Deep Visual-Semantic Alignments for Generating Image Descriptions

- Andrej Karpathy
- Li Fei-Fei

-TPAMI, 2017

- Dian Jia
- Anuja Tayal

Outline

Introduction

Related Works

Methodology

Results

Future Work/Improvement

Introduction

Background

Motivation

Challenge

Contribution

Background

Task: Generating Image Descriptions

Goal: Generating dense natural language descriptions of images and their regions.

Input: An RGB image

Output: The sets of regions and their corresponding descriptions

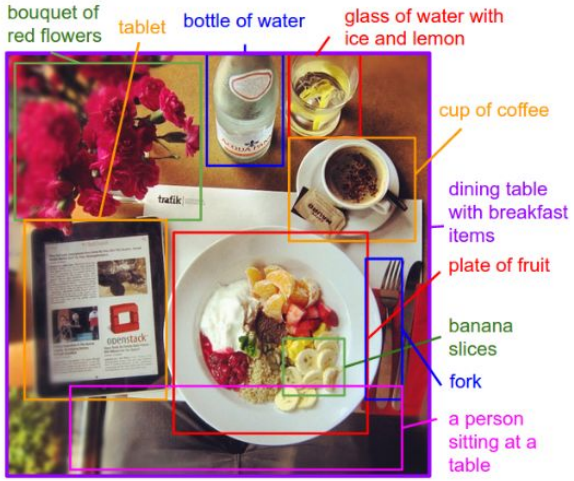


Figure 1. Motivation/Concept Figure: Our model treats language as a rich label space and generates descriptions of image regions.

Challenges

- Dataset of image caption available but these descriptions mention several entities whose locations in the images are unknown.

- How to generate descriptions without hard-code templates

Contribution

1

- Develop DNN that infer alignment between segments and regions
- Learn intermodal correspondence between language and visual data
- Alignment model produces sota results in retrieval experiments

2

- Develop Multimodal RNN Model to generate captions from image
- Generated captions produce sensible qualitative predictions.
- Evaluate performance on Region Caption- Visual Genome Dataset

Motivation

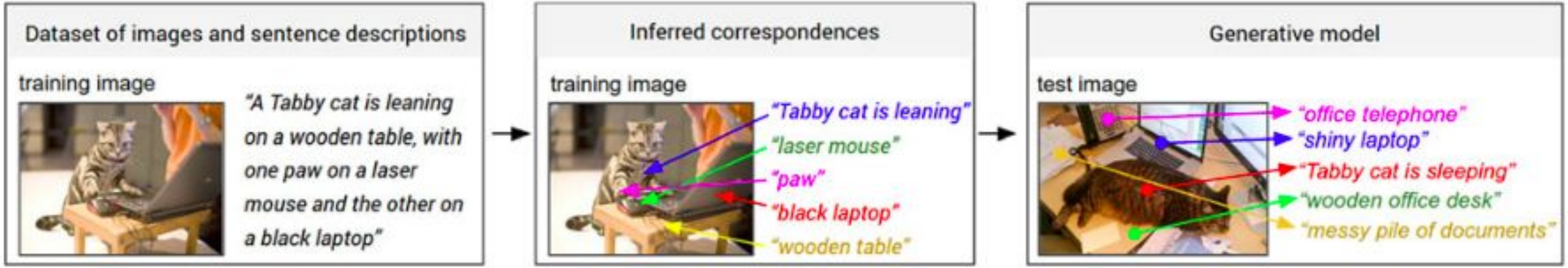
Previous methods have a few major drawbacks:

- Focused on labeling images with a fixed set of visual categories.
- Vastly restrictive compared to the enormous amount of rich descriptions that a human can compose.
- Previous models often rely on hard-coded visual concepts and sentence templates

Related Work

- Holistic scene understanding- correctly labeling scenes, objects and regions[3]
- Retrieval problem- Pick most compatible annotation [4]
- Combine most relevant annotations into meaningful sentence [5]
- Explicitly defined sentence templates [6] Single sentence
- Fixed Length [4][5][6], Closed vocabulary
- Relax fixed length- complex model [2]

Model



Dataset of images and sentence descriptions
training image
"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
training image
"Tabby cat is leaning"
"laser mouse"
"paw"
"black laptop"
"wooden table"
Generative model
test image
"office telephone"
"shiny laptop"
"Tabby cat is sleeping"
"wooden office desk"
"messy pile of documents"

Approaches

Aligning Sentences and Image Regions

- Representing images
- Representing sentences
- Alignment Objective
- Decoding text snippets

Generating Descriptions

Representing Images

Goal: Detect objects and encode the regions

Input: Pixels inside bounding boxes

Approach:
1.Region-based Convolutional Neural Network(R-CNN)
2.A pre-trained CNN is adopted to encode each bounding box into a 4096-dimensional vector.
3.A matrix Wm which has dimensions h × 4096 is used to compute the final h-dimensional representation as follows:

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

Output: Every image is thus represented as a set of h-dimensional vectors

Representing Sentences

Goal: Represent the words in the sentence into the same h-dimensional embedding space.

Input: A sequence of N words

Approach: Use a Bidirectional Recurrent Neural Network (BRNN) to compute the word representations as follows.

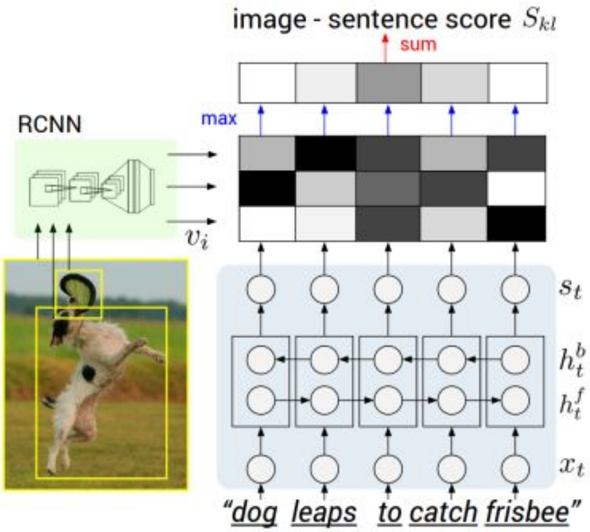
$$
\begin{aligned}
x_t &= W_w \mathbb{I}_t \\
e_t &= f(W_e x_t + b_e) \\
h_t^f &= f(e_t + W_f h_{t-1}^f + b_f) \\
h_t^b &= f(e_t + W_b h_{t+1}^b + b_b) \\
s_t &= f(W_d(h_t^f + h_t^b) + b_d).
\end{aligned}
$$

Output: Every word is thus represented as a set of h-dimensional vectors

Alignment Objective

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

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



Decoding text snippets

Goal: Generate contiguous sequences of words to a single bounding box.

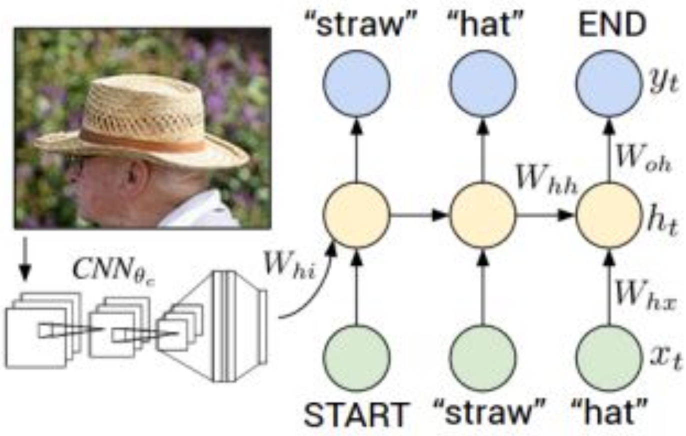
Input: Given a sentence with N words and an image with M bounding boxes, the input are latent alignment variables aj ∈ {1 . . . M } for j = 1 . . . N

Approach: Treat the true alignments as latent variables in a Markov Random Field (MRF)

$$E(\mathbf{a}) = \sum_{j=1 \ldots N} \psi_j^U(a_j) + \sum_{j=1 \ldots N-1} \psi_j^B(a_j, a_{j+1})$$
$$\psi_j^U(a_j = t) = v_i^T s_t$$
$$\psi_j^B(a_j, a_{j+1}) = \beta \mathbb{1}[a_j = a_{j+1}].$$

Output: A set of image regions annotated with segments of text.

Generating Descriptions



$$
\begin{aligned}
b_v &= W_{hi}[CNN_{\theta_c}(I)] \\
h_t &= f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1) \odot b_v) \\
y_t &= softmax(W_{oh}h_t + b_o).
\end{aligned}
$$

Datasets

- Flickr8K
- Flickr30K
- MSCOCO
- Visual Genome Dataset

Results

- Image sentence alignment evaluation
- Image Caption Generation
- Region Caption Generation
- Region Caption with Strong Supervision

Image Sentence Alignment Evaluation

Recall@k- a fraction of times an item was found within top K

Sorting on image sentence score Skl

Model	R@1	R@5	R@10
BRNN	22.2	48.2	61.4
Previous Model DeFrag[1]	19.2	44.5	58
Show n Tell[2]	23	-	63

Flickr30K Dataset

Sensitive to Compound Words and Modifiers

Decreasing Frequency Affects Results

Alignment Example

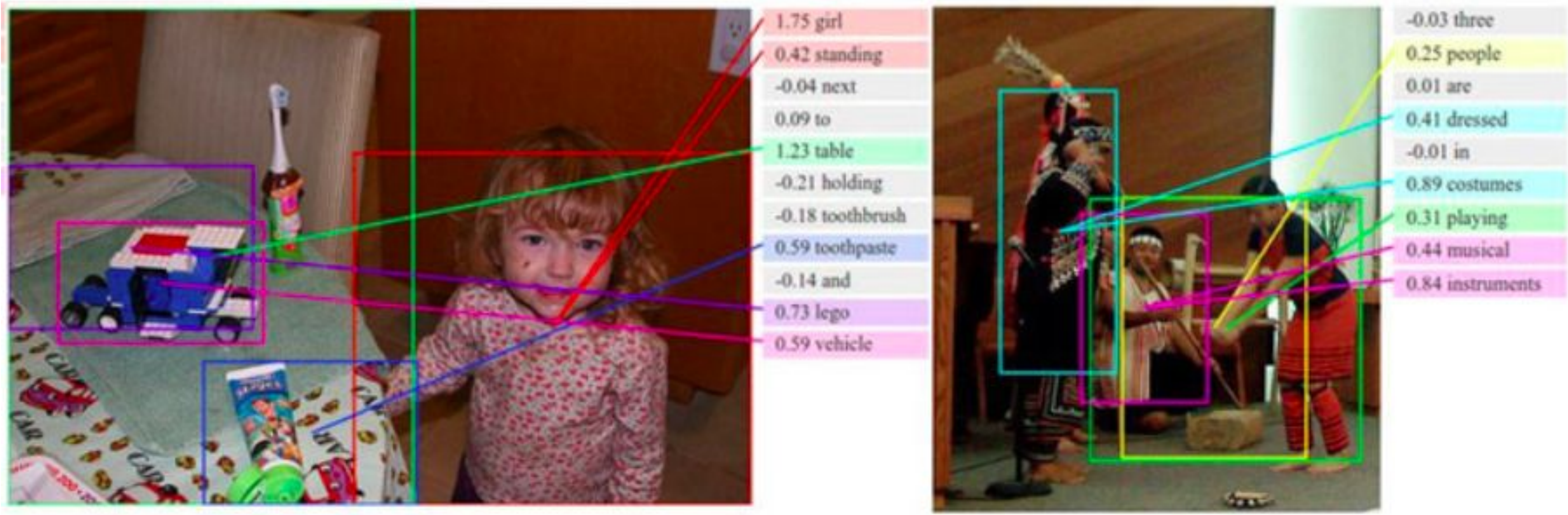


Image Caption Generation

<noun> in <noun> is <verb> in <noun>

man in black shirt is playing guitar.

woman in bikini is jumping over a hurdle.

Region Caption

Region Caption

Model	B-1	B-2	B-3	B-4
Human Agreement	61.5	45.2	30.1	22
Nearest Neighbour[7]	22.9	10.5	0	0
RNN- FullFrame	14.2	6.0	2.2	0
RNN-Region Level	35.2	23.0	16.1	14.8

Region Caption with Strong Supervision

Comparison on Region Level Model

Model	METEOR
RNN: FullFrame model	.209
RNN: Region level model	.272

Advantages

- Relax Description length
- Leveraged pre existing image caption dataset
- Alignment model- novel technique
- Used latent variable of MRF to take into account neighboring words and regions.
- Intrinsic and Extrinsic evaluation of models

Disadvantages

- The limitation of datasets.
- The image information fed into the RNN is only as a bias interaction.
- Not an end-to-end method from an image-sentence dataset.

Future Work

- Better dataset
- Word2vec can be substituted by random initializations
- Find a better interaction between images and the RNN
- Take into consideration the whole image as opposed to just regions to generate region descriptions.

Missing References

- 7. Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal Deep Learning. In Proceedings of the 2011 Conference International Conference on Machine Learning(ICML). 689–696.
- 8. Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human

References

1. Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14). MIT Press, 1889–1897.
2. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015), 3156–3164.
3. Stephen Gould, Richard Fulton, and Daphne Koller. 2009. Decomposing a scene into geometric and semantically consistent regions. In 2009 IEEE 12th International Conference on Computer Vision. 1–8.
4. Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every Picture Tells a Story: Generating Sentences from Images. In Computer Vision – ECCV 2010. 15–29.

Continued...

5. Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2012. Collective Generation of Natural Image Descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 359–368.
6. Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. BabyTalk: Understanding and Generating Simple Image Descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence (2013), 2891–2903.
7. Jacob Devlin, Saurabh Gupta, Ross B. Girshick, Margaret Mitchell, and C. Lawrence Zitnick. 2015. Exploring Nearest Neighbor Approaches for Image Captioning. ArXiv abs/1505.04467 (2015).

THANK YOU