Deep Visual-Semantic Alignments for Generating Image Descriptions

組別 : 15
0616023夏軒安, 0756029 邱肇珩

Deep Visual-Semantic Alignments for Generating Image Descriptions, 2015 CVPR,
Andrej Karpathy, Li Fei-Fei,
Department of Computer Science, Stanford University

Outline

- Introduction
- Alignment Model
  - RCNN
  - BRNN
- Similarity
- Generative Model
- Results
- Limitations

Concept

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

Introduction

- Generates natural language descriptions of images and their regions
- CNN over image regions, biderectional RNN over sentences
- Multimodal embedding
- Dataset
  - MSCOCO
  - Flickr8K
  - Flickr30K

Overview

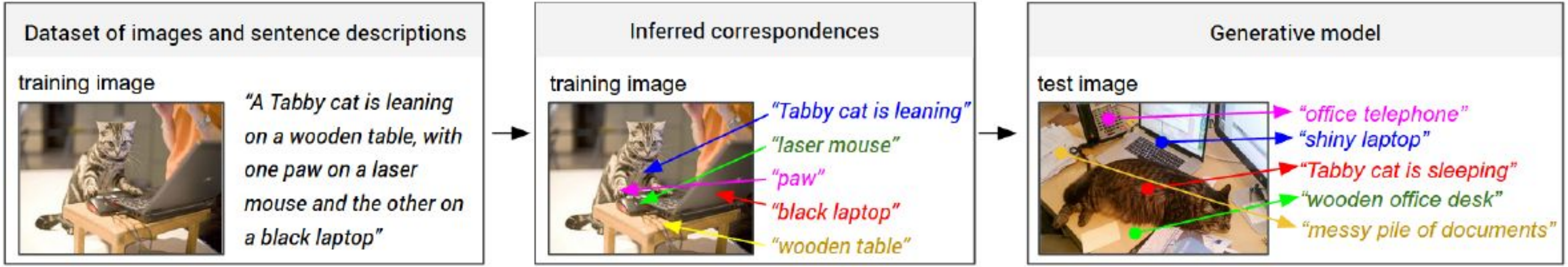
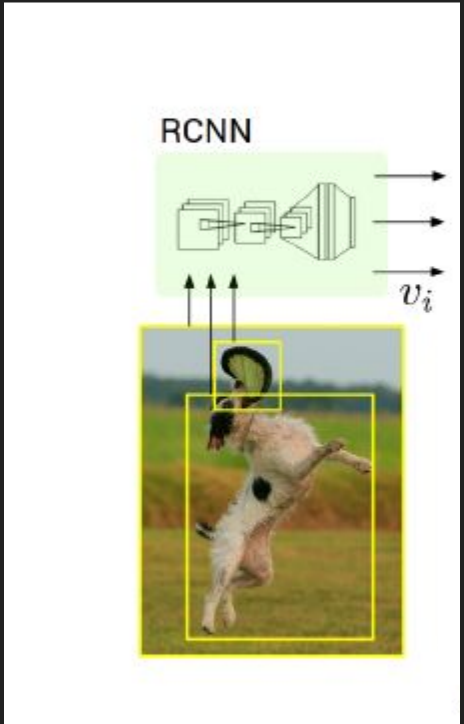


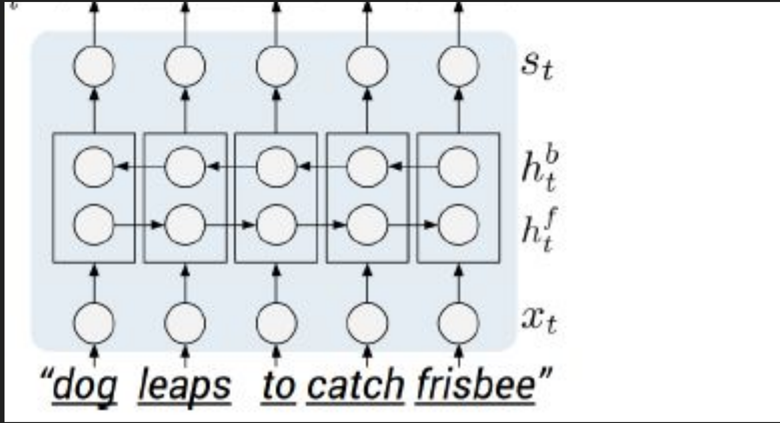
Figure 2. Overview of our approach. A dataset of images and their sentence descriptions is the input to our model (left). Our model first infers the correspondences (middle, Section 3.1) and then learns to generate novel descriptions (right, Section 3.2).

RCNN



$$v = W_m[CNN_{\theta_c}(I_b)] + b_m, \quad (1)$$

BRNN



$$x_t = W_w \mathbb{I}_t \tag{2}$$

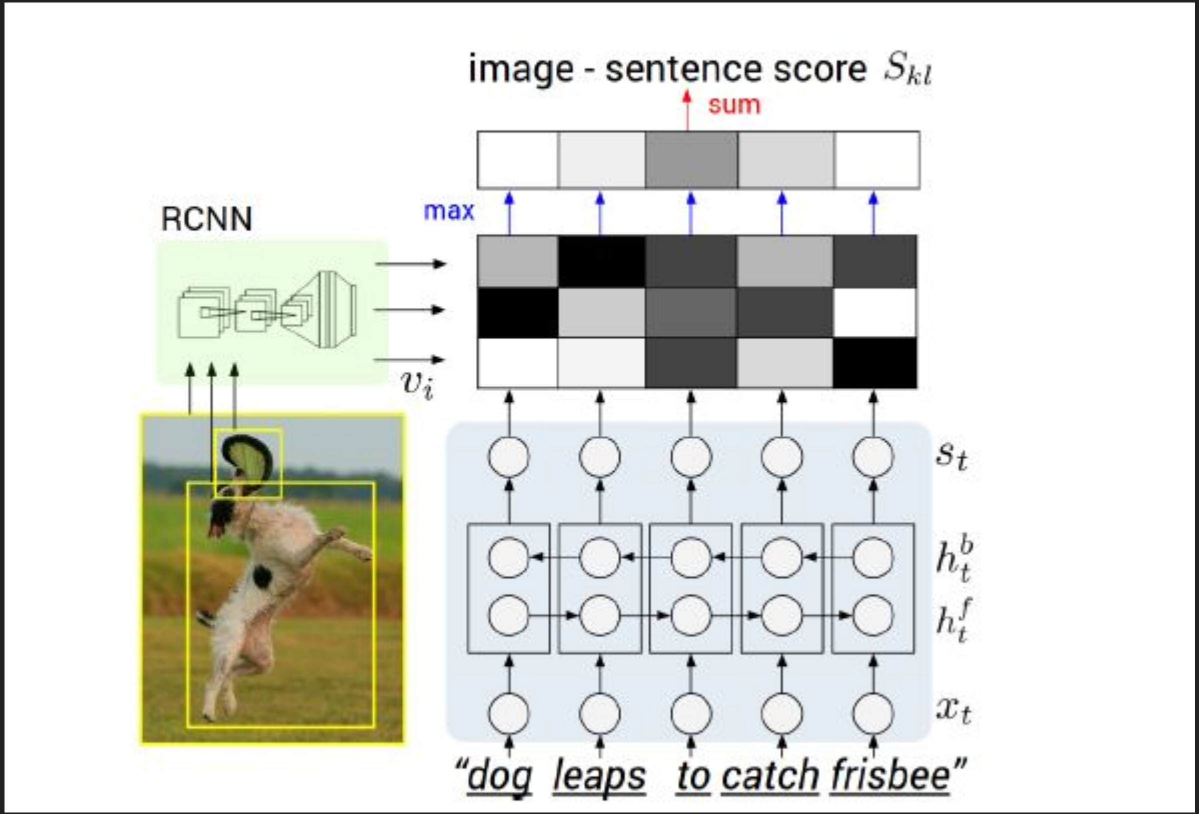
$$e_t = f(W_e x_t + b_e) \tag{3}$$

$$h_t^f = f(e_t + W_f h_{t-1}^f + b_f) \tag{4}$$

$$h_t^b = f(e_t + W_b h_{t+1}^b + b_b) \tag{5}$$

$$s_t = f(W_d(h_t^f + h_t^b) + b_d). \tag{6}$$

Alignment Model



Model Setting

- Region CNN is pretrained on Image Net, and use the top 19 detected locations and whole image projecting into the multimodal embedding space with dimension h (range from 1000-1600)
- BRNN maps each word to a the same embedding space
- The embedding matrix for input sentence was initialized with 300-dimensional word2vec weights and fixed due to overfitting concerns
- Activation function: ReLU

Similarity

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

Cost Function

$$\mathcal{C}(\theta) = \sum_k \Big[\underbrace{\sum_l max(0, S_{kl} - S_{kk} + 1)}_{\text{rank images}} \tag{9}$$

$$+ \underbrace{\sum_l max(0, S_{lk} - S_{kk} + 1)}_{\text{rank sentences}} \Big].$$

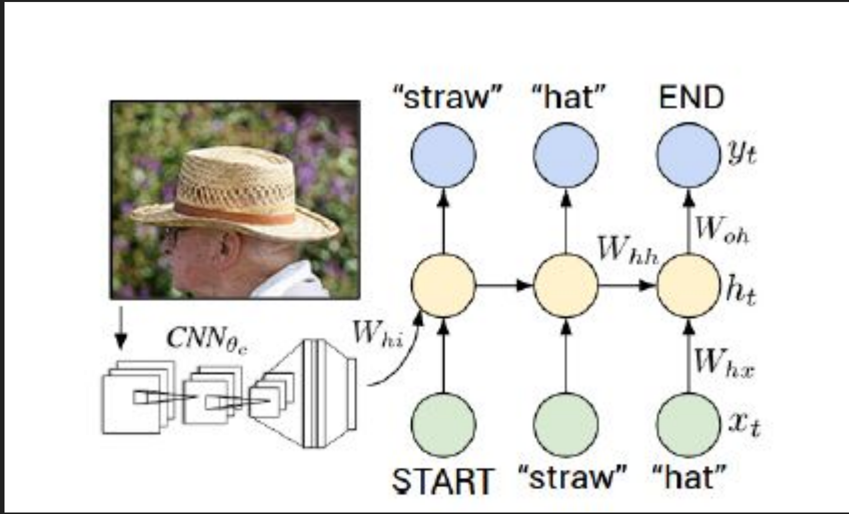
Magnitude

Magnitude	Word	Magnitude	Word
0.42	now	2.61	kayaking
0.42	simply	2.59	trampoline
0.43	actually	2.59	pumpkins
0.44	but	2.58	windsurfing
0.44	neither	2.56	wakeboard
0.45	then	2.54	acrobatics
0.45	still	2.54	sousaphone
0.46	obviously	2.54	skydivers
0.47	that	2.52	wakeboarders
0.47	which	2.52	skateboard
0.47	felt	2.51	snowboarder
0.47	not	2.51	wakeboarder
0.47	might	2.50	skydiving

Problem

- Multiple sentences may align to the same region
- Solved by Markov Random Field and control the length with hyperparameter beta

Generative Model



Model settings

- VGGNet + Simple RNN
- Hidden Size: 512
- SGD with Momentum 0.9 or RMSprop
- Dropout and Clipping

Question

- Why not LSTMs ?

→ Actually LSTMs consistently produced better results while took longer to train

Performance

Model	Image Annotation R@1	R@5	R@10	Med r	Image Search 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

Limitations

- Fixed Resolutions
- Pass image information only through bias terms
- No end-to-end training

reference

https://cs.stanford.edu/people/karpathy/cvpr2015.pdf

Thanks for listening