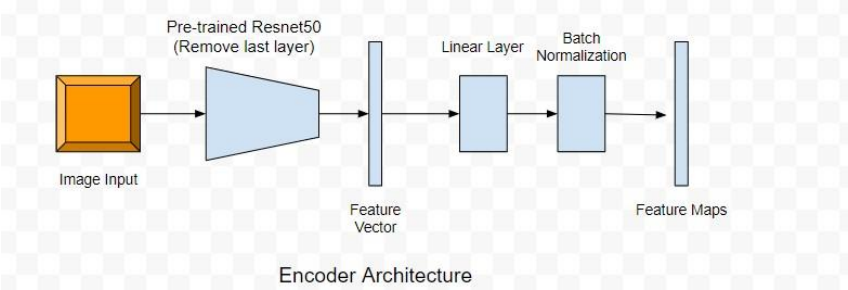
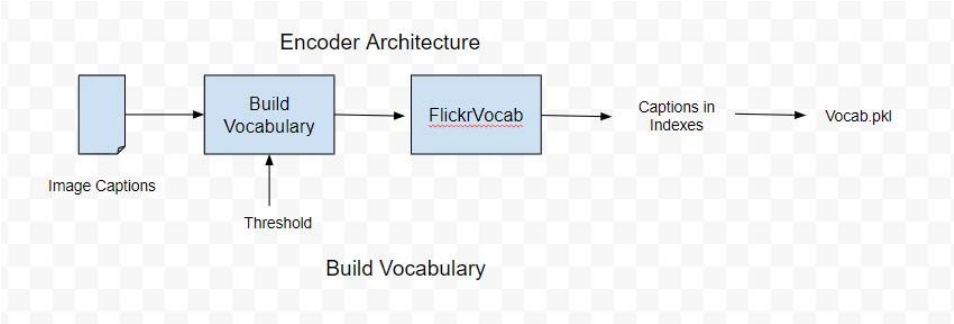
Encoder Decoder Model

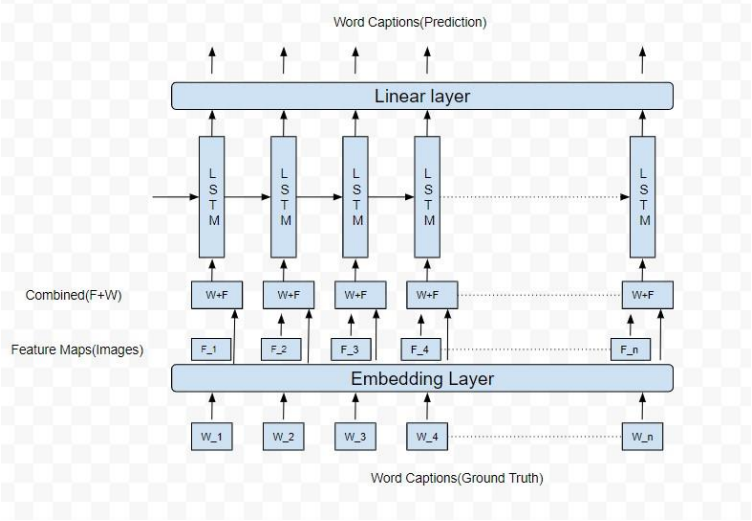
This Project is about image captioning which is combination of deep learning and Natural language processing. The goal is to caption input image providing description of what that input image is.

In this architecture, we use a Par-Inject approach where the LSTM/RNN uses two inputs at in every time step 1) a word and 2) an image.

At every step, we merge the word vector and the image vector into a similar-sized embedding space and pass it to be trained by the RNNs. the image feature vectors are generated by Resnet structure. In the Inject architectures, the hidden state of the RNN is affected by the image vector.







Flickr Dataset:

In this Project we use a Flickr 8k dataset where each image has multiple captions. All the ~8,000 images have together ~40,000 captions. We apply natural language processing techniques to preprocess the dataset into JSON files, tokenize the image captions to build vocabulary and utilize the tokens of each image captions to generate word2Vec embedding.

Dataset Preprocess:

- The Flickr 8k Image dataset is resized to 256 * 256 resolution.
- The Image captions in the file is parsed and vocabulary built based on that. The words with frequency less than 4 threshold are removed from the vocabulary. FlickrVocab.py
- Vocabulary generated is stored in vocab.pkl file.
- The Dataset is prepared using the Flickr8K dataset and corresponding tokenized captions. The tokenized captions are index based on the vocabulary generated.
- The Images are collated with tokenized word captions and split into Training, Testing and validation sets batchwise using FlickerDataLoader Class. The datasets of training, testing and validation are in 80%, 10% and10% ratio.

Encoder Model:

We use an encoder-decoder framework to achieve this task. An image encoder model is a convolution neural network which learns feature maps from the image. The decoder is a long short-term memory (LSTM) network. The image encoder is a convolutional neural network (CNN). To avoid training time and necessary effort to train the model with various data distribution samples, we use resnet-50 model pretrained on image classification dataset.

In this architecture, we delete the last layer of the resnet50 model and take the fully connected layer, which is before the last layer, pass it to fully connected linear layer and apply batch normalization on top of it.

The output of the model is feature maps of training images which are used in the decoder model to perform Image captioning.

Decoder Model:

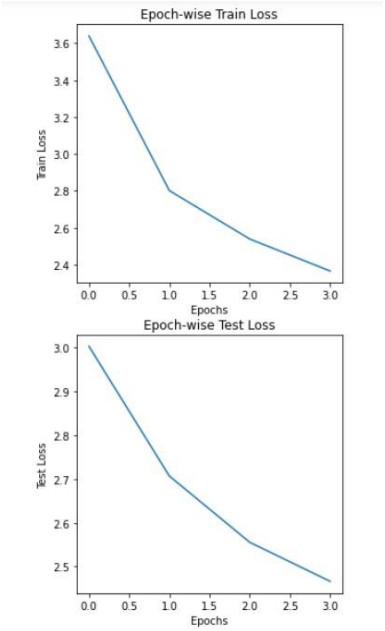
As we have seen in the Data Pre-preprocess step, The Image captions are pre-processed into tokenized word index vectors. The input to the Decoder model is the output of the Encoder model which is the feature maps and the tokenized captions.

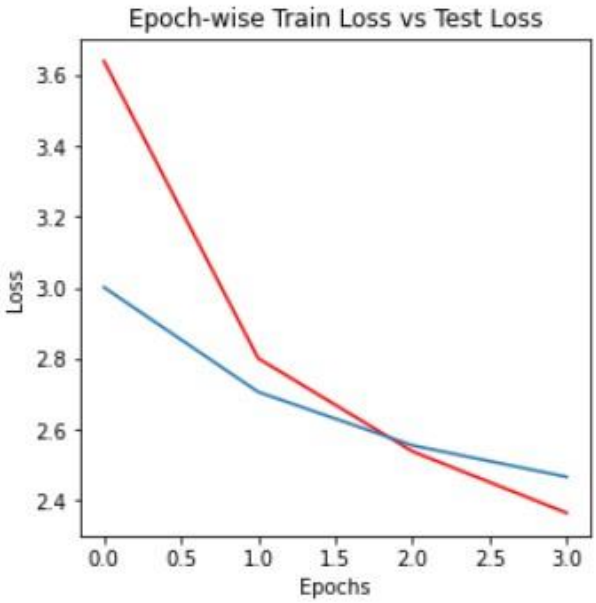
The Tokenized captions are further preprocessed to meet the design of the LSTM. We apply the pack padded sequence, which pads or packs data depending on the length of the input sequence. The Decoder model consists of the embedding layer to generate the embedding of the captions. The combination of the embedding and the features maps of the encoder model are passed to the LSTM model to generate the image captions. The output of the LSTM is sent to the linear layer to generate the image captions.

The Loss function used in the model is Cross Entropy loss and Adam Optimizer is used.

Evaluation:

Training Loss is calculated for each epoch. The training loss is plotted with respect to epochs. As the number of epochs increases the training loss decreases. There is overfitting noticed as the epochs are increasing.






```
Final Train Loss:  2.364564514914049 Final
Test Loss:  2.466470919549465
```

In each epoch, train dataset is processed. The training accuracy is calculated for each epoch. The training accuracy is calculated using the top-k approach.

```
Train Accuracy batch wise Top-K: 74.32584269662921
Validation Accuracy batch wise Top-K: 72.23131478450627
```

Sample Example Plot:

In this section, we plot 10 examples with their ground truth and predicted captions.

Predicted: <start> a man in a black shirt and jeans is standing in front of a large building . <end> .
Ground Truth: <start> there is a boy with a helmet and a stripe down the side of his pants riding on a skateboard . <end> <pad> <pad>

Predicted: <start> a woman in a black coat and a black hat is sitting on a bench . <end> . <end>
Ground Truth: <start> a woman in a headscarf and a boy wearing blue look to the right while sitting against a wall . <end> <pad> <pad> <pad>

Predicted: <start> a little girl in a pink shirt is jumping on a swing . <end> . <end> <end> . <end>
Ground Truth: <start> one boy wearing a red shirt riding on a swing while another boy wearing a blue shirt pushing him . <end> <pad> <pad> <pad>

Predicted: <start> a dog is jumping over a fence . <end> <end> . <end> <end> . <end> <end> . <end> <end>
Ground Truth: <start> a rodeo rider gets tossed up into the air by a black bull as fellow cowboys look on . <end> <pad> <pad> <pad> <pad>

Predicted: <start> a group of people are sitting on a bench . <end> . <end> <end> . <end> <end> . <end>
Ground Truth: <start> two men in black leather jackets , glasses , and hats , open them up for <unk> . <end> <pad> <pad> <pad> <pad> <pad>

Predicted: <start> a man in a black shirt and a hat is sitting on a bench . <end> . <end> <end>
Ground Truth: <start> small girl in a teal dress on the arm of a chair reaching for a floor lamp . <end> <pad> <pad> <pad> <pad> <pad>

Predicted: <start> a little girl in a pink bathing suit is jumping into the water . <end> <end> . <end> <end>
Ground Truth: <start> a child in rolled up jeans , a striped shirt , and a helmet frolics in sprinklers . <end> <pad> <pad> <pad> <pad> <pad>

Predicted: <start> a man in a black shirt and a black hat is standing in front of a crowd . <end>
Ground Truth: <start> the person in the black shirt <unk> a stunt in the street while the crowd watched . <end> <pad> <pad> <pad> <pad> <pad> <pad>

We aim to compute the **Bleu score** of the predicted caption of 10 images with associated multiple ground truth captions of the image.

image: 0 3374054694_fa56f29267.jpg
Predicted: <start> a black and white dog is jumping over a fence . <end> . <end> <end> . <end> <end> .
Groundtruth: a brown dog is jumping up at a woman in a black coat .
BLEU SCORE: 1.1200407237786664e-231
Groundtruth: a dog jumps for the woman 's treat .
BLEU SCORE: 1.2508498911928379e-231
Groundtruth: a large black and white dog jumps up to get something held by a woman wearing a black jacket . BLEU SCORE: 1.0244914152188952e-231
Groundtruth: a woman teaches a dog to jump for treats , outside on the grass .
BLEU SCORE: 1.1008876702055895e-231
Groundtruth: a woman with a vest and red shirt is holding her hand up above a black and white dog that is jumping . BLEU SCORE: 9.893133360884868e-232
image: 1 2322334640_d4d22619ff.jpg
Predicted: <start> a man in a black shirt and jeans is standing in front of a large building . <end> .
Groundtruth: a boy rides his ripstik on the street .
BLEU SCORE: 1.2508498911928379e-231
Groundtruth: a boy with a helmet skateboards across the street .
BLEU SCORE: 1.2183324802375697e-231
Groundtruth: a child in a blue jacket and helmet is on a skateboard . BLEU SCORE: 1.1409851298103347e-231
Groundtruth: a young boy wearing a helmet skateboards on the street .
BLEU SCORE: 1.1896457329133973e-231
Groundtruth: there is a boy with a helmet and a stripe down the side of his pants riding on a skateboard . BLEU SCORE: 1.012071042130996e-231 **image: 2 2635400219_2e1a984fd3.jpg**
Predicted: <start> a woman in a black coat and a black hat is sitting on a bench . <end> . <end>
Groundtruth: a middle eastern woman wearing a blue headscarf sits next to a boy in blue clothes .
BLEU SCORE: 1.066973992029681e-231
Groundtruth: a woman and a boy are both wearing blue traditional arab clothes
BLEU SCORE: 9.788429383461836e-232
Groundtruth: a woman and young man kneel beside a wall while dressed in blue themed clothing .
BLEU SCORE: 1.0832677820940877e-231

Groundtruth: a woman in a headscarf and a boy wearing blue look to the right while sitting against a wall . BLEU SCORE: 1.0244914152188952e-231
Groundtruth: a woman wearing a scarf on her head sits next to a young boy wearing blue .
BLEU SCORE: 1.0669733992029681e-231
image: 3 2512682478_b67cc525c7.jpg
Predicted: <start> a little girl in a pink shirt is jumping on a swing . <end> . <end> <end> . <end>
Groundtruth: a boy pushing another boy on the swing .
BLEU SCORE: 1.2508498911928379e-231
Groundtruth: a little boy in a blue shirt and blue jeans is pushing the swing that a little boy in a red shirt and blue shorts is sitting in .
BLEU SCORE: 9.336117803135294e-232
Groundtruth: one boy pushes another on a swing .
BLEU SCORE: 1.2882297539194154e-231
Groundtruth: one boy wearing a red shirt riding on a swing while another boy wearing a blue shirt pushing him .
BLEU SCORE: 1.0244914152188952e-231
Groundtruth: one young boy pushing another young boy on a swing .
BLEU SCORE: 1.1896457329133973e-231 **image: 4 3591462960_86045906bd.jpg**
Predicted: <start> a dog is jumping over a fence . <end> <end> . <end> <end> . <end> <end> . <end> <end>
Groundtruth: a man is roping a bull in a rodeo while others in cowboy hats watch .
BLEU SCORE: 1.0832677820940877e-231
Groundtruth: a man is up in the air with one hand on a bull .
BLEU SCORE: 1.1200407237786664e-231
Groundtruth: a rodeo rider gets tossed up into the air by a black bull as fellow cowboys look on .
BLEU SCORE: 1.0377133938315695e-231
Groundtruth: man getting thrown in the air while bull riding
BLEU SCORE: 0
Groundtruth: the man in black is flying off the full in front of a red fence .
BLEU SCORE: 1.0832677820940877e-231
image: 5 3143765063_a7761b16d3.jpg
Predicted: <start> a group of people are sitting on a bench . <end> . <end> <end> . <end> <end> . <end>
Groundtruth: a man in an elf hat holding a white umbrella is standing on the sidewalk with two other men .
BLEU SCORE: 1.0244914152188952e-231
Groundtruth: a man in a pointed red hat holds a white umbrella .
BLEU SCORE: 1.1640469867513693e-231
Groundtruth: a man is wearing a santa hat and holding a white umbrella while two men with orangish red boas look on in amusement .
BLEU SCORE: 9.788429383461836e-232
Groundtruth: people walk out of a building past two manequins .
BLEU SCORE: 1.2183324802375697e-231
Groundtruth: two men in black leather jackets , glasses , and hats , open them up for walkers .
BLEU SCORE: 8.844844403089352e-232 **image: 6 1410193619_13fff6c875.jpg**
Predicted: <start> a man in a black shirt and a hat is sitting on a bench . <end> . <end> <end>
Groundtruth: a baby stands on the side of a couch and knocks over a lamp .
BLEU SCORE: 1.1008876702055895e-231

Groundtruth: a small child climbs onto the arm of a red couch .
BLEU SCORE: 1.1640469867513693e-231
Groundtruth: baby indoors , climbing on red couch arm , reaching for lamp .
BLEU SCORE: 9.594503055152632e-232
Groundtruth: small girl in a teal dress on the arm of a chair reaching for a floor lamp .
BLEU SCORE: 1.0518351895246305e-231
Groundtruth: woman sitting on a red couch inside the house .
BLEU SCORE: 1.2183324802375697e-231 **image: 7**
1813266419_08bf66fe98.jpg
Predicted: <start> a little girl in a pink bathing suit is jumping into the water . <end> <end> . <end> <end>
Groundtruth: a child in a striped shirt gleefully plays among some water fountains .
BLEU SCORE: 1.1409851298103347e-231
Groundtruth: a child in rolled up jeans , a striped shirt , and a helmet frolics in sprinklers .
BLEU SCORE: 1.0518351895246305e-231
Groundtruth: a girl playing in a fountain
BLEU SCORE: 1.1640469867513693e-231
Groundtruth: a girls leaps as she runs through sprinklers .
BLEU SCORE: 1.2508498911928379e-231
Groundtruth: person with striped shirt is playing in the sprinklers .
BLEU SCORE: 1.0244914152188952e-231 **image: 8**
3359551687_68f2f0212a.jpg
Predicted: <start> a man in a black shirt and a black hat is standing in front of a crowd . <end>
Groundtruth: a girl does a cartwheel in the street while people watch from the sidewalk .
BLEU SCORE: 1.1008876702055895e-231
Groundtruth: a girl doing a cartwheel in front of a crowd
BLEU SCORE: 1.0244914152188952e-231
Groundtruth: a woman is doing a backwards flip in front of a crowd of other people sitting on the sidewalk .
BLEU SCORE: 1.0244914152188952e-231
Groundtruth: someone in shorts is somersaulting in front of a crowd .
BLEU SCORE: 1.1896457329133973e-231
Groundtruth: the person in the black shirt performed a stunt in the street while the crowd watched .
BLEU SCORE: 1.0669733992029681e-231
image: 9 701816897_221bbe761a.jpg
Predicted: <start> a girl in a pink bathing suit is jumping into a pool . <end> <end> . <end> <end> .
Groundtruth: a woman watching two kids playing at a pool .
BLEU SCORE: 1.2183324802375697e-231
Groundtruth: a young girl jumping into a pool while another stands on the side with an innertube .
BLEU SCORE: 1.0669733992029681e-231
Groundtruth: a young girl jumps feet first into a swimming pool while another little girl and woman watch .
BLEU SCORE: 1.0518351895246305e-231
Groundtruth: two little girls play at their family 's backyard swimming pool while a woman looks on .
BLEU SCORE: 1.0669733992029681e-231
Groundtruth: two little girls play by a hotel pool .
BLEU SCORE: 1.2508498911928379e-231