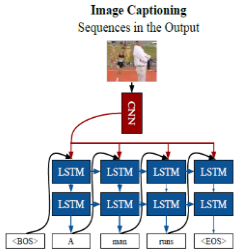
MATH 6038o Mini-Project 2: Image Captioning

XU SIAO, TAN JUNHONG, HE YUNJIE, LU JIHUA; Department of Mathematics

Introduction

The image description generator implemented by CNN/RNN/LSTM provides a useful framework for learning from image mapping to natural language image description. We obtain a model by training a collection of a large number of images and corresponding titles. This model can capture relevant semantic information from visual features.



Graph1:the procedure for image caption.

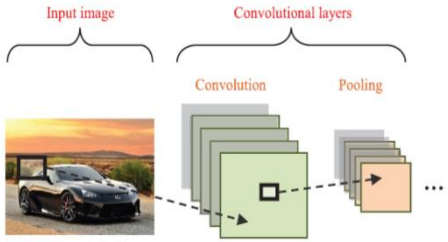
Data

The training and test sets are made up of 8,000 images and 5 captions for each. And have been divided into training and test sets already.

.

Feature extraction

A pre-trained CNN is used to encode an image to its features. In this implementation InceptionV3 is used as encoder and with it pretrained weights loaded. The last softmax layer of InceptionV3 is removed and the vector of dimension(4096) is obtained from the second last layer.



Graph 2: the feature extraction by CNN.

Word embedding

Word Embedding is a technique developed by Bengio from 2000. Now it has become the foundation in Natural Language Processing (NLP) field. Among all concepts of Word Embedding, word2vec is the most important and most frequently used method which transforms words from combination of letters into high-dimension vectors that computer can understand and process.

Training models of word2vec: Skip-gram and CBOW

(1)Skip-gram

Skip-gram: : P(Cw|v)

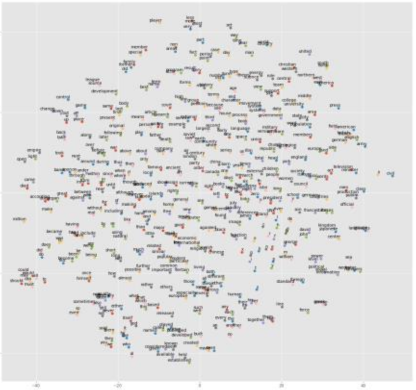
Use the center word to predict each of the context words

(2)CBOW

Continuous bag of words (CBOW): P(v|Cw)

Use the context words (average) to predict the center word

Graph 3: visualization of word embedding.



```
[[ 0.0117059  -0.07715672  0.04201284 ..., -0.07297134  0.04485371
   0.05419442]
 [-0.1991424  -0.10599716 -0.01000309 ...,  0.15442981  0.12517811
   0.08598424]
 [-0.07662097 -0.15854724  0.08641349 ...,  0.06110508  0.00803996
   0.08272742]
 ...,
 [ 0.14100078  0.13734384  0.00189618 ..., -0.02524431 -0.1101329
  -0.04445839]
 [-0.00695473  0.04827187 -0.03015917 ..., -0.0259588  -0.09127076
  -0.05680626]
 [-0.06476382  0.14940147  0.08061296 ...,  0.10035349 -0.14382018
   0.14347239]]
```

→ 'UNK'
→ 'the'
→ 'of'
→ 'george'
→ 'band'
→ 'together'

Graph 4: each vector stand for one specific word.

Methodology

Models

- **CNN**: We implement a CNN structure to do feature selection. We complete it with transfer learning by using InceptionV3 .
- **Word embedding** : We use the output before fully connected layer for feature extraction, and change the structure of fully connected layer for fine-turning.
- **RNN** : The specific recurrent neural network we use is long short-term memory which provide a solution by incorporating memory units that explicitly allow the network to learn when to “forget” previous hidden states and when to update hidden states given new Information.

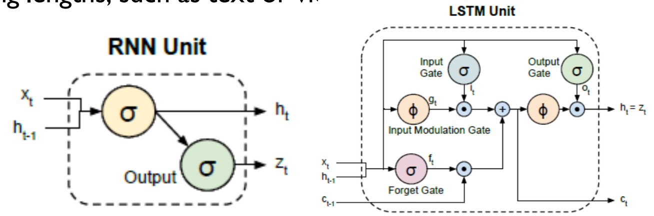
Evaluation

- BLEU(bilingual evaluation understudy**)**: an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another .

Long Short-Term Memory

The advantages of LSTMs for modeling sequential data in vision problems are twofold. First, when integrated with current vision systems, LSTM models are straightforward to fine-tune end-to-end. Second, LSTMs are not confined to fixed length inputs or outputs allowing simple modeling for sequential data of varying lengths, such as text or video

Graph 5 and 6: the comparison between traditional RNN and LSTM



Result

We used Beam search with k=3, 5, 7 and an Argmax search for predicting the captions of the images. The loss value of 1.5987 has been achieved which gives good results. After train the parameters of model by training data, we input test picture and get the following output. Below are two examples of output we obtain:

Graph 7 and 8: two examples for image captioning

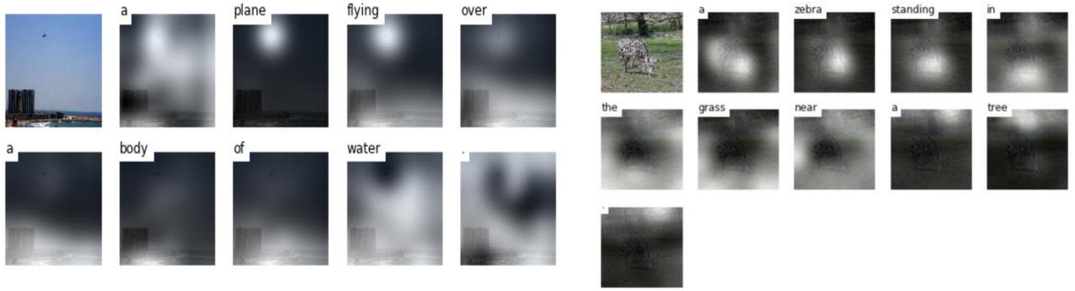
Normal Max search: A little girl in a red coat plays in snow .
Beam Search, k=3: A little kid plays in the snow in a brown jacket and red shorts on a harness .
Beam Search, k=5: Little girl in red coat going down a hill .
Beam Search, k=7: Little girl in red coat going down a hill .

Normal Max search: A snowboarder is riding down the ramp next to a hill .
Beam Search, k=3: A person on a snowboard jumps over a cliff in the snow .
Beam Search, k=5: A person on a snowboard jumps over a cliff in the snow .
Beam Search, k=7: A person on a snowboard jumps over a cliff in the snow .

Visual Attention

A plane flying over a body of water.

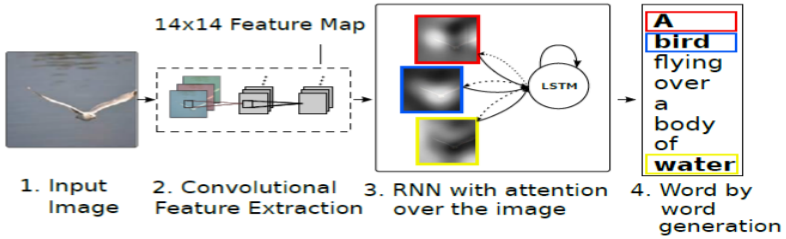
A zebra standing in the grass near a tree.



Graph 11 and 12: two examples of visual attention.

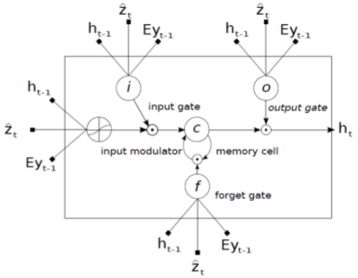
Visual Attention

Rather than compress an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed. This is especially important when there is a lot of clutter in an image.


14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Graph 9: procedure for visual attention which is similar to traditional image caption.

Using representations (such as those from the top layer of a convnet) that distill information in image down to the most salient objects is one effective solution that has been widely adopted in previous work. Unfortunately, this has one potential drawback of losing information which could be useful for richer, more descriptive captions. Using more low-level representation can help preserve this information. However working with these features necessitates a powerful mechanism to steer the model to information important to the task at hand.



Graph 10: LSTM used in visual attention which is a bit different from traditional LSTM.

BLEU

We report results with the frequently used BLEU metric which is the standard in the caption generation literature. We report BLEU from 1 to 4 without a brevity penalty. There has been, however, criticism of BLEU, so in addition we report another common metric METEOR, and compare whenever possible.

Chart 1: the test for BLEU and METEOR

ratio	1.01535300365
Bleu_1	0.65590395416
Bleu_2	0.446168005543
Bleu_3	0.30441521379
Bleu_4	0.211145033206
METEOR	0.215652854828

Contribution :
Every team member evenly contributes to all the tasks, including math modeling, data preprocessing, code writing and poster producing.

Reference

- Long-term Recurrent Convolutional Networks for Visual Recognition and Description. Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama,
- Show, Attend and Tell: Neural Image Caption Generation with Visual Attention Kelvin Xu KELVIN.
- ImageNet Classification with Deep Convolutional Neural Networks. Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton.