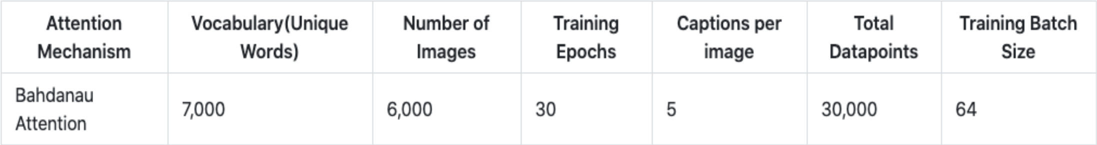
STUDY 1: IMAGE CAPTIONING USING ATTENTION MODEL

MODEL: ATTENTION MECHANISM + ResNet50 (Convolutional Network)

I. MODEL DESCRIPTION:



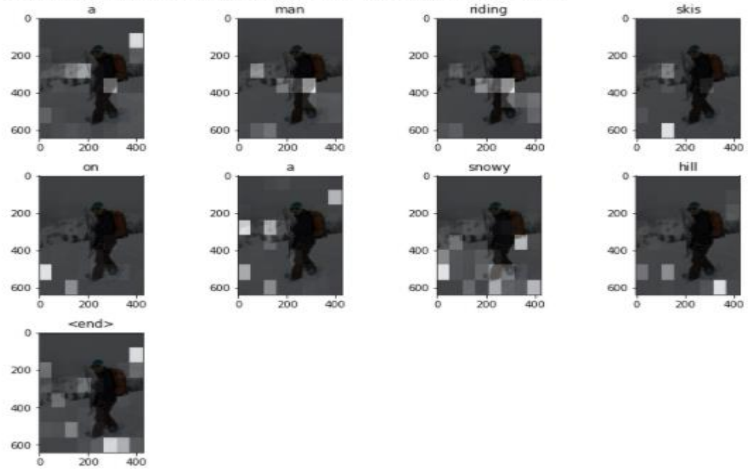
Attention Mechanism	Vocabulary(Unique Words)	Number of Images	Training Epochs	Captions per image	Total Datapoints	Training Batch Size
Bahdanau Attention	7,000	6,000	30	5	30,000	64

II. PREDICTION RESULTS:

In this study, we have analyzed the predicted results for 50 different images from the MS-COCO dataset and 3 other random images from the internet. We classify these image caption predictions as “Good”, “Fair” and “Bad”. Note that, the classifications into “Good”, “Fair” and “Bad” are based only on how close the predicted caption is to the real caption - based on human judgement. In the next section, we will analyze NLP metrics such as BLEU, GLEU and WER scores. We calculate the mean of these metrics from the 53 outcomes to make inferences about the model and the convolutional layer used. A few examples of “Good”, “Fair” and “Bad” are:

1) GOOD PREDICTION (example 1)

```
Real Caption: <start> a man in black jacket holding skis on a slope <end>
Prediction Caption: a man riding skis on a snowy hill <end>
```

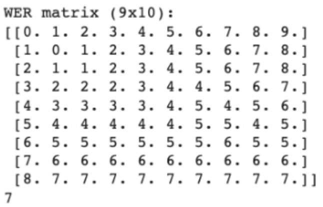


BLEU SCORE, GLEU SCORE and WER (Word Error Rate) metric comparison between sentences:

INDUVIDUAL N GRAM	
1 GRAM	0.111111
2 GRAM	1.000000
3 GRAM	1.000000
4 GRAM	1.000000

CUMMULATIVE N GRAM	
BLEU 1	0.111111
BLEU 2	0.333333
BLEU 3	0.484284
BLEU 4	0.577350

GLEU SCORE	
Sentence Level Frequency	
1 to 4 grams	0.2647
1 to 2 grams	0.4215

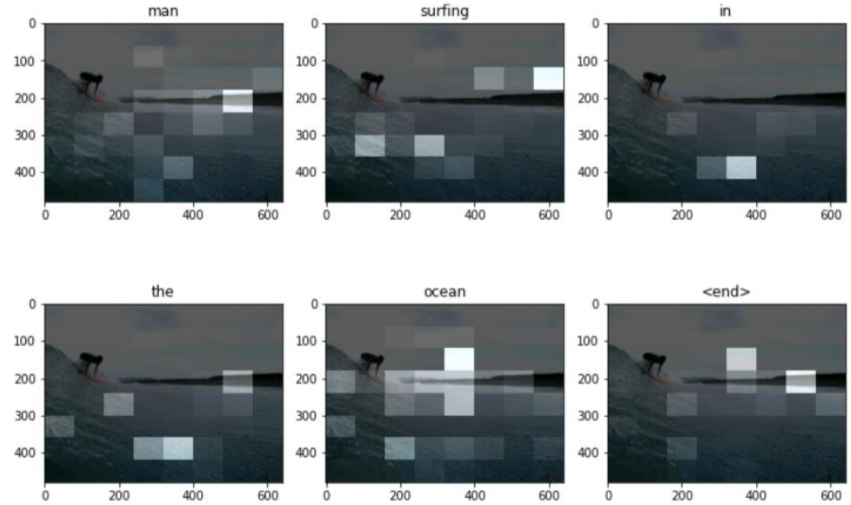


```
WER matrix (9x10):
[[0. 1. 2. 3. 4. 5. 6. 7. 8. 9.]
 [1. 0. 1. 2. 3. 4. 5. 6. 7. 8.]
 [2. 1. 1. 2. 3. 4. 5. 6. 7. 8.]
 [3. 2. 2. 2. 3. 4. 4. 5. 6. 7.]
 [4. 3. 3. 3. 3. 4. 5. 4. 5. 6.]
 [5. 4. 4. 4. 4. 4. 5. 5. 4. 5.]
 [6. 5. 5. 5. 5. 5. 5. 6. 5. 5.]
 [7. 6. 6. 6. 6. 6. 6. 6. 6. 6.]
 [8. 7. 7. 7. 7. 7. 7. 7. 7. 7.]]
7
```

[Next Page]

2) GOOD PREDICTION (example 2)

```
Real Caption: <start> the surfer is riding on a wave in the ocean <end>
Prediction Caption: man surfing in the ocean <end>
```

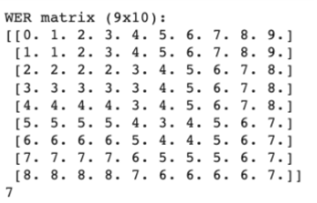


BLEU SCORE, GLEU SCORE and WER (Word Error Rate) metric comparison between sentences:

INDUVIDUAL N GRAM	
1 GRAM	0.111111
2 GRAM	1.000000
3 GRAM	1.000000
4 GRAM	1.000000

CUMMULATIVE N GRAM	
BLEU 1	0.111111
BLEU 2	0.333333
BLEU 3	0.484284
BLEU 4	0.577350

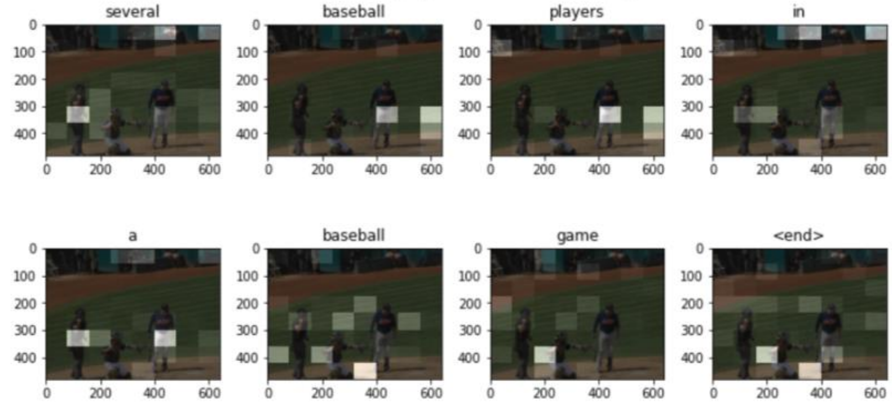
GLEU SCORE	
Sentence Level Frequency	
1 to 4 grams	0.1176
1 to 2 grams	0.2105



```
WER matrix (9x10):
[[0. 1. 2. 3. 4. 5. 6. 7. 8. 9.]
 [1. 1. 2. 3. 4. 5. 6. 7. 8. 9.]
 [2. 2. 2. 2. 3. 4. 5. 6. 7. 8.]
 [3. 3. 3. 3. 3. 4. 5. 6. 7. 8.]
 [4. 4. 4. 4. 3. 4. 5. 6. 7. 8.]
 [5. 5. 5. 5. 4. 3. 4. 5. 6. 7.]
 [6. 6. 6. 6. 5. 4. 4. 5. 6. 7.]
 [7. 7. 7. 7. 6. 5. 5. 5. 6. 7.]
 [8. 8. 8. 8. 7. 6. 6. 6. 6. 7.]]
7
```

3) GOOD PREDICTION (example 3)

```
Real Caption: <start> group of baseball players preparing next move in a game <end>
Prediction Caption: several baseball players in a baseball game <end>
```

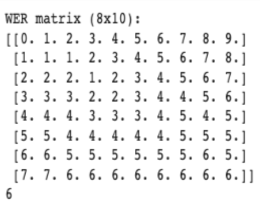


BLEU SCORE, GLEU SCORE and WER (Word Error Rate) metric comparison between sentences:

INDUVIDUAL N GRAM	
1 GRAM	0.125000
2 GRAM	1.000000
3 GRAM	1.000000
4 GRAM	1.000000

CUMMULATIVE N GRAM	
BLEU 1	0.125000
BLEU 2	0.353553
BLEU 3	0.503478
BLEU 4	0.594604

GLEU SCORE	
Sentence Level Frequency	
1 to 4 grams	0.2058
1 to 2 grams	0.3684

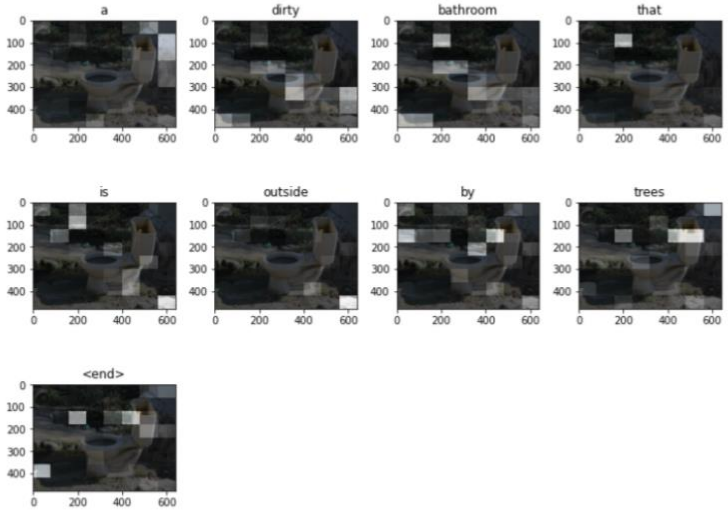


```
WER matrix (8x10):
[[0. 1. 2. 3. 4. 5. 6. 7. 8. 9.]
 [1. 1. 1. 2. 3. 4. 5. 6. 7. 8.]
 [2. 2. 2. 1. 2. 3. 4. 5. 6. 7.]
 [3. 3. 3. 2. 2. 3. 4. 4. 5. 6.]
 [4. 4. 4. 3. 3. 3. 4. 5. 4. 5.]
 [5. 5. 4. 4. 4. 4. 4. 5. 5. 5.]
 [6. 6. 5. 5. 5. 5. 5. 5. 6. 5.]
 [7. 7. 6. 6. 6. 6. 6. 6. 6. 6.]]
6
```

[Next Page]

4) FAIR PREDICTION (example 1)

```
Prediction Caption: a dirty bathroom that is outside by trees <end>
```



BLEU SCORE, GLEU SCORE and WER (Word Error Rate) metric comparison between sentences:

INDUVIDUAL N GRAM	
1 GRAM	0.111111
2 GRAM	1.000000
3 GRAM	1.000000
4 GRAM	1.000000

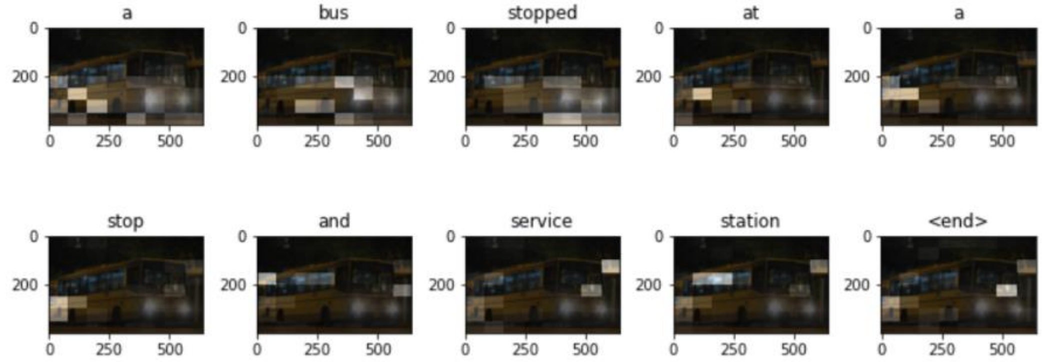
CUMMULATIVE N GRAM	
BLEU 1	0.111111
BLEU 2	0.333333
BLEU 3	0.484284
BLEU 4	0.577350

GLEU SCORE	
Sentence Level Frequency	
1 to 4 grams	0.0789
1 to 2 grams	0.1428

```
WER matrix (9x11):
[[ 0.  1.  2.  3.  4.  5.  6.  7.  8.  9. 10.]
 [ 1.  1.  2.  3.  4.  5.  6.  7.  8.  9. 10.]
 [ 2.  2.  2.  3.  4.  5.  6.  7.  8.  9. 10.]
 [ 3.  3.  3.  3.  4.  5.  6.  7.  8.  9. 10.]
 [ 4.  4.  3.  4.  4.  5.  6.  7.  8.  9. 10.]
 [ 5.  5.  4.  3.  4.  5.  6.  7.  8.  9. 10.]
 [ 6.  6.  5.  4.  4.  5.  6.  7.  8.  9. 10.]
 [ 7.  7.  6.  5.  5.  5.  6.  7.  8.  9. 10.]
 [ 8.  8.  7.  6.  6.  6.  6.  7.  8.  9. 10.]]
10
```

5) FAIR PREDICTION (example 2)

```
Real Caption: <start> a commuter bus parked at a stop at night time <end>
Prediction Caption: a bus stopped at a stop and service station <end>
```



BLEU SCORE, GLEU SCORE and WER (Word Error Rate) metric comparison between sentences:

INDUVIDUAL N GRAM	
1 GRAM	0.100000
2 GRAM	1.000000
3 GRAM	1.000000
4 GRAM	1.000000

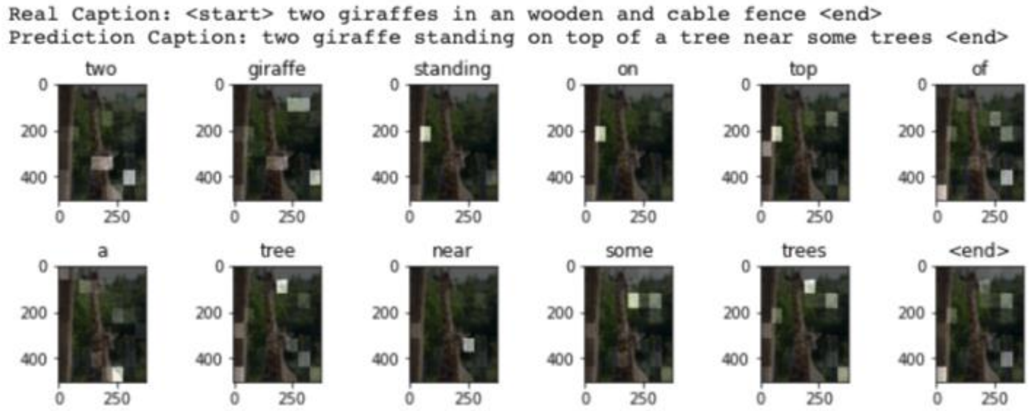
CUMMULATIVE N GRAM	
BLEU 1	0.100000
BLEU 2	0.316288
BLEU 3	0.467735
BLEU 4	0.562341

GLEU SCORE	
Sentence Level Frequency	
1 to 4 grams	0.2352
1 to 2 grams	0.3684

```
WER matrix (10x10):
[[0. 1. 2. 3. 4. 5. 6. 7. 8. 9.]
 [1. 1. 1. 2. 3. 4. 5. 6. 7. 8.]
 [2. 2. 2. 2. 3. 4. 5. 6. 7. 8.]
 [3. 3. 3. 3. 2. 3. 4. 5. 6. 7.]
 [4. 4. 4. 4. 3. 2. 3. 4. 5. 6.]
 [5. 5. 5. 5. 4. 3. 2. 3. 4. 5.]
 [6. 6. 6. 6. 5. 4. 3. 3. 4. 5.]
 [7. 7. 7. 7. 6. 5. 4. 4. 4. 5.]
 [8. 8. 8. 8. 7. 6. 5. 5. 5. 5.]
 [9. 9. 9. 9. 8. 7. 6. 6. 6. 6.]]
6
```

[Next Page]

6) FAIR PREDICTION (example 3)



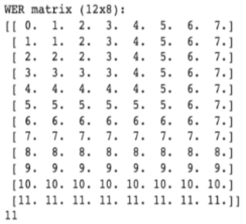
```
Real Caption: <start> two giraffes in an wooden and cable fence <end>
Prediction Caption: two giraffe standing on top of a tree near some trees <end>
```

BLEU SCORE, GLEU SCORE and WER (Word Error Rate) metric comparison between sentences:

INDUVIDUAL N GRAM	
1 GRAM	0.083333
2 GRAM	1.000000
3 GRAM	1.000000
4 GRAM	1.000000

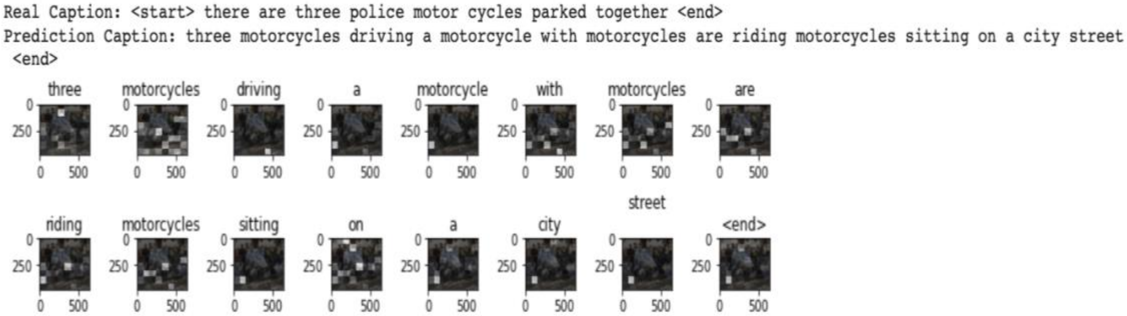
CUMMULATIVE N GRAM	
BLEU 1	0.083333
BLEU 2	0.288675
BLEU 3	0.440423
BLEU 4	0.537285

GLEU SCORE	
Sentence Level Frequency	
1 to 4 grams	0.0238
1 to 2 grams	0.0434



```
WER matrix (12x8):
[[ 0.  1.  2.  3.  4.  5.  6.  7.]
 [ 1.  1.  2.  3.  4.  5.  6.  7.]
 [ 2.  2.  2.  3.  4.  5.  6.  7.]
 [ 3.  3.  3.  3.  4.  5.  6.  7.]
 [ 4.  4.  4.  4.  4.  5.  6.  7.]
 [ 5.  5.  5.  5.  5.  5.  6.  7.]
 [ 6.  6.  6.  6.  6.  6.  6.  7.]
 [ 7.  7.  7.  7.  7.  7.  7.  7.]
 [ 8.  8.  8.  8.  8.  8.  8.  8.]
 [ 9.  9.  9.  9.  9.  9.  9.  9.]
 [10. 10. 10. 10. 10. 10. 10. 10.]
 [11. 11. 11. 11. 11. 11. 11. 11.]]
11
```

7) BAD PREDICTION (example 1)



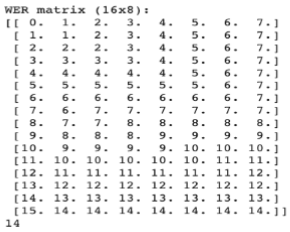
```
Real Caption: <start> there are three police motor cycles parked together <end>
Prediction Caption: three motorcycles driving a motorcycle with motorcycles are riding motorcycles sitting on a city street
 <end>
```

BLEU SCORE, GLEU SCORE and WER (Word Error Rate) metric comparison between sentences:

INDUVIDUAL N GRAM	
1 GRAM	0.062500
2 GRAM	1.000000
3 GRAM	1.000000
4 GRAM	1.000000

CUMMULATIVE N GRAM	
BLEU 1	0.062500
BLEU 2	0.250000
BLEU 3	0.400535
BLEU 4	0.500000

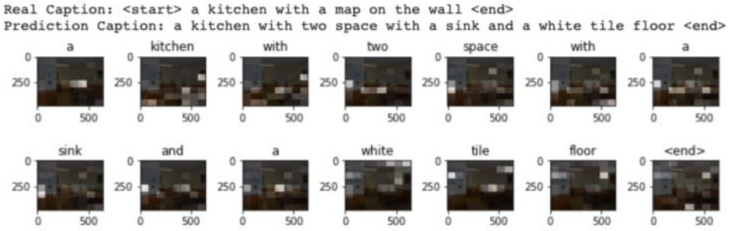
GLEU SCORE	
Sentence Level Frequency	
1 to 4 grams	0.0344
1 to 2 grams	0.0645



```
WER matrix (16x8):
[[ 0.  1.  2.  3.  4.  5.  6.  7.]
 [ 1.  1.  2.  3.  4.  5.  6.  7.]
 [ 2.  2.  2.  3.  4.  5.  6.  7.]
 [ 3.  3.  3.  3.  4.  5.  6.  7.]
 [ 4.  4.  4.  4.  4.  5.  6.  7.]
 [ 5.  5.  5.  5.  5.  5.  6.  7.]
 [ 6.  6.  6.  6.  6.  6.  6.  7.]
 [ 7.  6.  7.  7.  7.  7.  7.  7.]
 [ 8.  7.  7.  8.  8.  8.  8.  8.]
 [ 9.  8.  8.  8.  9.  9.  9.  9.]
 [10.  9.  9.  9.  9. 10. 10. 10.]
 [11. 10. 10. 10. 10. 10. 11. 11.]
 [12. 11. 11. 11. 11. 11. 11. 12.]
 [13. 12. 12. 12. 12. 12. 12. 12.]
 [14. 13. 13. 13. 13. 13. 13. 13.]
 [15. 14. 14. 14. 14. 14. 14. 14.]]
14
```

[Next Page]

8) BAD PREDICTION (example 2)



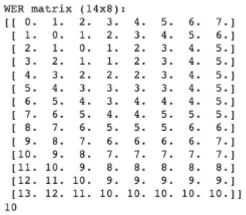
```
Real Caption: <start> a kitchen with a map on the wall <end>
Prediction Caption: a kitchen with two space with a sink and a white tile floor <end>
```

BLEU SCORE, GLEU SCORE and WER (Word Error Rate) metric comparison between sentences:

INDUVIDUAL N GRAM	
1 GRAM	0.071429
2 GRAM	1.000000
3 GRAM	1.000000
4 GRAM	1.000000

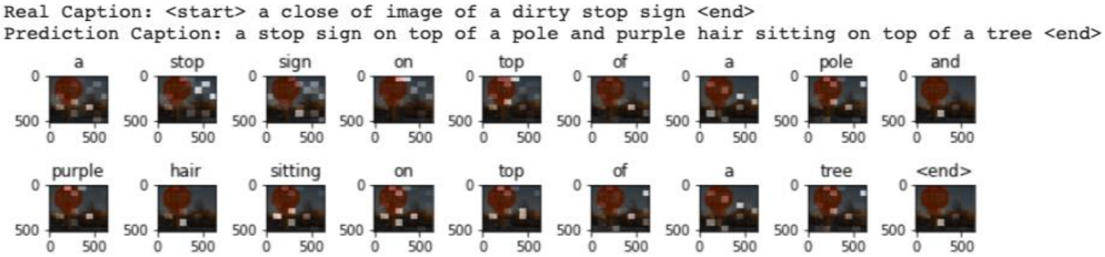
CUMMULATIVE N GRAM	
BLEU 1	0.071429
BLEU 2	0.267261
BLEU 3	0.418579
BLEU 4	0.516973

GLEU SCORE	
Sentence Level Frequency	
1 to 4 grams	0.1600
1 to 2 grams	0.2592



```
WER matrix (14x8):
[[ 0.  1.  2.  3.  4.  5.  6.  7.]
 [ 1.  0.  1.  2.  3.  4.  5.  6.]
 [ 2.  1.  0.  1.  2.  3.  4.  5.]
 [ 3.  2.  1.  1.  2.  3.  4.  5.]
 [ 4.  3.  2.  2.  2.  3.  4.  5.]
 [ 5.  4.  3.  3.  3.  3.  4.  5.]
 [ 6.  5.  4.  3.  4.  4.  4.  5.]
 [ 7.  6.  5.  4.  4.  5.  5.  5.]
 [ 8.  7.  6.  5.  5.  5.  6.  6.]
 [ 9.  8.  7.  6.  6.  6.  6.  7.]
 [10.  9.  8.  7.  7.  7.  7.  7.]
 [11. 10.  9.  8.  8.  8.  8.  8.]
 [12. 11. 10.  9.  9.  9.  9.  9.]
 [13. 12. 11. 10. 10. 10. 10. 10.]]
10
```

9) BAD PREDICTION (example 3)



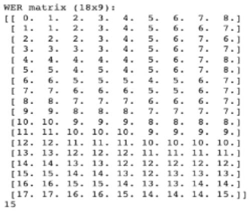
```
Real Caption: <start> a close of image of a dirty stop sign <end>
Prediction Caption: a stop sign on top of a pole and purple hair sitting on top of a tree <end>
```

BLEU SCORE, GLEU SCORE and WER (Word Error Rate) metric comparison between sentences:

INDUVIDUAL N GRAM	
1 GRAM	0.055556
2 GRAM	1.000000
3 GRAM	1.000000
4 GRAM	1.000000

CUMMULATIVE N GRAM	
BLEU 1	0.055556
BLEU 2	0.235702
BLEU 3	0.385265
BLEU 4	0.485492

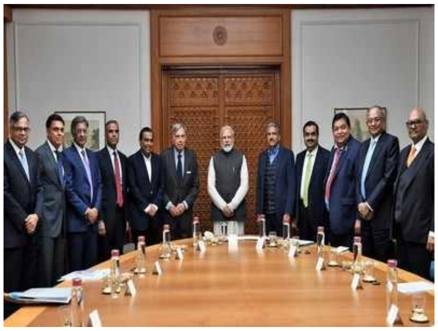
GLEU SCORE	
Sentence Level Frequency	
1 to 4 grams	0.1212
1 to 2 grams	0.2285



```
WER matrix (18x9):
[[ 0.  1.  2.  3.  4.  5.  6.  7.  8.]
 [ 1.  1.  2.  3.  4.  5.  6.  6.  7.]
 [ 2.  2.  2.  3.  4.  5.  6.  7.  6.]
 [ 3.  3.  3.  3.  4.  5.  6.  7.  7.]
 [ 4.  4.  4.  4.  4.  5.  6.  7.  8.]
 [ 5.  5.  4.  5.  4.  5.  6.  7.  8.]
 [ 6.  6.  5.  5.  5.  4.  5.  6.  7.]
 [ 7.  7.  6.  6.  6.  5.  5.  6.  7.]
 [ 8.  8.  7.  7.  7.  6.  6.  6.  7.]
 [ 9.  9.  8.  8.  8.  7.  7.  7.  7.]
 [10. 10.  9.  9.  9.  8.  8.  8.  8.]
 [11. 11. 10. 10. 10.  9.  9.  9.  9.]
 [12. 12. 11. 11. 11. 10. 10. 10. 10.]
 [13. 13. 12. 12. 12. 11. 11. 11. 11.]
 [14. 14. 13. 13. 12. 12. 12. 12. 12.]
 [15. 15. 14. 14. 13. 12. 13. 13. 13.]
 [16. 16. 15. 15. 14. 13. 13. 14. 14.]
 [17. 17. 16. 16. 15. 14. 14. 14. 15.]]
15
```

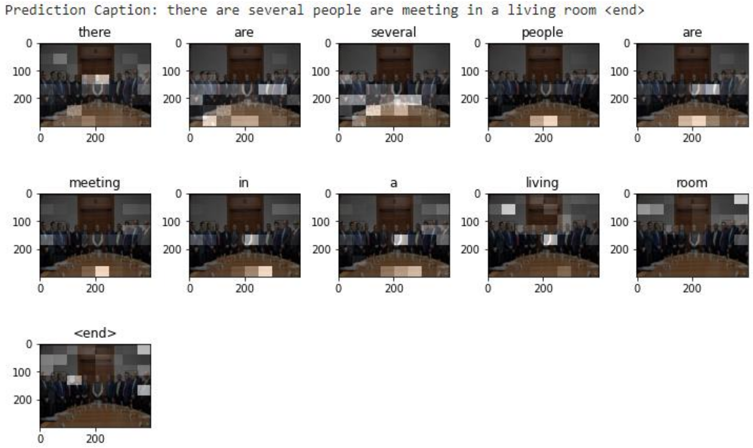
10) RANDOM IMAGES (example 1)

REAL IMAGE



PREDICTIONS BY MODEL

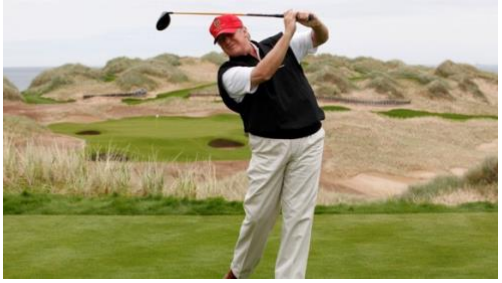
“There are several people are meeting in a living room”



```
Prediction Caption: there are several people are meeting in a living room <end>
```

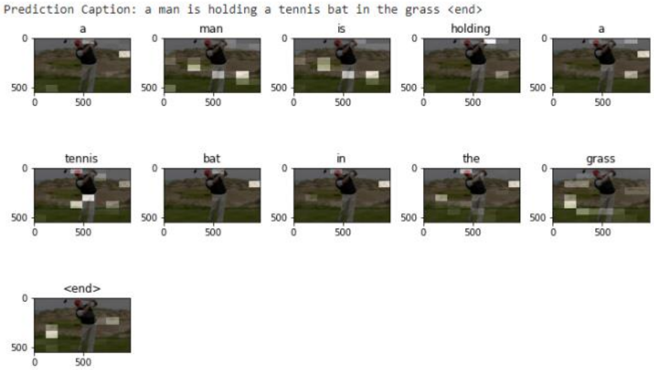
11) RANDOM IMAGES (example 2)

REAL IMAGE



PREDICTIONS BY MODEL

“A man is holding a tennis bat in the grass”

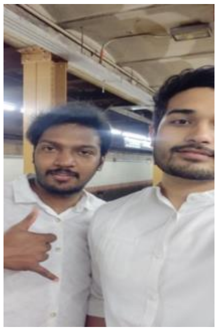


```
Prediction Caption: a man is holding a tennis bat in the grass <end>
```

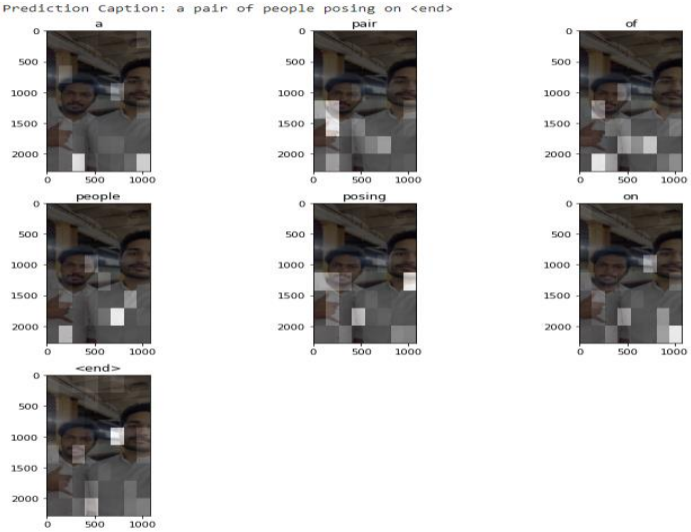
12) RANDOM IMAGES (example 2)

This is an image of myself (on the right) and my teammate!!
The model predicted the following for my picture:

REAL IMAGE:

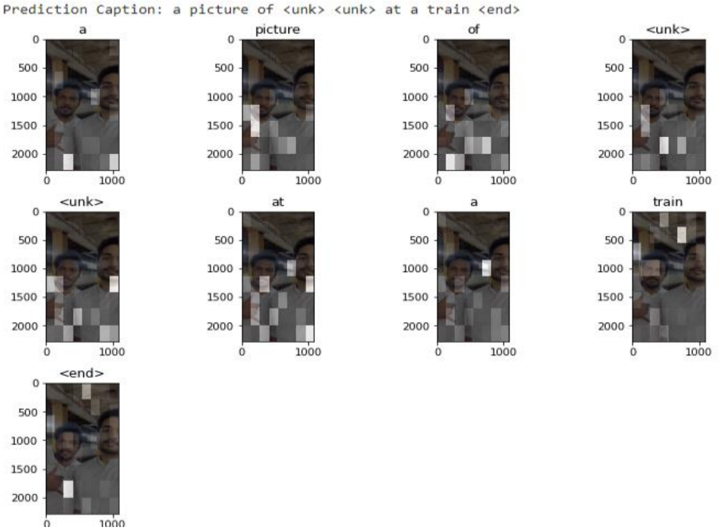


PREDICTION 1: “A pair of people posing on”



```
Prediction Caption: a pair of people posing on <end>
```

PREDICTION 2: “A picture of <unk> <unk> at a train”



```
Prediction Caption: a picture of <unk> <unk> at a train <end>
```