Image Captioning using Flickr8k Dataset

Group 6

Hemanth K., Madhuri Y., Sharmin K.

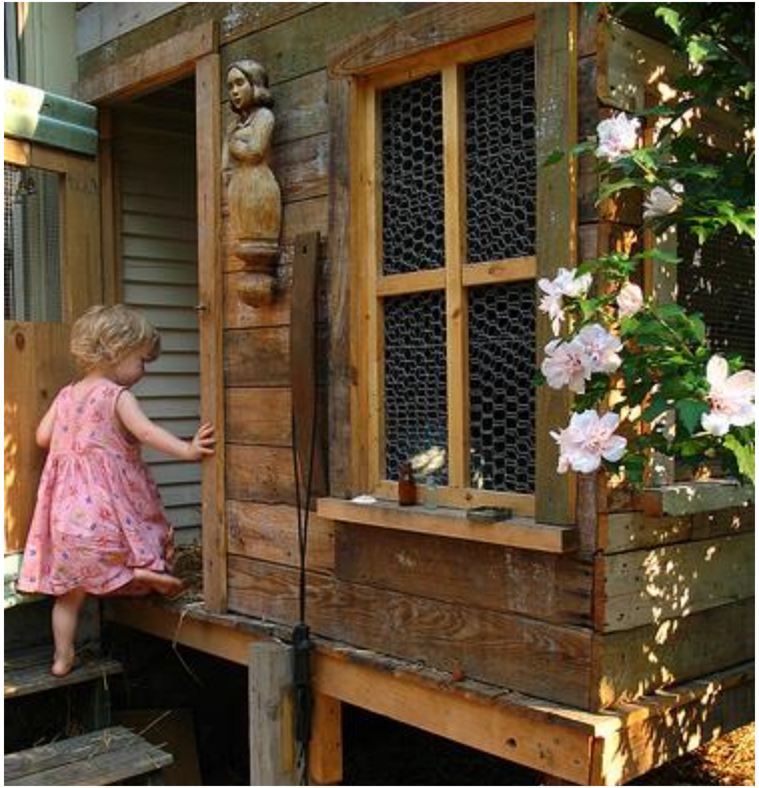
Introduction

Image Captioning is a combination of AI + Deep Learning + Natural Language Processing.

Applications include:

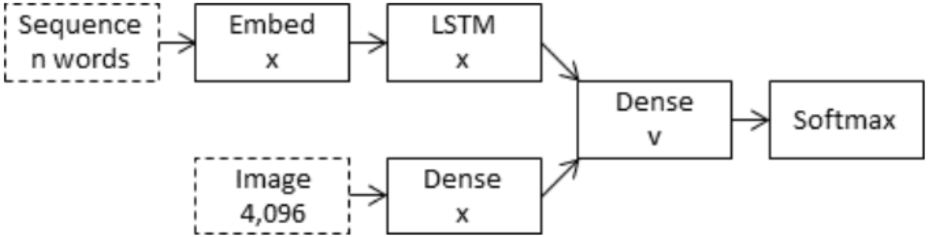
- ❏ Image Search in Search Engines
- ❏ Image Segregation and Classification
- ❏ Text-to-Speech to aid visually impaired individuals
- ❏ Automatic Image annotation in Facial Recognition, E-commerce, etc.

The Flickr8k dataset is taken from Kaggle and has 8092 images, each having 5 captions. It has an image folder and a text file containing captions.



```
image,caption
1000268201_693b08cb0e.jpg,A child in a pink dress is climbing up a set of stairs in an entry way .
1000268201_693b08cb0e.jpg,A girl going into a wooden building .
1000268201_693b08cb0e.jpg,A little girl climbing into a wooden playhouse .
1000268201_693b08cb0e.jpg,A little girl climbing the stairs to her playhouse .
1000268201_693b08cb0e.jpg,A little girl in a pink dress going into a wooden cabin .
```

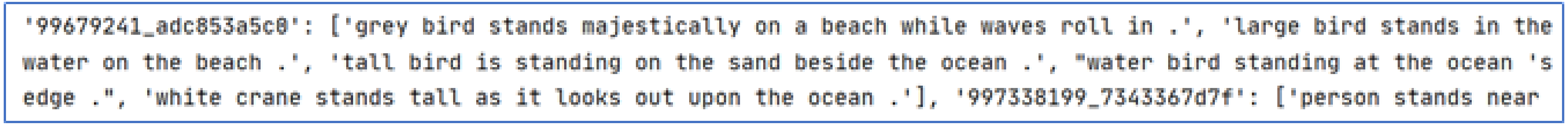
Project Overview



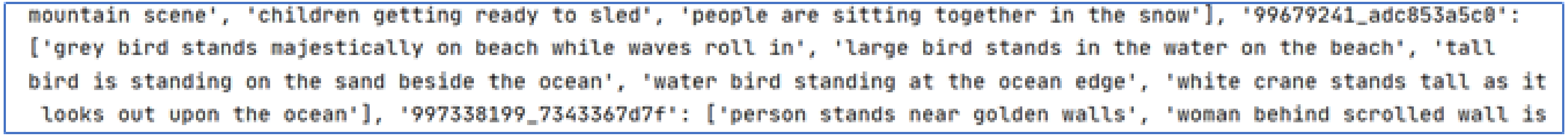
The project is divided into 5 main components:

1. Data Splitting – generate train and test ids of the images
2. Text Preprocessing – clean captions
3. Image Preprocessing – extract image features using pre-trained models (VGG16, InceptionV3, ResNet50)
4. LSTM Model – processes sequences
5. Model Evaluation – using BLEU score performance metric

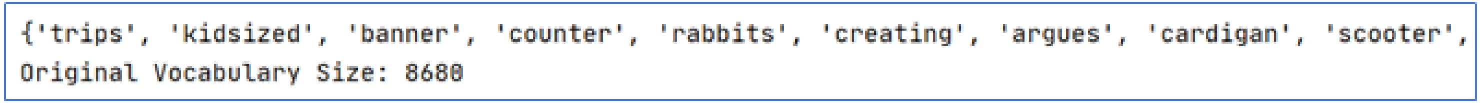
Text Pre-processing



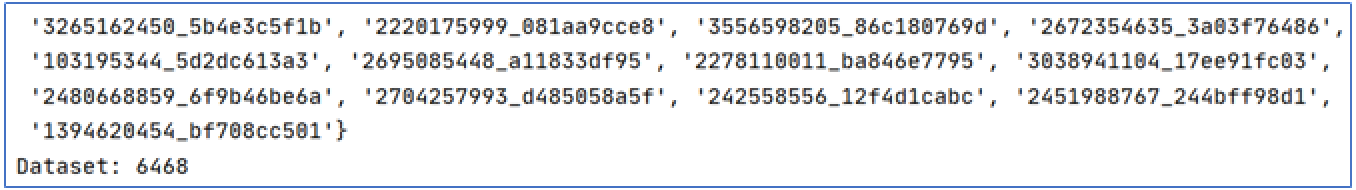
```
'99679241_adc853a5c0': ['grey bird stands majestically on a beach while waves roll in .', 'large bird stands in the
water on the beach .', 'tall bird is standing on the sand beside the ocean .', "water bird standing at the ocean 's
edge .", 'white crane stands tall as it looks out upon the ocean .'], '997338199_7343367d7f': ['person stands near
```



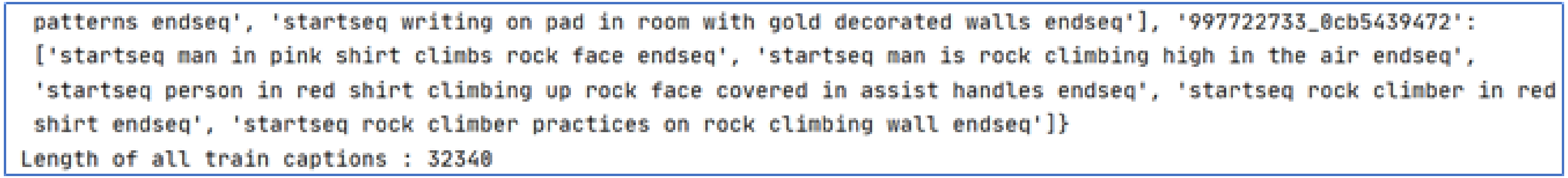
```
mountain scene', 'children getting ready to sled', 'people are sitting together in the snow'], '99679241_adc853a5c0':
['grey bird stands majestically on beach while waves roll in', 'large bird stands in the water on the beach', 'tall
bird is standing on the sand beside the ocean', 'water bird standing at the ocean edge', 'white crane stands tall as it
 looks out upon the ocean'], '997338199_7343367d7f': ['person stands near golden walls', 'woman behind scrolled wall is
```



```
{'trips', 'kidsized', 'banner', 'counter', 'rabbits', 'creating', 'argues', 'cardigan', 'scooter',
Original Vocabulary Size: 8680
```

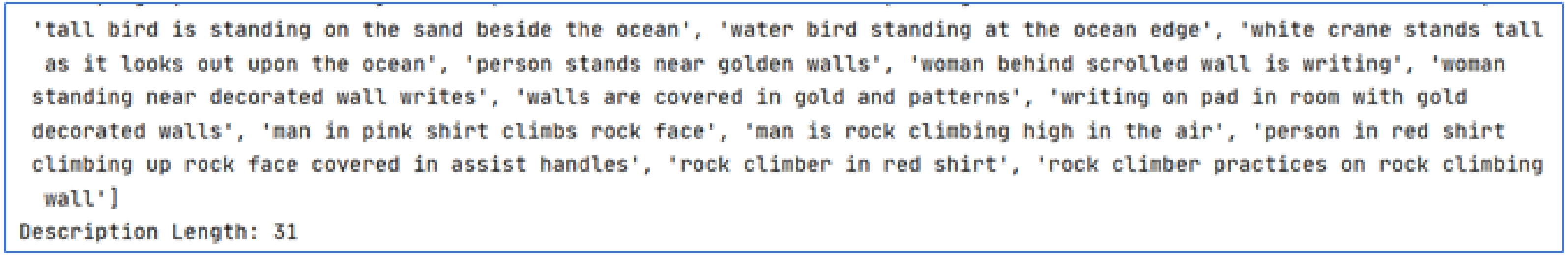


```
 '3265162450_5b4e3c5f1b', '2220175999_081aa9cce8', '3556598205_86c180769d', '2672354635_3a03f76486',
 '103195344_5d2dc613a3', '2695085448_a11833df95', '2278110011_ba846e7795', '3038941104_17ee91fc03',
 '2480668859_6f9b46be6a', '2704257993_d485058a5f', '242558556_12f4d1cabc', '2451988767_244bff98d1',
 '1394620454_bf708cc501'}
Dataset: 6468
```



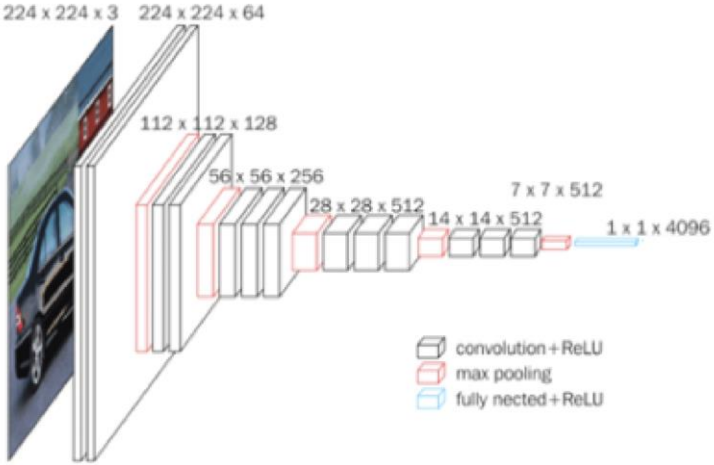
```
 patterns endseq', 'startseq writing on pad in room with gold decorated walls endseq'], '997722733_0cb5439472':
 ['startseq man in pink shirt climbs rock face endseq', 'startseq man is rock climbing high in the air endseq',
 'startseq person in red shirt climbing up rock face covered in assist handles endseq', 'startseq rock climber in red
 shirt endseq', 'startseq rock climber practices on rock climbing wall endseq']}
Length of all train captions : 32340
```

Input = Image_1 + ‘startseq’; Output = ‘the’



```
 'tall bird is standing on the sand beside the ocean', 'water bird standing at the ocean edge', 'white crane stands tall
  as it looks out upon the ocean', 'person stands near golden walls', 'woman behind scrolled wall is writing', 'woman
 standing near decorated wall writes', 'walls are covered in gold and patterns', 'writing on pad in room with gold
 decorated walls', 'man in pink shirt climbs rock face', 'man is rock climbing high in the air', 'person in red shirt
 climbing up rock face covered in assist handles', 'rock climber in red shirt', 'rock climber practices on rock climbing
  wall']
Description Length: 31
```

Image Pre-processing



VGG-16 CNN Model Architecture

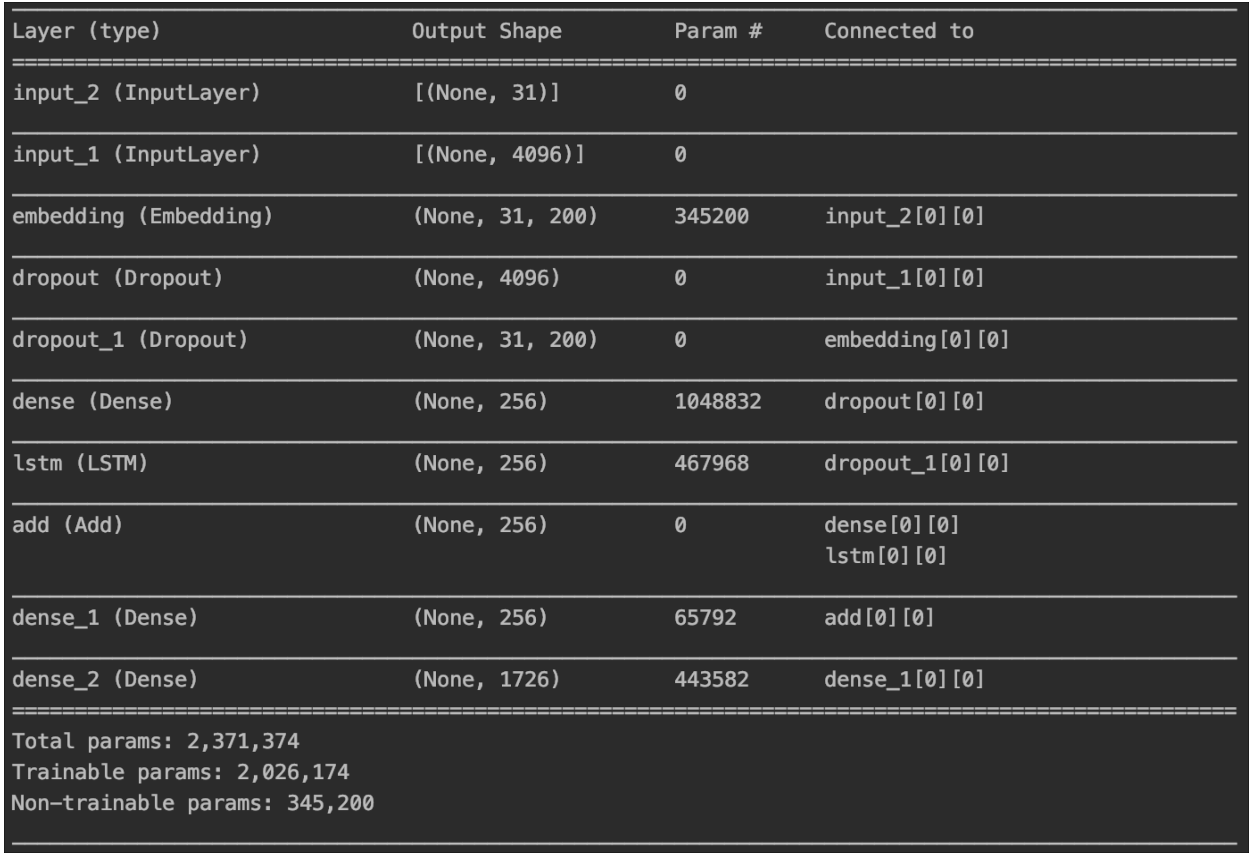
- Input Size: 224X224 RGB Images

- 16 Hidden Layers:
    - 13 Convolution Layers
    - 3 Fully Connected Dense Layers
- Parameters: 138,357,544

- Output Size: 4096 (Excluded last 2 dense layers)

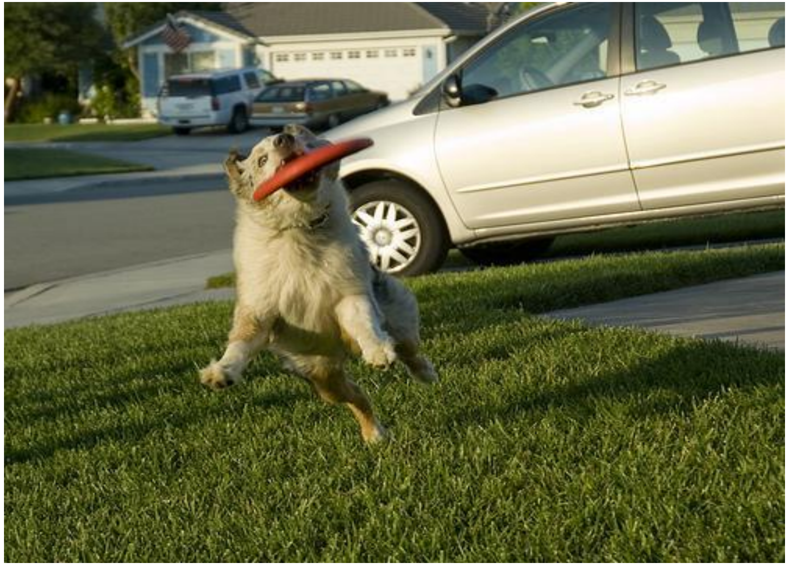
Image Courtesy:https://neurohive.io/en/popular-networks/vgg16/

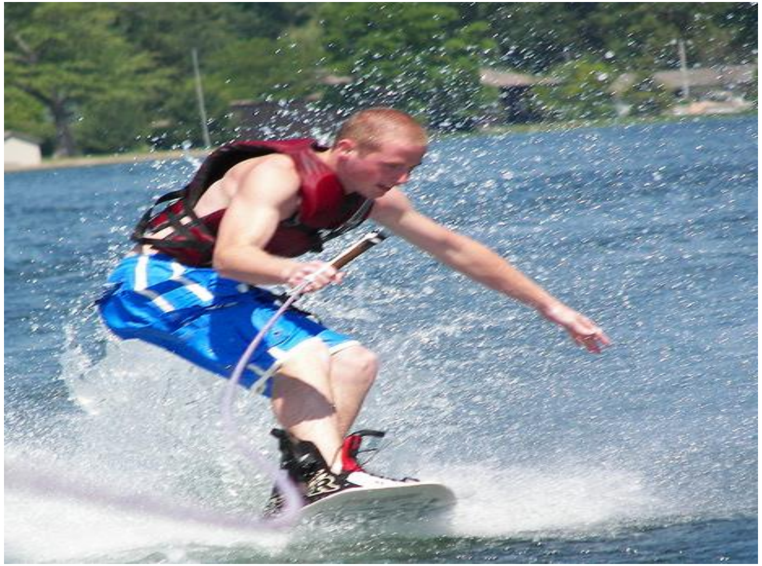
Modeling

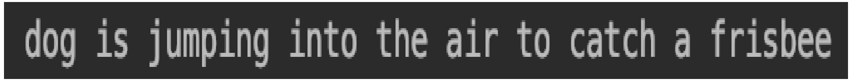


```
________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
================================================================================================
input_2 (InputLayer)            [(None, 31)]         0

________________________________________________________________________________________________
input_1 (InputLayer)            [(None, 4096)]       0

________________________________________________________________________________________________
embedding (Embedding)           (None, 31, 200)      345200      input_2[0][0]

________________________________________________________________________________________________
dropout (Dropout)               (None, 4096)         0           input_1[0][0]

________________________________________________________________________________________________
dropout_1 (Dropout)             (None, 31, 200)      0           embedding[0][0]

________________________________________________________________________________________________
dense (Dense)                   (None, 256)          1048832     dropout[0][0]

________________________________________________________________________________________________
lstm (LSTM)                     (None, 256)          467968      dropout_1[0][0]

________________________________________________________________________________________________
add (Add)                       (None, 256)          0           dense[0][0]
                                                                 lstm[0][0]

________________________________________________________________________________________________
dense_1 (Dense)                 (None, 256)          65792       add[0][0]

________________________________________________________________________________________________
dense_2 (Dense)                 (None, 1726)         443582      dense_1[0][0]
================================================================================================
Total params: 2,371,374
Trainable params: 2,026,174
Non-trainable params: 345,200

________________________________________________________________________________________________
```

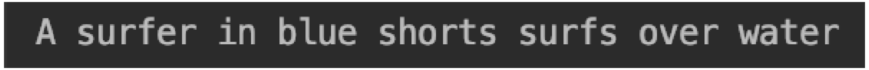
Results





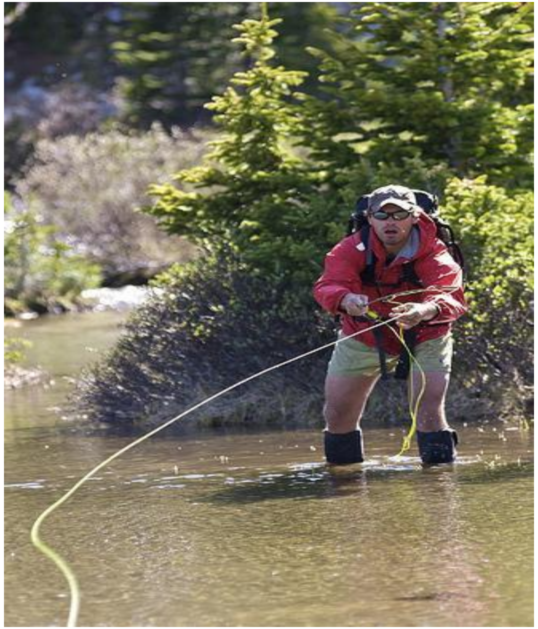


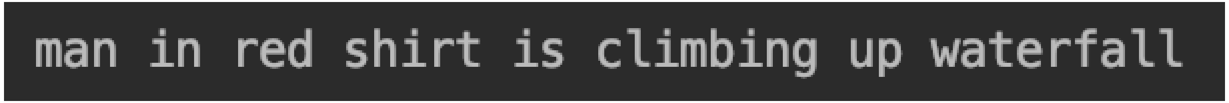
```
dog is jumping into the air to catch a frisbee
```



```
A surfer in blue shorts surfs over water
```

Results

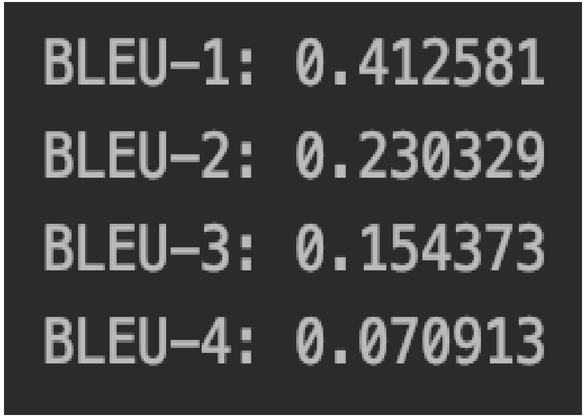




```
man in red shirt is climbing up waterfall
```

Evaluation

Bilingual Evaluation Understudy (BLEU) score is a metric for evaluating a generated sentence to a reference sentence.



```
BLEU-1: 0.412581
BLEU-2: 0.230329
BLEU-3: 0.154373
BLEU-4: 0.070913
```

Conclusion & Future Work

- Conclusion: The image vector extracted by our VGG-16 and partial caption vector extracted using LSTM, fed into the decoder with greedy search gives the best results and a BLEU-1 Score of 0.41.
- Challenges & Future Work:
    - Flickr 30K Dataset
    - Beam Search
    - Other Evaluation metrics such as METEOR, SPICE etc.