Final Project

Generation of Captions from Images

Cai, Tianyuan

CSCI S-89a Deep Learning, Summer 2019

Harvard University Extension School

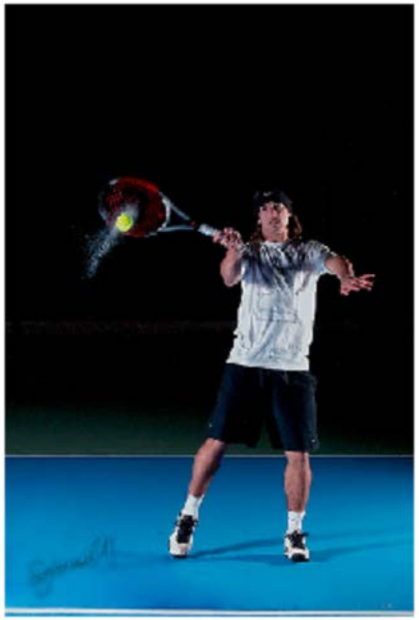
Prof. Zoran B. Djordjević

Goal

Use neural network to generate sensible captions for images

Presentation Structure

- Data
- Embeddings
  - Image
  - Text
- Model
- Discussion



```
Predicted Caption: a young man is jumping to catch a tennis ball
Actual Caption:  a man is swinging a racket at a ball
```

Data

- **Images**
  Training image data set from Common Objects in Context (COCO) data set produced by Microsoft. Due to the size of this data set, I obtain a subset of the downloaded data set and split it into train and validation set.

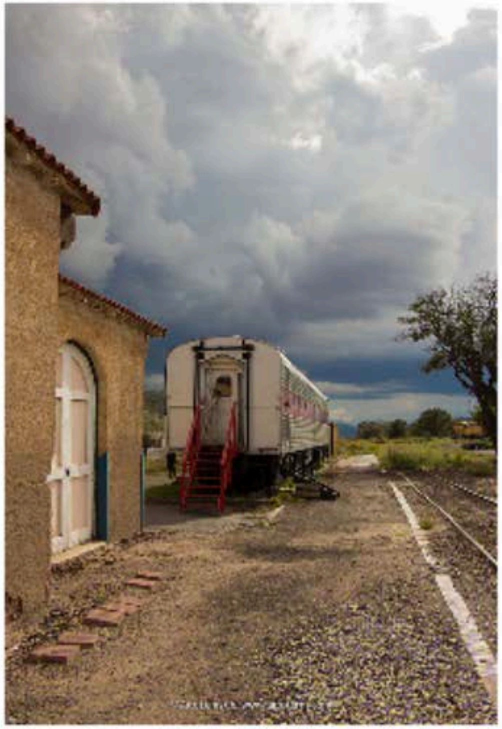
- **Captions**
  Captions of the training images from COCO. This data set contain captions that correspond to the images in the training images data set.

Image Embeddings

- Use the **Inception V3** model as an image encoder by removing the fully connected layers in the end.
  - The weights of the model is trained on ImageNet data.
- Data manipulations
  - Convert images to array
  - Resize images to 299 by 299
  - Keras preprocess

Sample Output from Inception V3

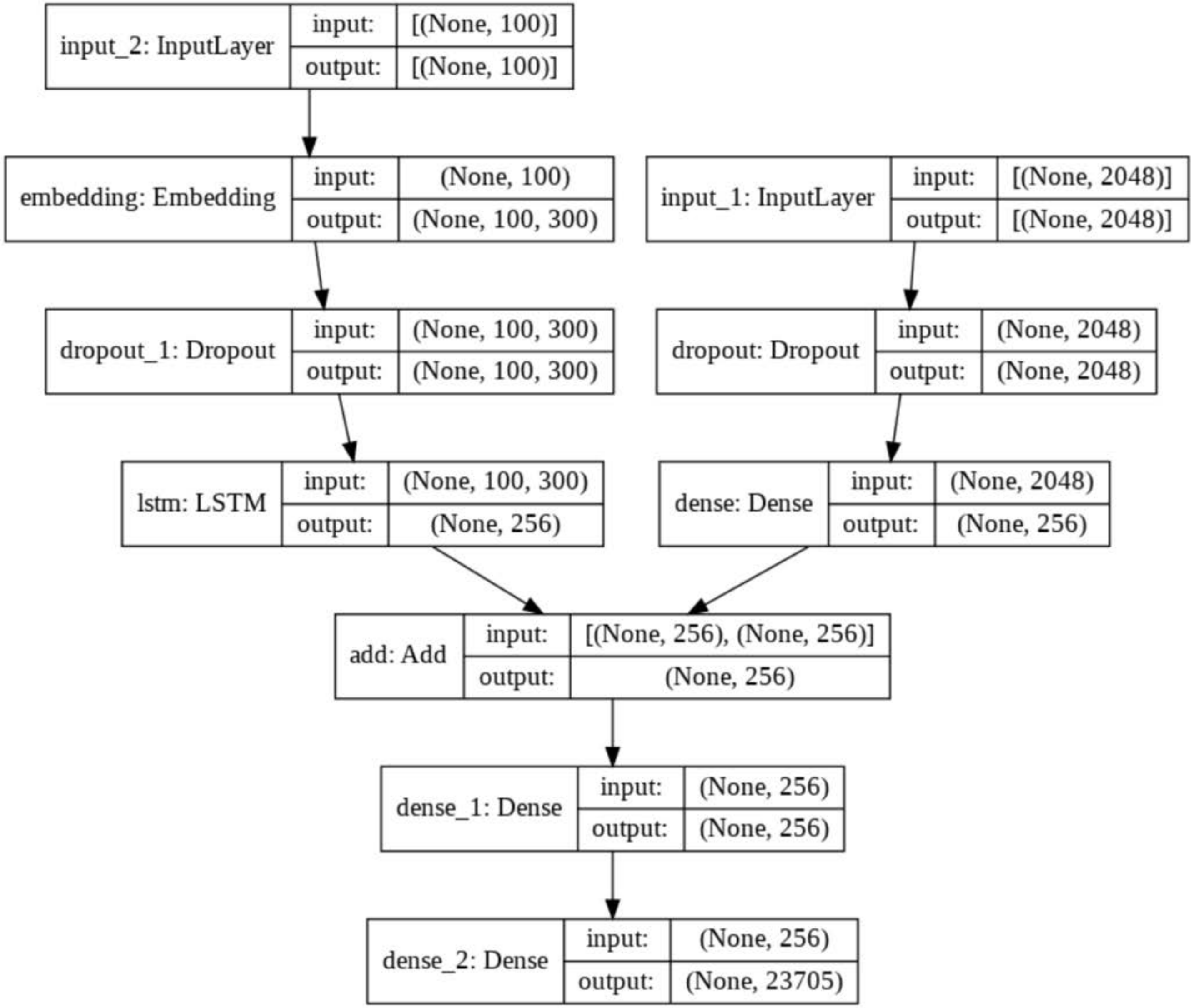
- Freight car: 40.83%
- Passenger car: 29.35%
- Electric locomotive: 5.97%
- Steam locomotive: 1.17%
- Mobile home: 0.59%



Word Embeddings

- Use Global Vectors for Word Representation (GloVe) to obtain embedding vectors for words in the captions.
  - The position of a word within the vector space is learned based on the words that surround the word in the training data, in this case, on the Wikipedia 2014 and Gigaword 5 data (glove.6B.300d.txt)
- Data manipulations
  - Remove special characters
  - Padded to a length of 100 words
  - Set the maximum number of words to 6,000
  - Adding tags “start_sentence” and “end_sentence”

Model



input_2: InputLayer | input: [(None, 100)] | output: [(None, 100)]

embedding: Embedding | input: (None, 100) | output: (None, 100, 300)

dropout_1: Dropout | input: (None, 100, 300) | output: (None, 100, 300)

lstm: LSTM | input: (None, 100, 300) | output: (None, 256)

input_1: InputLayer | input: [(None, 2048)] | output: [(None, 2048)]

dropout: Dropout | input: (None, 2048) | output: (None, 2048)

dense: Dense | input: (None, 2048) | output: (None, 256)

add: Add | input: [(None, 256), (None, 256)] | output: (None, 256)

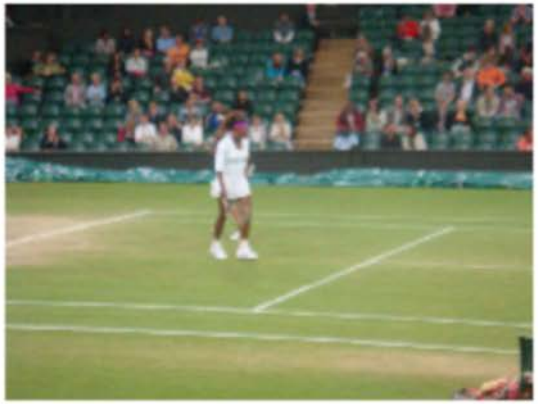
dense_1: Dense | input: (None, 256) | output: (None, 256)

dense_2: Dense | input: (None, 256) | output: (None, 23705)

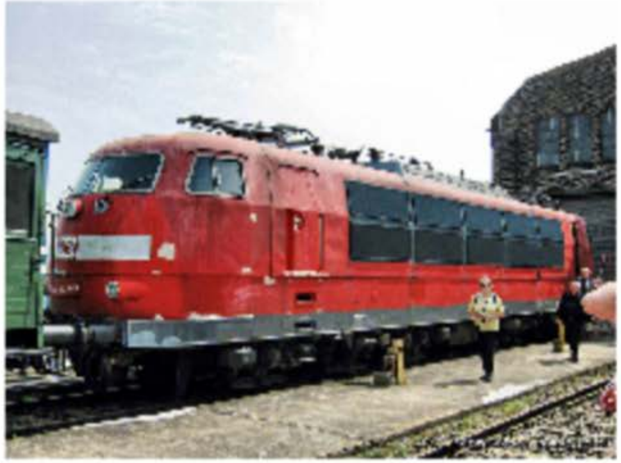
Model Overview

- Input layers from images and captions vectors are passed into the model
  - Dropout layers are applied to the respective inputs
  - The text input is passed through an LSTM layer
  - The image input through a fully connected layer to ensure the output has the same dimension as the LSTM output.
- The input tensors are then added together and passed through fully connected layers for next word prediction.
- To train and predict the caption word-by-word,
  - the model starts by using the current image and the starting tag “start_sentence” to predict the first actual word of the caption.
  - After this iteration, two words are in the input.
  - Based on the “start_sentence” and the last word predicted, the model is then trained to predict the second actual word of the caption.
  - This cycle continues until the model has predicted the “end_sentence” word.
- Final model: batch size = 500, epochs = 50.

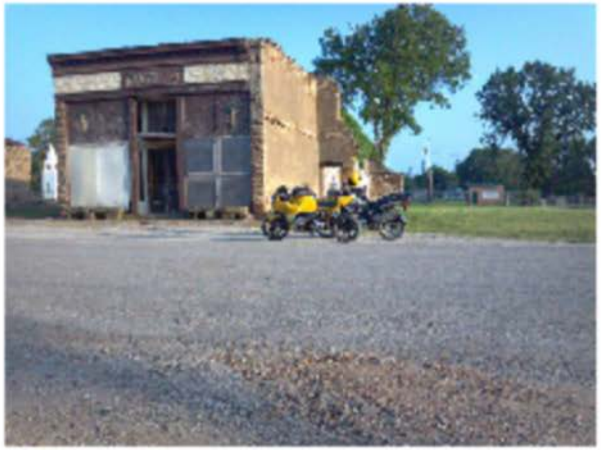
Good Results



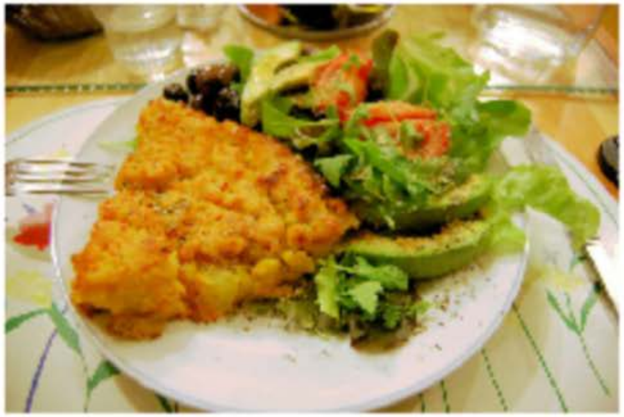
A professional tennis game with a lot of spectators.



A train comes down the tracks and enters the tunnel

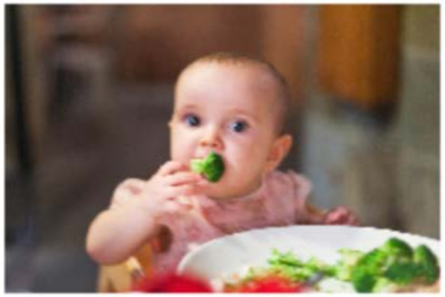


A green motorcycle parked in a parking space.

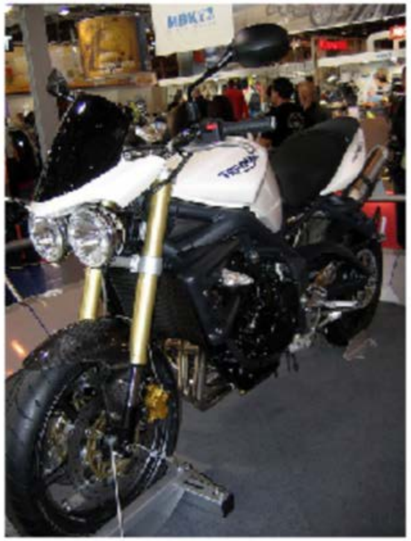


A buffet with lots of clutter and vegetables on a table.

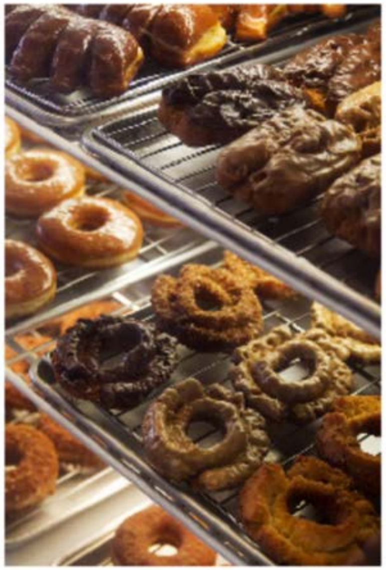
Sensible Results



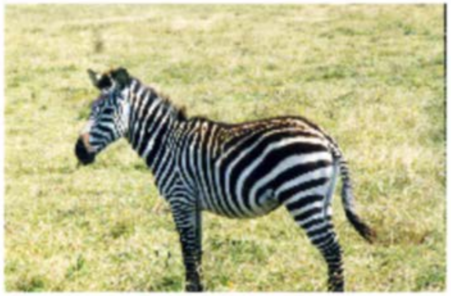
A lot of buckets of fruits including red and green apples.



A golden motorcycle driving along a street near a truck space.

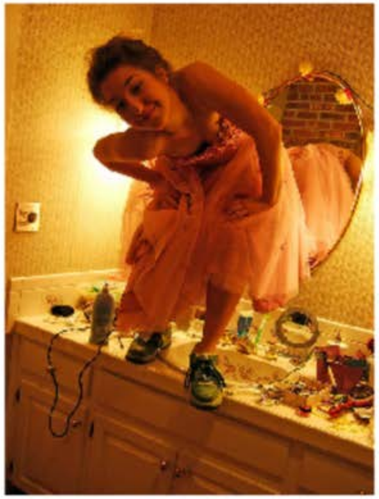


Grouped fruits in boxes with handwritten price signs.

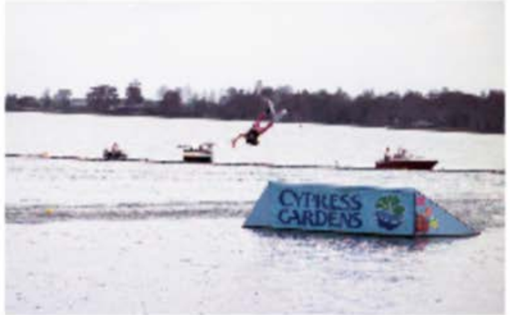


This is a image of an zoo outdoor.

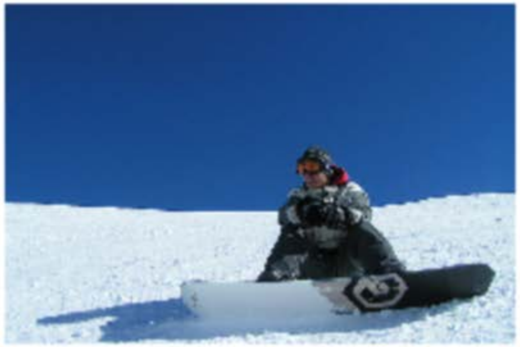
Surprising Results



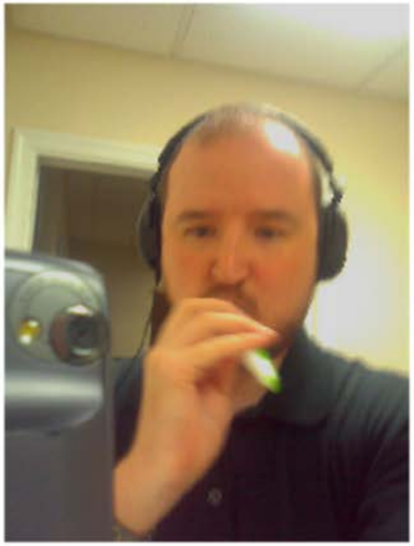
A cookie and orange on a table next to a tablet computer.



A man is standing next to others talking on a cell phone.



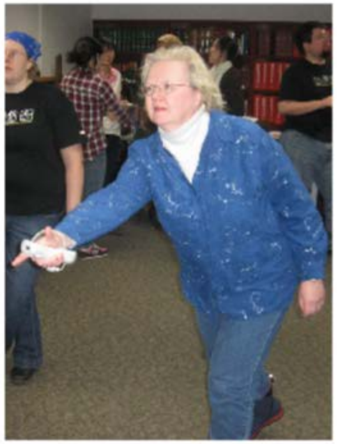
A man flies a kite on the beach.



A bearded man is wearing a tank top tie and a hat.

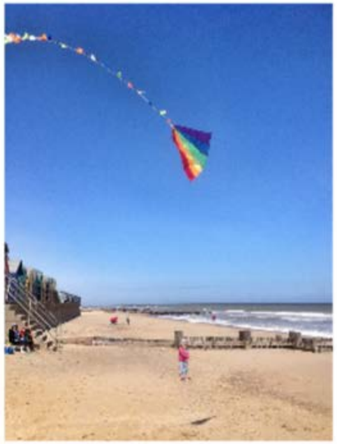
Areas of Improvement

- The model seems to have trouble distinguishing between objects that have subtle differences, such as sky and ocean, snowboards and surf board, etc.
- Class imbalance in training data cause bias in captioning



```
The actual caption is:
an older woman playing nintendo wii near other people

The predicted captions is:
a man about a <unk> while hit a skateboard
```



```
The actual caption is:
a child flying a kite on the beach.

The predicted captions is:
a man about a kite on the air
```

- Bad BLEU score does not mean bad caption
  - “a baby giraffe eating leaves on a meadow
  - “A giraffe that is eating some leaves off of a tree”.
  - BLEU score of $1.39e{-}231$

Thank you

- **Topic**
  - Generation of Captions from Images
- **Name**
  - Tianyuan Cai
- **Two minute (short) video**:
  - https://youtu.be/_IFIf4Gn4i0
- **Reference Links:**
  - Data:
    - Coco
      - http://cocodataset.org
    - Glove
      - https://nlp.stanford.edu/projects/glove/