Deep Visual-Semantic Alignments for Generating Image Descriptions

Andrej Karpathy, Fei-Fei Li

Overview

Goal

"To generate dense, free-form descriptions of images"

Not to generate Flickr-like descriptions!

Figure from (Karpathy and Li 2014)

Motivation

- It's hard
- Humans can do it

Challenges

- Big Challenge: build a model that reasons jointly about vision and language

- Deep Learning Manifesto: "The model should be free of assumptions about specific hard-coded templates, rules or categories, and instead rely primarily on training data"

- Immediate challenge: how do we learn a model that generates dense, region-level descriptions from training data of sparse, image-level descriptions?

- Training data is *extremely* noisy

"trampolines are fun way to exercise"

Contributions

1. Infer region-word alignments
(R-CNN + BRNN + MRF)

2. Generative model of image descriptions
(new RNN architecture)

3. Generate region-level descriptions

Technical Approach

Inferring Alignments

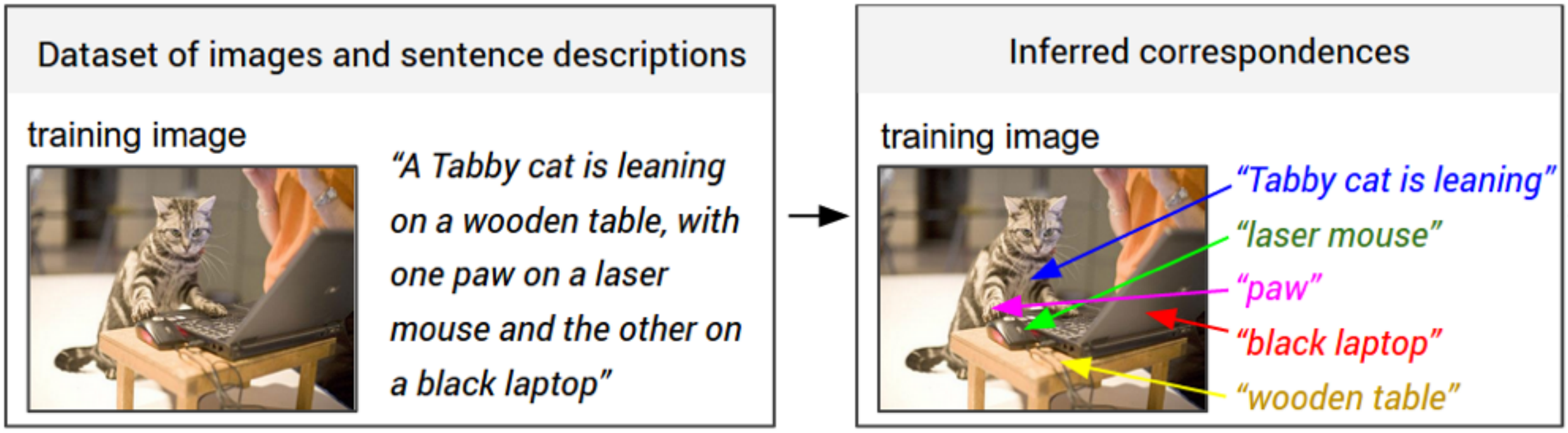

Figure from (Karpathy and Li 2014)

R-CNN: *Regions with CNN features*

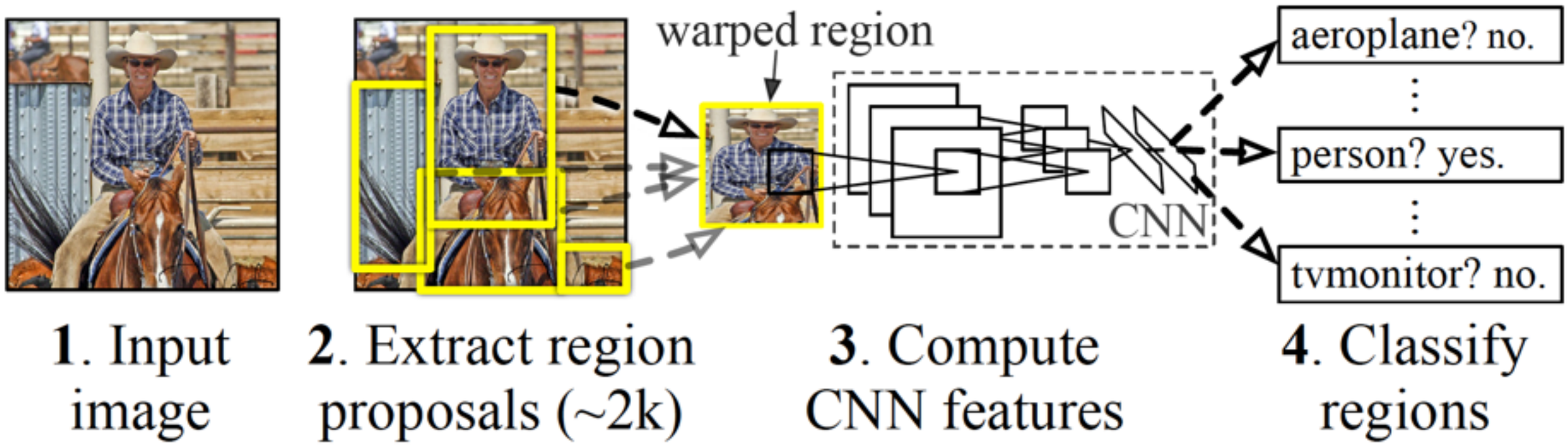

Figure from (Girshick et al 2014) - http://www.cs.berkeley.edu/~rbg/papers/r-cnn-cvpr.pdf

Inferring Word Alignments

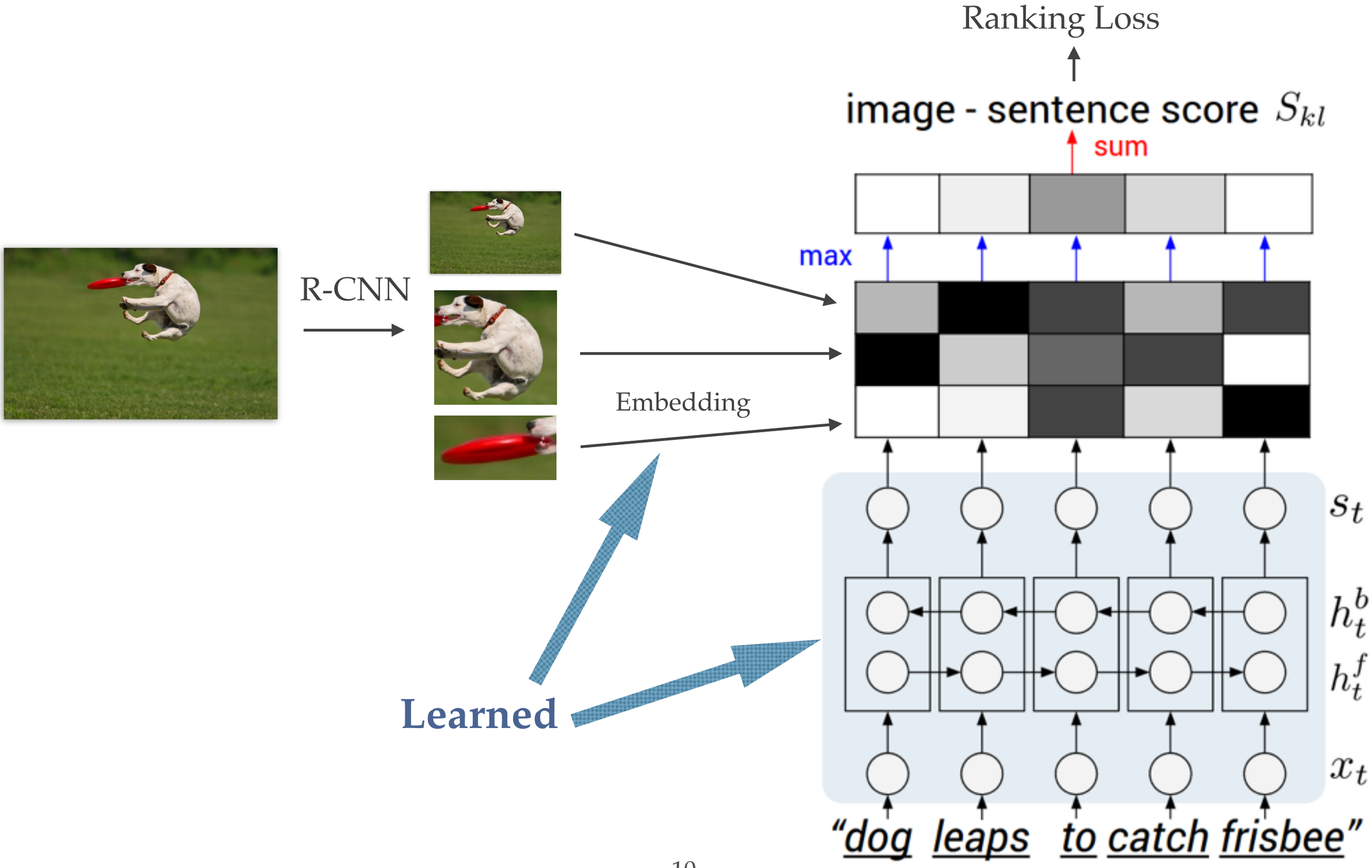

Figure adapted from (Karpathy and Li 2014)

Inferring Segment Alignments

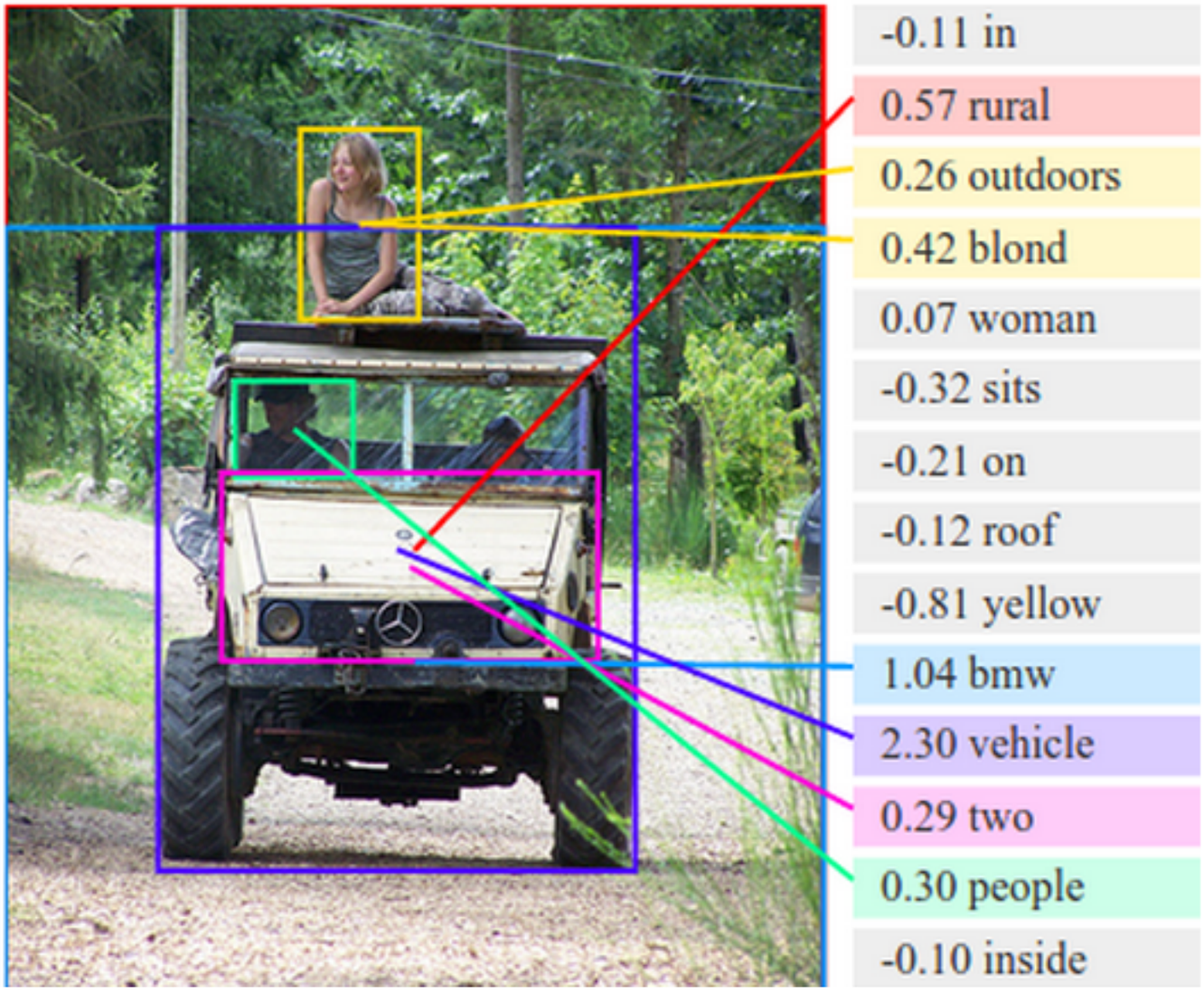

Figure generated by
http://cs.stanford.edu/people/karpathy/deepimagesent/rankingdemo/

Smoothing via MRF

?

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.

Generating Descriptions

- RNN architecture
- doesn't try to embed images and descriptions into a common space
- uses a "context" and the previous word to determine the probability distribution over the next word
- *adds* image features to the initial context

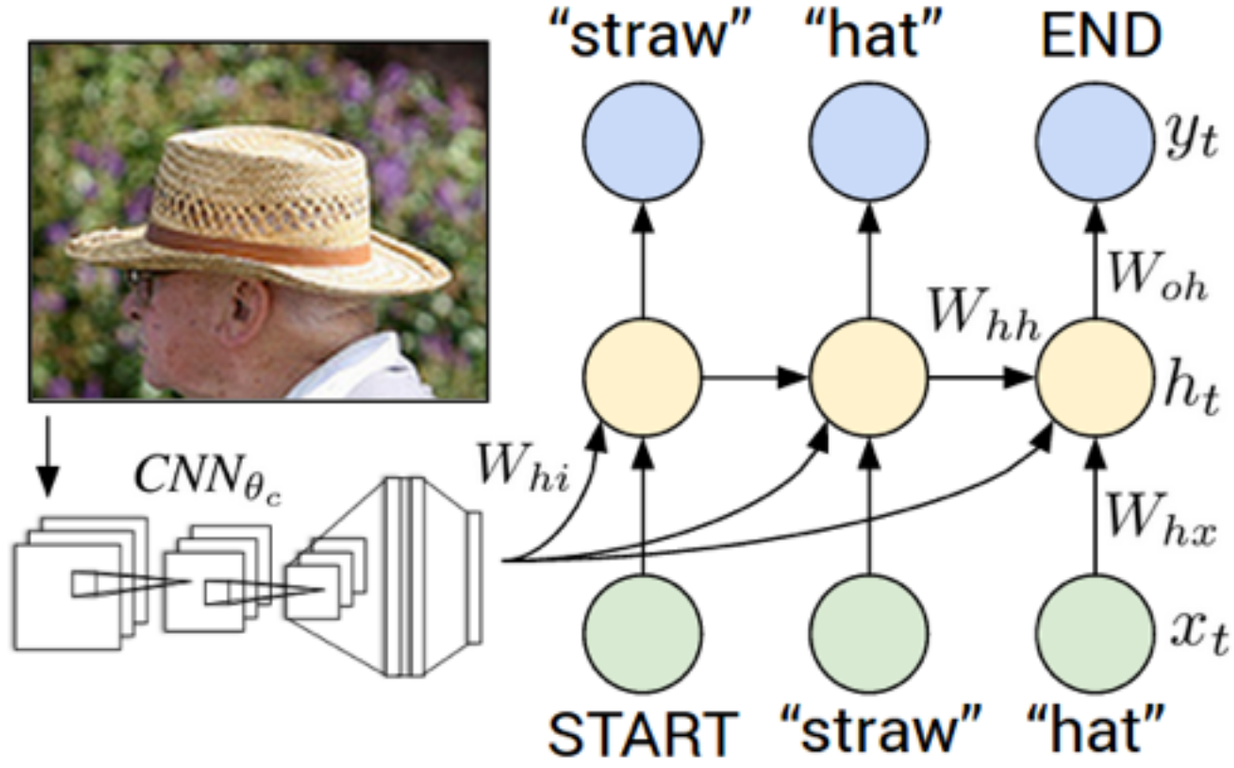

Figure from (Karpathy and Li 2014)

Region-Level Descriptions

- Train the description generating RNN on the aligned regions + sentence fragments!
- At test time, run on many bounding boxes separately

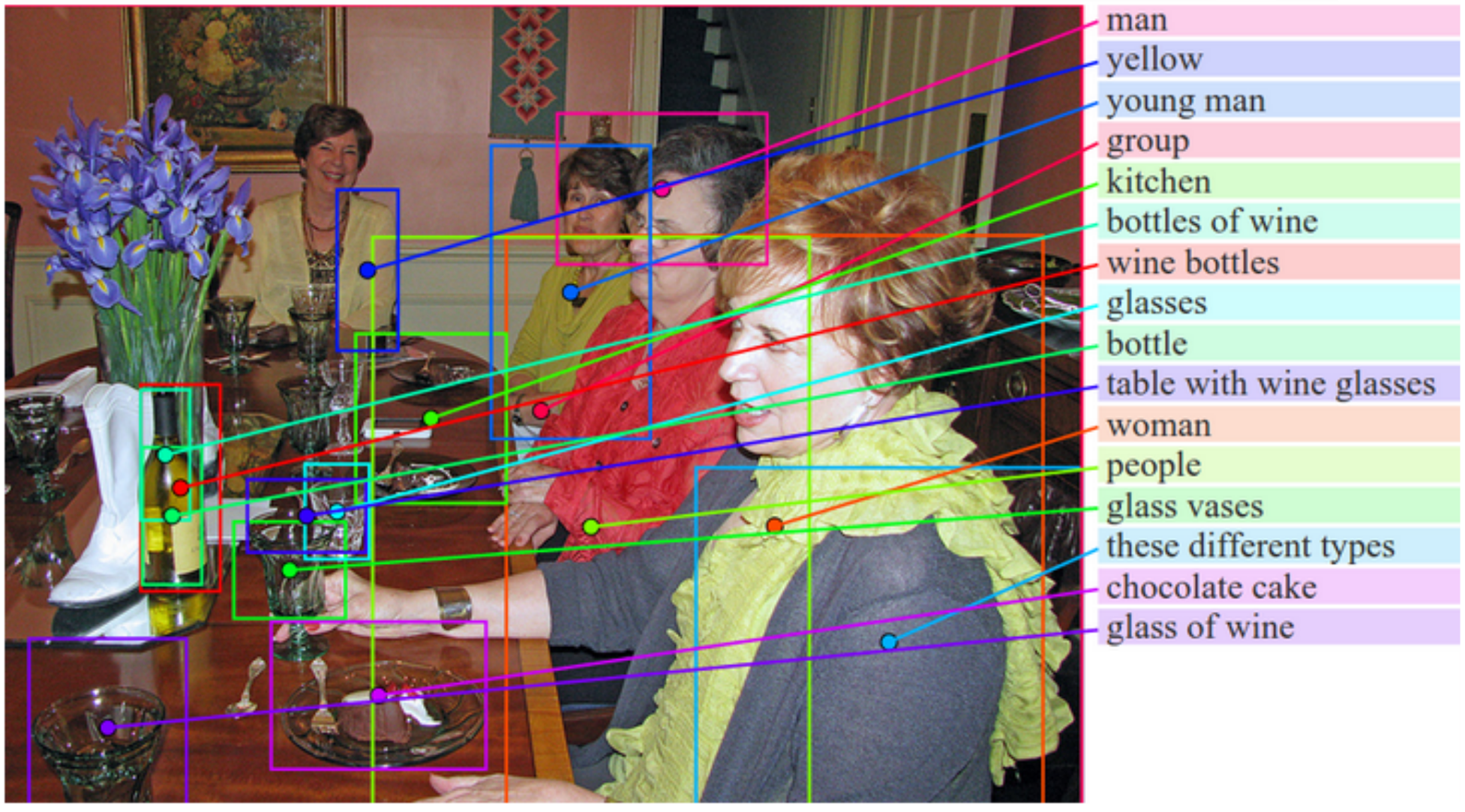

Figure from (Karpathy and Li 2014)

Evaluation

Evaluation

1. Infer region-word alignments
   (R-CNN + BRNN + MRF)

2. Generative model of image descriptions
   (new RNN architecture)

3. Generate region-level descriptions

evaluated with

Image-Sentence Ranking
(Flickr8k, Flickr30k, MSCOCO)

BLEU on new region dataset

BLEU, Perplexity
(Flickr8k, Flickr30k, MSCOCO)

Alignment Evaluation

Model	Image Annotation R@1	R@5	R@10	Med *r*	Image Search R@1	R@5	R@10	Med *r*
Flickr30K								
DeViSE (Frome et al. [10])	4.5	18.1	29.2	26	6.7	21.9	32.7	25
SDT-RNN (Socher et al. [42])	9.6	29.8	41.1	16	8.9	29.8	41.1	16
Kiros et al. [19]	14.8	39.2	50.9	10	11.8	34.0	46.3	13
Mao et al. [31]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
DeFrag (Karpathy et al. [18])	14.2	37.7	51.3	10	10.2	30.8	44.2	14
Our implementation of DeFrag [18]	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**

Figure from (Karpathy and Li 2014)

- ranking evaluation as a proxy for alignment quality
  - Image Annotation: given image, find closest training description
  - Image Search: given description, find closest image

- uses image-sentence alignment score (sum of word-region alignments), instead of distance in multimodal space.

- outperform across the board, but is this a fair comparison?

Description Evaluation

	Flickr8K				Flickr30K				MSCOCO			
Method of generating text	$\mathcal{PPL}$	B-1	B-2	B-3	$\mathcal{PPL}$	B-1	B-2	B-3	$\mathcal{PPL}$	B-1	B-2	B-3
4 sentence references												
Human agreement	-	0.63	0.40	0.21	-	0.69	0.45	0.23	-	0.63	0.41	0.22
5 sentence references												
Generating: RNN	-	0.45	0.21	0.09	-	0.47	0.21	0.09	-	0.53	0.28	0.15
Mao et al. [31]	24.39	**0.58**	0.28	**0.23**	35.11	**0.55**	0.24	**0.20**	-	-	-	-
Generating: RNN (OxfordNet CNN [40])	**22.66**	0.51	**0.31**	0.12	**21.20**	0.50	**0.30**	0.15	19.64	0.57	0.37	0.19

Figure from (Karpathy and Li 2014)

- BLEU scores surprisingly close to human agreement

- Perplexity (how well the model predicts test descriptions) is surprisingly low at 20 bits per sentence
  - vs ~8 bits per word for the best language models

Region-Level Evaluation

Method of generating text	B-1	B-2	B-3
Human agreement	0.54	0.33	0.16
Ranking: Nearest Neighbor	0.14	0.03	**0.07**
Generating: Full frame model	0.12	0.03	0.01
Generating: Region level model	**0.17**	**0.05**	0.01

Figure from (Karpathy and Li 2014)

- Used new dataset of region-level annotations
  - 1469 annotations in 237 images
  - average length of annotation: 4.13 words

- Model performs much worse relative to humans
  - brevity means BLEU penalties sharper
  - model trained on *extremely* noisy data

Discussion

Priors (Deep Learning Heresies)

- descriptions are object-centric (hence use R-CNN detections)

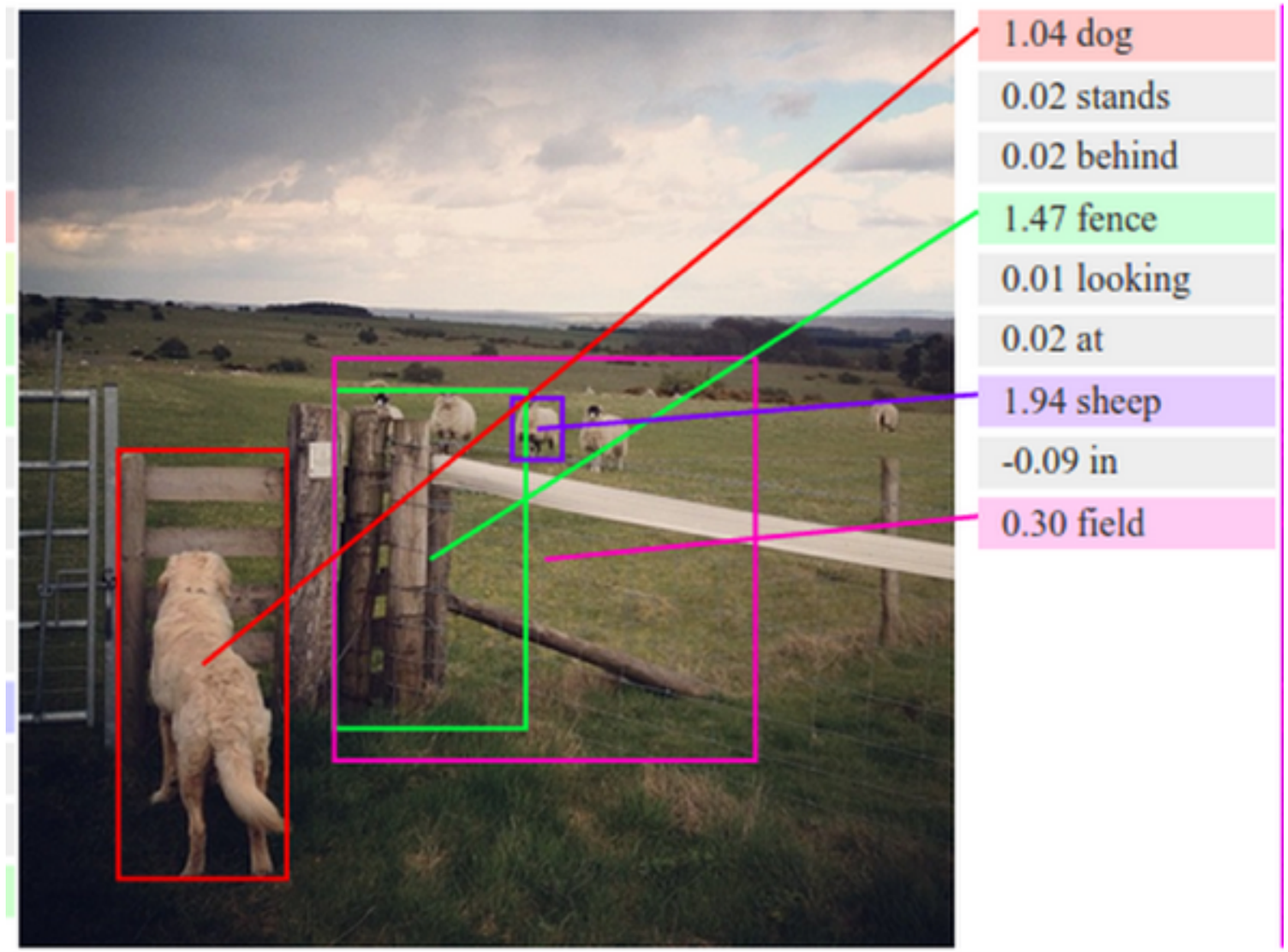

Figure generated by
http://cs.stanford.edu/people/karpathy/deepimagesent/rankingdemo/

Priors (Deep Learning Heresies)

- image-sentence scores are sums of word-region scores

Best image result for "small dog sleeping in living room chair"

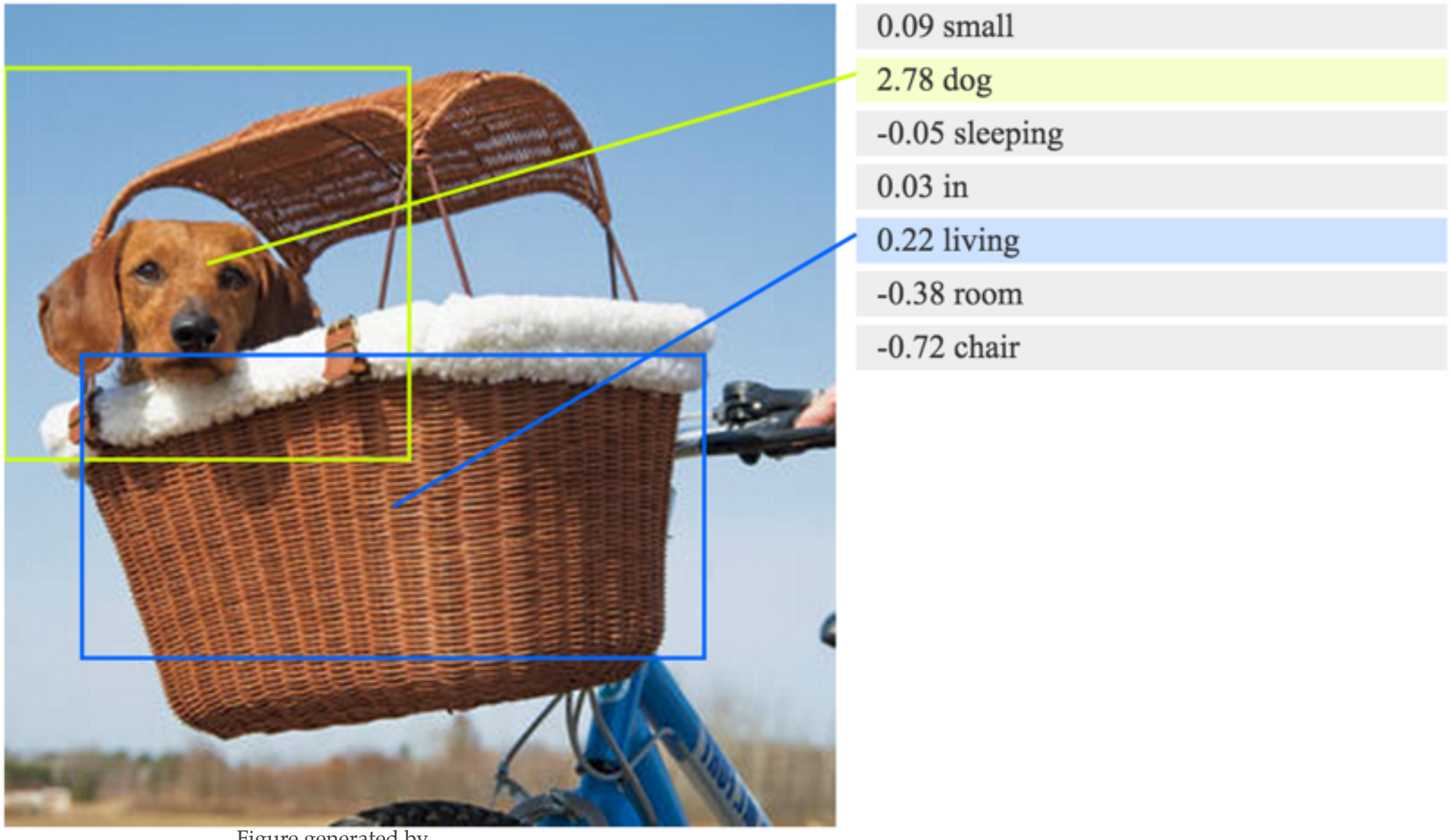

Figure generated by
http://cs.stanford.edu/people/karpathy/deepimagesent/rankingdemo/

Summary

- New goal: dense descriptions

- New model to infer word-region alignments
  - BRNN to embed words
  - R-CNN to embed images
  - Trained using ranking loss over alignment score

- New description generation model
  - RNN with additive image context

- New region-level annotation dataset