Capstone Project Proposal

1. Objective

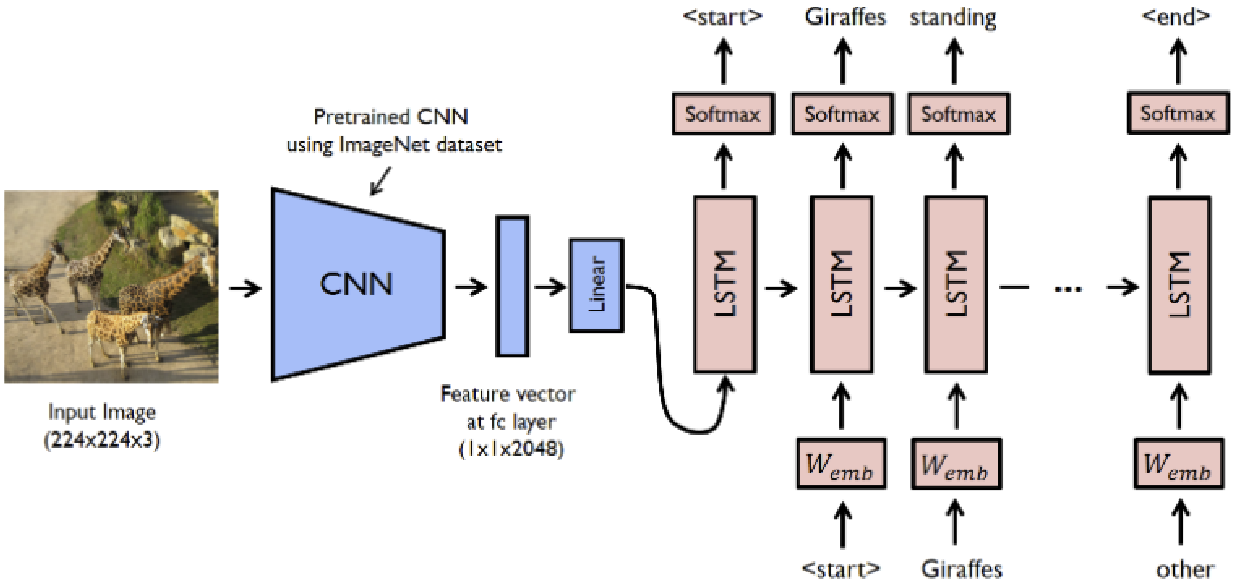
Image captioning

2. Dataset

Flickr 30k image dataset: 30,000+ images with 5 captions per image
https://www.kaggle.com/hsankesara/flickr-image-dataset

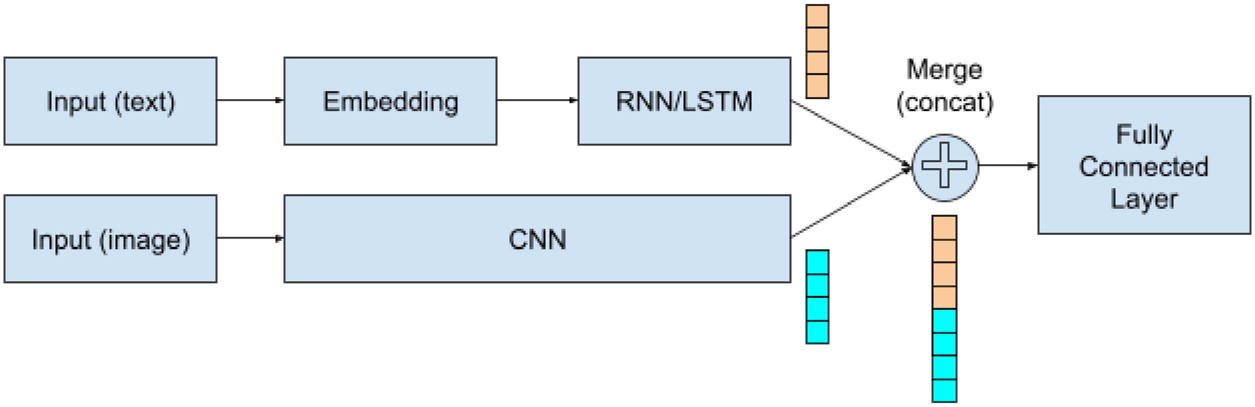
3. Learning Models

a. Encoder-Decoder Model

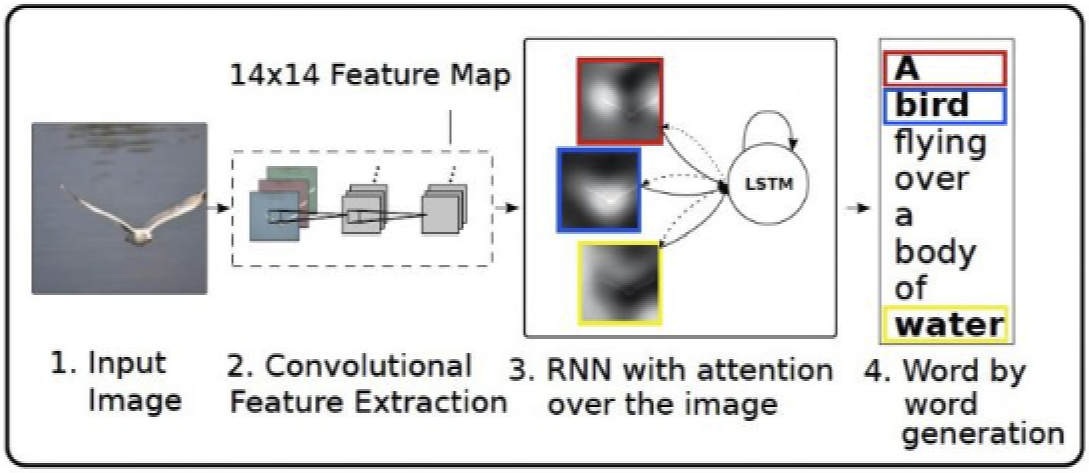


* CNN layer can be implemented from scratch or leverage pre-trained models such as InceptionV3, Resnet, VGG, EfficientNet etc.
* Word2Vec, GloVe, FastText can be used for word embedding.

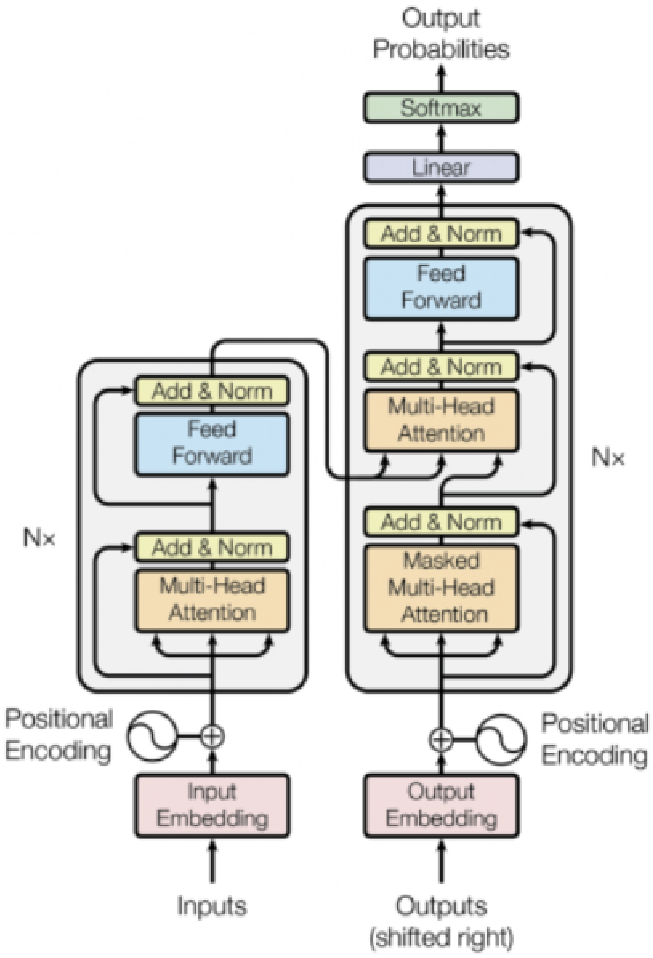
b. RNN(LSTM) and CNN Merged Model



c. Encoder-Decoder Model with Attention Mechanism


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

d. Attention Mechanism on Transformers



4. Prediction Methods

- Greedy search
- Beam search

5. Performance Evaluation

- Categorical cross entropy error
- BLEU(Bilingual Evaluation Understudy)

- ROGUE(Recall-Oriented Understudy for Gisting Evaluation)
- METEOR(Metric for Evaluation of Translation with Explicit ORdering, https://en.wikipedia.org/wiki/METEOR)