

CAP 6412 Advanced Computer Vision

<http://www.cs.ucf.edu/~bgong/CAP6412.html>

Boqing Gong

Feb 02, 2016

Today

- Administrivia
- R-CNN Review & Project I
- Image Captioning, by Harish
- Neural networks & Backpropagation (Part V)

Past due (02/02 Tuesday, 12pm)

- Assignment 3: Review the following paper

{**Major**} Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." *arXiv preprint arXiv:1412.2306* (2014).

Template for paper review:

<http://www.cs.ucf.edu/~bgong/CAP6412/Review.docx>

Upcoming due (02/04 Tuesday, 12pm)

- Assignment 4: Review the following paper

{**Major**} Xu, Kelvin, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. “Show, attend and tell: Neural image caption generation with visual attention.” arXiv preprint arXiv:1502.03044 (2015).

Template for paper review:

<http://www.cs.ucf.edu/~bgong/CAP6412/Review.docx>

Next week

Week 2	CNN visualization & object recognition
Week 3	CNN & object localization
Week 4	CNN & transfer learning
Week 5	CNN & segmentation, super-resolution
Week 6	CNN & videos (optical flow, pose)
Week 7	Image captioning & attention model
Week 8	Visual question answering
Week 9	Attention model, aligning books with movies
Week 10--16	Video: tracking, action, surveillance Human-centered CV 3D CV Low-level CV, etc.

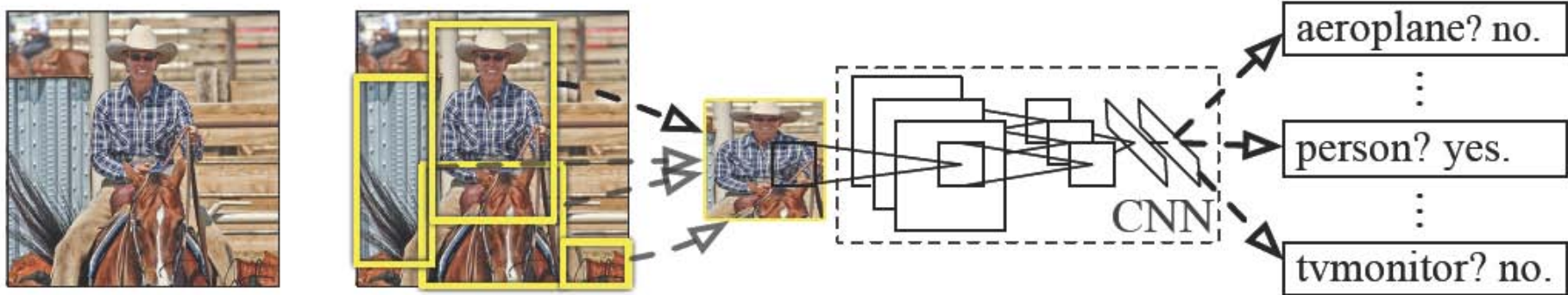
Next week: CNN & Segmentation and super-resolution

- ⌘
Tuesday (02/09) **[Super-resolution]** Dong, Chao, Chen Change Loy, Kaiming He, and Xiaoou Tang. “Learning a deep convolutional network for image super-resolution.” In Computer Vision–ECCV 2014, pp. 184-199. Springer International Publishing, 2014. (Extended version on ArXiv) [& Secondary papers](#)
- Jose Sanchez
- Thursday (02/11) **[Edge detection]** Xie, Saining, and Zhuowen Tu. “Holistically-Nested Edge Detection.” In Proceedings of the IEEE International Conference on Computer Vision, 2015. [& Secondary papers](#)
- Goran Igic

Today

- Administrivia
- **R-CNN Review & Project I**
- Image Captioning, by Harish
- Neural networks & Backpropagation (Part V)

R-CNN: Regions with CNN features



Input
image

Extract region
proposals (~2k / image)

Compute CNN
features

Classify regions
(linear SVM)

Project I: R-CNN at test time

- *INPUT: an image*
- **1.** Extract detection proposals (cf. Samer's presentation on 01/26)
- **2.** Warp proposals to 227-by-227
- **3.** Extract CNN features for each proposal (region) by Caffe
- For class $c=1, 2, \dots, 20$
 - **4.** Output a detection score for each proposal by SVM(proposal, class c)
 - **5.** Nonmaximum suppression using the scores of class c
 - **6.** Regression for the survived proposals
- *OUTPUT: bounding boxes each with a class label & a detection score*

Project I: R-CNN at training time (*bonus*)

- *INPUT: an image*
- 1. **Extract detection proposals** (*10 pts*)
- 2. Warp proposals to 227-by-227
- 3. Extract **CNN** features for each proposal (region) by Caffe (*30 pts*)
- For class $c=1, 2, \dots, 20$
 - 4. Output a detection score for each proposal by **SVM(proposal, class c)** (*10pts*)
 - 5. Nonmaximum suppression using the scores of class c
 - 6. **Regression** for the survived proposals (*10 pts*)
- *OUTPUT: bounding boxes each with a class label & a detection score*

Project I: Grading criteria

- Total: 100 points + *60 bonus points* + x points to promote innovation
- Quantitative results (*65 pts*)
 - Detection average precision on VOC 2012 validation (*40 pts*)
 - Detection average precision on VOC 2012 validation before regression (*10 pts*)
 - Detection average precision on VOC 2012 validation with 1000 proposals (*15 pts*)
- Qualitative results (*35 pts*)

Project I: Resources

- Technical report at <http://arxiv.org/abs/1311.2524>
- Ross' Github repository: <https://github.com/rbgirshick/rcnn>

Project I: Objective

- Get familiar with the state-of-the-art object detection pipeline
- Learn about PASCAL VOC
- Know how to benchmark different algorithms
 - Benchmark datasets
 - Task specification
 - Evaluation procedure and metrics
- Benefit future research / R&D

Today

- Administrivia
- R-CNN Review & Project I
- Image Captioning, by Harish
- Neural networks & Backpropagation (Part V)

Upload slides after class

- See “Paper Presentation” on UCF webcourse
- Sharing your slides
 - **Refer to the original sources of images, figures, etc. in your slides**
 - Convert them to a PDF file
 - Upload the PDF file to “Paper Presentation” after your presentation



Deep Visual-Semantic Alignments for Generating Image Descriptions

Andrej Karpathy & Li Fei-Fei
Stanford University

Presented by Harish RaviPrakash
harishr@knights.ucf.edu

Motivation

- Humans can do it!
- “Build a bridge between natural language & images” – *Karpathy*

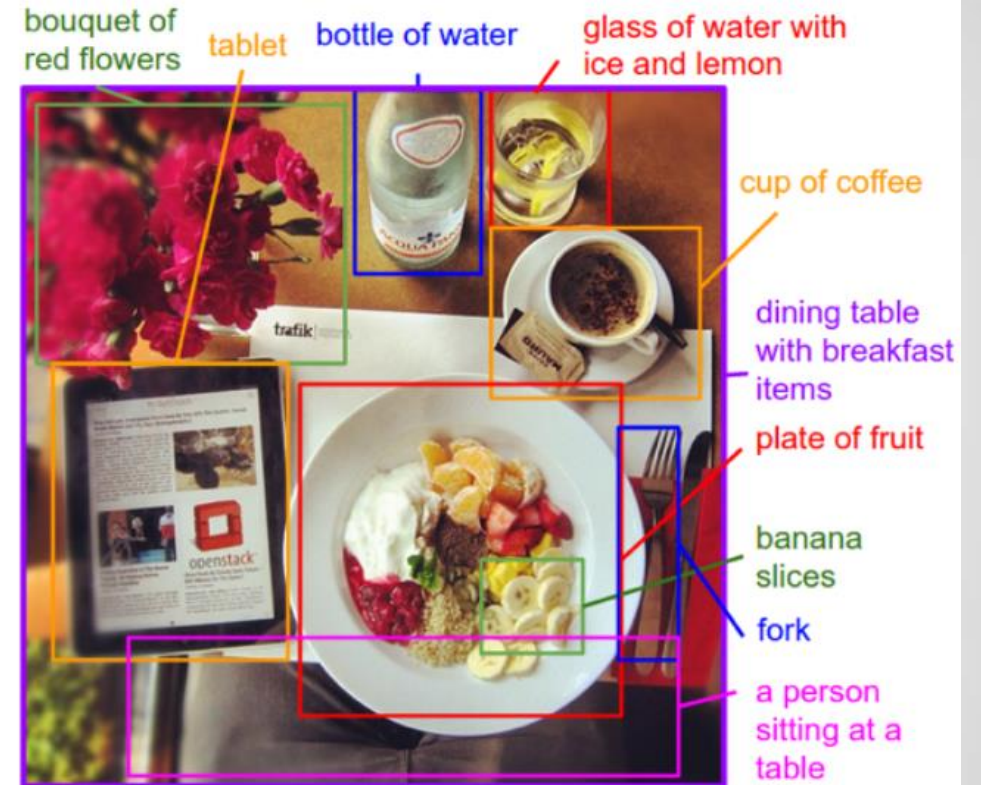
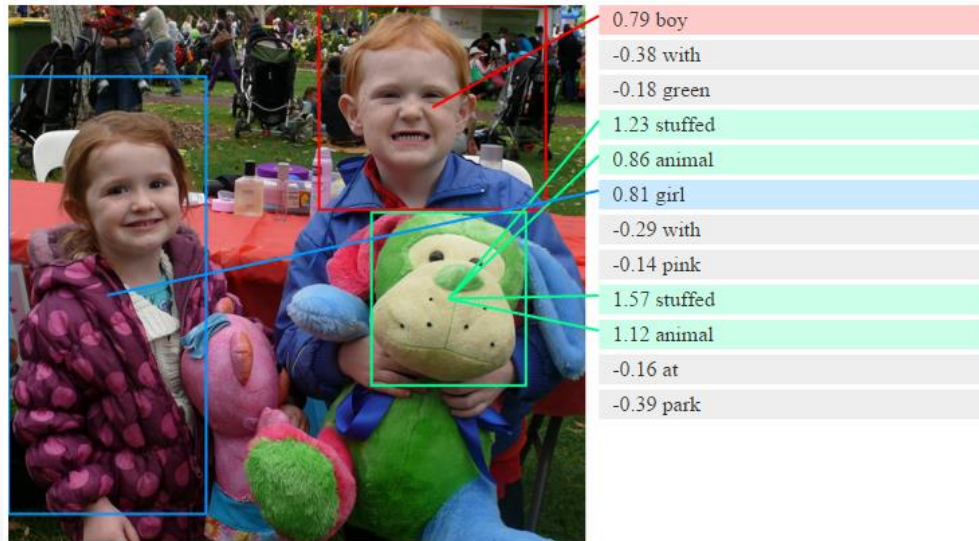


Figure from (Karpathy and Li 2014)

Problem Statement

- Generate Dense Image Descriptions
- Build a better correspondence between image and their sentence descriptions



Main Contributions

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

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

3. Generate region-level descriptions



Approach Outline

- Alignment Inference Model
 - R-CNN
 - BRNN (Bidirectional Recurrent Neural Network)
 - MRF
- Multimodal RNN

R-CNN Stage

- Use whole image + top 19 detected locations (total 20) from RCNN
- CNN pre-trained on ImageNet & fine-tuned

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

- I_b - pixels inside bounding box
- $CNN_{\theta_c}(I_b)$ – FC7 output
- b_m - bias (to be learned)
- W_m - Weight Matrix (to be learned)

BRNN

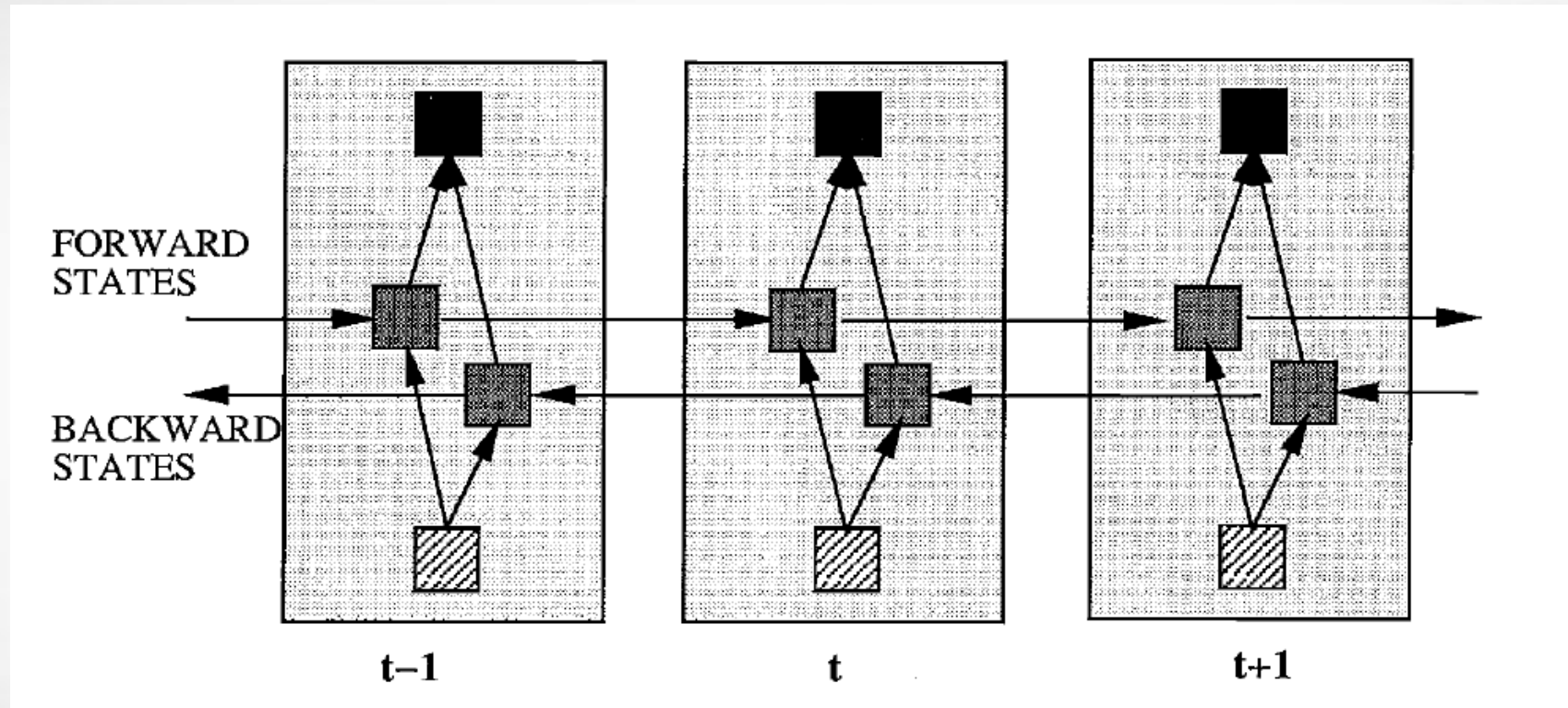


Figure from M. Schuster and K. K. Paliwal. Bidirectional recurrent neural



BRNN Training

1) FORWARD PASS

Run all input data for one time slice $1 \leq t \leq T$ through the BRNN and determine all predicted outputs.

- a) Do forward pass just for forward states (from $t = 1$ to $t = T$) and backward states (from $t = T$ to $t = 1$).
- b) Do forward pass for output neurons.

2) BACKWARD PASS

Calculate the part of the objective function derivative for the time slice $1 \leq t \leq T$ used in the forward pass.

- a) Do backward pass for output neurons.
- b) Do backward pass just for forward states (from $t = T$ to $t = 1$) and backward states (from $t = 1$ to $t = T$).

3) UPDATE WEIGHTS



- BRNN input – sequence of N words
- BRNN output – N h-dimensional vectors

$$x_t = W_w \mathbb{I}_t$$

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

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

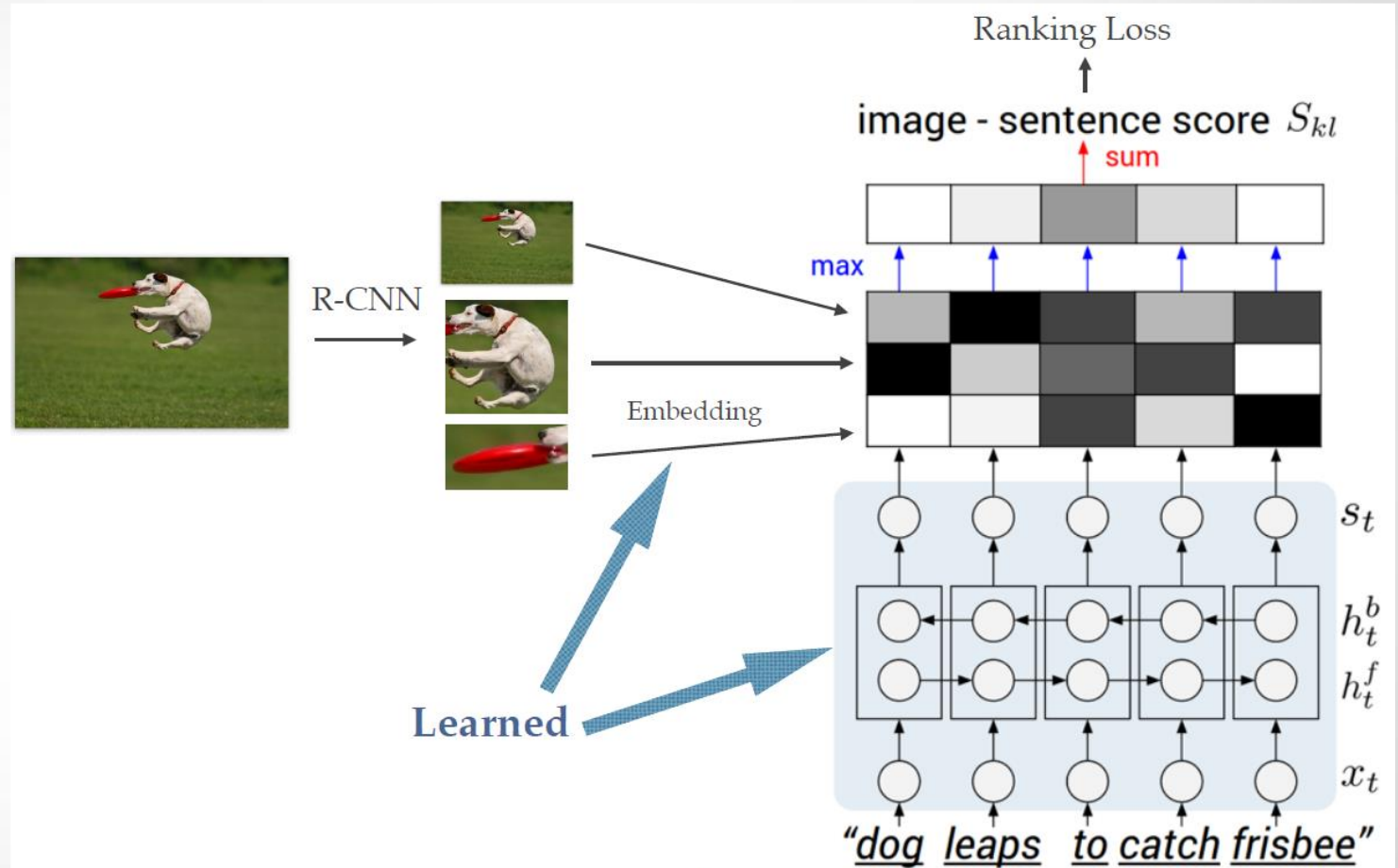
$$h_t^b = f(e_t + W_b h_{t+1}^b + b_b)$$

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

Inferring Word Alignments

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

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



MRF (Markov Random Field)

- Purpose – Smoothing
- Encourage nearby words to point to the same region

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

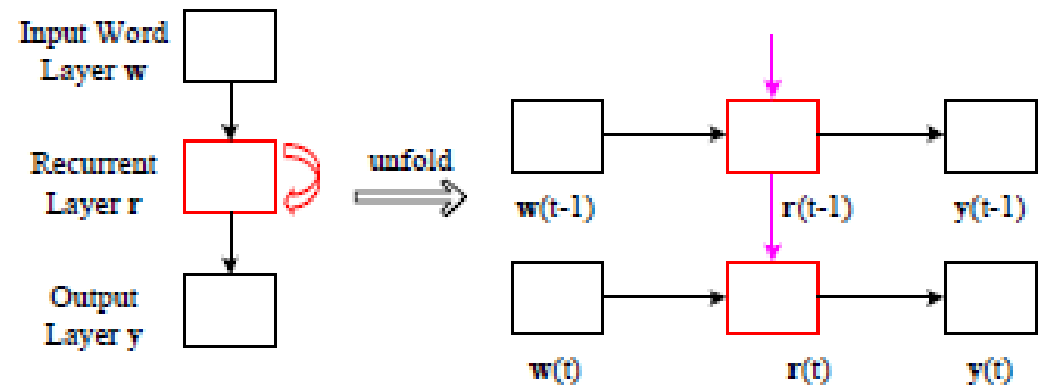
$$E(a_1..a_N) = \sum_{a_j=t} -\textit{similarity}(w_j, r_t) + \sum_{j=1..N-1} \beta[a_j = a_{j+1}]$$

Simple RNN

$w(t)$ – one hot representation of current word

$f_1()$ – sigmoid function

$g_1()$ – softmax function



$$\mathbf{x}(t) = [\mathbf{w}(t) \ \mathbf{r}(t - 1)]; \quad \mathbf{r}(t) = f_1(\mathbf{U} \cdot \mathbf{x}(t)); \quad \mathbf{y}(t) = g_1(\mathbf{V} \cdot \mathbf{r}(t));$$

Multimodal RNN

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

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

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

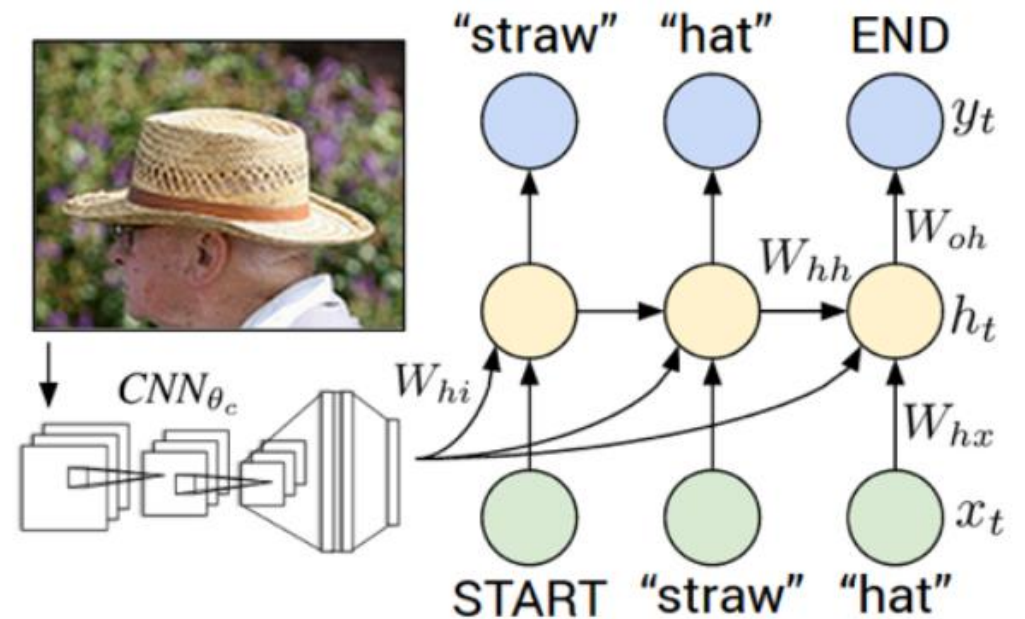


Figure from (Karpathy and Li 2014)



Experiments

- Datasets
 - Flickr8K
 - Flickr30K
 - MSCOCO
- Preprocessing
 - Convert to lowercase
 - Eliminate OoV (Out of Vocabulary)

Generated Descriptions – Full Frame



a group of people sitting at a table with wine glasses



a man riding a horse on a city street



a pizza with toppings on a white plate



a man sitting on a bench with a large umbrella



a woman holding a surfboard in the sand



a group of people standing around a table with a cake

Generated Descriptions – Region

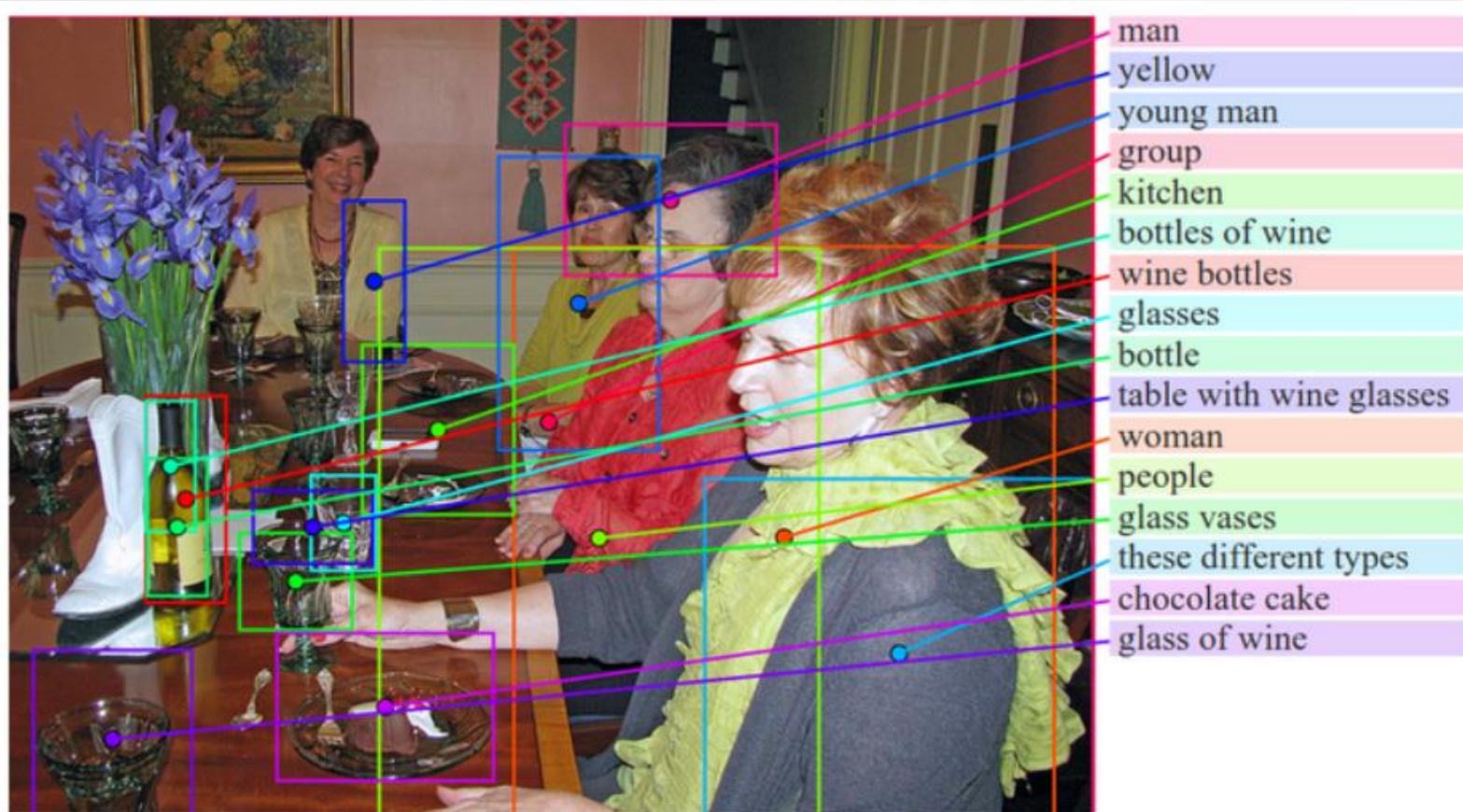


Figure from (Karpathy and Li 2014)



man in graduation robes riding bicycle
cyclist giving thumbs up poses with his bicycle by right
of way sign at park
man riding motorcycle on street



one man and two women sitting in living room
man and woman are playing wii game while woman
sits on couch with wine glass in her hand
group of people sitting on couch with their laptops



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.

Related Work

Explain Images With Multimodal Recurrent Neural Networks

Junhua Mao^{1,2}, Wei Xu¹, *Yi Yang*¹, *Jiang Wang*¹, *Alan L. Yuille*²

¹Baidu Research ²University of California, Los Angeles

- Goal : Generate novel sentence descriptions to explain the contents of images

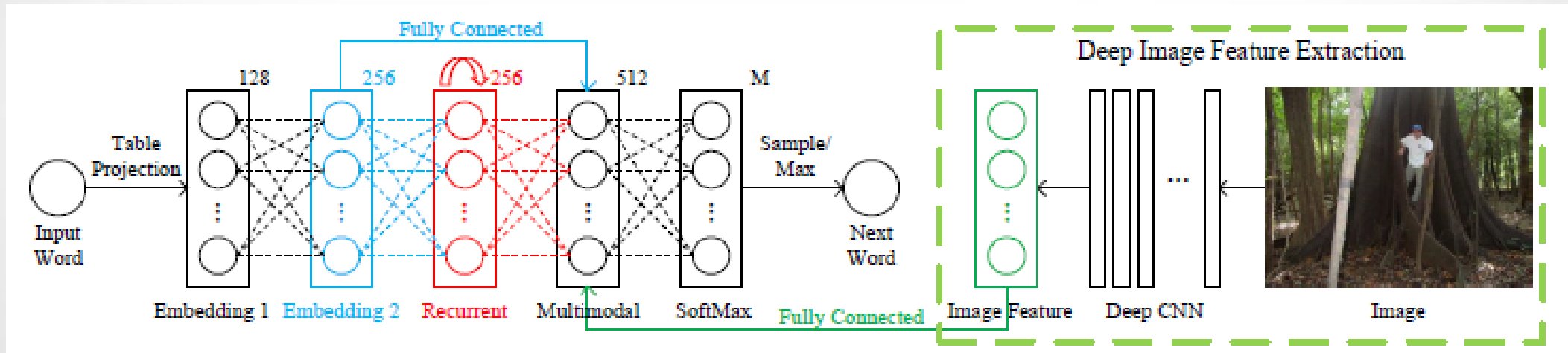
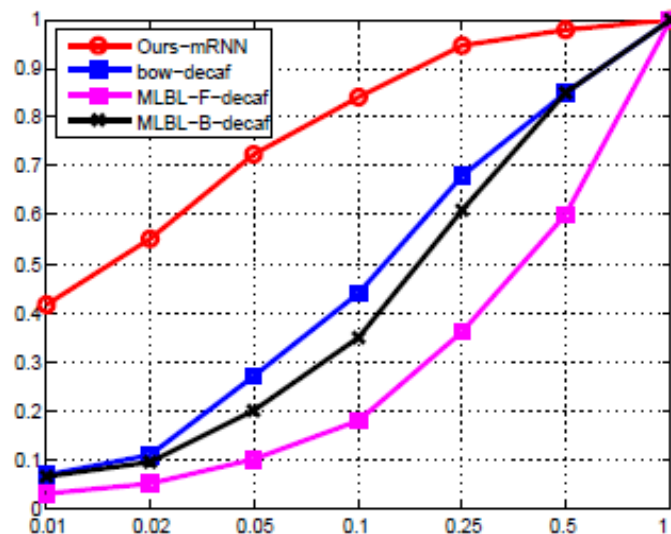


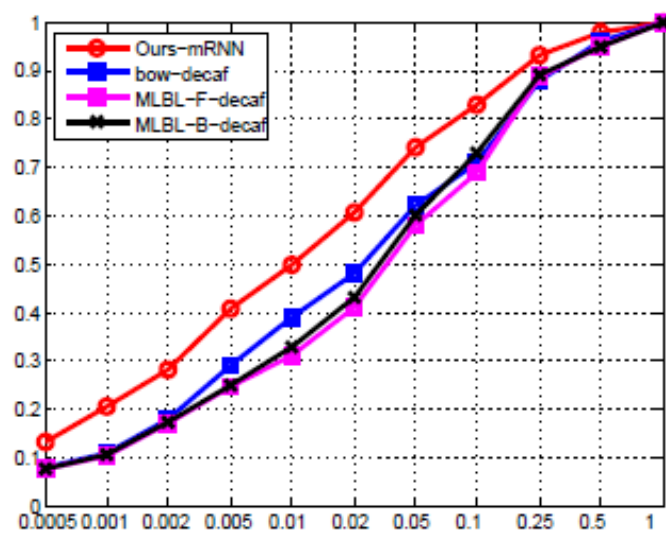
Figure from Mao et. Al : Explain Images with Multimodal Recurrent Neural Networks



- Tasks
 - Sentence generation
 - Sentence retrieval
 - Image retrieval



(a) Image to Text Curve



(b) Text to Image Curve

Figure 3: Retrieval recall curve for (a). Sentence retrieval task (b). Image retrieval task on IAPR TC-12 dataset. The behavior on the far left (i.e. top few retrievals) is most important.

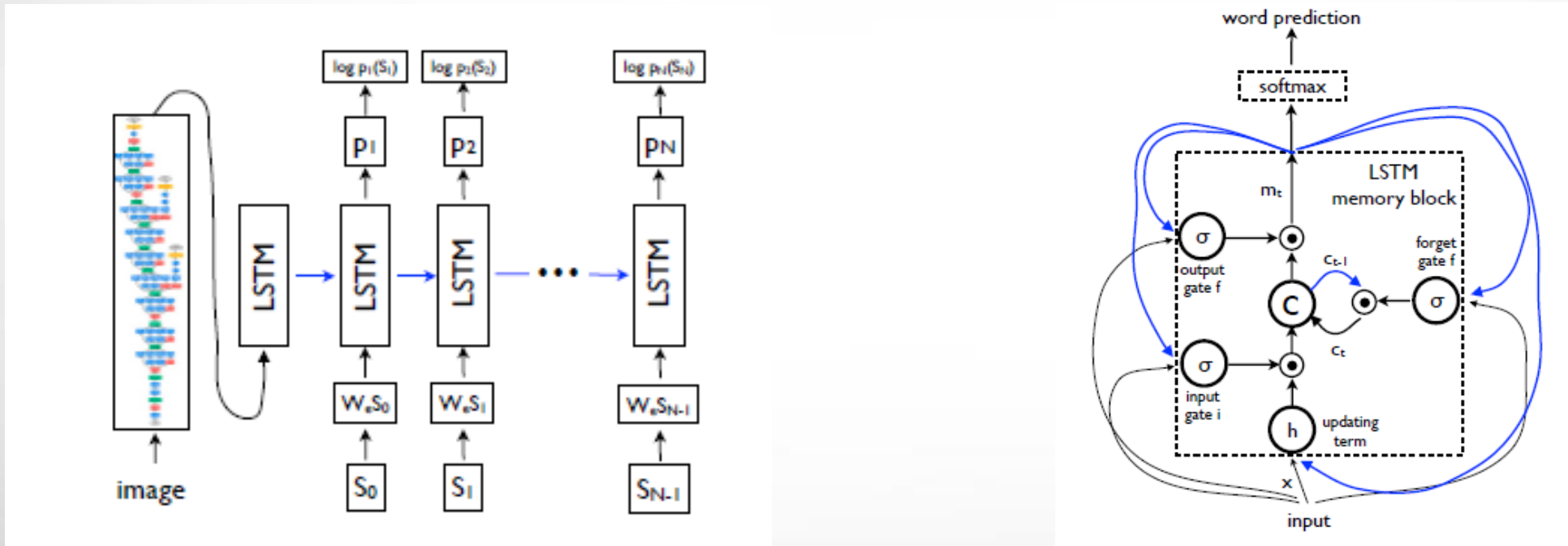
	Sentence Retrieval (Image to Text)				Image Retrieval (Text to Image)			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Ours-m-RNN	20.9	43.8	54.4	8	13.2	31.2	40.8	21

Table 2: R@K and median rank (Med r) for iaprtc-12 dataset.

Show and Tell : A Neural Image Caption Generator

Oriol Vinyals, Alexander Toshev, Samy Bengio & Dumitru Erhan
Google

- Goal : Generate novel sentence descriptions to explain the contents of images





Metric	BLEU-4	METEOR	CIDER
NIC	27.7	23.7	85.5
Random	4.6	9.0	5.1
Nearest Neighbor	9.9	15.7	36.5
Human	21.7	25.2	85.4

Table 1. Scores on the MSCOCO development set.

A person riding a motorcycle on a dirt road.



Two hockey players are fighting over the puck.

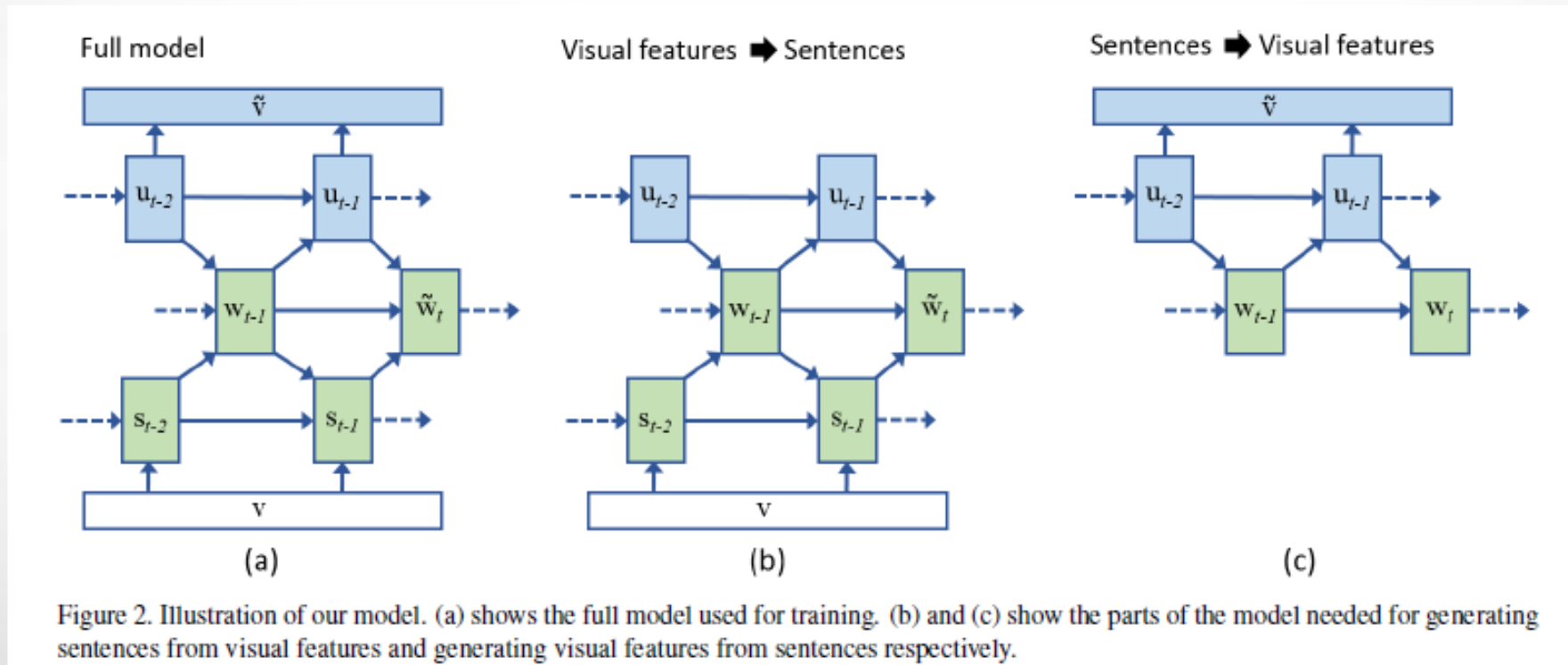


Mind's Eye: A Recurrent Visual Representation for Image Caption Generation

Xinlei Chen¹, C. Lawrence Zitnick²

¹Carnegie Mellon University ²Microsoft Research

- Goal : Generate novel captions, reconstructing image features given an image description





	Flickr 8K			Flickr 30K			MS COCO Val			MS COCO Test		
	PPL	BLEU	METEOR	PPL	BLEU	METEOR	PPL	BLEU	METEOR	BLEU	METEOR	CIDEr
RNN	17.5	4.5	10.3	23.0	6.3	10.7	16.9	4.7	9.8	-	-	-
RNN+IF	16.5	11.9	16.2	20.8	11.3	14.3	13.3	16.3	17.7	-	-	-
RNN+IF+FT	16.0	12.0	16.3	20.5	11.6	14.6	12.9	17.0	18.0	-	-	-
RNN+VGG	15.2	12.4	16.7	20.0	11.9	15.0	12.6	18.4	19.3	18.0	19.1	51.5
Our Approach	16.1	12.2	16.6	20.0	11.3	14.6	12.6	16.3	17.8	-	-	-
Our Approach+FT	15.8	12.4	16.7	19.5	11.6	14.7	12.0	16.8	18.1	16.5	18.0	44.8
Our Approach+VGG	15.1	13.1	16.9	19.1	12.0	15.2	11.6	18.8	19.6	18.4	19.5	53.1
Human	-	20.6	25.5	-	18.9	22.9	-	19.2	24.1	21.7	25.2	85.4

Table 2. Results for novel sentence generation for Flickr 8K, FLickr 30K, MS COCO Validation and MS COCO Test. Results are measured using perplexity (PPL), BLEU (%) [35], METEOR (%) [1] and CIDEr-D (%) [40]. Human agreement scores are shown in the last row. See the text for more details.



A table topped with plates of food and bowls of food. This table is filled with a variety of different dishes.



A group of baseball players playing a game of baseball. A group of baseball players is crowded at the mound.



A person that is flying a kite in the snow .

A person up in the air, upside down while outside.



A stuffed teddy bear sitting on top of a piece of luggage.

This wire metal rack holds several pairs of shoes and sandals.



Comparative 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

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.

Model	Image Annotation				Image Search			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Flickr8K								
DeViSE (Frome et al. [16])	4.5	18.1	29.2	26	6.7	21.9	32.7	25
SDT-RNN (Socher et al. [49])	9.6	29.8	41.1	16	8.9	29.8	41.1	16
Kiros et al. [25]	13.5	36.2	45.7	13	10.4	31.0	43.7	14
Mao et al. [38]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
DeFrag (Karpathy et al. [24])	12.6	32.9	44.0	14	9.7	29.6	42.5	15
Our implementation of DeFrag [24]	13.8	35.8	48.2	10.4	9.5	28.2	40.3	15.6
Our model: DepTree edges	14.8	37.9	50.0	9.4	11.6	31.4	43.8	13.2
Our model: BRNN	16.5	40.6	54.2	7.6	11.8	32.1	44.7	12.4

Table 5. Ranking experiment results for the Flickr8K dataset.



Model	Flickr8K				Flickr30K				MSCOCO 2014					
	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.



Conclusion

- Region based dense descriptions
- Multimodal RNN
- Novel model to infer alignments



Future Directions

- Use LSTM in the m-RNN model
- Try different CNNs – VGGNet, GoogLeNet
- Changing the RNN hidden layer function from Sigmoid to ReLU
- Adding Mind's Eye paper approach – will it work?



Some Useful Videos

- Recurrent Neural Networks and LSTM
<https://www.youtube.com/watch?v=56TYLaQN4N8>
- Automated Image Captioning with ConvNets and Recurrent Nets
<https://www.youtube.com/watch?v=xKt21ucdBY0>



THANK
YOU!