Final Project Proposal

Wen Ge
Zhaoyu Sun
Haoxiang Ma

What we want to accomplish:

It is very easy for human to recognize and describe an immense amount of details in an huge image.However this task is pretty exclusive for computer to accomplish.

There remains many great previous works focusing on labeling images with a fixed set of visual categories. However, these works usually rely on hard-coded visual concepts and sentence templates, which are vastly restrictive compared to the enormous amount of rich descriptions that a human can compose.

We want to accomplish a model which is rich enough to simultaneously detect objects of images and generate their descriptions in text. Moreover, the description of objects are free of assumptions about specific hard-core tempates,rules and categories and instead rely on training data which means our description are not simply specific labels.

Dataset :

We use Flickr8K and MSCOCO (if time permits), each dataset is annotated with 5 sentences using Amazon Mechanical Turk. For Flickr8K, we use 1000 images for validating, 1000 for testing and the rest for training, and for MSCOCO, we use 5000 images for both validating and testing.

We convert all sentences to lower-case, discard non-alphanumeric characters, and filter words to those occur at least 5 times in the training set, which results in 2538, 7414 and 8971 words for Flickr8K and MSCOCO datasets respectively.

The Flickr dataset is built by forming links between images sharing common metadata from Flickr. Edges are formed between images from the same location, submitted to the same gallery, group, or set, images sharing common tags, images taken by friends, etc.

MSCOCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

1. 1.Object segmentation
2. 2.Recognition in context
3. 3.Superpixel stuff segmentation
4. 4.330K images (>200K labeled)
5. 5.1.5 million object instances
6. 6.80 object categories
7. 7.91 stuff categories
8. 8.5 captions per image
9. 9.250,000 people with keypoints

Evaluation:

We use recall@K, which measures the fraction of times a correct item was found among the top K results to evaluate how the image-sentence alignment works.

Besides that, we use B-n, which is BLEU score between 0 and 100 that uses up to n-gram, to evaluate how well is our result (which is candidate sentence in BLEU) matching a set of five ground truth sentence (which is reference sentence in BLEU) given by our dataset.