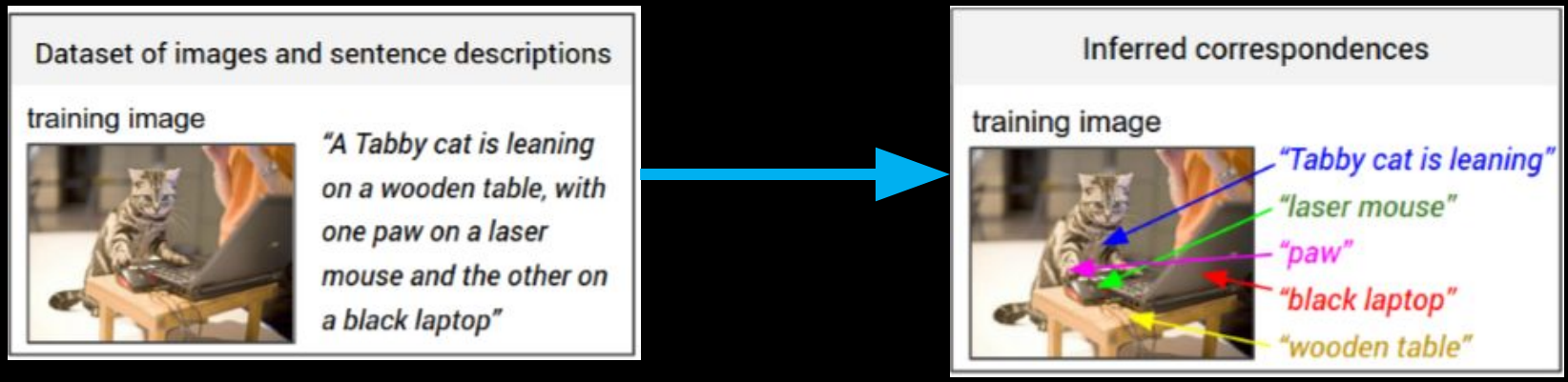
RELATED WORK: GENERATING DESCRIPTIONS

- Retrieval Solution:
  - Match most applicable training description to the test image
  - Stitch together segments of training descriptions
- Fixed templates
  - Fill templates based on image contents
- Full image description generation
  - Uses fixed window approach, generated words don't depend on previous words

PAPER CONTRIBUTION

- First paper to combine previous two concepts
  - Associates object location in image and sentence segments through a **multimodal embedding**
  - Generate descriptions for test images that significantly outperforms baselines
- Other papers were in pre-publish state that use a similar approach

STEP 1: EMBED IMAGE DATA & TEXT DESCRIPTION



Align sentence snippets to the visual regions they describe through **multimodal embedding**

https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

STEP 2: GENERATE NEW IMAGE DESCRIPTIONS USING IMAGE/TEXT CORRESPONDENCE

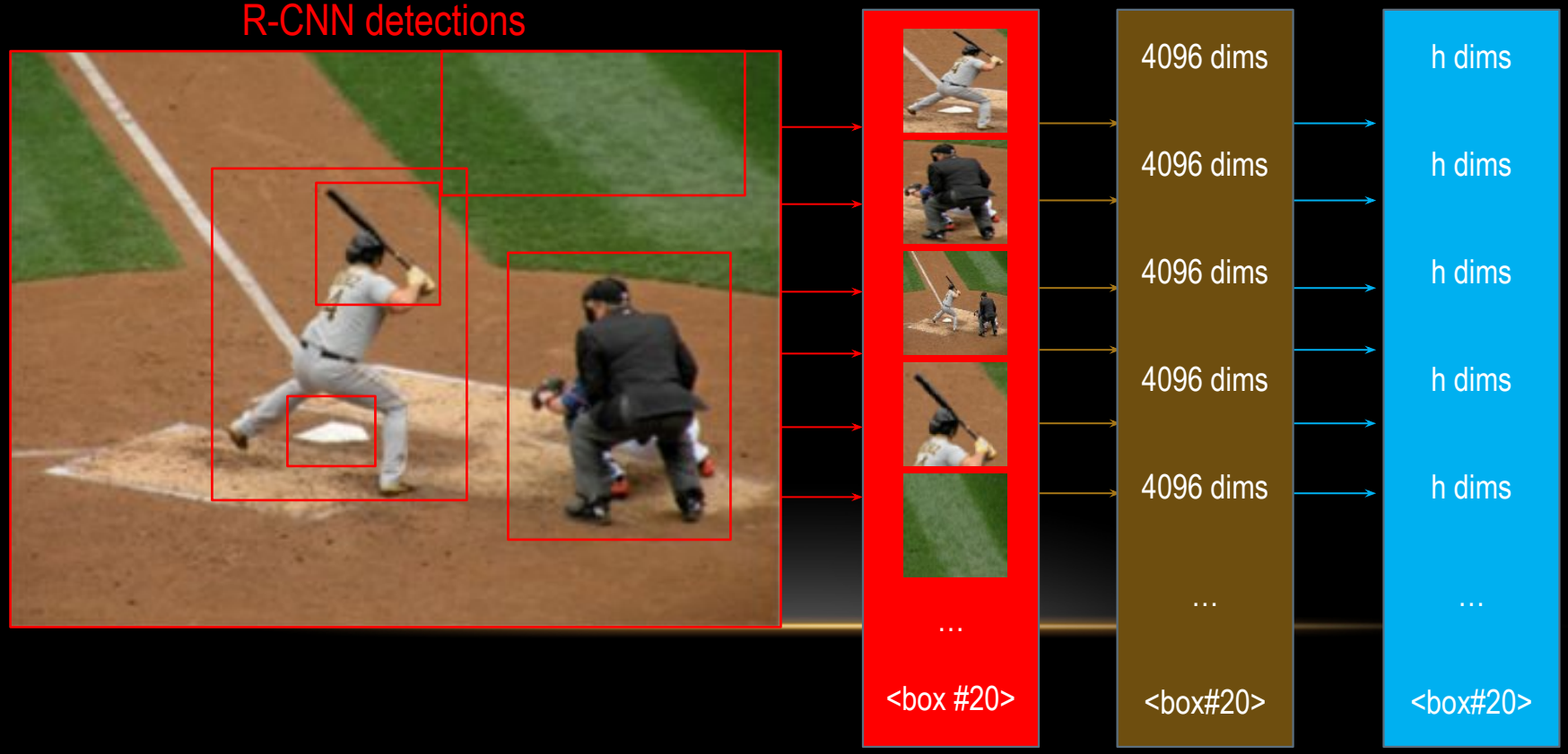
Use previous correspondence as input to multi-modal RNN which learns to generate novel descriptions

https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

EMBED IMAGE DATA

$$v = W_m[CNN_{\theta_c}(I_b)] + b_m,$$

Learnable Projection



EMBED TEXT DATA

The man at bat readies to swing at the pitch while the umpire looks on.

Index	Word
1	the
2	man
3	at
4	bat
5	readies
6	to
7	swing
8	at
9	the
10	pitch
11	while
12	the
13	umpire
14	looks
15	on
16	.

Thousands of dims

Index	at	the	bat	…
1	0	1	0	0
2	0	0	0	0
3	**1**	0	0	0
4	0	0	**1**	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	**1**	0	0
10	0	0	0	0
11	0	0	0	0
12	0	**1**	0	0
13	0	0	0	0
14	0	0	0	0
15	0	0	0	0
16	0	0	0	0

h-dimensions

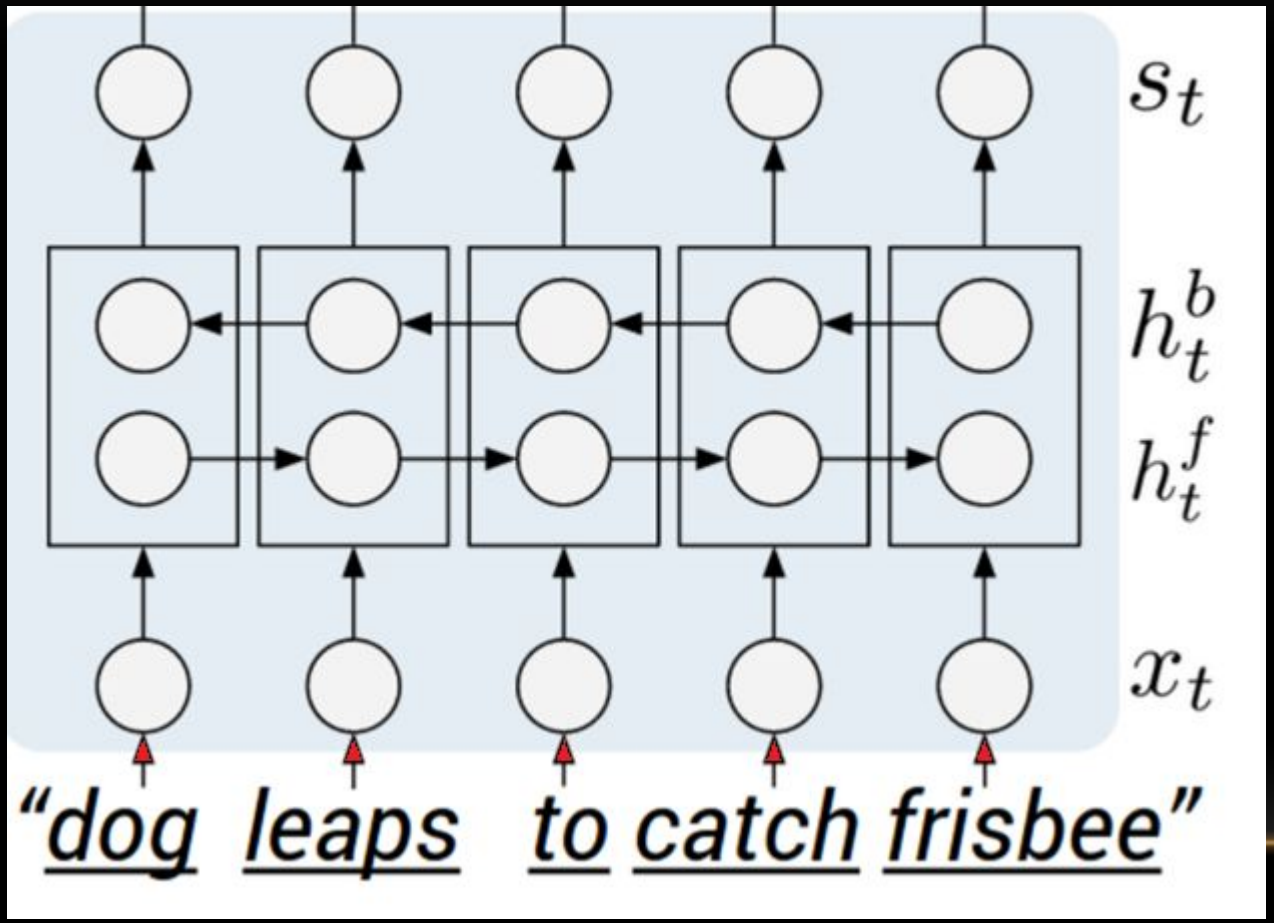
Index	0	1	2	…300
1	0.17	0.20	0.21	0.04
2	0.68	0.68	0.28	0.55
3	0.33	0.89	0.12	0.92
4	0.49	0.46	0.41	0.41
5	0.15	0.24	0.79	0.96
6	0.68	0.99	0.46	0.91
7	0.38	0.11	0.34	0.74
8	0.13	0.13	0.46	0.54
9	0.24	0.24	0.96	0.31
10	0.43	0.59	1.00	0.83
11	0.46	0.26	0.33	0.59
12	0.80	0.14	0.74	0.61
13	0.06	0.04	0.98	0.93
14	0.15	0.56	0.74	0.76
15	0.88	0.49	0.62	0.87
16	0.14	0.78	0.85	0.30

x_t

word2vec
autoencoding

EMBED TEXT DATA

BRNN



$$s_t = f(W_d(h_t^f + h_t^b) + b_d)$$

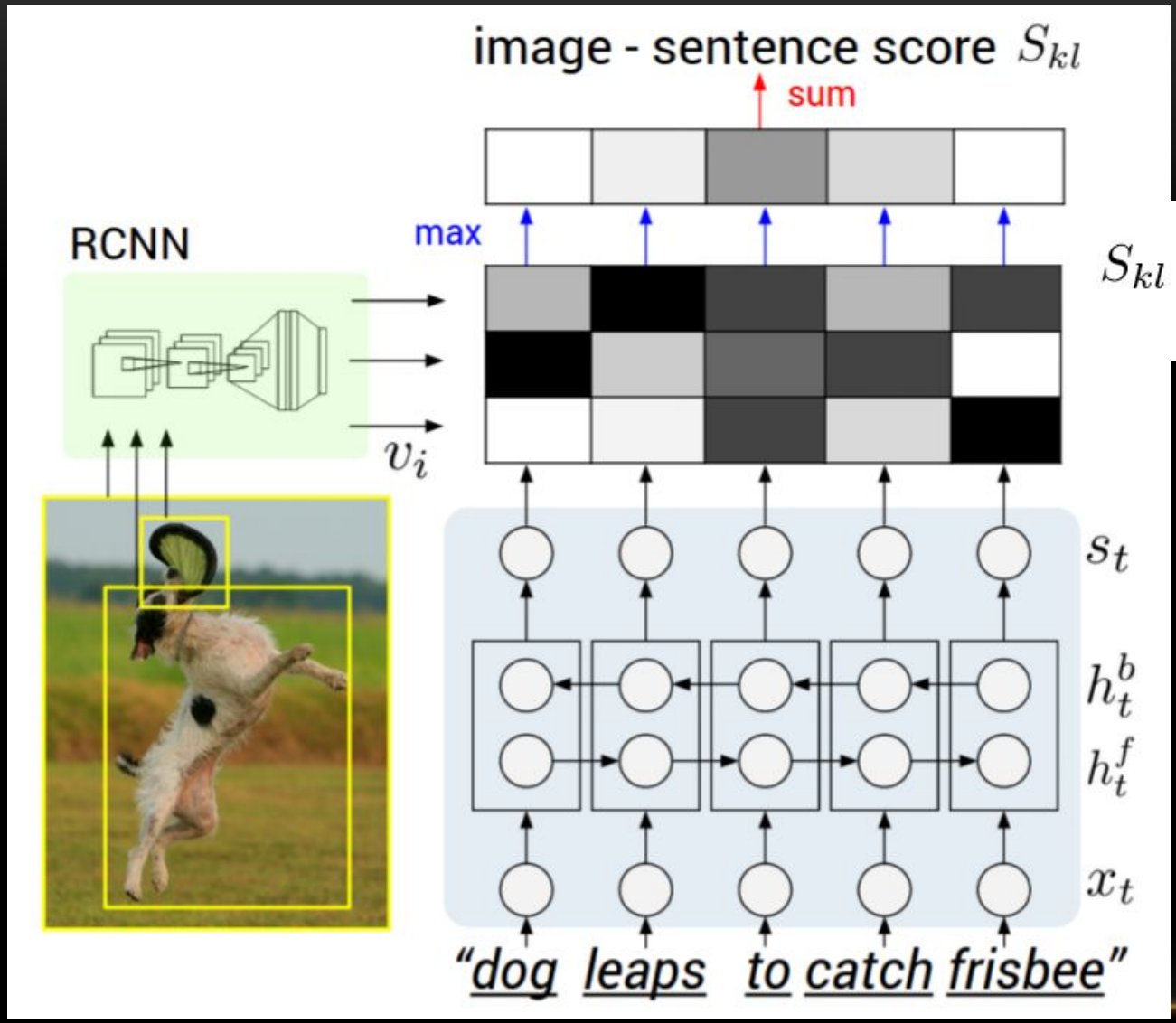
$$h_t^b = f(e_t + W_b h_{t+1}^b + b_b)$$

$$h_t^f = f(e_t + W_f h_{t-1}^f + b_f)$$

$$e_t = f(W_e x_t + b_e)$$

https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

IMAGE-SENTENCE SCORE (CONT)



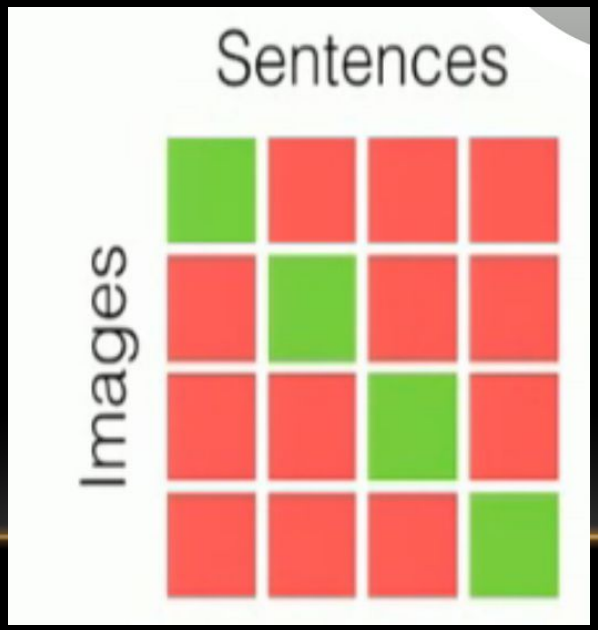
$$S_{kl} = \sum_{t \in g_l} max_{i \in g_k} v_i^T s_t.$$

https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

LOSS FUNCTION

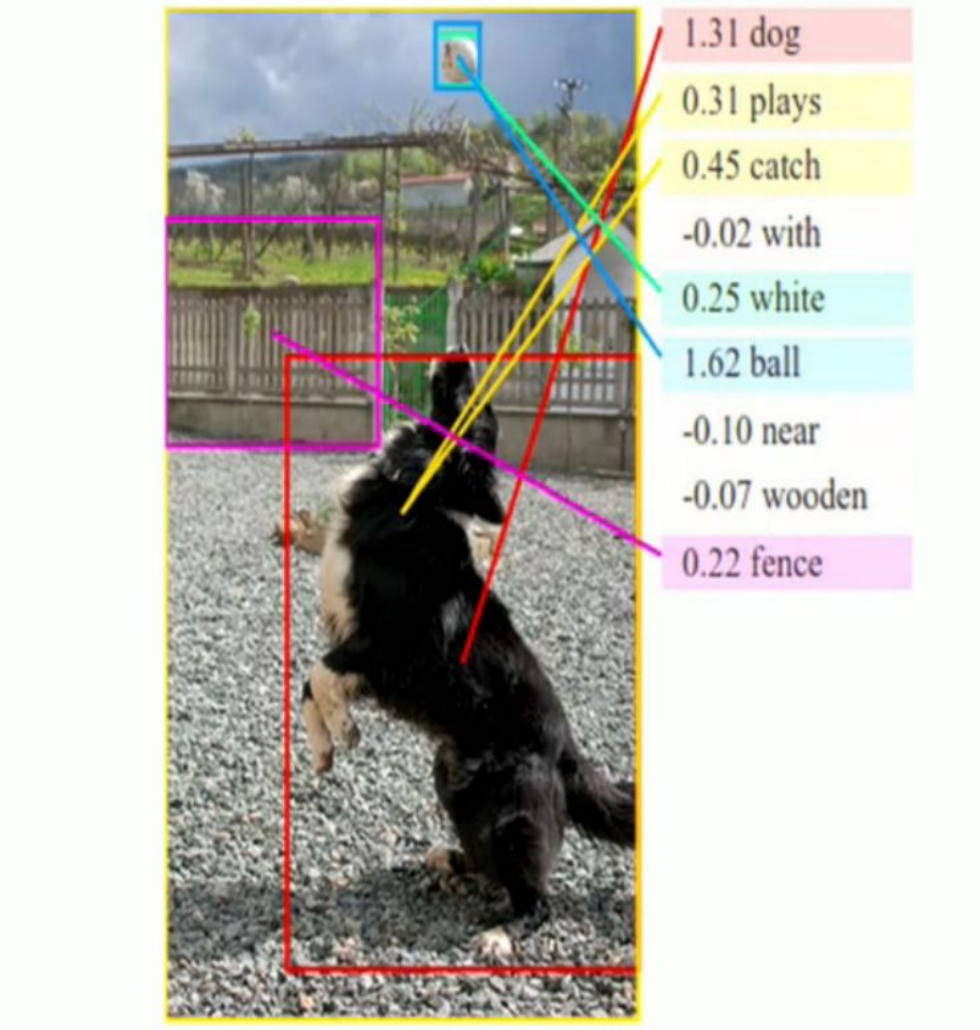
- Applied over a batch of training examples (set of images, sentances pair)
- Punishes wrong image, sentence pair for having high alignment score

$$\mathcal{C}(\theta) = \sum_k \Big[\underbrace{\sum_l max(0, S_{kl} - S_{kk} + 1)}_{\text{rank images}} + \underbrace{\sum_l max(0, S_{lk} - S_{kk} + 1)}_{\text{rank sentences}} \Big]$$



http://techtalks.tv/talks/deep-visual-semantic-alignments-for-generating-image-descriptions/61593/

ALIGNMENT MODEL: EXAMPLE



https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

Smoothing with an MRF

Let $a_j = t$ mean that the jth word w_j is aligned to the tth region r_t.

Then to independently align each word to the best region, minimize

$$E(a_1..a_N) = \sum_{a_j=t} -similarity(w_j, r_t)$$

But to encourage nearby words to point to the same region, add a penalty β when nearby words point to different regions:

$$E(a_1..a_N) = \sum_{a_j=t} -similarity(w_j, r_t) + \sum_{j=1..N-1} \beta[a_j = a_{j+1}]$$

The argmin can be found with dynamic programming.

From presentation by Ivan Vendrov, University of Toronto

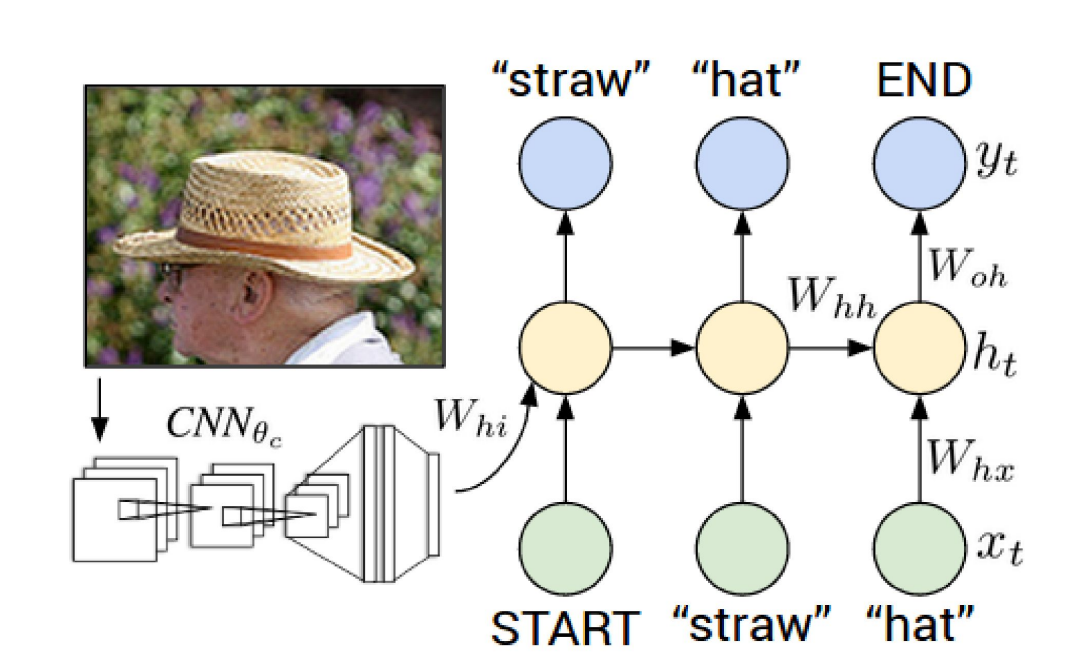
STEP 2: GENERATE NEW IMAGE DESCRIPTIONS USING IMAGE/TEXT CORRESPONDENCE

Use previous correspondence as input to multi-modal RNN which learns to generate novel descriptions

https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

GENERATE REGION DESCRIPTIONS

- Generative model is TRAINED on the bounding boxes + sentence snippets generated from the alignment model

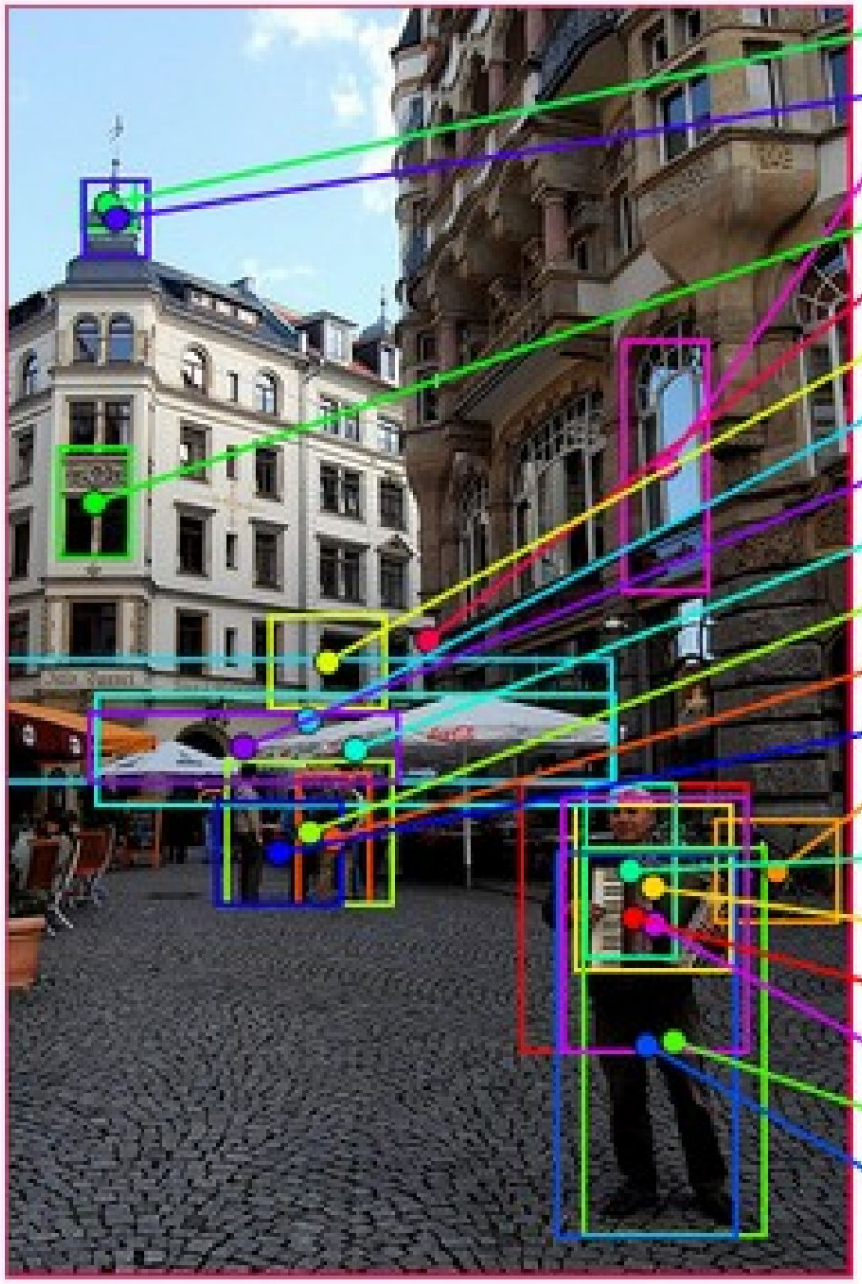


$$b_v = W_{hi}[CNN_{\theta_c}(I)] \quad (13)$$

$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1) \odot b_v) \quad (14)$$

$$y_t = softmax(W_{oh}h_t + b_o). \quad (15)$$

https://cs.stanford.edu/people/karpathy/cvpr2015.pdf



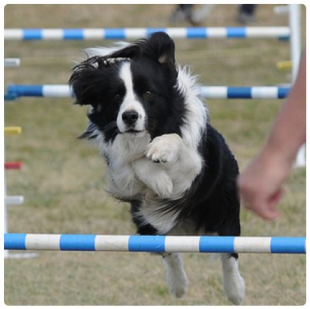
person is taking pictures
large white statue
building
front atm
front building
subway
guitar
red white crane
red umbrella
group people are walking
people walking
street
bicycle
man in suit
man in plaid shirt plays accordion
man playing musical instrument
band is playing music
man in black shirt jeans pants
man in black shirt is standing

Full-Frame Model

No image region to word embedding step, more like vanilla image captioning with RNN.

Input is full-frame image with text description

"girl in pink dress is jumping in air."



"black and white dog jumps over bar."

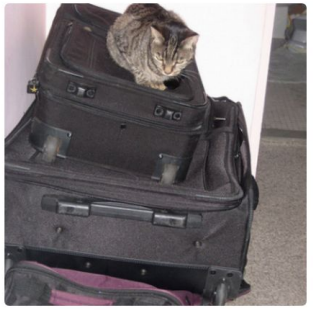
"young girl in pink shirt is swinging on swing."

"man in blue wetsuit is surfing on wave."

"little girl is eating piece of cake."

"baseball player is throwing ball in game."

"woman is holding bunch of bananas."



"black cat is sitting on top of suitcase."

"a young boy is holding a baseball bat."

"a cat is sitting on a couch with a remote control."

"a woman holding a teddy bear in front of a mirror."

"a horse is standing in the middle of a road."

Model Evaluation

1) Evaluating Image - Word Alignments

2) Evaluating Descriptions generated on a full image

3) Evaluating a densely annotated image

$$S_{kl} = \sum_{t \in g_l} max_{i \in g_k} v_i^T s_t.$$

Evaluating the Alignment Model

Test Image

Sentence	Image-Sentence Score S_{kl}	Rank
“test sentence 1”	4.5	1
“test sentence 2”	0.6	4
“test sentence 3”	**3.5**	**2**
“test sentence 4”	2.8	3
...	...	...

Ground Truth

Metrics Used for Alignment Model

Recall@K: the fraction of times a correct item was found among the top K results

Med r: Given a caption OR image, rank image-sentence scores

Metric is the median rank of ground truth

Alignment Model Results

Model	Image Annotation				Image Search			
	R@1	**R@5**	**R@10**	**Med *r***	**R@1**	**R@5**	**R@10**	**Med *r***
Flickr30K								
SDT-RNN (Socher et al. [49])	9.6	29.8	41.1	16	8.9	29.8	41.1	16
Kiros et al. [25]	14.8	39.2	50.9	10	11.8	34.0	46.3	13
Mao et al. [38]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
Donahue et al. [8]	17.5	40.3	50.8	9	-	-	-	-
DeFrag (Karpathy et al. [24])	14.2	37.7	51.3	10	10.2	30.8	44.2	14
Our implementation of DeFrag [24]	19.2	44.5	58.0	6.0	12.9	35.4	47.5	10.8
Our model: DepTree edges	20.0	46.6	59.4	5.4	15.0	36.5	48.2	10.4
Our model: BRNN	**22.2**	**48.2**	**61.4**	**4.8**	**15.2**	**37.7**	**50.5**	**9.2**
Vinyals et al. [54] (more powerful CNN)	23	-	63	5	17	-	57	8
MSCOCO								
Our model: 1K test images	38.4	69.9	80.5	1.0	27.4	60.2	74.8	3.0
Our model: 5K test images	16.5	39.2	52.0	9.0	10.7	29.6	42.2	14.0

https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

Metrics Used for Generating Image Descriptions

BLEU, METEOR and CIDEr scores. scale: 0 (worst) - 100 (best)

- Evaluate a candidate sentence by measuring how well it matches a set of five reference sentences written by humans

BLUE score explanation

- Measures number of words in input that are matched divided by the length of the output
- B-n matches n-grams from the input
  - B-1 - unigram score, how much information is retained
  - Higher n-gram count (B-3, B-4) relates more to fluency of translation

Full Image Description Results

	Flickr8K				Flickr30K				MSCOCO 2014					
Model	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	METEOR	CIDEr
Nearest Neighbor	—	—	—	—	—	—	—	—	48.0	28.1	16.6	10.0	15.7	38.3
Mao et al. [38]	58	28	23	—	55	24	20	—	—	—	—	—	—	—
Google NIC [54]	63	41	27	—	66.3	42.3	27.7	18.3	66.6	46.1	32.9	24.6	—	—
LRCN [8]	—	—	—	—	58.8	39.1	25.1	16.5	62.8	44.2	30.4	—	—	—
MS Research [12]	—	—	—	—	—	—	—	—	—	—	—	21.1	20.7	—
Chen and Zitnick [5]	—	—	—	14.1	—	—	—	12.6	—	—	—	19.0	20.4	—
Our model	57.9	38.3	24.5	16.0	57.3	36.9	24.0	15.7	62.5	45.0	32.1	23.0	19.5	66.0

https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

Evaluating Relevance of Text Snippets at Regions

Used AMT to created new test-set by drawing bounding boxes in images and annotating regions

Compared labelled regions from alignment model to this test set

Region Snippets

Model	B-1	B-2	B-3	B-4
Human agreement	61.5	45.2	30.1	22.0
Nearest Neighbor	22.9	10.5	0.0	0.0
RNN: Fullframe model	14.2	6.0	2.2	0.0
RNN: Region level model	**35.2**	**23.0**	**16.1**	**14.8**

https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

Paper Weaknesses

Challenges/Difficulties faced

- No region-annotated training data
- Learn local visual semantics (even rare & small objects) through global descriptions
- No established evaluations method for region-annotated training data

Issues with Paper

- Limitation: Includes 2 separate models - not trained end-to-end
- No LSTM used (only BRNN and RNN)
- No Model Ensemble
- Object Detection was vanilla R-CNN extracting region features with Alexnet
- Lost to Google's Show and Tell

Demo

Backup slides

Evaluation

In this paper, three evaluations were performed:

- Full Model - “BRNN” Alignment Evaluation
- Full-Frame Description Evaluation
- Local Description Evaluation