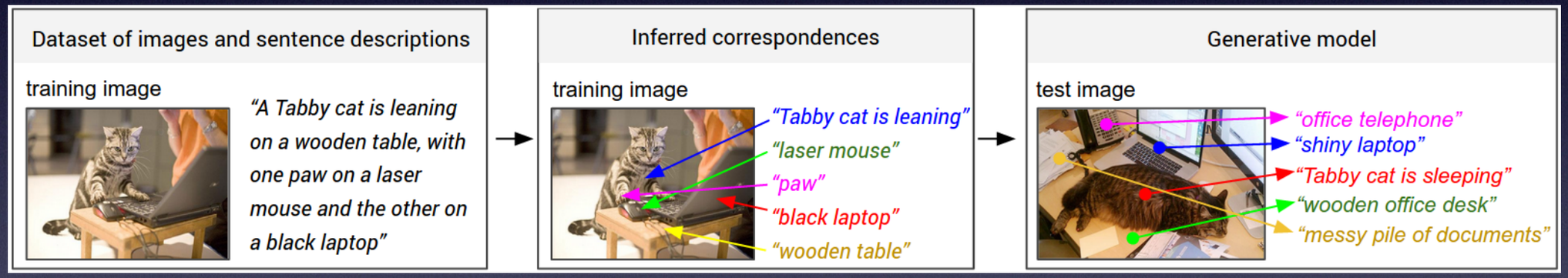
Deep Visual-Semantic Alignments for Generating Image Descriptions

Andrej Karpathy, Fie-Fie (2015)

Slides by Samir

Goal

- Generate a model free-from natural language descriptions of image region



Approach

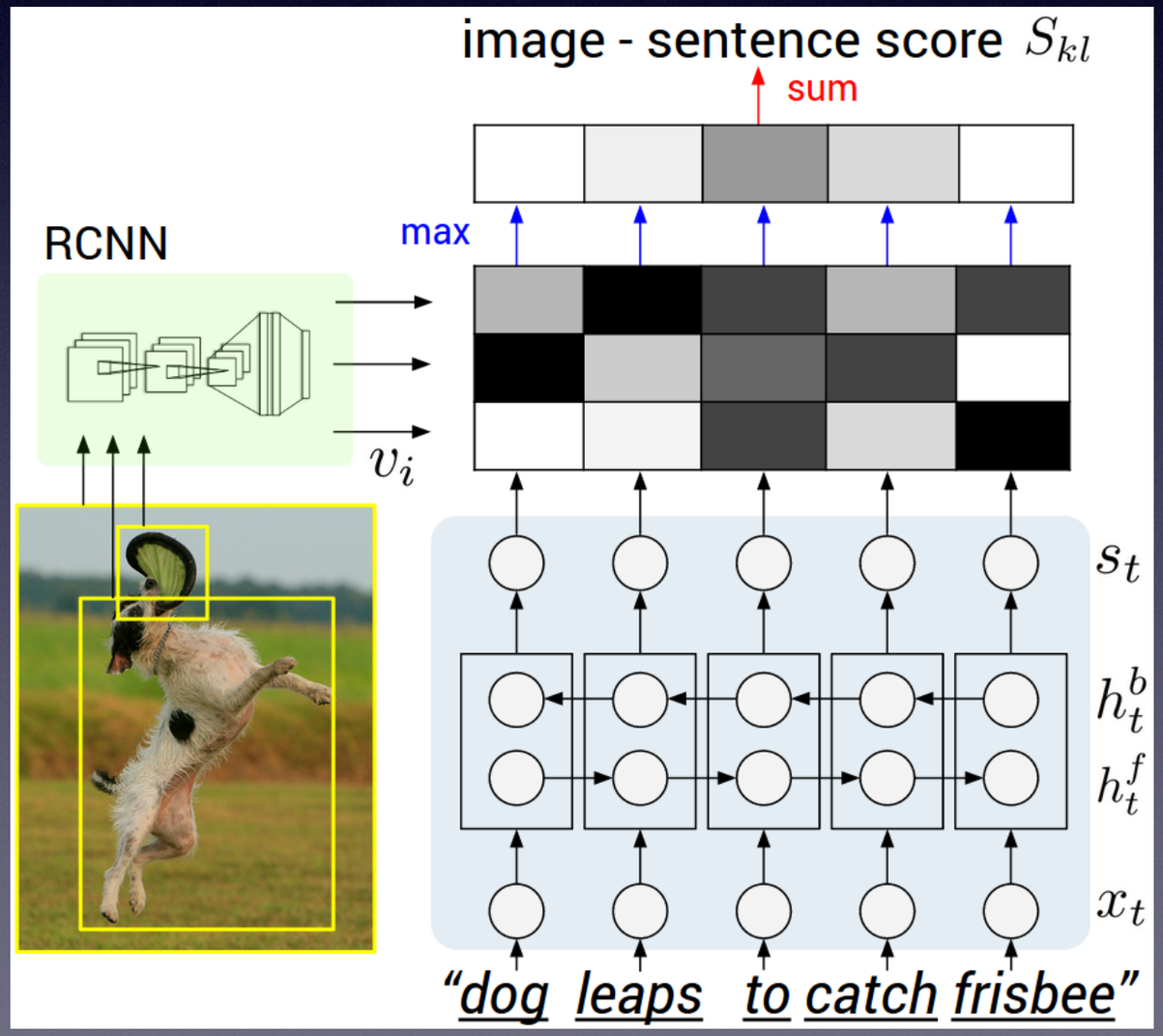
- Learn about the inter-modal correspondences between text and visual data from the dataset of images and their sentence descriptions
- **Learning Architecture** - Combination of CNN, biRNN and multimodal embedding
  - CNN over image region
  - Bidirectional RNN over sentences
  - Multimodal embedding to align the two modalities
- **Generating Architecture** - RNN uses the inferred alignments to learn to generate novel descriptions of image regions

Related Work

- Dense Image Annotations
- Generating textual descriptions
- Grounding natural language in images
- Neural networks in visual and language domains

Training or Learning Architecture

Aligns visual and language data



1. CNN for representing image

- Method of Girshick
    - To detect objects in every image with a Region Convolution Neural Network
- Pre-trained
    - ImageNet
    - Fine-tuned on 200 classes of the ImageNet Detection Challenge
    - Used the top 19 detected location and the whole image to compute the representation based on the pixel I_b inside each bounding box

$$v = Wm[CNN\theta_c\,(Ib)] + bm$$

2. biRNN for representing sentences

- To establish inter-modal relationship, the representation of the words in the sentence in the same *h*-dimensional embedding space that the image region occupy
- biRNN is used to compute the word representations
  - Takes a sentence of N words and tranforms each one into a *h*-dimensional vector
  - Representation of each word is enrich by the variably-sized context around the word
  - Using index $t = 1....N$ to denote position of the word in the sentence

Cont…

Bidirectional Recurrent Neural Network

$$x_t = W_w I_t$$

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

$$h^f_t = f(e_t + W_f\, h^f_{t-1} + b_f)$$

$$h^b_t = f(e_t + W_b\, h^b_{t+1} + b_b)$$

$$s_t = f(W_d(h^f_t + h^b_t) + b_d)$$

3. Alignment objective

- Every image and sentence into a set of vectors in the common *h*-dimensional space

- Labels are at all level of entire images and sentences

- Image-sentence score S_{kl} is calculated, it is a function of individual scores that measures how well a word aligns to a region of a image

- Sentence-image pair should have high matching score if its words have confidence support in the image

Generating Architecture

Multimodal Recurrent Neural Network for generating descriptions

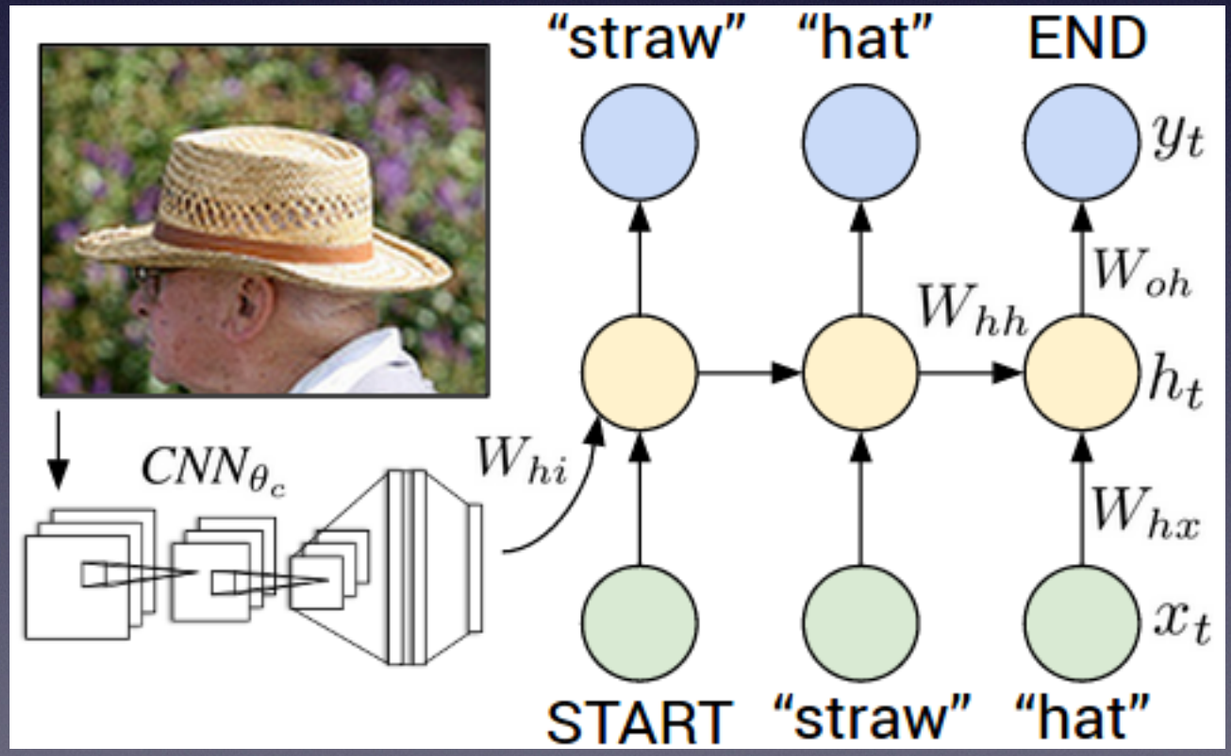
- Based on the previous RNN model, which is defined as the probability distribution of the next word in the sequence, given the current word and the context from previous time steps

- Extension to above method an image context vector b_v was pass to the RNN only at the first iteration (which was lead to better performance than providing at every step)

- It was also discovered that giving b_v, $(W_{hx}x_t)$ through activation function was helpful

RNN Training

- Combine the word (x_t), the previous context (h_{t-1}) and the image information (b_v) to predict the next word (y_t)



RNN test (predict)

- Compute the representation of the image b_v
- $h_0 = 0$
- x_1 to the embedding of the word ‘the’
- Compute the distribution over the first word y_1
- Sample from the distribution, set its embedding vector as x_2 and repeat till the END token is generated

Dataset

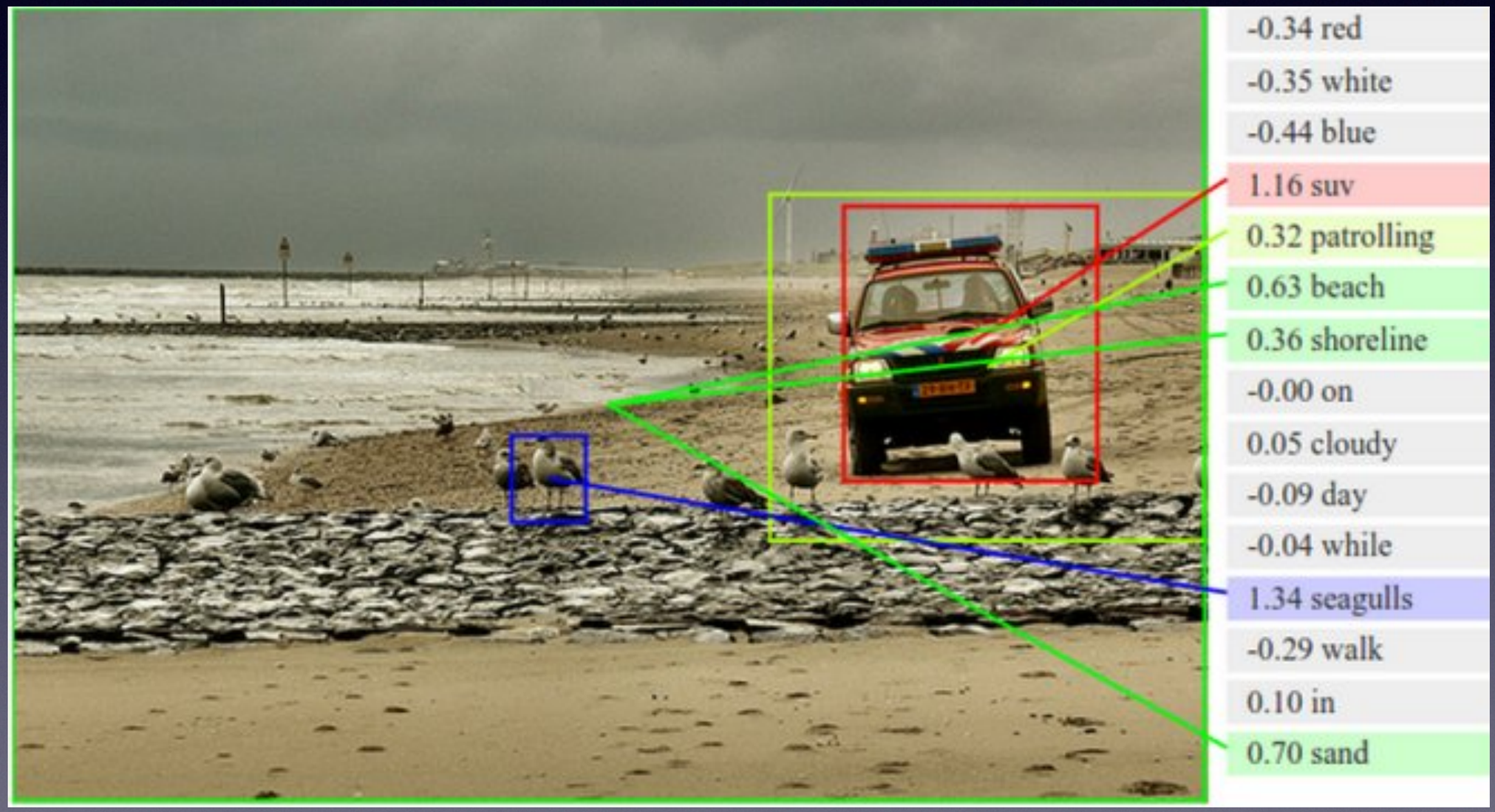
- Flickr8K
    - 8000 images
    - 1000 for validation, 1000 for testing and rest for training
- Flickr30K
    - 31000 images
    - 1000 for validation, 1000 for testing and rest for training
- MSCOCO
    - 123000 images
    - 5000 for both validation and testing, remaining for training
- Each of this images were annotated with 5 sentences using Amazon Mechanical Turk

Experiment

- Image-sentence alignment evaluation
- Outperform previous work
  - DeViSE is a model that learns a score between the words and images
  - SDT-RNN is trained with similar objective, but instead of averaging the word representations
  - Kiros used LSTM to encode sentences
  - DeFrag - averaged the word and image region representations to obtain a single vector with modality

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. [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]	13.5	36.2	45.7	13	10.4	31.0	43.7	14
Mao et al. [31]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
DeFrag (Karpathy et al. [18])	12.6	32.9	44.0	14	9.7	29.6	42.5	15
Our implementation of DeFrag [18]	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**
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**
MSCOCO								
Our model: 1K test images	29.4	62.0	75.9	2.5	20.9	52.8	69.2	4.0
Our model: 5K test images	11.8	32.5	45.4	12.2	8.9	24.9	36.3	19.5

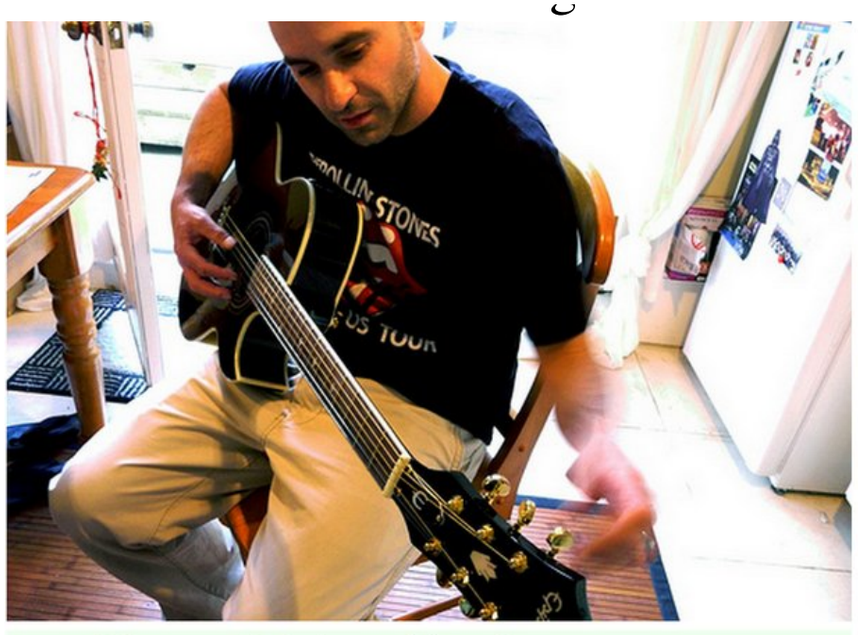
Alignment predicted by the model



Evaluation of generated Descriptions

- BLEU score
  - Evaluate the prediction of the image region as it is considered to be the standard metric of evaluation in this setting
  - Evaluate a candidate sentence by measuring the fraction of n-grams that appear in a set of references
- Multimodal RNN outperform the retrieval baseline
  - Used first 4 out of 5 sentences as references and the fifth one to evaluate human agreement
  - Annotate each test image with highest scoring sentences from the training set

	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
Ranking: Nearest Neighbor	-	0.29	0.11	0.03	-	0.27	0.08	0.02	-	0.32	0.11	0.03
Generating: RNN	-	0.42	0.19	0.06	-	0.45	0.20	0.06	-	0.50	0.25	0.12
Generating: RNN (OxfordNet CNN [40])	-	**0.49**	**0.28**	**0.11**	-	**0.49**	**0.28**	**0.12**	-	**0.54**	**0.34**	**0.16**
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



guy sitting on chair tunes his guitar

orchestra conductor is conducting orchestra

man in black shirt is playing guitar

Limitations

- Model can only generate description of one input array of pixel at a fixed resolution
- RNN couples the visual and language domains in the hidden representation only through additive interactions, which are less expressive than more complicated multiple interactions
- Going directly from image-sentence dataset to region-level annotations as part of a single model that is trained end-to-end with a single objective

Conclusions

- Model that generates free-from description of image region based on weak labels in form of dataset of images and sentences
- Aligned the visual and textual modalities through a common multimodal embedding
- Approach leads to consistent state of art performance on ranking experiments across three datasets
- Multimodal RNN architecture to generate description which outperform the retrieval baseline