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:

\[
\begin{align*}
    h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \\
    \tilde{h}_t &= g(W_x h_{t-1} + U (r_t \odot h_{t-1}) + b) \\
    z_t &= \sigma(W_x h_{t-1} + U z_t + b) \\
    r_t &= \sigma(W_x h_{t-1} + U r_t + b)
\end{align*}
\]

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

\[
\begin{align*}
    z_t &= \sigma(W_x h_{t-1} + U z_t + b) \\
    r_t &= \sigma(W_x h_{t-1} + U r_t + b)
\end{align*}
\]

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

\[
\begin{align*}
    h_{i+1} &= (1 - z_t) \odot h_{i+1} + z_t \odot \tilde{h}_t \\
    \tilde{h}_t &= g(W_x h_{i+1} + U (r_t \odot h_{i+1}) + b) \\
    z_t &= \sigma(W_x h_{i+1} + U z_t + b) \\
    r_t &= \sigma(W_x h_{i+1} + U r_t + b)
\end{align*}
\]

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:

\[
\sum_t \log P(w^i_t | w^0_t, h_t) + \sum_t \log P(w^{i+1}_t | w^i_t, h_t) + \sum_t \log P(w^i_t | w^{i+1}_t, h_t)
\]

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 model’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.

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.