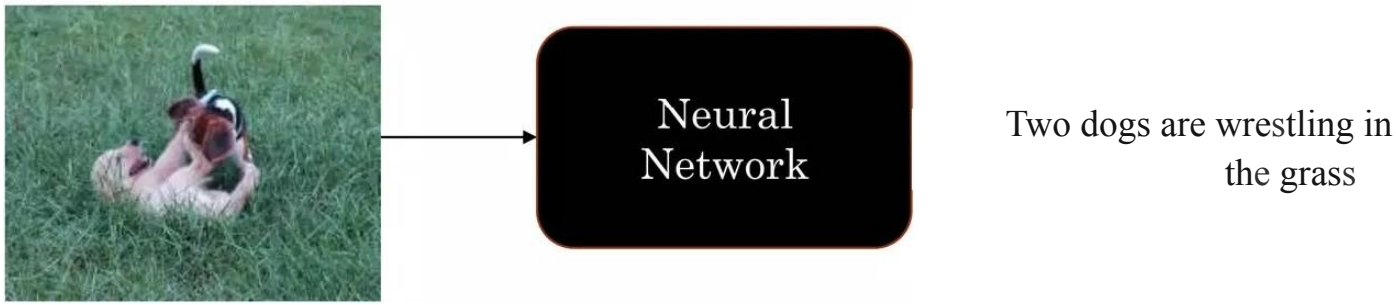
Image captioning using CNN and LSTM

DESCRIPTION

- Goal is to generate a descriptive sentence of an image
- Project was inspired by the works of Andrej Karpathy and Marc Tanti et al.(2017)



- Potential Applications:
  - Aiding visually impaired
  - Generating video summary using individual frames

DATASET

SUGGESTED CAPTIONS

- A man riding his bike on a hill
- A man with helmet and backpack standing on dirt bike in a hilly grassy area
- A person rides a motorbike through a grassy field
- Man on motorcycle riding in dry field wearing a helmet and backpack
- The biker is riding through a grassy plain .

- We used Flickr8K dataset for this project
- Flickr8K dataset contains a variety of images depicting scenes and situations
- The dataset consists of 8000 images and each image has 5 corresponding descriptions
- The images are of different dimensions

TEXT PRE-PROCESSING

- Each description is tokenized and converted to lowercase
- Removed alphanumeric characters and punctuation marks
- We use startseq and endseq as prefix and postfix for each caption respectively
- Filtered out unique words from the corpus and represented each word by an integer
- To generate a fixed length word vector we calculated the maximum length caption

IMAGE FEATURE GENERATION

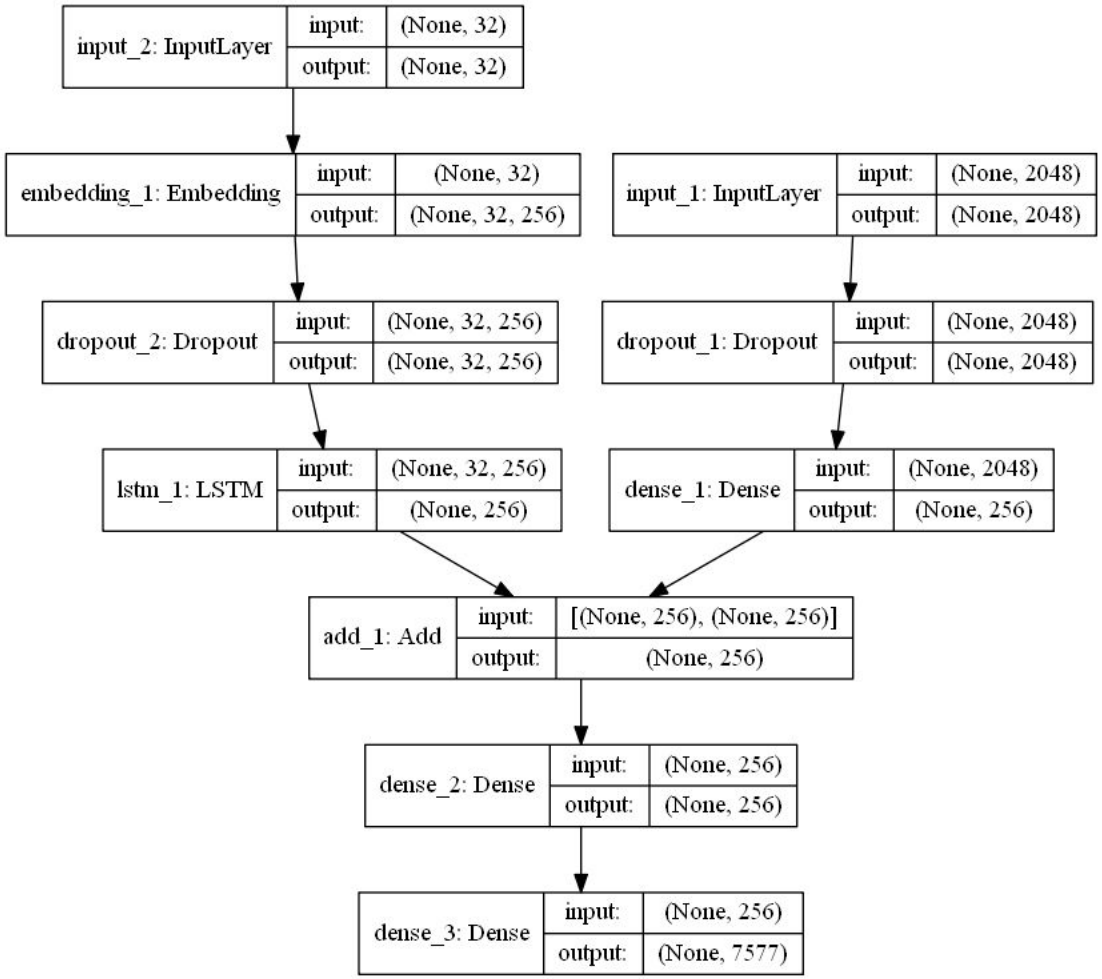
- Resized all images to a fixed size of 299x299x3 using OpenCV
- Employed transfer learning using pre-trained Xception CNN model to encode images
- We removed the last softmax layer from the Xception network to extract 2048 image vector

SEQUENTIAL CAPTION INJECTION

- For each image we will train the model by temporally injecting incremental sequences of the description
- In this phase, we essentially create labels in our training data

Image	Partial Caption	Target Word
Image	startseq	a
Image	startseq a	young
Image	startseq a young	boy
Image	startseq a young boy wearing a helmet and riding a bike in a park	endseq

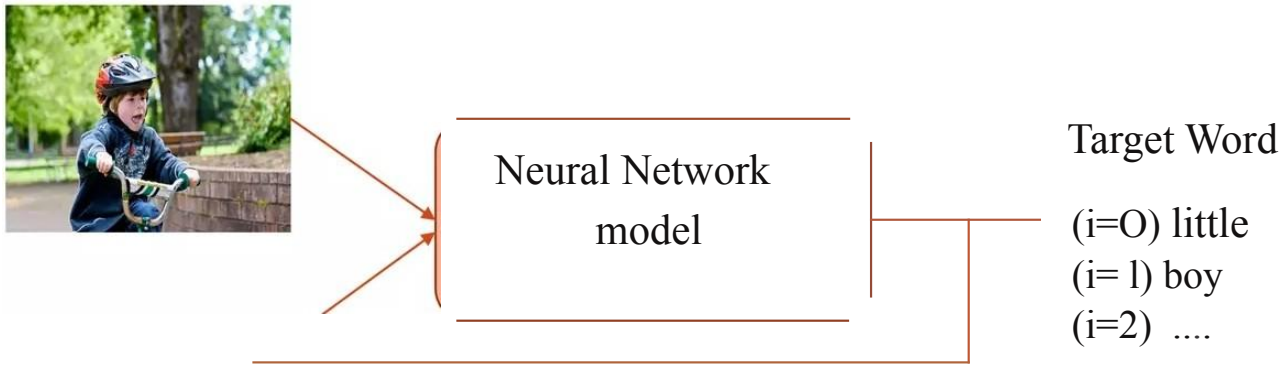
ARCHITECTURE



- We used an encoder-decoder architecture
- 2048 image vector is fed to a Dense layer to generate 256 length image vector
- 32 length word vector is fed to LSTM/RNN to output 256 length word vector
- Decoder model adds both the encoder outputs and is fed to Dense 256 layer
- The last Dense layer will have as many nodes as in the vocabulary
- The last softmax layer predicts the next word present in the output vocabulary

GENERATING PREDICTIONS

- Caption is predicted word by Word
- Image is fed along with the first word(startseq) to the RNN to predict the second word
- Again the same image along with first word + second word is fed to the RNN to predict the third word and so on until the last word(endseq) is encountered



EVALUATION METRICS

Bilingual Evaluation Understudy Score (BLEU)

- BLEU is a metric for evaluating a generated sentence to a reference sentence
- BLEU score lies between 0 and 1

LSTM (Long Short Term Memory)

BLEU N-GRAM	SCORE
BLEU-1	0.572214
BLEU-2	0.339204
BLEU-3	0.237129
BLEU-4	0.116733

Simple RNN (Recurrent Neural Network)

BLEU N-GRAM	SCORE
BLEU-1	0.364472
BLEU-2	0.181942
BLEU-3	0.103185
BLEU-4	0.085675

RESULTS

Correct Predictions

Actual Caption:
a boy with a blue helmet is riding a bike

Predicted Caption:
little boy rides bike with helmet

Actual Caption:
white fluffy dog running in the dirt

Predicted Caption:
white dog runs across the sand

Actual Caption:
a boy dribbles a basketball in the gymnasium

Predicted Caption:
boy in white shirt is playing basketball

RESULTS

Funny Predictions ??

Actual Caption:
man fly fishing in a small river with steam in the background
Predicted Caption:
Man is swinging on a swing

Actual Caption:
a woman wearing a black and white outfit ·while holding her sunglasses
Predicted Caption:
man in pink dress is holding her head

Actual Caption:
a group of different people are walking in all different directions in a city
Predicted Caption: group of people walking ocean

RESULTS

Predictions that went really wrong!

Actual Caption:
A man wearing a red life jacket is holding a purple rope while waterskiing

Predicted Caption:
man in white and white and white shorts leash on swing

Actual Caption:
A dog is chewing on a metal pole

Predicted Caption:
dog is standing in its mouth

Actual Caption:
a young hockey player playing in the ice rink

Predicted Caption:
chasing player in motorcycle is playing chasing

FUTURE WORK

- We can enhance the predictions by using more training examples. For example using Flickr32k dataset which has 32000 images

- Implement visual attention techniques, which focuses on interesting parts of the image

- Creating an application for visually impaired to convert the generated caption into voice output