Captioning ImageNet

Songtuan Lin

u6162630
Motivation and Contributions

• Motivations: Available datasets use to construct image captioning model are limited to MSCOCO and Flickr. Therefore, we want to extend available dataset.

• Contributions:
  • Caption images of ImageNet dataset in a semi-autonomous way and hence, extend the available dataset in community.
  • Provide essential analysis which lead us to find some clues to further filter the good and bad generated captions.
Outline

• Structure of ImageNet Dataset.

• Definition of Image Captioning Task.

• Model

• Decoding Algorithm

• Implementation
ImageNet Dataset
ImageNet

• A large collection of images (about 14 millions).

• Group images which contain the same entity into a set called synset, e.g. all images contain dog will be in the same synset.

• A tag to describe the entity, e.g. {dog, domestic dog}
Image Captioning
Definition

• Generate natural language description for a given image.

A large elephant that is standing in grass.
Mathematical Definition

• Each word is a random variable and takes value from a pre-defined vocabulary $\mathcal{V}$, e.g. the $i$th word in caption is a random variable $w_i$ and $w_i \in \mathcal{V}$, then, for a given image $\mathcal{I}$, we define the image captioning task as:

$$\max_{w_1 w_2 \cdots w_n} \mathcal{P}(w_1 w_2 \cdots w_n | \mathcal{I})$$

• E.g. in previous example:

$$w_1 = a, w_2 = \text{large}, w_3 = \text{elephant}, \cdots, w_8 = \text{grass}$$
Model
Encoder-decoder

\[ P(w_i | w_1 w_2 \ldots w_{i-1}, I) \]

\[ w_i = v_i \]
Decoding Algorithm
Greedy

• Pick $w_i$ which maximize $\mathcal{P}(w_i|w_1w_2\cdots w_{i-1}, \mathcal{I})$, i.e.

$$w_i = \arg\max_{w_i} \mathcal{P}(w_i|w_1w_2\cdots w_{i-1}, \mathcal{I})$$

• Drawback:

$$\max\log\mathcal{P}(w_1w_2\cdots w_n|\mathcal{I}) \geq \sum_{i=1}^{n} \max\log\mathcal{P}(w_i|w_1w_2\cdots w_{i-1}, \mathcal{I})$$

• E.g. a caption $w_1w_2$ and vocabulary $\mathcal{V} = \{v_1, v_2\}$, i.e.

$$w_1, w_2 \in \mathcal{V}$$
Example

\[
\begin{array}{|c|c|c|}
\hline
   & w_1 = v_1 & w_1 = v_2 \\
\hline
\mathcal{P}(w_1 | I) & 0.45 & 0.55 \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|c|}
\hline
   & w_2 = v_1 & w_2 = v_2 \\
\hline
\mathcal{P}(w_2 | w_1 = v_1, I) & 0.1 & 0.9 \\
\hline
\mathcal{P}(w_2 | w_1 = v_2, I) & 0.6 & 0.4 \\
\hline
\end{array}
\]

\[
\max \log \mathcal{P}(w_1, w_2 | I) = \log \mathcal{P}(w_1 = v_1, w_2 = v_2 | I)
\]
\[
> \max \log \mathcal{P}(w_1 | I) + \max \log \mathcal{P}(w_2 | w_1, I)
\]
\[
= \log \mathcal{P}(w_1 = v_2 | I) + \log \mathcal{P}(w_2 = v_1 | w_1 = v_2, I)
\]
Beam Search

- Choose $k$ candidates who have top $k$ joint probability values.
Constrained Beam Search

• We hope our generated captions can satisfy some constrains, e.g. contain a specific token $v_s \in \mathcal{V}$.

• Need two beams and assign different update rules to them.

• E.g.
Implementation
Implementation

• Target: Caption images in ImageNet and force one of the phrase in ImageNet tag appear in generated captions.
• ImageNet tag: e.g. \{dog, hunting dog, domestic dog\}
• Approach: regular expression $\rightarrow$ finite automata
• Regular expression: $*$dog|hunting dog|domestic dog*$

• Finite automata:
Outcomes
Outcomes

**Constrained:** A picture of a fish in the water.

**Original:** A close up of a black toaster on the ground.

**Constrained:** A fish that is swimming in the water.

**Original:** A bird that is swimming in the water.
Analysis
Experiment Setting Up

• Pick 100 images and manually categorized them into two sets:
  • One set in which constrained beam search improves the quality of captions.
  • One set that constrained beam search does not work so well.

• For each image in each set, we compare the log-probability of the captions generated by both constrained beam search and original beam search.
Analysis Result

Case that constrained beam search improve the outcome

Avg log-probability difference: 3.8821

Case that constrained beam search does not improve the outcome

Avg log-probability difference: 5.9515
Example

**Constrained**: A close up of a *snake* in the dirt. ($p = -14.405$)

**Original**: A close up of a piece of wood on the ground ($p=-10.221$)

**Constrained**: A chocolate *snake* sitting on top of a wooden table. ($p = -15.852$)

**Original**: A chocolate cake sitting on top of a wooden table ($p=-5.138$)
Result 2

Case that constrained beam search improve the outcome

Avg log-probability difference: 6.4097

Case that constrained beam search does not improve the outcome

Avg log-probability difference: 7.1695
Thank You