Deep Visual-Semantic Alignments for Generating Image Descriptions

[Andrej Karpathy, Li Fei-Fei, 14.04.2015]

Доклад подготовил
Дубов Дмитрий

Content

1. Task / Objective
2. Motivation and related work
3. New Model
4. Experiments
5. More examples

Task / Objective

Model that generates **natural language descriptions** of images and their regions.

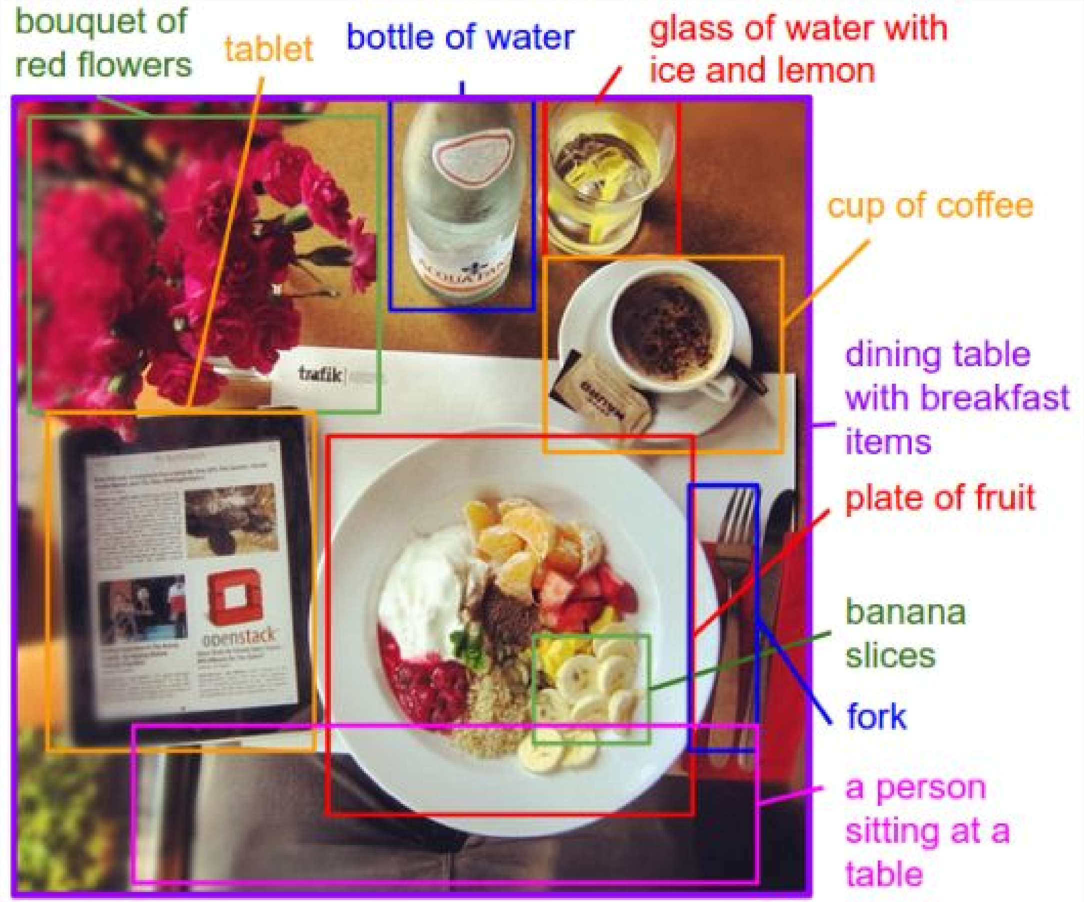
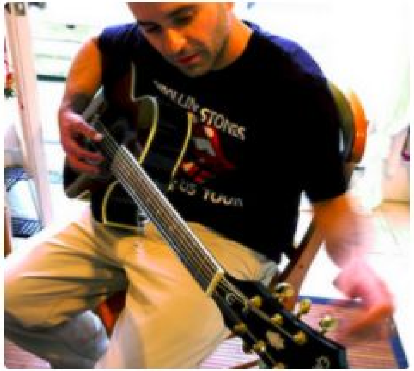
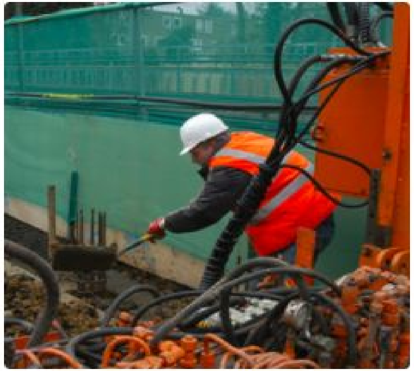

bouquet of red flowers
tablet
bottle of water
glass of water with ice and lemon
cup of coffee
dining table with breakfast items
plate of fruit
banana slices
fork
a person sitting at a table

Figure 1. Motivation/Concept Figure: Our model treats language as a rich label space and generates descriptions of image regions.

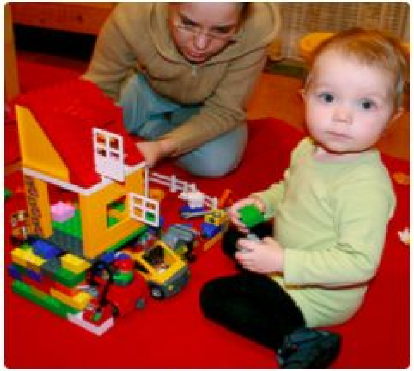
!SPOILERS!



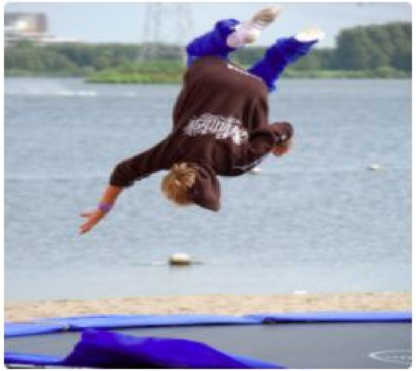
"man in black shirt is playing guitar."



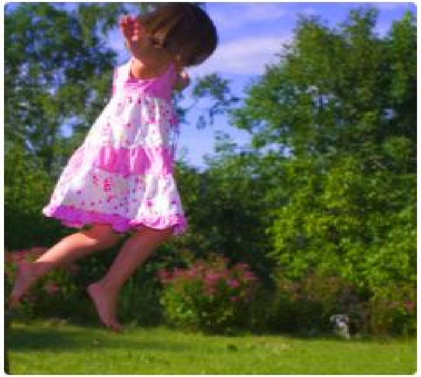
"construction worker in orange safety vest is working on road."



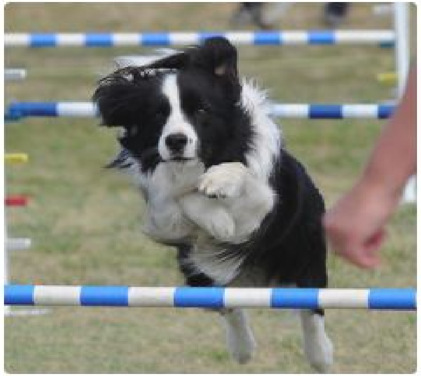
"two young girls are playing with lego toy."



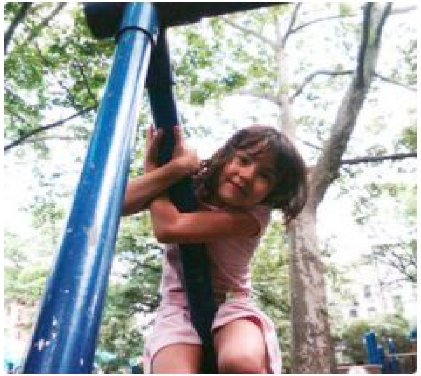
"boy is doing backflip on wakeboard."



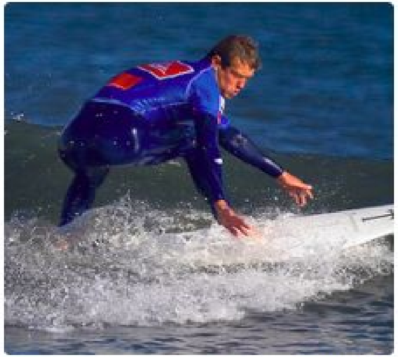
"girl in pink dress is jumping in air."



"black and white dog jumps over bar."



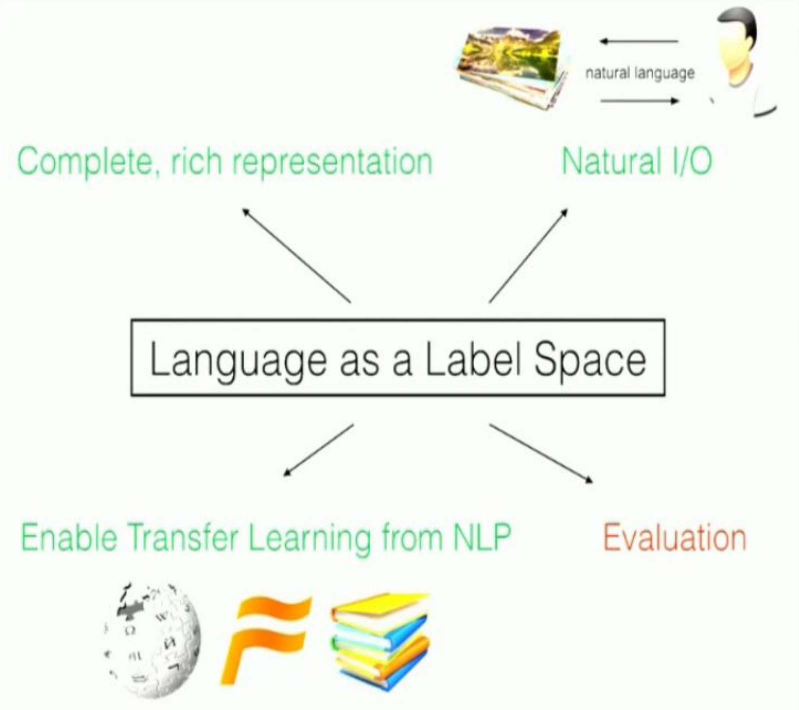
"young girl in pink shirt is swinging on swing."



"man in blue wetsuit is surfing on wave."

Motivation

- Most previous works are focused on fixed set of categories
- We want to use all the variety of the language to genereate dense image description
- Treat language - as a label space



natural language

Complete, rich representation

Natural I/O

Language as a Label Space

Enable Transfer Learning from NLP

Evaluation

Related Work

- Retrievial solutions:
    - 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

Our Model

- Associates object location in image and sentence segments
- Generate descriptions for test images that significantly outperforms baselines

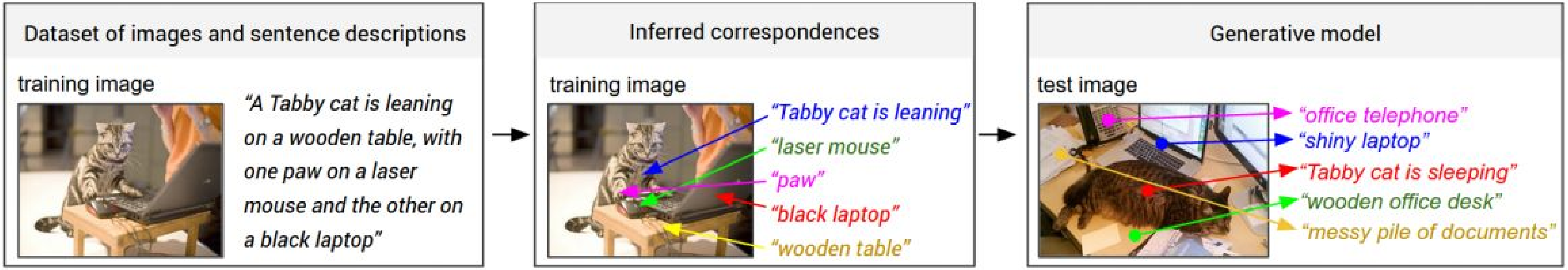
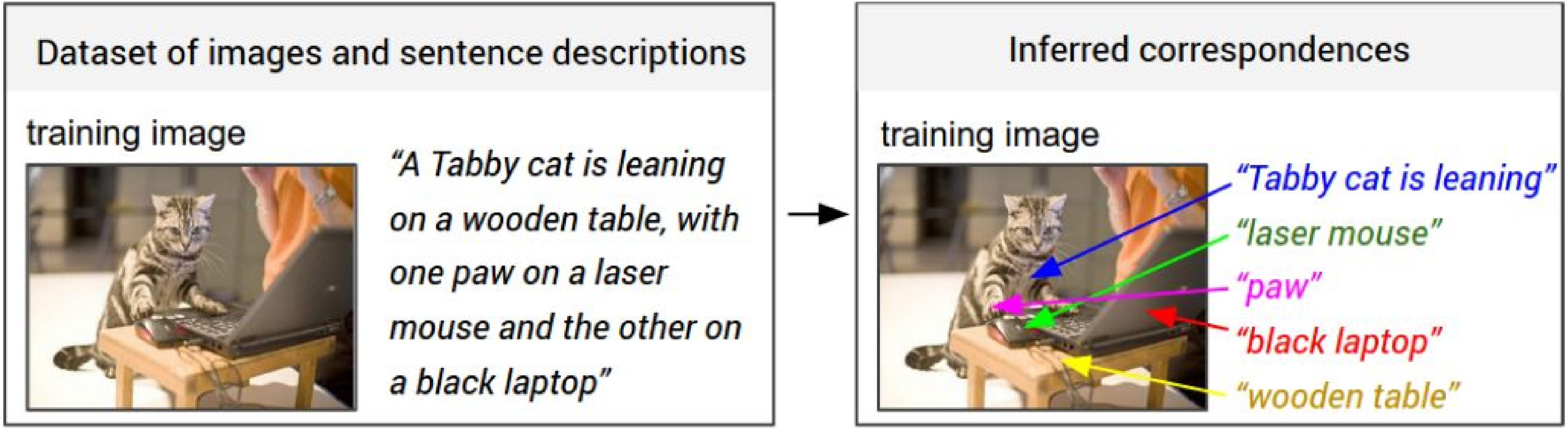


Figure 2. Overview of our approach. A dataset of images and their sentence descriptions is the input to our model (left). Our model first infers the correspondences (middle, Section 3.1) and then learns to generate novel descriptions (right, Section 3.2).

STEP1:
Learning to align visual and language data



Dataset of images and sentence descriptions

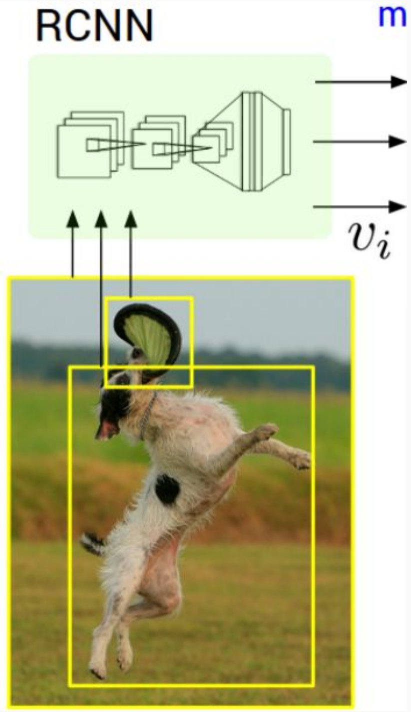
“A Tabby cat is leaning on a wooden table, with one paw on a laser mouse and the other on a black laptop”

Inferred correspondences

Learning to align visual and language data: **Representing Images**

- Detect objects in every image with a Region Convolutional Neural Network.
- Top 19 detected locations in addition to the whole image.
- For every bounding box I_b get vector:

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

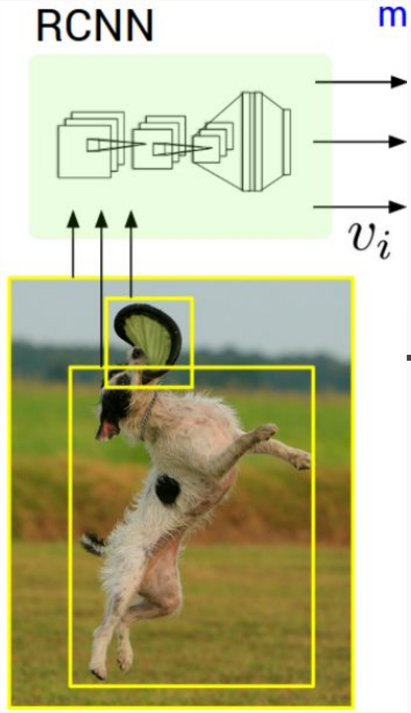


$CNN_{\theta_c}(I_b)$ - transforms the pixels inside bounding box into 4096-dimensional activations of the fully connected layer immediately before the classifier. [pre-trained on ImageNet]

W_m - *h × 4096 dimension; h is the size of the embedding space*

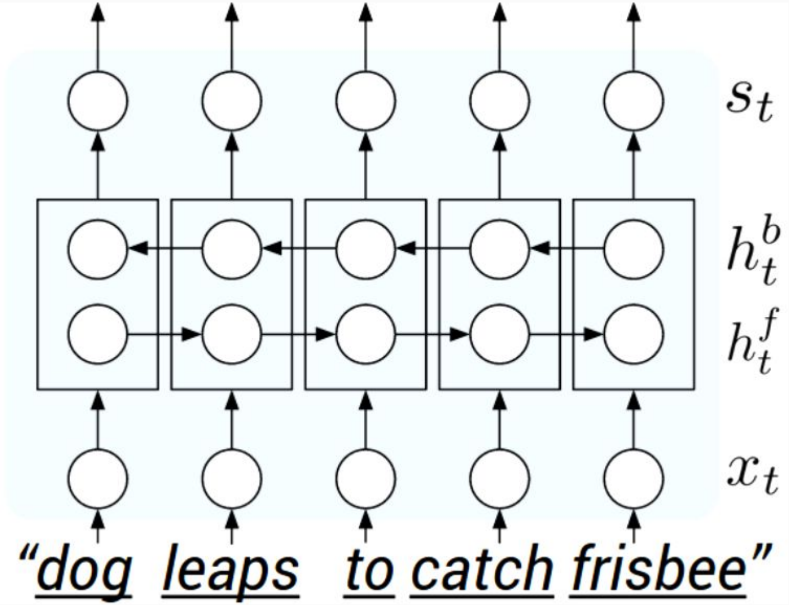
b_m - bias

Learning to align visual and language data: **Representing Images**



$$\{v_i \mid i = 1 \ldots 20\}$$

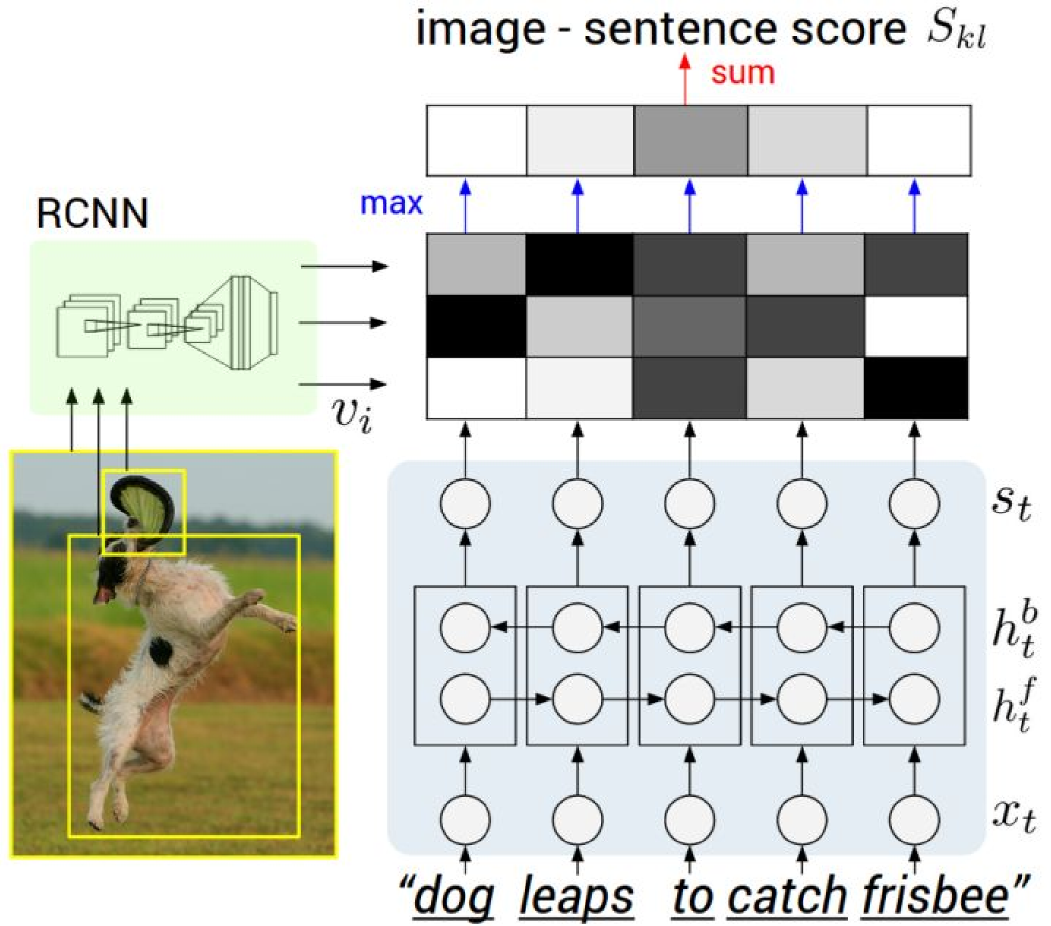
Learning to align visual and language data: **Representing Sentences**



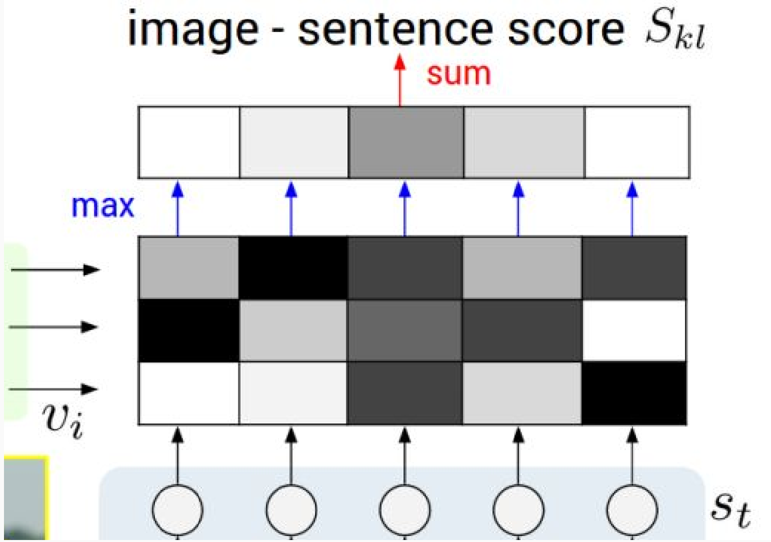
$$
\begin{aligned}
x_t &= W_w \mathbb{I}_t \\
e_t &= f(W_e x_t + b_e) \\
h_t^f &= f(e_t + W_f h_{t-1}^f + b_f) \\
h_t^b &= f(e_t + W_b h_{t+1}^b + b_b) \\
s_t &= f(W_d(h_t^f + h_t^b) + b_d).
\end{aligned}
$$

$$f : x \mapsto max(0, x)$$

Alignment objective



Alignment objective



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

g_k - is the set of image fragments in image k

g_l - is the set of sentence fragments in sentence l.

Here, every word aligns to the single best image region.

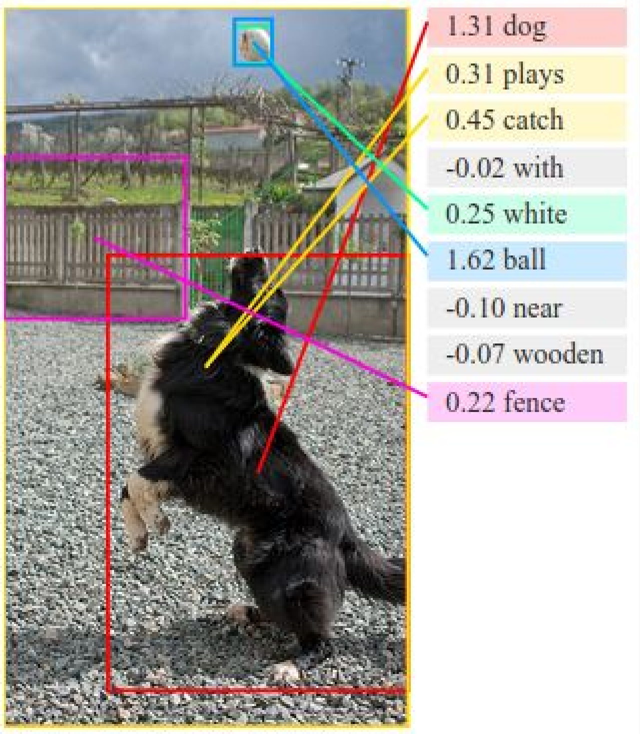
Loss

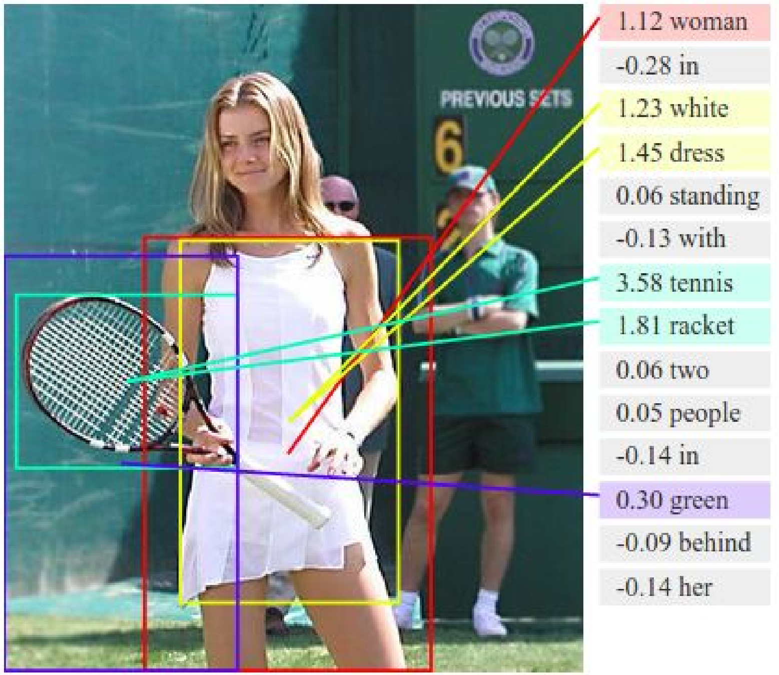
$$\mathcal{C}(\theta) = \sum_k \Big[\underbrace{\sum_l max(0, S_{kl} - S_{kk} + 1)}_{\text{rank images}} + \underbrace{\sum_l max(0, S_{lk} - S_{kk} + 1)}_{\text{rank sentences}} \Big]$$

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

This objective encourages aligned image-sentences pairs to have a higher score than misaligned pairs.

Examples of alignment





Smoothing the aligned words with bounding boxes using MRF

Given a sentence with N words and an image with M bounding boxes, we introduce the latent alignment variables: $a_j \in \{1 \ldots M\}$, $j = 1 \ldots N$

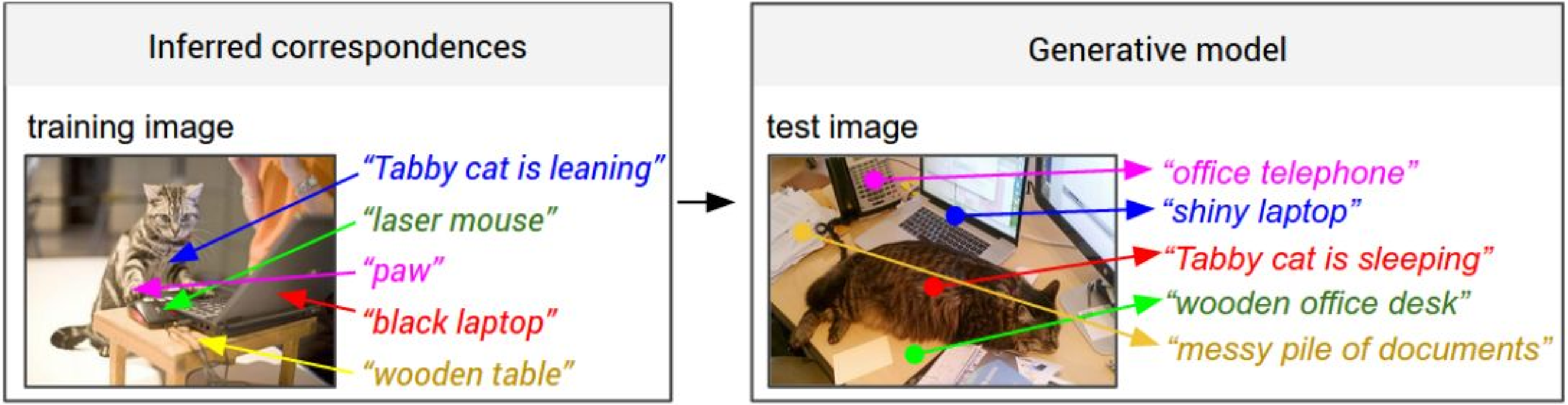
$$E(\mathbf{a}) = \sum_{j=1 \ldots N} \psi_j^U(a_j) + \sum_{j=1 \ldots N-1} \psi_j^B(a_j, a_{j+1})$$

$$\psi_j^U(a_j = t) = v_i^T s_t$$

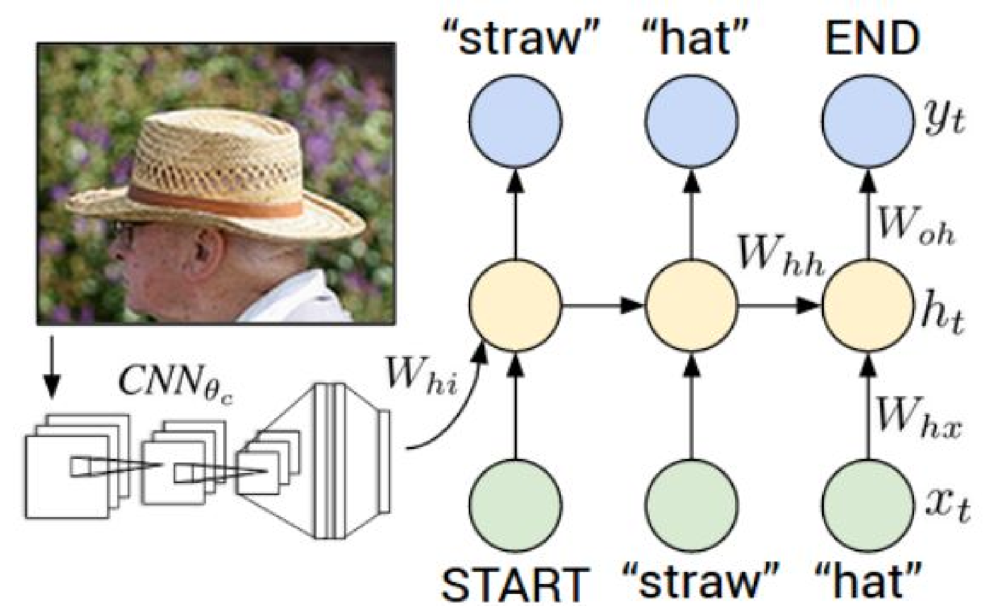
$$\psi_j^B(a_j, a_{j+1}) = \beta \mathbb{1}[a_j = a_{j+1}].$$

β is a hyperparameter that controls the affinity towards longer word phrases.

STEP 2:
Multimodal RNN for generating descriptions



Multimodal Recurrent Neural Network for generating descriptions



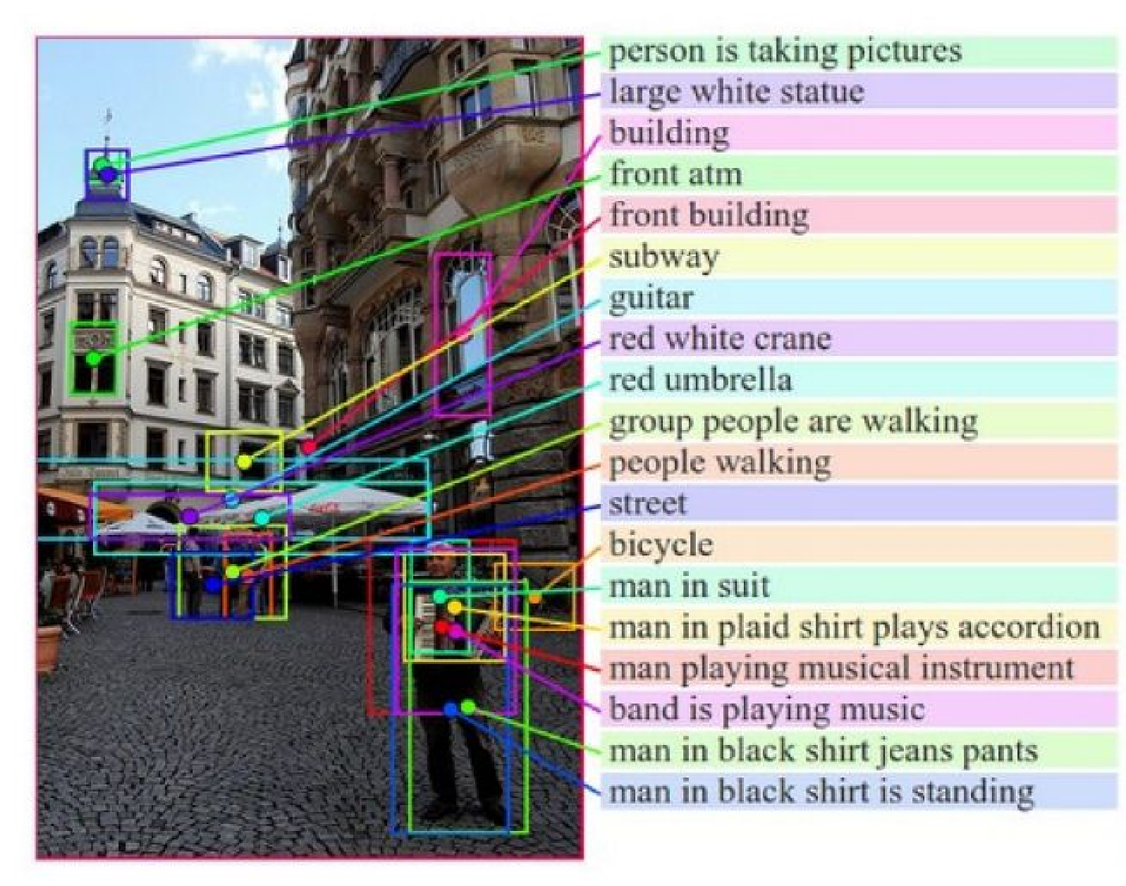
The cost function is to maximize the log probability assigned to the target labels (i.e. Softmax classifier).

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

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

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

Multimodal Recurrent Neural Network for generating descriptions, **Example**



Experiments: **Alignment Model**

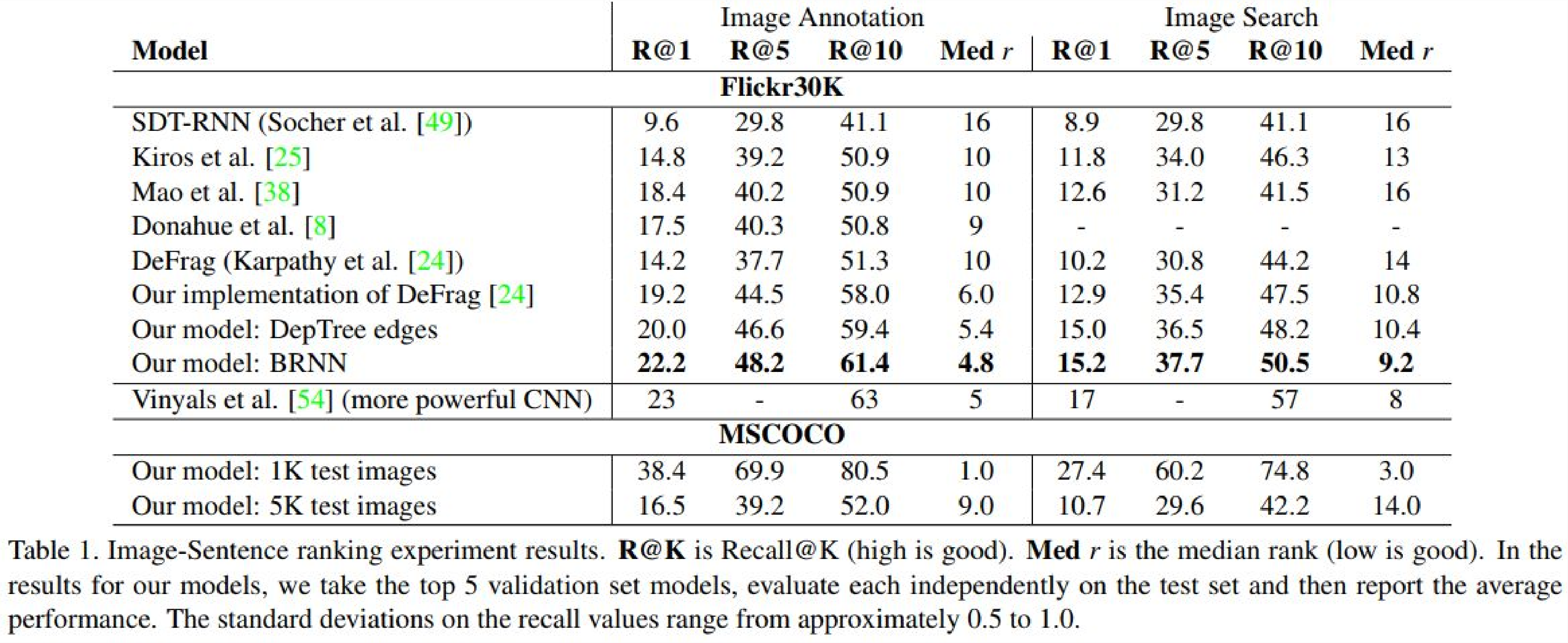
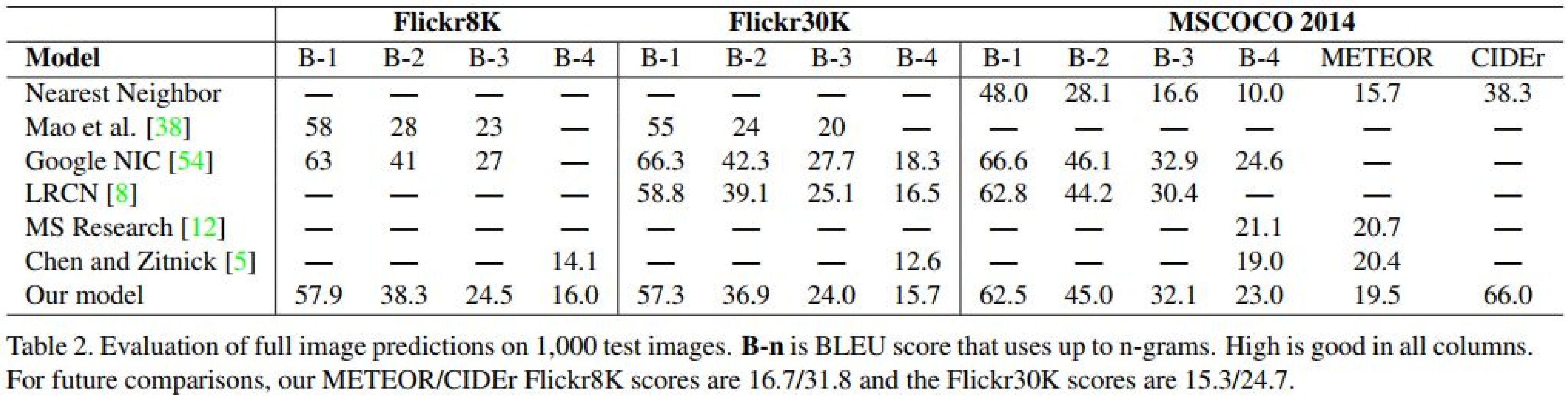


	Image Annotation				Image Search			
Model	**R@1**	**R@5**	**R@10**	**Med** *r*	**R@1**	**R@5**	**R@10**	**Med** *r*
Flickr30K								
SDT-RNN (Socher et al. [49])	9.6	29.8	41.1	16	8.9	29.8	41.1	16
Kiros et al. [25]	14.8	39.2	50.9	10	11.8	34.0	46.3	13
Mao et al. [38]	18.4	40.2	50.9	10	12.6	31.2	41.5	16
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

Table 1. Image-Sentence ranking experiment results. **R@K** is Recall@K (high is good). **Med** *r* is the median rank (low is good). In the results for our models, we take the top 5 validation set models, evaluate each independently on the test set and then report the average performance. The standard deviations on the recall values range from approximately 0.5 to 1.0.

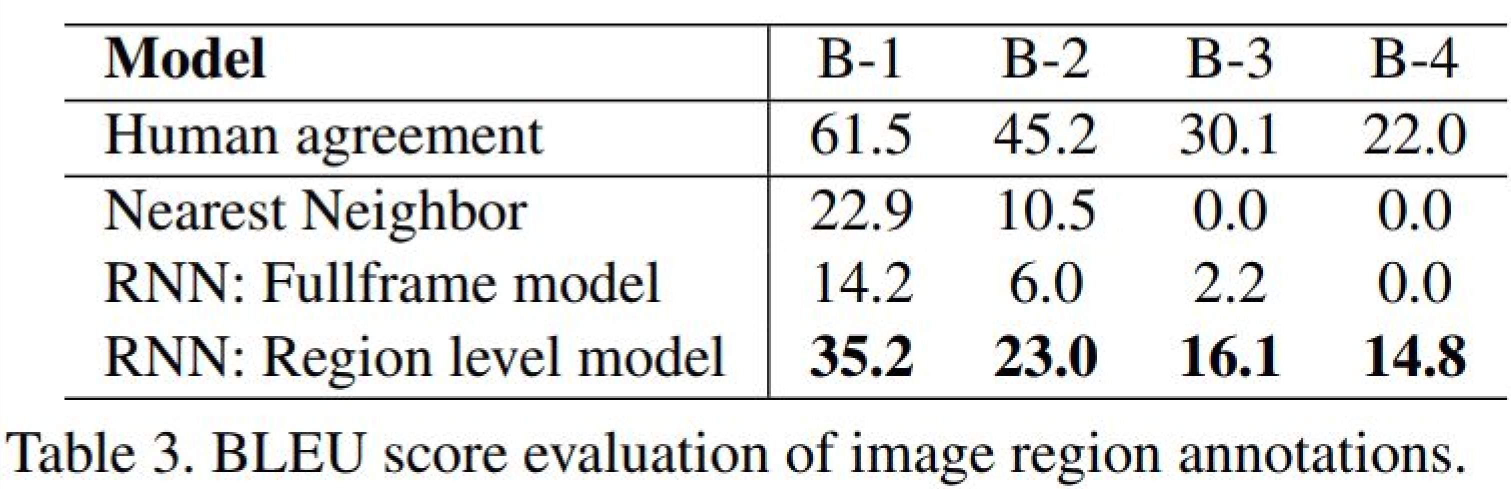
Experiments: **Generating Image Descriptions**



	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

Table 2. Evaluation of full image predictions on 1,000 test images. **B-n** is BLEU score that uses up to n-grams. High is good in all columns. For future comparisons, our METEOR/CIDEr Flickr8K scores are 16.7/31.8 and the Flickr30K scores are 15.3/24.7.

Experiments: **Relevance of Text Snippets at Regions**



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**

Table 3. BLEU score evaluation of image region annotations.

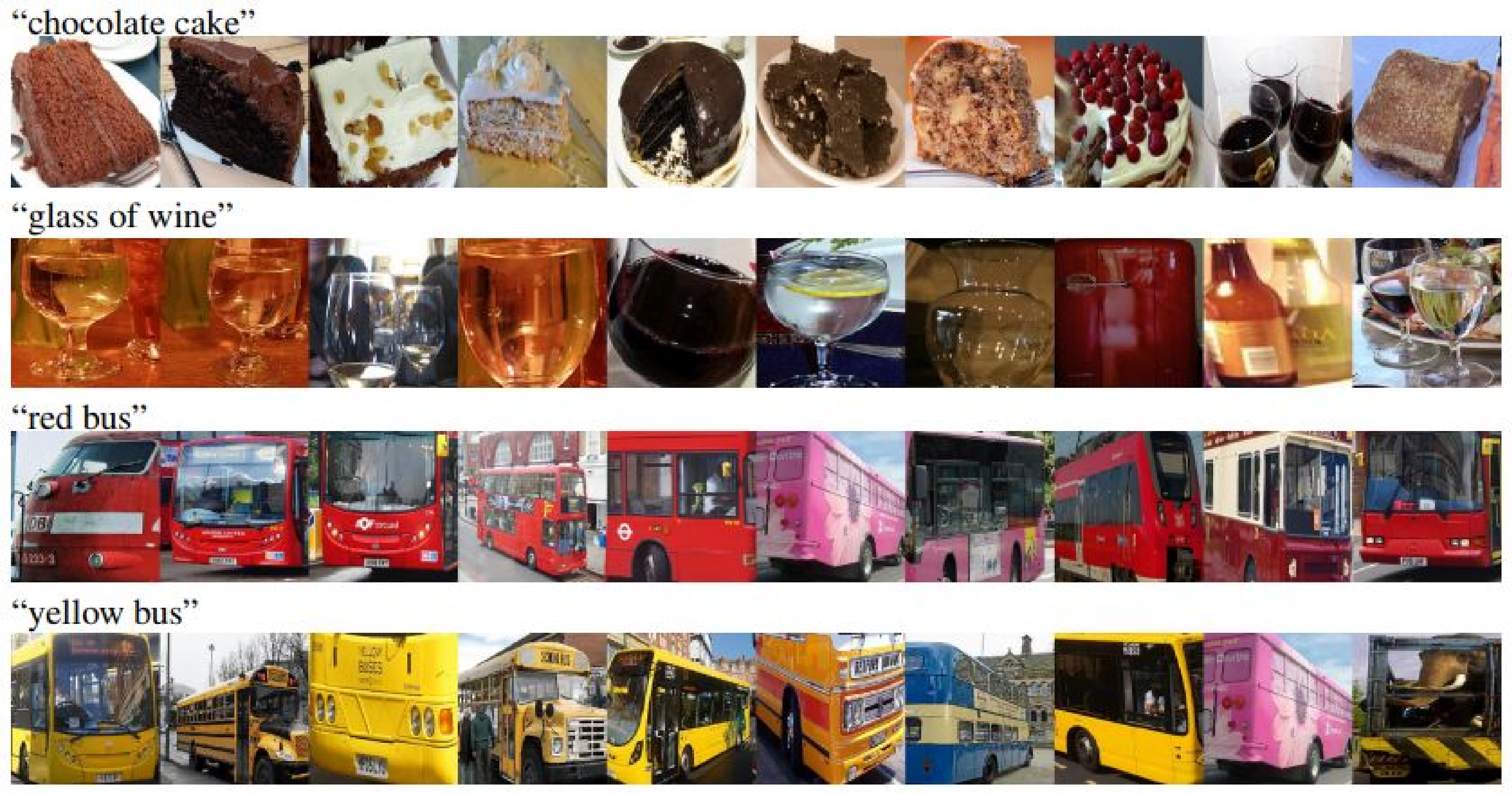
Conclusion

❑ Introduced a model that generates natural language descriptions of image regions based on weak labels in form of a dataset of images and sentences, and with very few hardcoded assumptions.

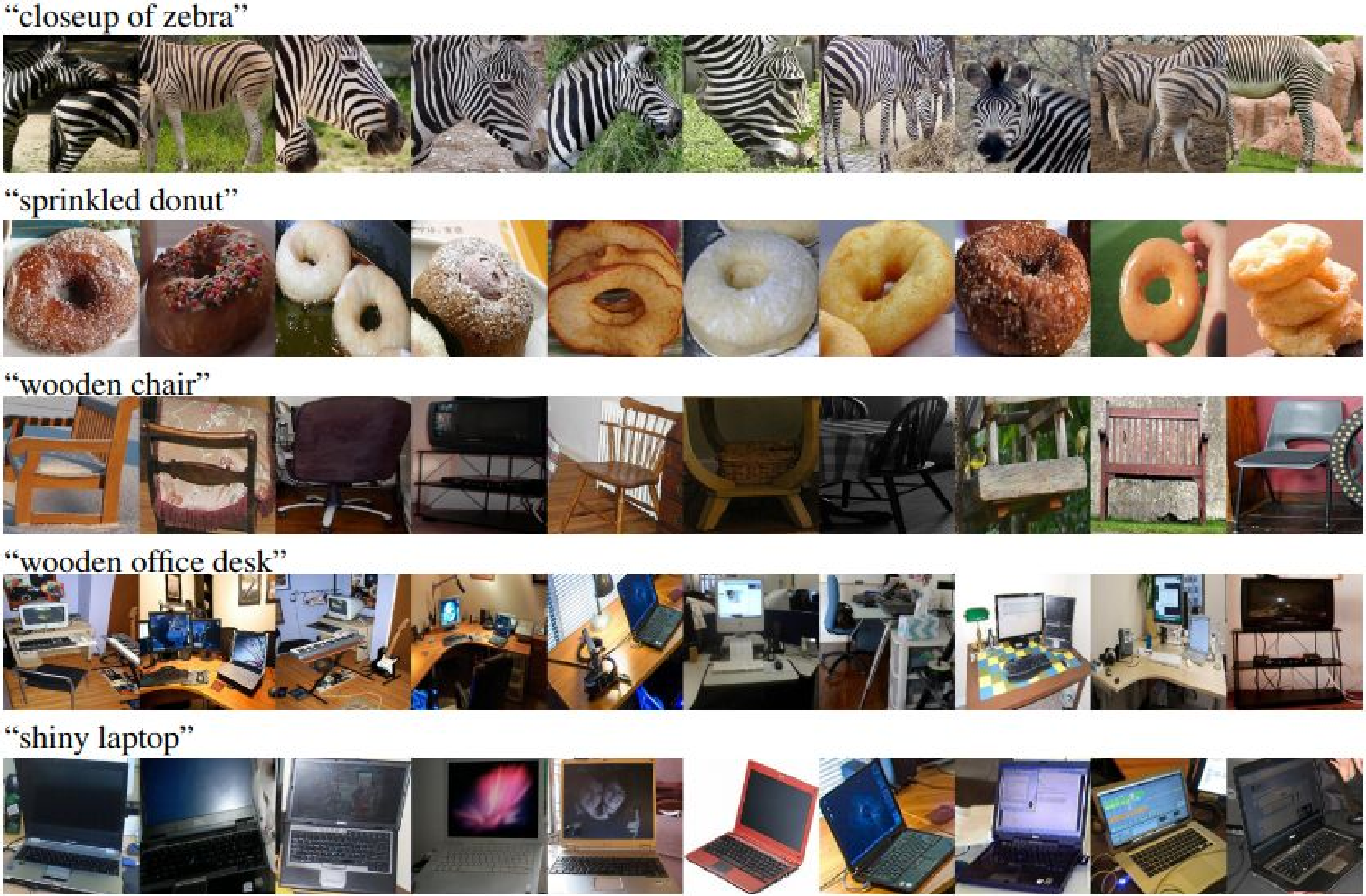
❑ Our approach features a novel ranking model that aligned parts of visual and language modalities through a common, multimodal embedding.

❑ We showed that this model provides state of the art performance on image-sentence ranking experiments.

❑ Second, we described a Multimodal Recurrent Neural Network architecture that generates descriptions of visual data. We evaluated its performance on both full frame and region-level experiments and showed that in both cases the Multimodal RNN outperforms retrieval baselines.



. Examples of highest scoring regions for queried snippets of text



“closeup of zebra”

“sprinkled donut”

“wooden chair”

“wooden office desk”

“shiny laptop”

. Examples of highest scoring regions for queried snippets of text

Demo

http://cs.stanford.edu/people/karpathy/deepimagesent/generationdemo/

Sources

1. Andrej Karpathy, Li Fei-Fei, **Deep Visual-Semantic Alignments for Generating Image Descriptions** [https://arxiv.org/abs/1412.2306]
2. http://cs.stanford.edu/people/karpathy/deepimagesent/
3. http://techtalks.tv/talks/deep-visual-semantic-alignments-for-generating-image-descriptions/61593/