Surveillance Alert System

Image Captioning using Deep Learning Approaches

Team 4

MEET THE TEAM
TEAM 4

Parthiv Borgohain

Kavya Angara

Rudraksh Garg

Pratik Gawli

Saurabh Arora

How good are you at multitasking?

Problem Statement

Analyse activities from real-time surveillance videos frame and alert the user if suspicious activity is encountered.

Semantic segmentation helps us determine what objects are in the image but fails to explain the relationship between these objects using verbs or contextual information. Our model output can help us identify these events and be used to alert users to take action.

Data

Dataset Name:

- Flickr 8k Images
- Flickr 8k Text (contains captions)
- Custom images for surveillance training

Source:
https://www.kaggle.com/datasets/kunalgupta2616/flickr-8k-images-with-captions

Description:

- 8092 image files, Additional 568 surveillance images
- 5 captions for each image
- The text file is structured as a csv file and has columns of images and captions

Methodology

01 TEXT PREPROCESSING

02 IMAGE FEATURE EXTRACTION (CNN)

03 SEQUENTIAL CAPTION GENERATION (LSTM)

04 FINAL MODEL ARCHITECTURE

05 SENTIMENT ANALYSIS

06 RESULTS

Text Preprocessing

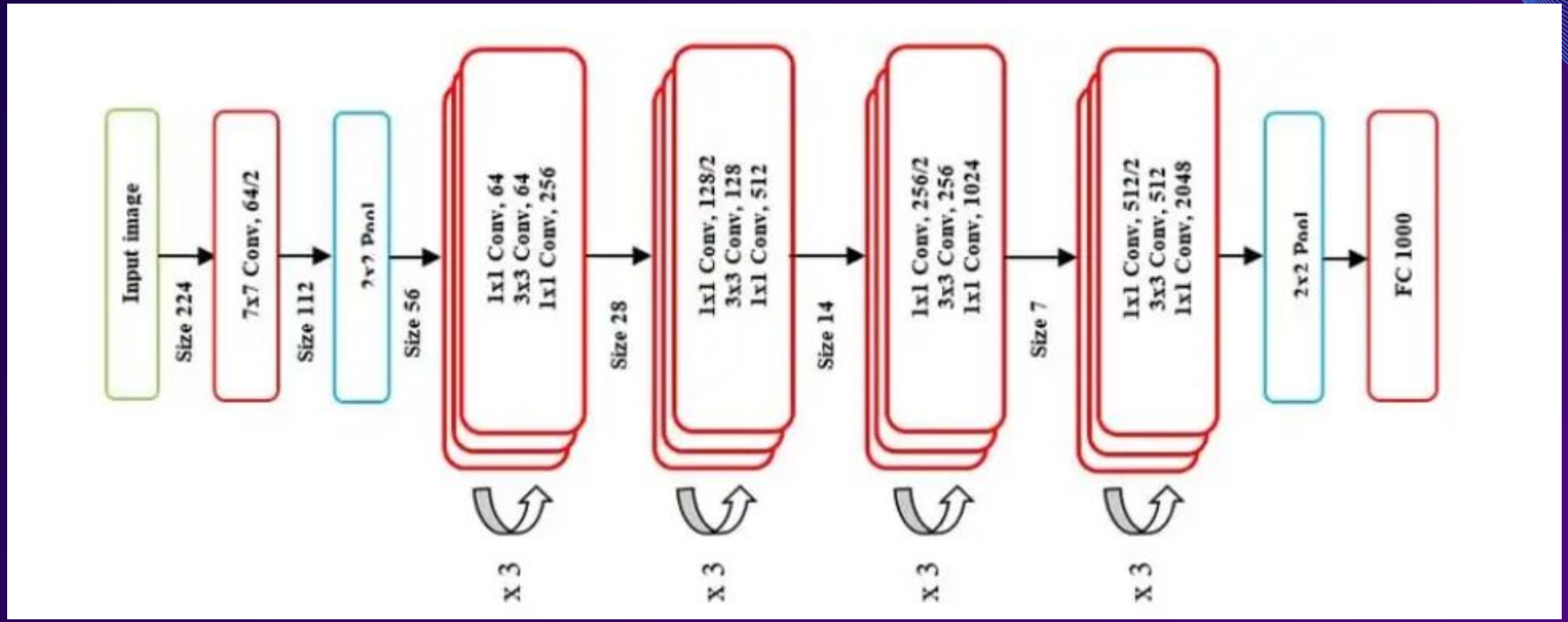
- Every caption has been tokenized and converted to lowercase letters
- Removed special characters and punctuation marks from the captions
- Removed captions that had only one word
- Added START, END, UNKNOWN and PAD tokens in our vocabulary of words required for LSTM training
- Created word embeddings for each caption by converting to a tensor of integers of a fixed length using dynamic padding

```
tensor([3269, 1094,  977, 2579, 1641,  389,  633, 2594, 3270,    0,    0,    0,
           0,    0,    0,    0,    0])
```

```
man leaning on ledge of a balcony
tensor([3269, 1138, 1698, 2579,  633, 3271, 1693,  440,  633,  838, 1420,  633,
         797, 3108,  633,  140, 3270])
```

```
woman sitting on a UNKNOWN bench with a pillow and a bag near a restaurant
```

Image feature extraction



Sequential Caption Injection

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

Image	Partial Caption	Target Word
Image	startseq	a
Image	startseq a	man
Image	startseq a man	in
…….	…….	…….
Image	startseq a man in ski mask holding a gun	endseq

Final Model Architecture

CNN (resnet)

1 X 2048 FEATURE VECTOR

Captions Dataset

Text Preprocessing

Word Embedding

<start> dog running in water <end>

[1,4,6,2,3,5,0,0,0]

<start>: 1
In: 2
Water: 3
Dog: 4
<end>: 5
<unknown>: 6
<pad>: 0

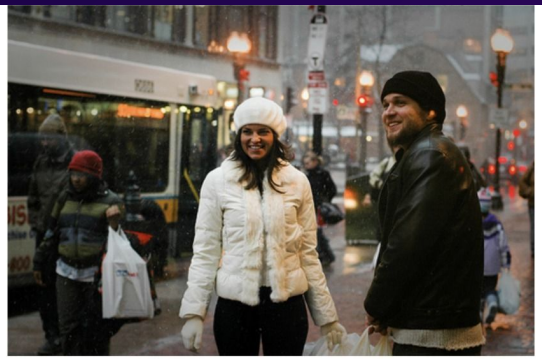
Dog <unknown> <in> .... <end>

Softmax Softmax Softmax Softmax

LSTM layer LSTM layer LSTM layer LSTM layer

Predictions

Actual and predicted captions for general images from Flickr8k dataset



a man in a black leather jacket is standing in a crowd
[['a couple standing outside a', 'a man and a woman standing on a city street', white are smiling', 'a woman and a man standing on a busy city street smiling', winter']]

a dog runs through the grass
[['a brown dog chasing after', 'a brown dog is gallopsing through the grass', brown dog runs fast outside on green grass', 'a german sheperd breed dog is r n']]

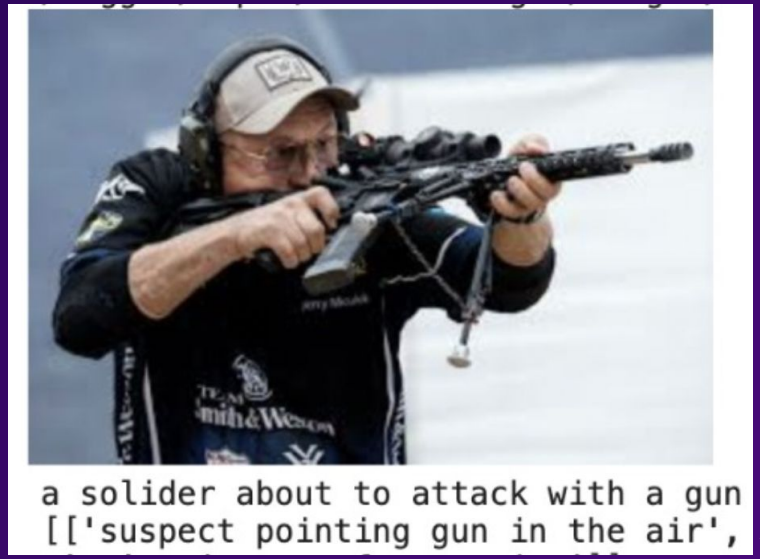
PREDICTED: a dog swimming in the water END
ACTUAL: a brown dog swimming in the water with a tennis ball in his mouth

PREDICTED: a dog is running through a grassy area END
ACTUAL: dogs on grass

Predictions

Actual and predicted captions for surveillance images from Custom dataset:

(Actual) - From the custom dataset



Predicted

Actual

RESULTS

Minimum training batch loss(cross entropy): 0.72

Minimum validation batch loss(cross entropy) : 5.51

Test set predictions improve as number of epochs completed increase

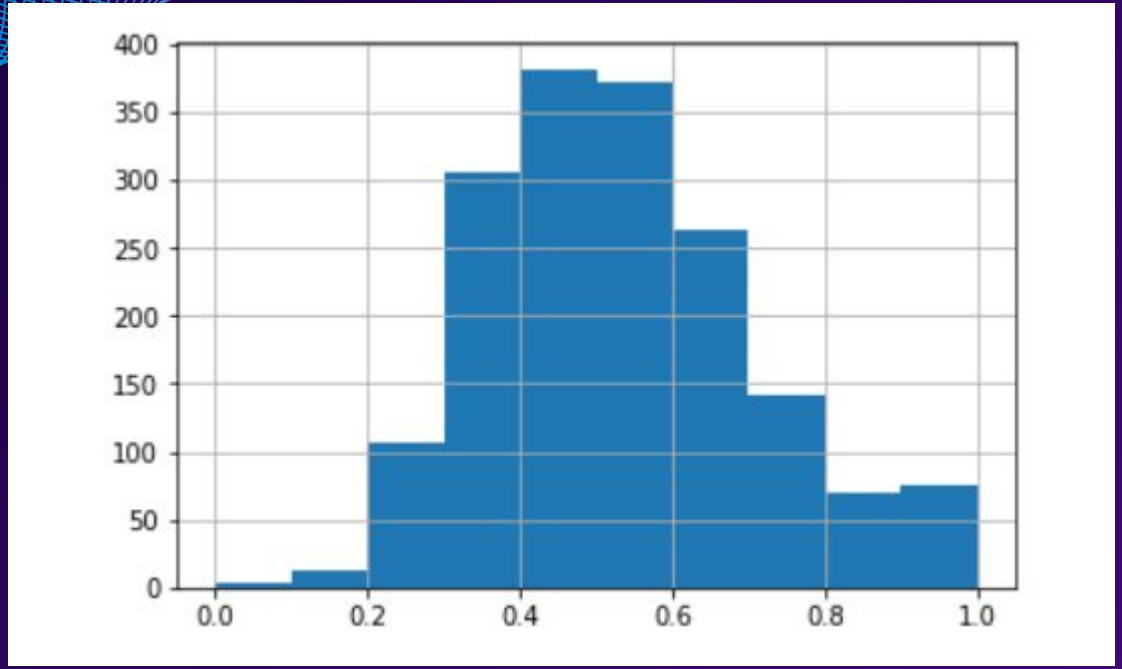
BLEU score a better metric to evaluate the predictions on the validation and test set

Explored other custom loss functions such as a differential BLEU loss function and mixed cross entropy to better model loss

Evaluation Metrics

- BLEU (Bilingual Evaluation Understudy) Score is a metric for evaluating a system generated sentence to a reference sentence.
- A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0
- The approach works by counting matching n-grams in the candidate translation to n-grams in the reference text, where 1-gram or unigram would be each token and a bigram comparison would be each word pair
- It is quick, computationally inexpensive, easy to understand and has been widely adopted

BLEU Score on holdout set



Majority of the captions generated had a BLEU score of 0.4 to 0.6

More images had a higher BLEU score than a lower BLEU score, implying our model is generating reasonable captions in our holdout set for majority of the images

BLEU score can be improved with more training and a richer dataset - the case with any Deep Learning model

Sentiment Analysis

Model output

“Man pointing gun <end>”

Alert Based on Sentiment Analysis using VADER

VIOLENT!

Thank you !