Image Captioning System

General Assembly DSI-10 2019 Capstone
Kelvin Kong

Problem Statement

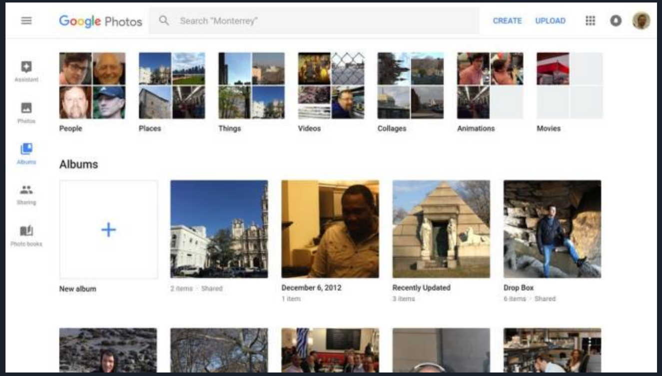
With an onslaught of photos on the web today:

- 350 Million photos uploaded daily on FB
- 1.2 billion photos uploaded to Google Photos daily

It is challenge to search across all these photos because most photos are not tagged with keywords.

Without specific keywords being associated to the photos, the photos are not searchable.

If millions of photos are not easily searchable, they are essentially useless because no person can manually analyze that amount of photos to locate a single picture of interest



Data Science Problem

Build a model that could automatically describe any image that is given to it.

The model is to:

- Take in an image of any size and dimension as input
- Output a human readable sentence that describes what is in the image

A baseball player is swinging a bat at a baseball game .

A group of cows grazing in a field .

Data Science Problem

- The generated descriptions for the photos can now be:
    - Stored in a database with a reference to the original photo
    - Indexed in search engines such as ElasticSearch or Apache Solr
- When this is run against a large number of photos that have accumulated and untouched over many years, it makes all the photos searchable.
- There is a big competitive advantage for companies that could rapidly search and analyze vast amounts of photos/videos in a short time. Eg. Google and Instagram are already implementing these tools

Visualizing the Training Data


a fire hydrant with graffiti next to some flowers


a large pizza is shown before it is cut


a group of people having a meal together .

Common Objects in Context (Microsoft COCO dataset):
- large-scale object detection, segmentation, and captioning dataset.

Preprocessing

For Images:

- Resized
- Randomly cropped
- Randomly Flipped
- Normalized

For Captions:

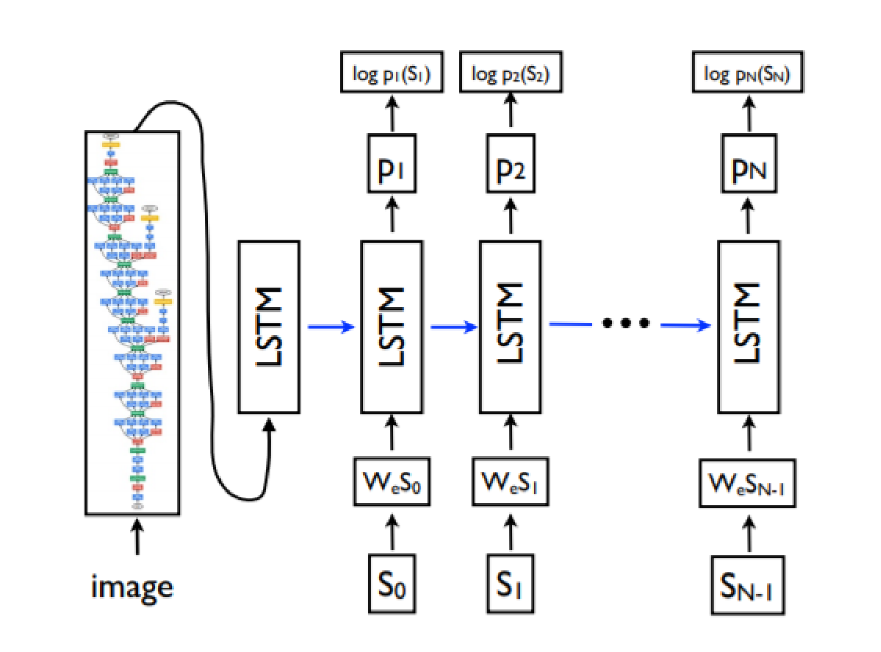
- <Start> and <End> tokens added to captions
- Words below threshold count are converted to <Unknown> token
- All Captions must have the same length per batch. The sampler is programmed to only obtain sentences with the same length when sampling every batch
- Goes into the usual NLP pipeline. Tokenized, Vectorized, Word Embeddings.

Modelling

An end to end Neural Network model which consist of:

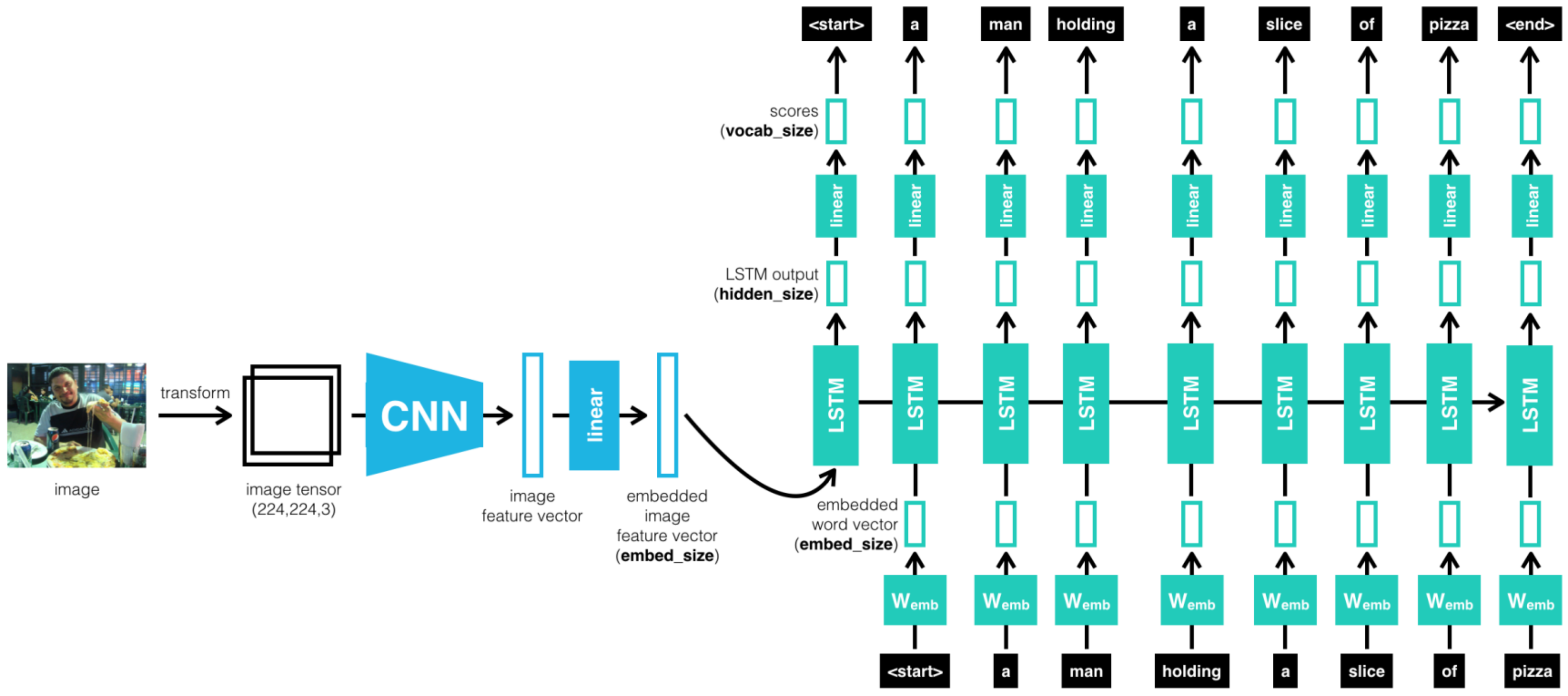
A CNN encoder which translates the image into a fixed length vector representation that is passed in as the initial step for the RNN

A RNN 'decoder' which generates the target sentence, one word at a time. LSTM is used in this model.



Ref: [A Neural Image Caption Generator]

Architecture



Training and Evaluation

Training is performed on:

- Tesla P100 GPUs, 16GB
- 6 Epochs for about ~6 hours due to large dataset (19GB) and RNNs

Evaluation:

- How to evaluate the performance of a generated sentence is still an ongoing research as of now
- The best evaluation metric so far is BLEU, originally developed for machine translation tasks
- It has some major drawbacks, especially when applied to tasks that it was never intended to evaluate.

The Bilingual Evaluation Understudy

A string-matching algorithm that provides basic quality metrics for Machine Translation
BLEU Downsides includes:

- Only measures direct word-to-word similarity and similar word clusters in two sentences
- There are no consideration of paraphrases or synonyms
    - "wander" doesn't get partial credit for "stroll," nor does "sofa" for "couch."

N-grams parameter can be set while evaluating BLEU score. Typically 1-4 grams are considered

Example:

Actual: "A large airplane is taking off from the runway."

Generated: "A large airplane is flying through the air on a sunny day ."

Score:

'BLEU-1 Score: 0.38'
'BLEU-2 Score: 0.31'
'BLEU-3 Score: 0.26'
'BLEU-4 Score: 0.20'

Inference - Good predictions on Val Set

A man is skiing down a snowy hill .

Inference - Good predictions

A bus is parked on the side of a road .

Inference - Good predictions

A man is standing on a street with a red and white bus .

Inference - Good predictions

A group of elephants standing in a field .

Inference - Good predictions

A computer desk with a laptop , keyboard , and a keyboard .

Inference - Good predictions

A large jet flying through the air on a cloudy day .

A large plane flying in the sky with a sky background .

Inference - Good predictions

A bear is standing in a field with a tree .

A cat laying on a desk with a laptop computer .

Inference - Good Predictions

A man is standing in front of a large pizza .

A group of people standing around a table with a cake and a glass of wine .

Inference - Average predictions

A bathroom with a toilet , sink , and a mirror .

A bathroom with a toilet , sink , and a mirror .

Predictions Not Related to image

A train traveling down train tracks next to a building .

Predictions Not Related to image

A train traveling down a train track next to a building .

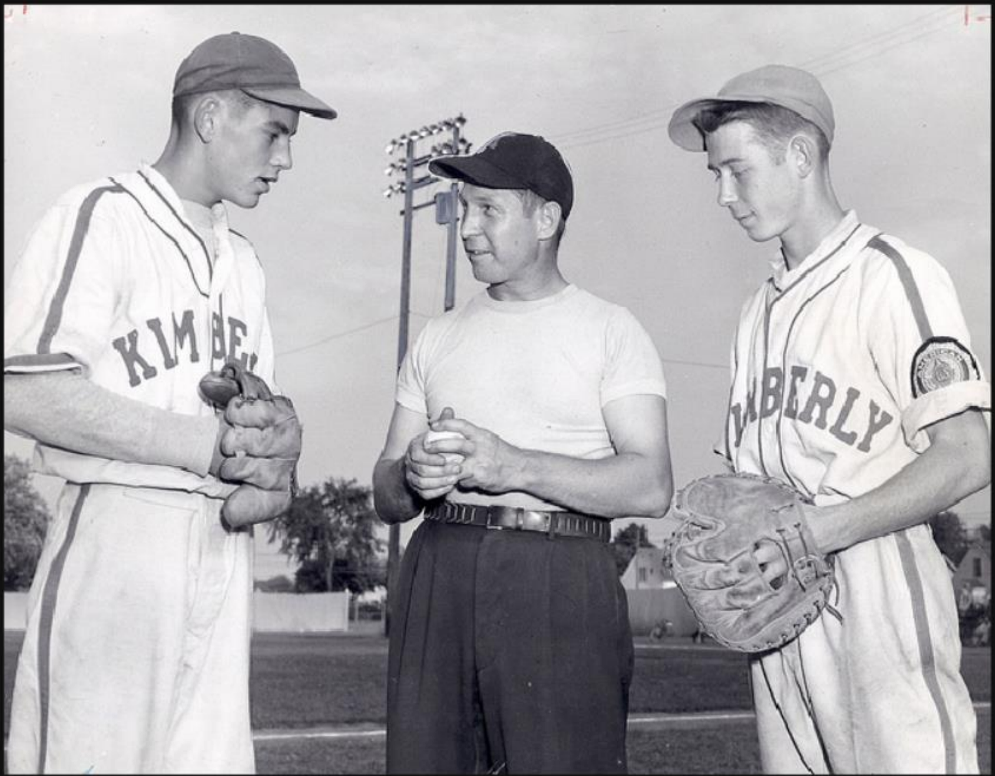
Predictions Not Related to image

A giraffe standing in a field with a tree in the background .

Predictions Not Related to image

A man is riding a horse on a beach .

Predictions Not Related to image



A group of people standing around a table with a large amount of bananas .

Test Images

A man is holding a slice of pizza on a table .

Test Images

A man is holding a glass of wine .

Test Images

A man in a suit and tie is standing in a room .

Test Images

A group of people standing around a table with a laptop .

Test Images

A man in a suit and tie is standing in front of a building .

Test Images

A man is standing in front of a building with a clock on it .

Test Images

A man is standing in front of a large building .

Conclusion & Improvements

- The model is able to train and works well on images that contains:
  - Planes, cats, giraffes, laptops, tennis, surfing, skiing, buses, trains, pizza, food, toilet, kitchen, Clocks
- Room for improvements:
  - Gender neutral or provide more examples of different gender
  - Improve the training captions to be more comprehensive
- Model Architecture Improvements
  - Beam Search
  - Experimenting with adding more layers after CNN
  - Try different CNN model architectures (Resnet-50 used here)
  - Adding Attention

Thank you DSI-10 for the great time!

A man is holding a cell phone in his hand .