10.2. [Learning notes] Deep Visual-Semantic Alignments for Generating Image Descriptions. ☆

CVPR-2015 Andrej Karpathy, Li Fei-Fei.

Goal: To generate ~~g~~dense, free-form Region-level descriptions of images

not to generate Flicker-Like descriptions.

challenges: ① build a model that's rich enough to simultaneously reason about contents of images and their descriptions in the domain of natural language.

② how to learn a model that generates dense, region-level descriptions from training data of sparse, image-level descriptions.

Experiment:

1. Image-Sentence Alignment Evaluation
2. Generated Descriptions: Fulframe evaluation
3. : Region evaluation

Limitations

1. only for image: one input array of pixels at a fixed resolution.
2. RNN receive only bv. image represntation, less expressive
3. two separate models, not combine them together.

Contribution architecture

1. Infer region-word alignments, ~~image description~~ → (R-CNN (image) + BRNN (description) + MRF) learning to align visual & language data

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

combine

3. Generate region-level descriptions.

Model Overview:

During Training, Input: a set of images and their corresponding sentence descriptions.

1. present a model that aligns sentences snippets to the visual regions that they describe through a multimodel embedding.
2. Then, treat these correspondenses as training data for a second, multimodel RNN that learns to generates the snippets.

Technical Approach:

1. Learning to align visual and language data

1.1 Representing Images. (RCNN).

Input: Images

Output: a set of h-dimensional vectors.

$\{v_i | v_1, v_2, \cdots, v_{20}\}$. (top 20 detected locations)

※ h: size of multimodel embedding space. [$h \in (1000, 1600)$]

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

1.2 Representing Sentenses. (BRNN).

Input: Sentences. (N words, encoded in a 1-of-k representation)

Output: a set of h-dimensional vectors.

$\{s_t | s_1, s_2, \ldots, s_N\}$. (N: amount of words in sentense)

1.3 Alignment region & word

using an image-sentence score ⟶ region-word score.

(a image-sentence pair should have high score if words fit well in image)

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

Input: v_i, s_t.

Output: scores.

v_i: i-th region in image k

s_t: t-th word in sentence l.

g_k: a set of image fragments in image k

g_l: a set of sentence fragments in sentence l.

→ every word s_t aligns to the single best image region.

1.4. Alignment region & snippet (MRF).

we are interested in generating snippets of text instead of single words, so we align a sequence of word → a single bounding box.

Output: a set of image regions annoted with segments of text.

2. Multimodel RNN for generating descript

key challenge: predict a variable-sized sequence of outputs given an image.

RNN training:

combine: a word (x_t), previous context (h_{t-1}).

image information (b_v).

to predict next word (y_t).

RNN testing

compute: b_v, set $h_0 = 0$, x_1 to the START vector.

to predict the first word y_1.

Until the "END" token is generated.