Optimized Neural Network Story Generator

Cong Ye and David Wang CS229 2017 Stanford University

Introduction

The goal of this project is to combine visual and language processing by building an intelligent storyteller based on input images. We will focus on entertainment purpose first and same model with appropriate dataset and feature adjustments could also be used for early Childhood Education, medical science and geography research.

To achieve our goal, we first try to employ unsupervised learning of a generic, distributed sentence encoder. Then, we will leverage the continuity of text from novel and movie scripts as training dataset. The only source of supervision in our models is from Microsoft COCO images to captions. That is, we did not collect any new training data to directly predict stories given images.

Data Collection

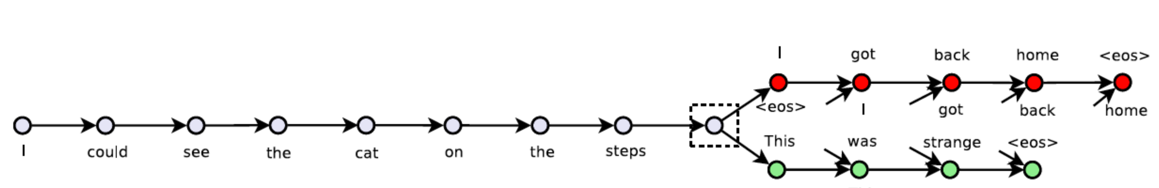
- For training visual-semantic embedding modal, we use Microsoft COCO dataset.
- For training sentence semantic "vector", we use BookCorpus dataset from University of Toronto
- To evaluate our smart vector performance, we use SICK dataset.

Research

We divided the potential areas that could help our project into the following two parts, image captioning and sentence semantics.

For the image captioning, we found some early stage work. For example some researchers used CRF Labeling method to present a system to automatically generate natural language descriptions from images. And a system with better performance and accuracy was built with CNN. Similar goals were also achieved by m-RNN.

The paper inspired us most describes skip-thought vector algorithm to track the sentence semantics. Sentences that share semantic and syntactic properties are thus mapped to similar vector representations.



Model and Algorithms

The entire systems includes three parts, encoder, decoder and objective function. We focused on optimizing the first two parts.

Encoder Optimization

The Gated Recurrent Neural Network have shown success in applications involving sequential or temporal data but increase parameterization and is expensive. We experiment with different variation of GRU and reduce the parameters in the network without compromising the performance.

The encoder definition:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$
$$\tilde{h}_t = g(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h)$$

With the two gates to the variant gate presented as update gate and reset gate:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad \Rightarrow \quad z_t = \sigma(U_z h_{t-1} + b_z)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad \Rightarrow \quad r_t = \sigma(U_r h_{t-1} + b_r)$$

Reduced the number of parameters in the GRU RNN from $3 \times (n^2 + nm + n)$ by $2 \times mn$

Decoder Optimization

We defined "Smart Vector" algorithm to better extract each sentence's semantic with unsupervised learning.

one decoder for the first sentence of each paragraph, one for the next sentence and one for the previous sentence. A number of linguistics researches have shown first sentence of each paragraph have significant semantic relatedness with the rest of the sentence in this same paragraph.

$$h_{t+1} = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$
$$\tilde{h}_t = g(W^d x_t + U^d (r_t \odot h_{t-1}) + b_h)$$
$$z_t = \sigma(W_z^d x_{t-1} + U_z^d h_{t-1} + b_z)$$
$$r_t = \sigma(W_r^d x_{t-1} + U_r^d h_{t-1} + b_r)$$

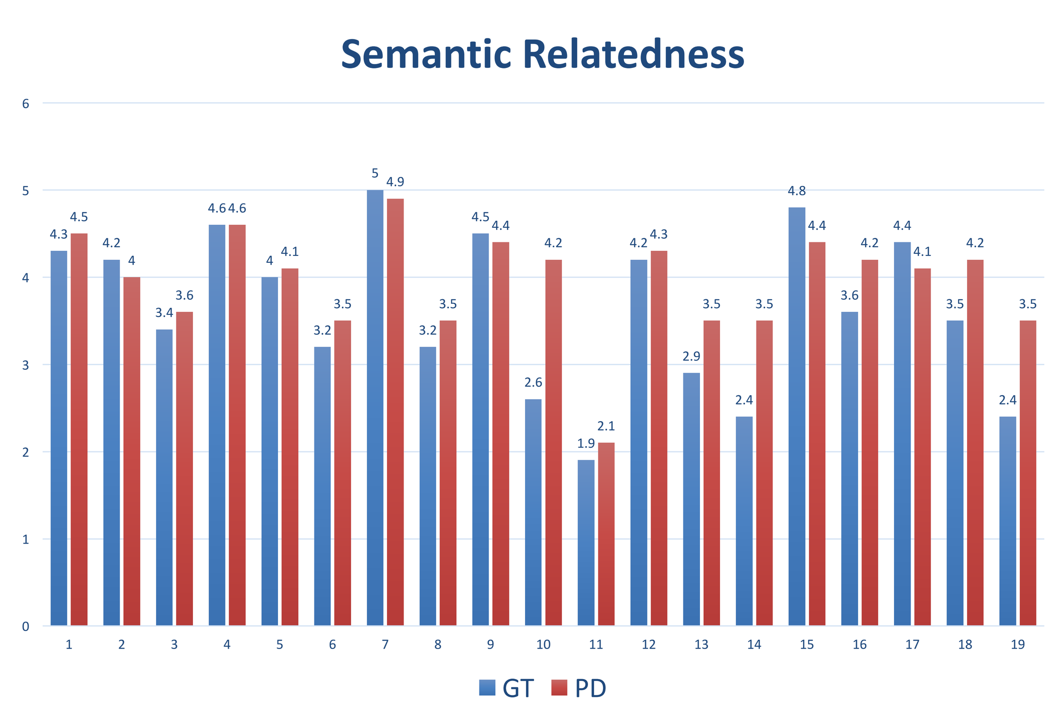
Objective: given a tuple, where S_0 is the first sentence of the paragraph

$$(s_0, s_{i-1}, s_i, s_{i+1})$$

The objective optimized is the sum of the log-probabilities for the first sentence of each paragraph, the forward and backward sentence conditioned on the encoder representation:

$$\sum_t \log P(w_0^t \mid w_0^{<t}, h_i) + \sum_t \log P(w_{i+1}^t \mid w_{i+1}^{<t}, h_i) + \sum_t \log P(w_{i-1}^t \mid w_{i-1}^{<t}, h_i)$$

Evaluation



Predictions from the SICK test set. GT is the ground truth relatedness marked in dataset, scored between 1 and 5 and PD is our modal's judgment.

The table shows nearest neighbors of sentences from a smart vector model trained on the BookCorpus dataset. These results show that smart vectors learn to accurately capture semantics and syntax of the sentences they encode.

Query and nearest sentence
if he had a weapon, he could maybe take out their last imp , and then beat up errol and vanessa . if he could ram them from behind, send them sailing over the far side of the levee , he had a chance of stopping them .
then, with a stroke of luck , they saw the pair head together towards the portaloos . then, from out back of the house , they heard a horse scream probably in answer to a pair of sharp spurs digging deep into its flanks .
I'll take care of it , " goodman said , taking the phonebook . I'll do that ," julia said , coming in .

Result

Test with a online picture

Nearest Captions:
A woman on the beach has a pink hat and umbrella .
A woman is standing on the beach with a red umbrella .
A women on a beach holding a pink umbrella .
A woman walking across a beach in her panties and a blue shirt .
The woman in a dress stands by the shoreline and waves streamers all around

Story Generated:

I woman was on the beach , holding her breath . She gave me a quick hug , and she had no idea what to do . In fact , it seemed as if I had never seen her come out of the surf . In fact , I was going to be the only woman in the world for the past twenty-four hours . She shook her head and bowed her head over my shoulder . In fact , it was so much easier for him to go on a tropical beach at The Shade . She felt as if I were the only woman in the world , standing naked on a sandy beach .

Future Work

- We will find a way to improve training efficiency in further and reduce computational complexity in network parameterization.
- In future, instead of a simple description of the picture, we want to optimized our module to understand the meaning of a picture.