Deep Visual-Semantic Alignments for Generating Image Descriptions *(https://cs.stanford.edu/people/karpathy/cvpr2015.pdf)*

1. AIM

1.1 Generating natural language descriptions of images and their regions

1.2 Leveraging datasets of images and their sentence descriptions to learn about the inter-modal correspondences between language and visual data

2. PROBLEMS WITH PREVIOUS WORKS IN IMAGE CAPTIONING

These models often rely on hard-coded visual concepts and sentence templates, which imposes limits on their variety.

Moreover, ***the focus of these works has been on reducing complex visual scenes into a single sentence***, which we consider to be an unnecessary restriction.

3. INSPIRATION / RELATED WORK

Dense image annotations:

Barnard et al. [2] and Socher et al. [48] : Annotation of segments of images by studying multimodal correspondences between words and images

[34, 18, 15, 33] : inferring scene type, objects and their spatial support

Generating descriptions.

[21, 49, 13, 43, 23] : most compatible annotation in the training set is transferred to a test image

[30, 35, 31] :training annotations are broken up and stitched together

[19, 29, 13, 55, 56, 9, 1] generate image captions based on fixed templates that are filled based on the content of the image

[38, 54, 8, 25, 12, 5] : use RNNs to generate image descriptions

Grounding natural language in images.

Frome et al. [16] :associating words and images through a semantic embedding

Karpathy et al. [24] : decomposing images and sentences into fragments and infer their inter-modal alignment using a ranking objective. (instead based on grounding dependency tree relations, our model aligns contiguous segments of sentences)

Neural networks in visual and language domains.

CNN's for image classification and object detection [32, 28,45]

Pretrained word vectors for sentence generation [41, 22, 3]

RNN's for language modelling [40, 50]

4. CONTRIBUTIONS

1. Developing a deep neural network model that infers the latent alignment between segments of sentences and the region of the image that they describe this is done by treating the sentences as weak labels, in which contiguous segments of words correspond to some particular, but unknown location in the image

2. Associating the two modalities through a common, multimodal embedding space and a structured objective

3. Introducing a multimodal Recurrent Neural Network architecture that takes an input image and generates its description in text.

4. Training the model on the inferred cor- respondences and evaluate its performance on a new dataset of region-level annotations.

5. THE PROPOSED MODEL MUST

1. Generate dense descriptions of images

2. Be rich enough to simultaneously reason about contents of images and their representation in the domain of natural language.

3. Only rely on learning from the training data. No rules/ templates.

4. Datasets available in large quantities but locations of various entities in an image unknown.

6. USING

The alignment model is based on a novel combination of **Convolutional Neural Networks** over image regions, **bidirectional Recurrent Neural Networks** over sentences, and a **structured objective that aligns** the two modalities through a multimodal embedding

A **Multimodal Recurrent Neural Network** architecture that uses the inferred alignments to learn to generate novel descriptions of image regions.

7. MODEL / METHODOLOGY (uses 2 models)

DURING TRAINING TIME:

1. **Input** is a set of images and their corresponding descriptions.

2. **ALIGNMENT MODEL** : Aligning sentence snippets to the visual regions that they describe through a multimodal embedding

3. **M-RNN** : treat these correspondences as training data for a second, multi- modal Recurrent Neural Network model that learns to generate the snippets.

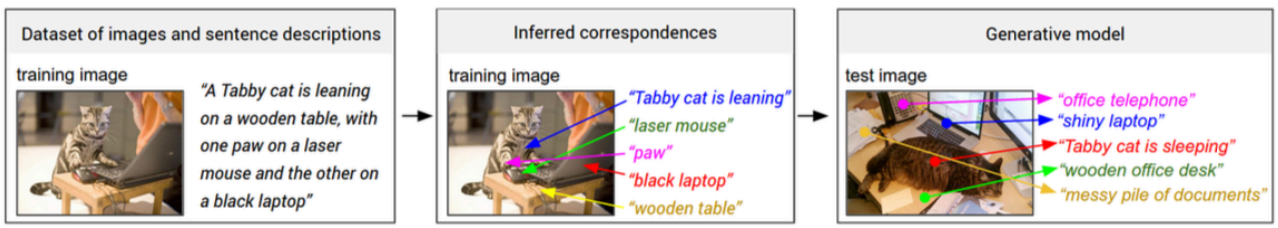

Figure 2. Overview of our approach. A dataset of images and their sentence descriptions is the input to our model (left). Our model first infers the correspondences (middle, Section 3.1) and then learns to generate novel descriptions (right, Section 3.2).

Explanation of above 3 steps:

ALIGNMENT MODEL uses

Regional Convolutional Neural Network (RCNN) to detect objects in every image.

(In this paper : The CNN is pre-trained on ImageNet [6] and fine tuned on the 200 classes of the ImageNet Detection Challenge [45]. They use the top 19 detected locations in addition to the whole image and compute the representations based on the pixels I_b inside each bounding box)

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

(*CNN*(I_b) transforms the pixels inside bounding box I_b into 4096-dimensional activations of the fully connected layer immediately before the classifier. The CNN parameters ✓$_c$ contain approximately 60 million parameters. **Every image is thus represented as a set of h-dimensional vectors**)

It also uses **Bidirectional recurrent neural network (BRNN)** to compute word representations in the sentence. (inferring the latent correspondences eg. “wooden table” for table)

(The BRNN takes a sequence of N words (encoded in a 1-of-k representation) and trans- forms each one into an h-dimensional vector)

Image Sentence Score carries the interpretation that a sentence fragment aligns to a subset of the image regions whenever the dot product is positive. Here, **every word s_t aligns to the single best image region.**

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

But we want to associate snippets of text instead of single word to each bounding box. Therefore we use the concept of Markov Random Field(MRF) and latent alignment variables to generate a set of image regions annotated with segments of text

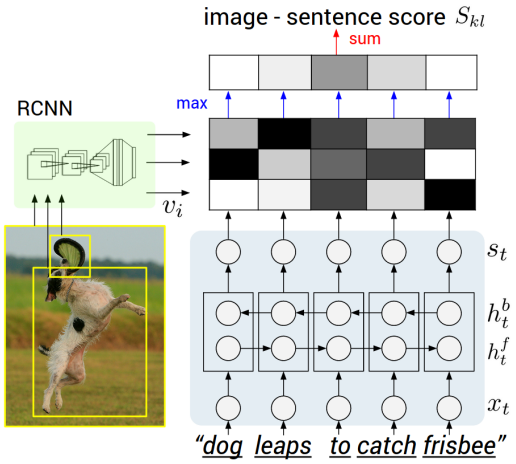

Figure 3. Diagram for evaluating the image-sentence score S_{kl}. Object regions are embedded with a CNN (left). Words (enriched by their context) are embedded in the same multimodal space with a BRNN (right). Pairwise similarities are computed with inner products (magnitudes shown in grayscale) and finally reduced to image-sentence score with Equation 8.

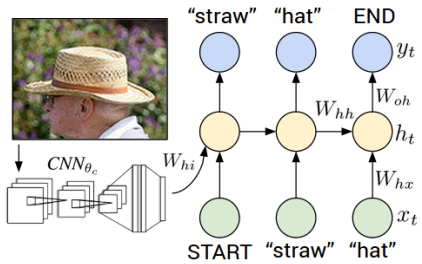

Figure 4. Diagram of our multimodal Recurrent Neural Network generative model. The RNN takes a word, the context from previous time steps and defines a distribution over the next word in the sentence. The RNN is conditioned on the image information at the first time step. START and END are special tokens.

Explained By Shivali Goel

during training our Multimodal RNN takes the image pixels I and a sequence of input vectors (x_1, . . . , x_T). It then computes a sequence of hidden states (h_1 , . . . , h_t) and a sequence of outputs (y_1 , . . . , y_t) by iterating a recurrence relation. (refer figure 4)

(Note that we provide the image context vector b_v to the RNN only at the first iteration, which we found to work better than at each time step. A typical size of the hidden layer of the RNN is 512 neurons.)

8. OPTIMISING TECHNIQUES

Used SGD with mini-batches of 100 image-sentence pairs and momentum of 0.9 to optimize the alignment model.

Achieved the best results using RMSprop [52], which is an adaptive step size method that scales the update of each weight by a running average of its gradient norm.

9. EVALUATION TECHNIQUES

B-n, **BLEU score** , which is a form of precision of word n-grams between generated and reference sentences.

METEOR and Cider

Proxy task of ranking a set of available descriptions with respect to a given image. **ranking metrics like recall@k.**

10. DATASETS / RESULTS

	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.

	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.

11. CONCLUSION

Natural language descriptions of image regions based on weak labels in form of a dataset of images and sentences, and with very few hard- coded assumptions.

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.

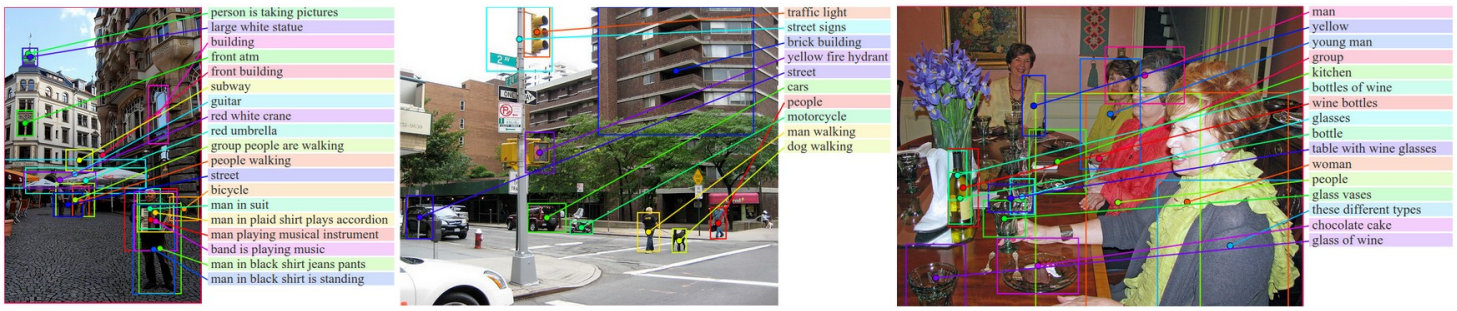

Figure 7. Example region predictions. We use our region-level multimodal RNN to generate text (shown on the right of each image) for some of the bounding boxes in each image. The lines are grounded to centers of bounding boxes and the colors are chosen arbitrarily.

12. LIMITATIONS

First, the model can only generate a description of one input array of pixels at a fixed resolution.

Additionally, the RNN receives the image information only through additive bias interactions, which are known to be less expressive than more complicated multiplicative interactions

Lastly, this approach consists of two separate models

13. FUTURE SCOPE

Going directly from an image- sentence dataset to region-level annotations as part of a single model trained end-to-end remains an open problem

Explained By Shivali Goel