CAPSTONE PROJECT
Image Caption Generator

Sing Ee Shawn

ABOUT ME

Through the 12-week DSI program, I went from learning the basics of Python to exploring machine learning using libraries like scikit-learn, but I also excited to try more.

Deep learning with TensorFlow and Keras seemed like a good gateway into working with computer vision and big data, which seemed like a good milestone at this point.

PROBLEM STATEMENT

MACHINE LEARNING

A machine can be trained to generate a sentence in a natural language based on the features identified within an image.

MY HOBBY

As a photography hobbyist, it would be interesting to train a model that can automatically generate captions for the images I share.

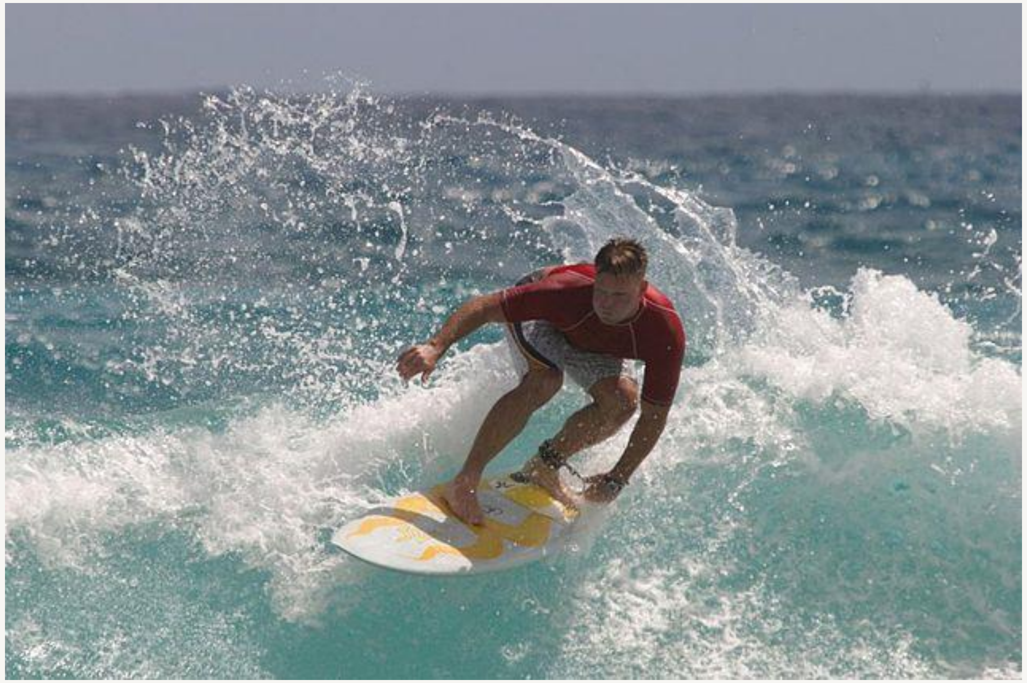
STEPS INVOLVED

Before we add on additional complexities, the base model needs to correctly identify the features, then generate a simple caption describing the features.

GOAL

To train an attention mechanism-based caption generator that is able to generate a descriptive caption of an image with a BLEU-1 score of at least 0.5.

IMAGE CAPTION GENERATOR



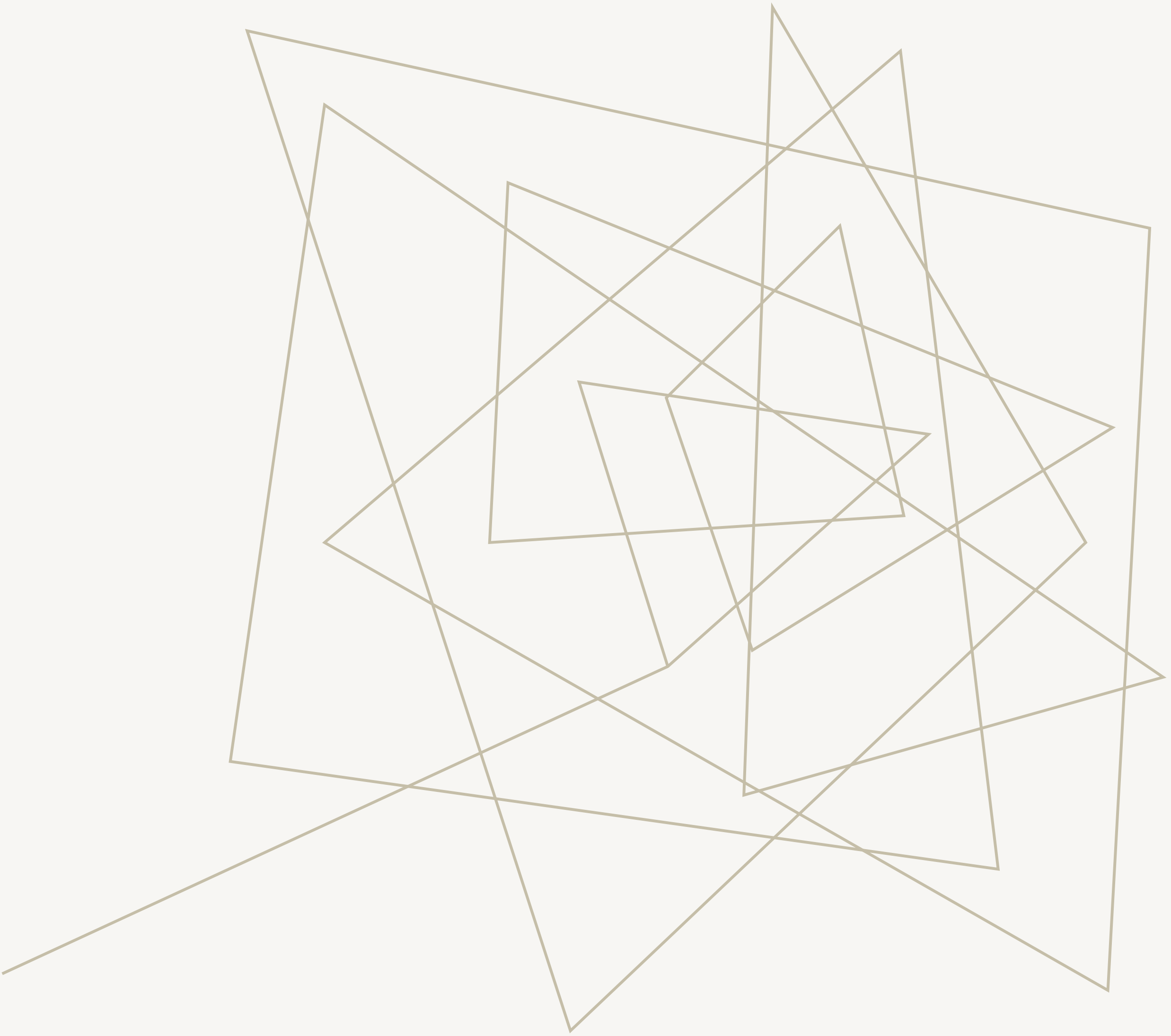
STEP 1: Caption Generator

Man in red shirt is surfing in the ocean

Goal

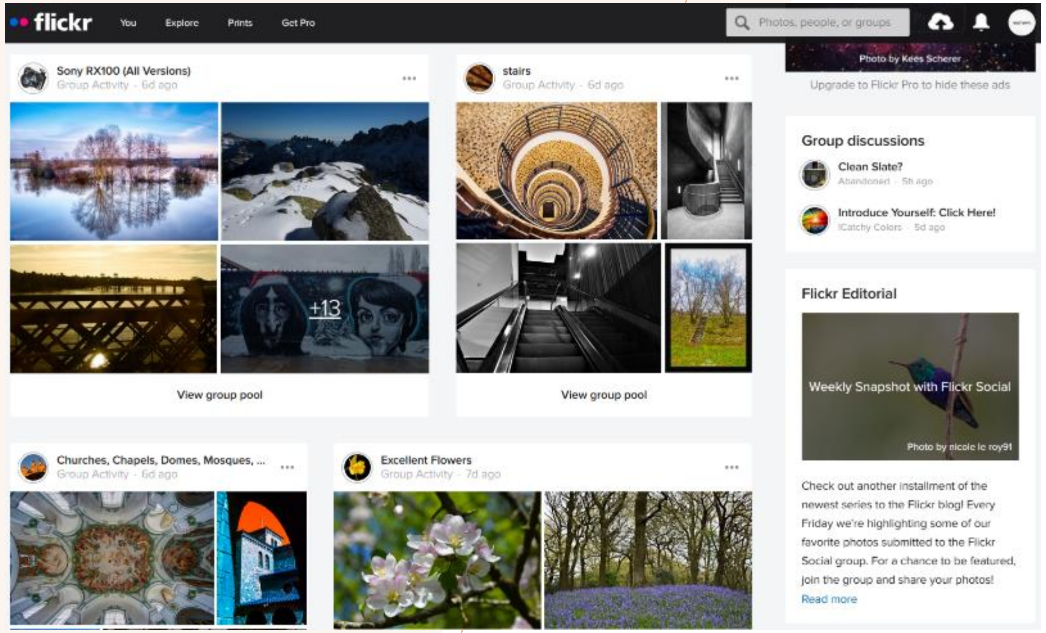
STEP 2: Text Generation AI

Man in a red shirt is surfing in the ocean off Australia’s Gold Coast, one of the world’s top surfing destinations.



DATASET

Flickr30k



Check out another installment of the newest series to the Flickr blog! Every Friday we're highlighting some of our favorite photos submitted to the Flickr Social group. For a chance to be featured, join the group and share your photos!

Flickr

Online photo hosting service and community started in 2004

SOURCE

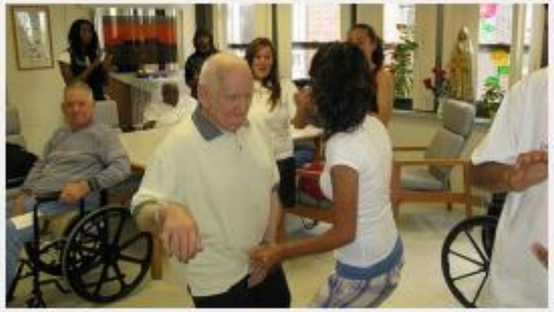
From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions

Peter Young and Alice Lai and Micah Hodosh and Julia Hockenmaier

DATA

158,915 crowd sourced **captions** describing **31,783 images** sourced from Flickr, mostly focusing on people involved in everyday activities and events.

DATA SAMPLES



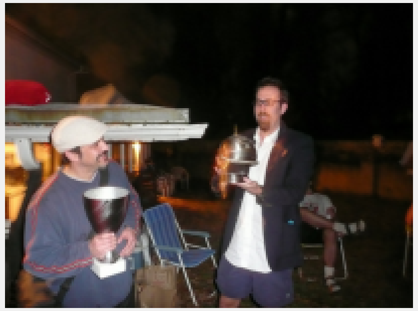
A child is dancing with an elderly man .

An old man is just dancing with a woman .

An elderly man is dancing with a young lady .

An elderly man is dancing with a young girl .

A young woman dancing with an old man while other people watch .



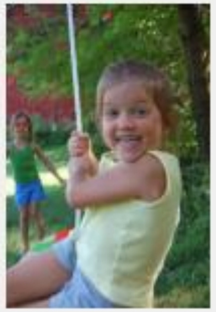
Two friends hold trophies .

Two men enjoy the weather outside .

Two friends are comparing trophy 's .

Two men enjoying drinks at an outdoor event .

Two men outside looking at metalwork objects .



A girl swings on a rope swing .

The small girl is swinging on the rope

A girl rides a swing while another girl watches .

A young girl wearing a yellow shirt swings on a tree rope .

A little girl with a yellow shirt swings while a little girl in green watches .

CAPTIONS CORPUS

1,957,129 TOTAL WORDS

Total number of words across all the captions

18,293 UNIQUE WORDS

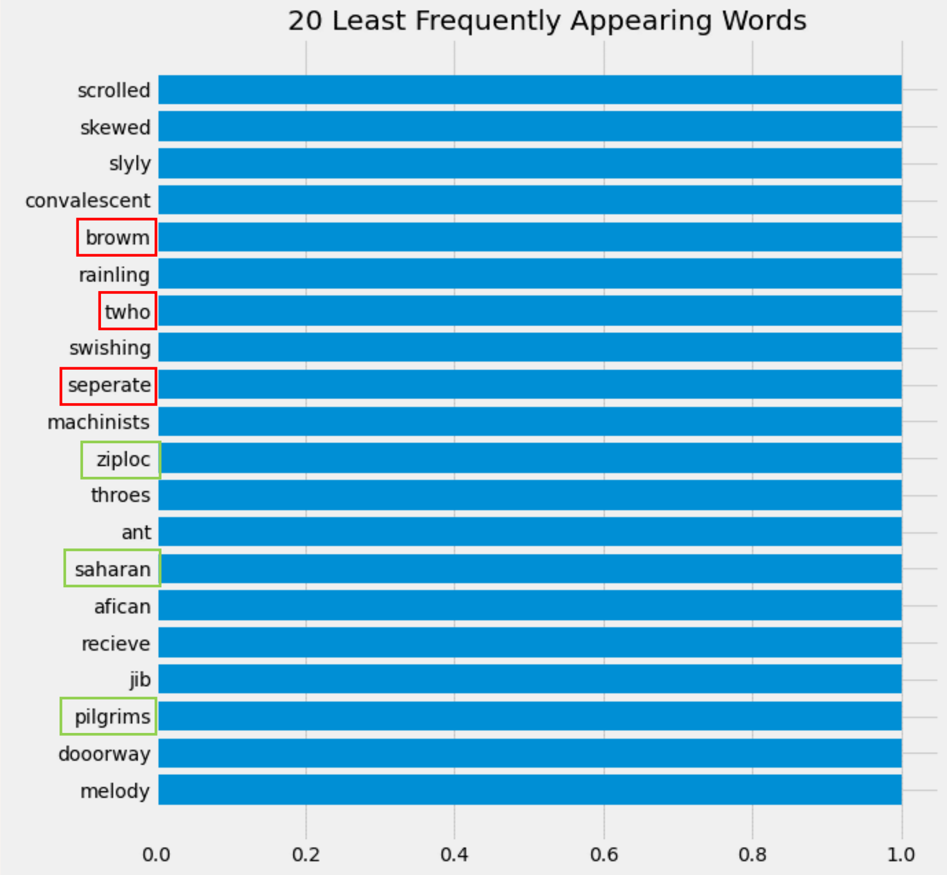
Words that appear at least once throughout all the captions

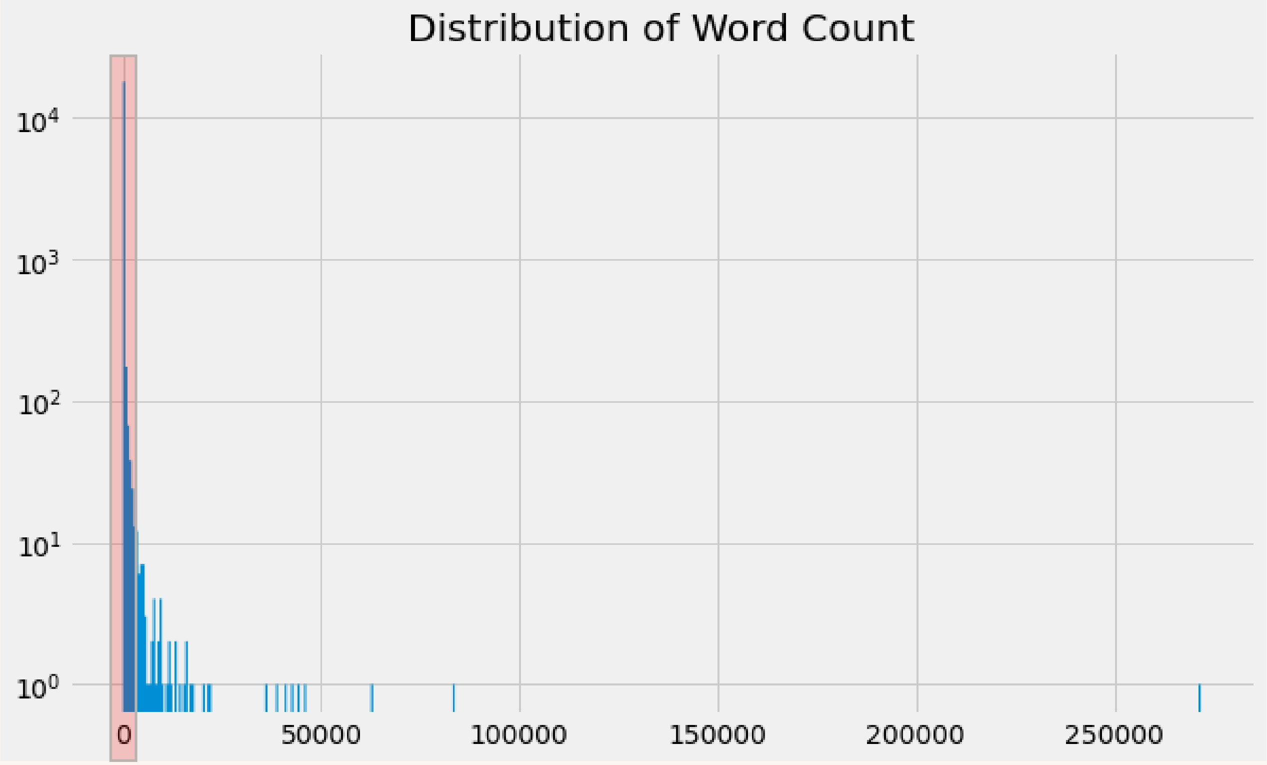
BETWEEN 2 TO 78 WORDS

Length of the captions

TOP 20 LEAST COMMON WORDS

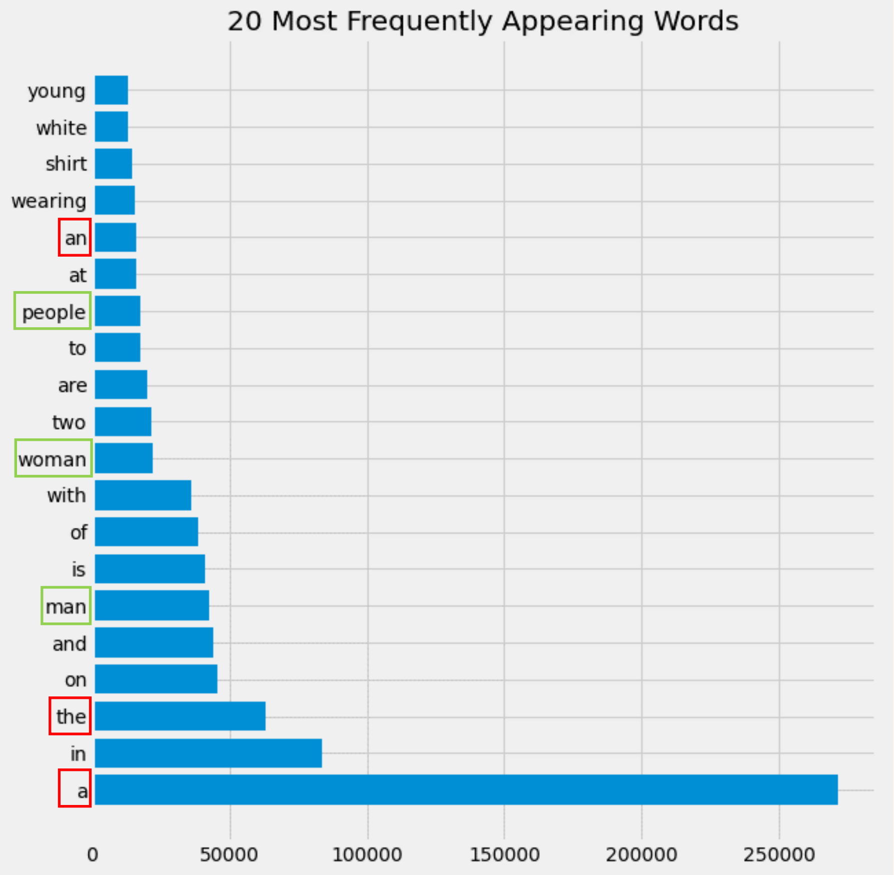
Often misspelt or very specific words

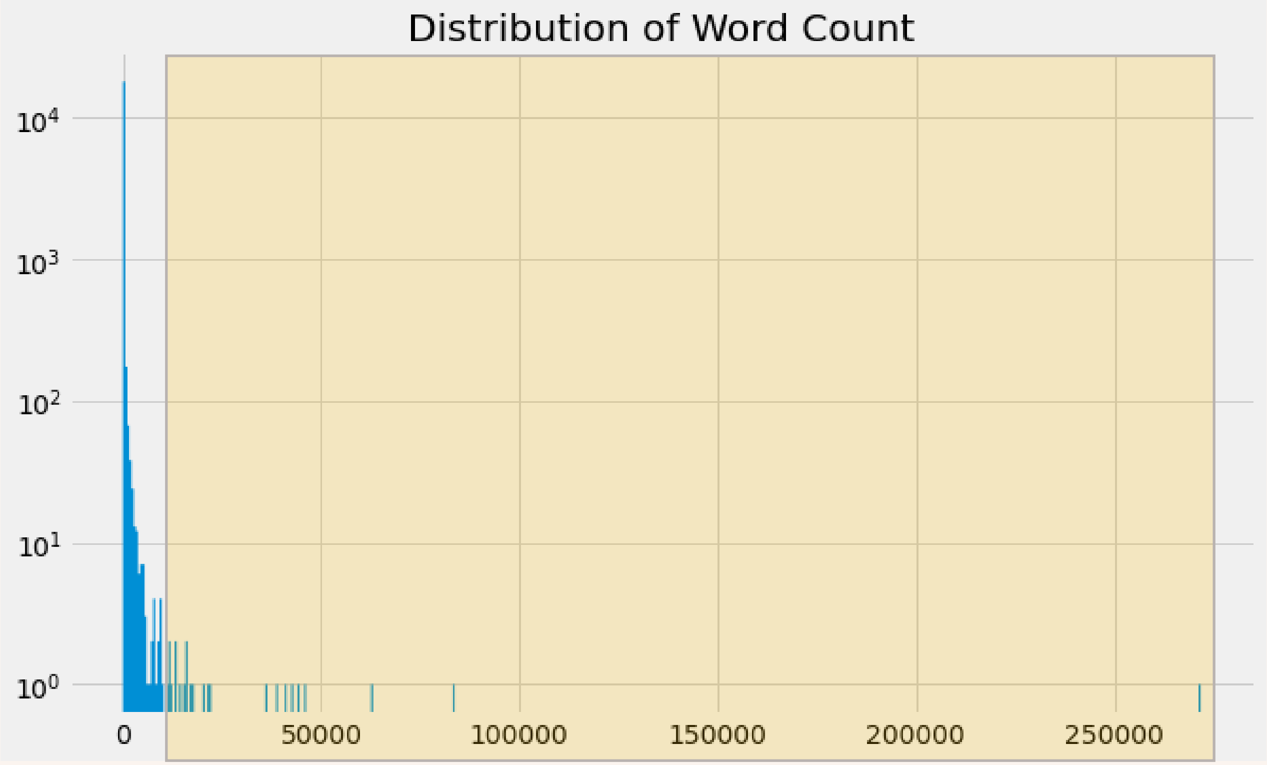




TOP 20 MOST COMMON WORDS

Broad terms, prepositions, articles, adjectives





CAPTION CLEANING

ORIGINAL CAPTION

Man reads newspaper in a park while drinking Starbuck 's coffee .

CONVERT TO LOWER CASE

Man reads newspaper in a park while drinking Starbuck 's coffee .

REMOVE PUNCTUATION

man reads newspaper in a park while drinking starbuck ‘s coffee .

REMOVE SINGLE CHARACTER WORDS

man reads newspaper in a park while drinking starbuck s coffee

CLEANED CAPTION

man reads newspaper in park while drinking starbuck coffee

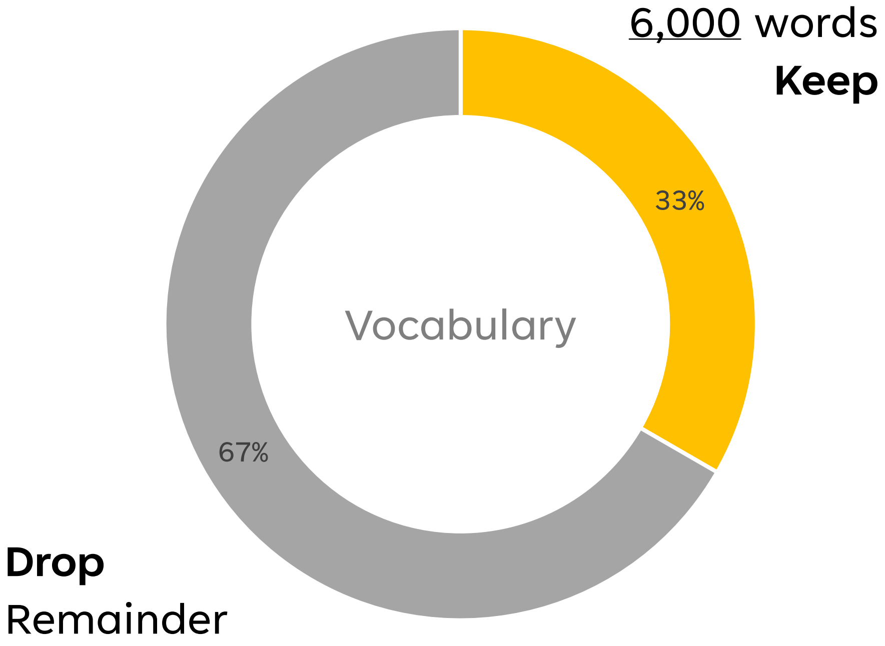
AFTER CLEANING

Shortest Caption

1 word

Longest Caption

72 words



DATASET SPLIT

87%

TRAIN

27,648 images

138,240 captions

10%

VALIDATION

3,072 images

15,360 captions

3%

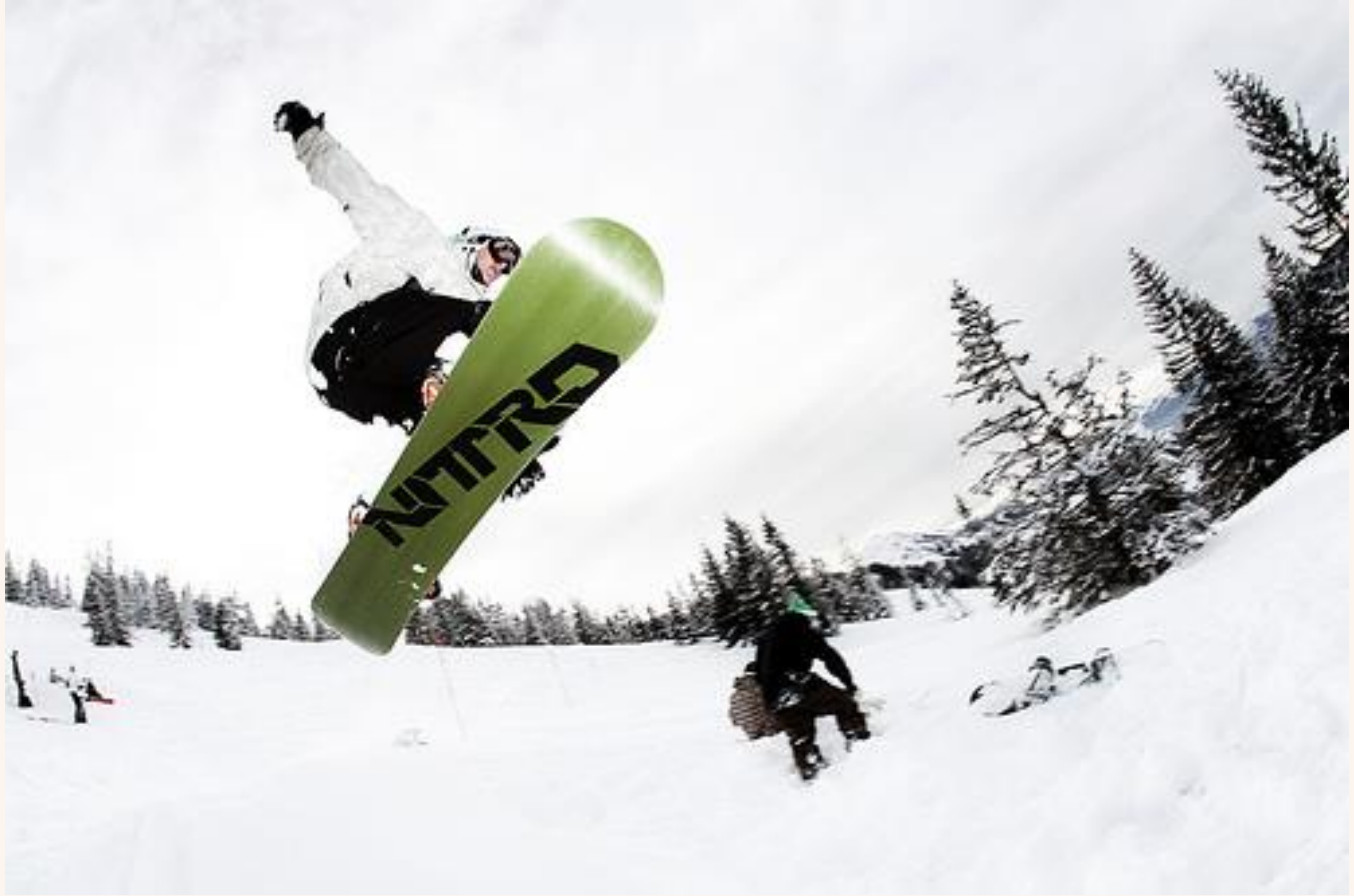
TEST

1,063 images

5,315 captions

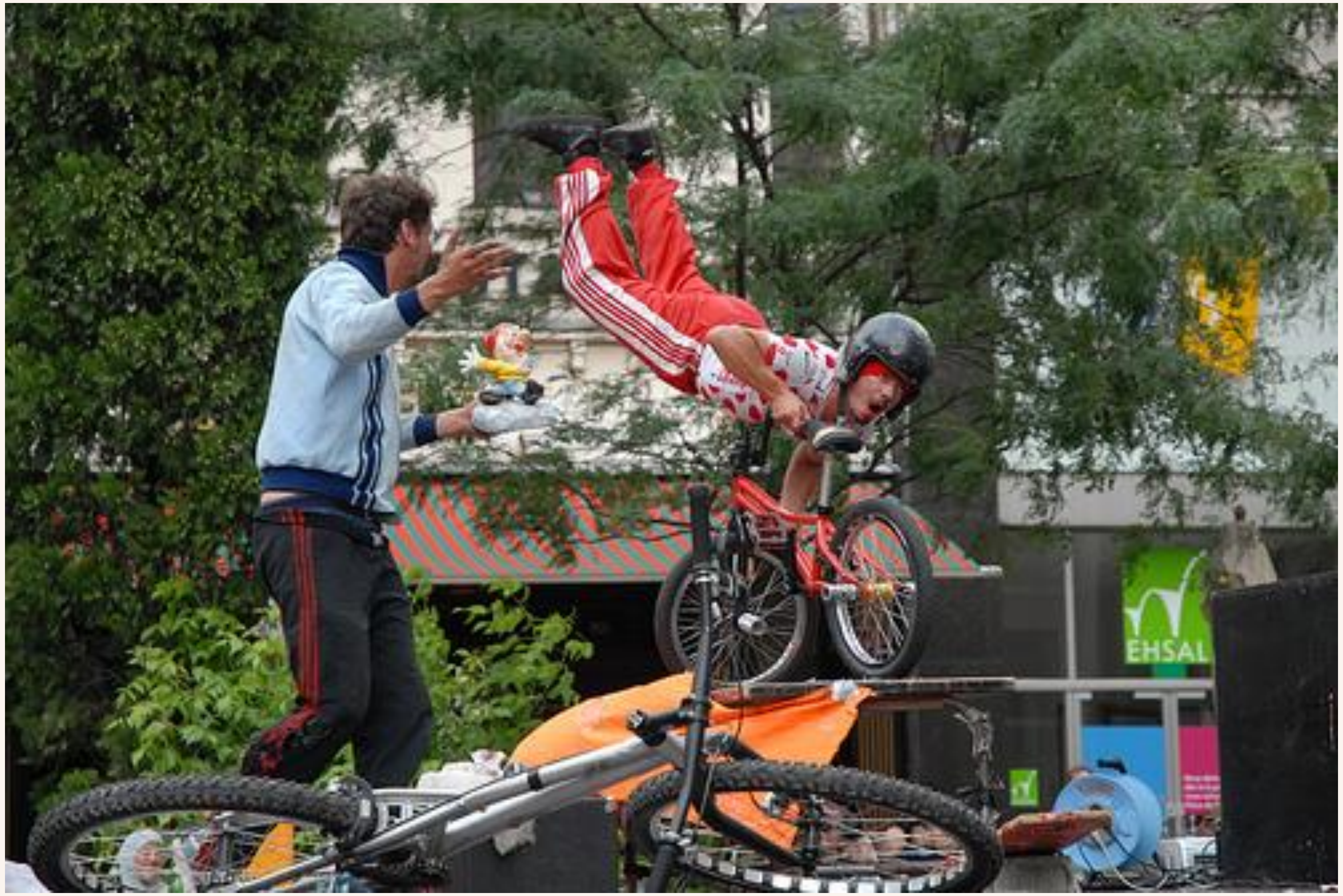
SHORTEST CAPTION

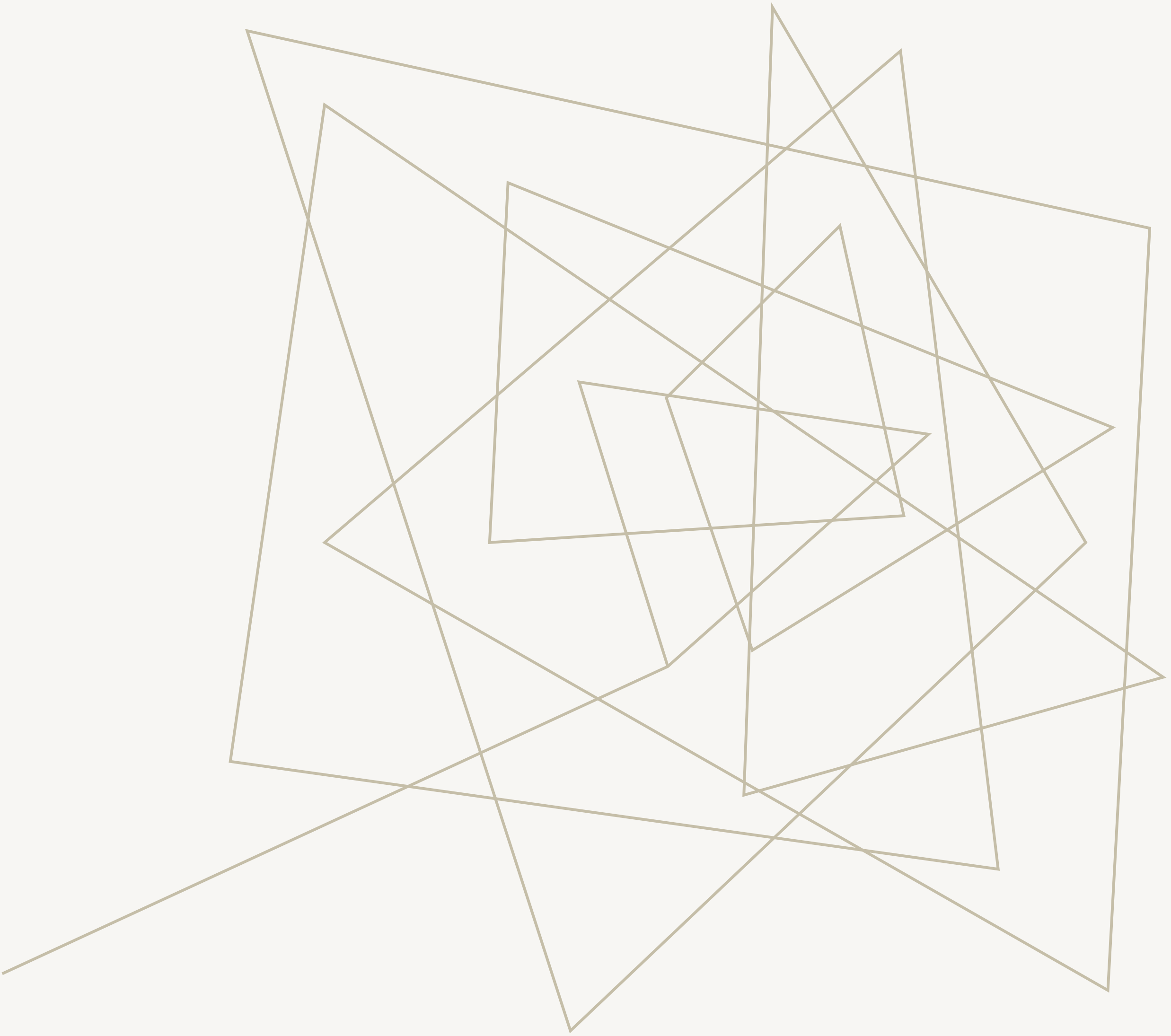
“snowboarder”



LONGEST CAPTION

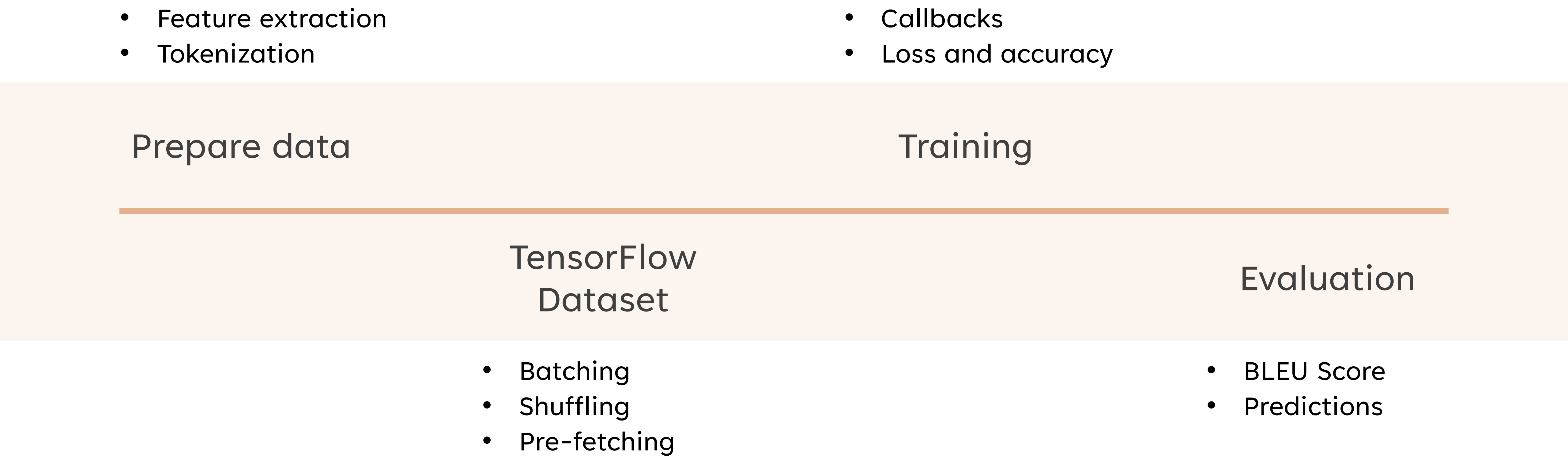
“man wearing helmet red pants with white stripes going down the sides and white and red shirt is on small bicycle using only his hands while his legs are up in the air while another man wearing light blue shirt with dark blue trim and black pants with red stripes going up the sides is standing nearby gesturing toward the first man and holding small figurine of one of the seven dwarves





MODEL

MODEL WORKFLOW



Prepare data

- Feature extraction
- Tokenization

TensorFlow Dataset

- Batching
- Shuffling
- Pre-fetching

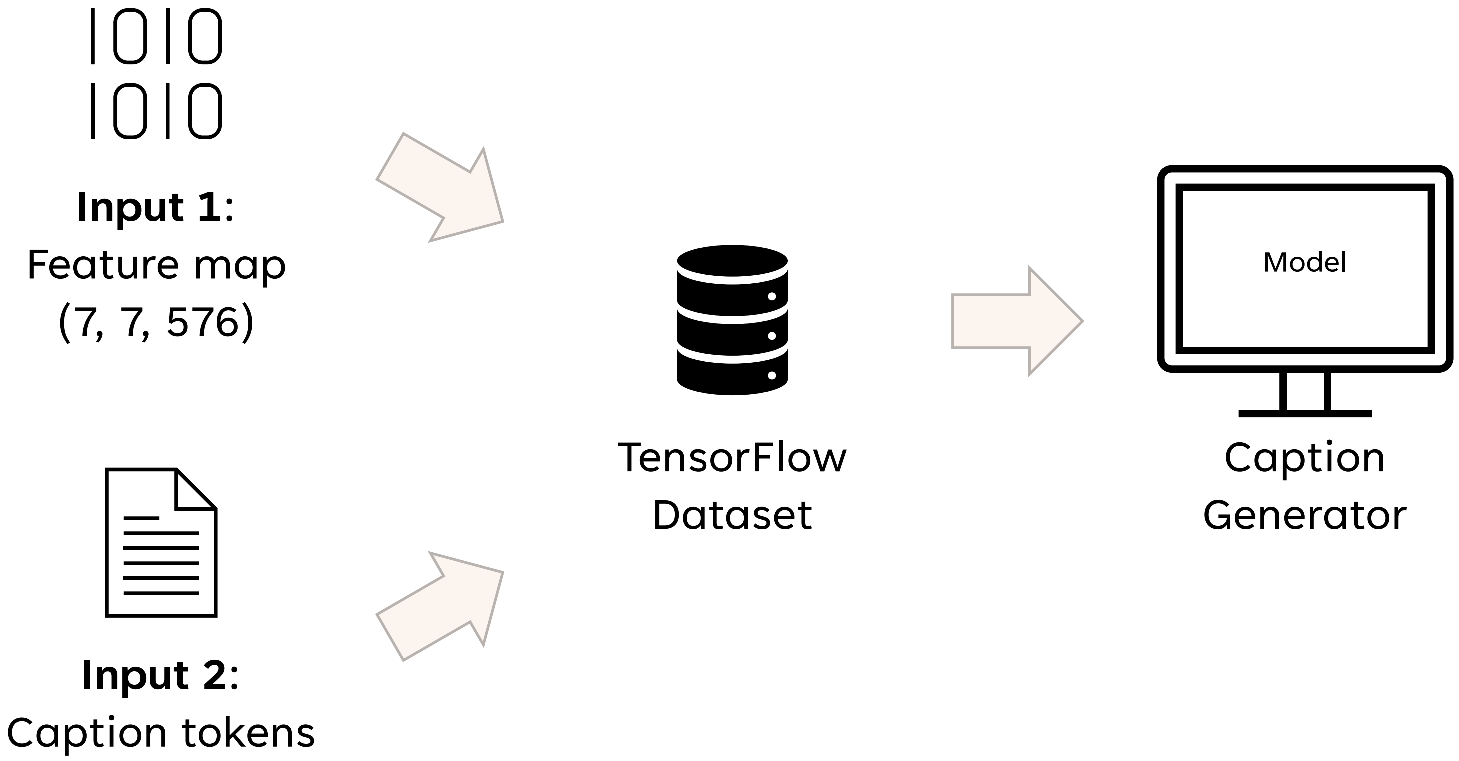
Training

- Callbacks
- Loss and accuracy

Evaluation

- BLEU Score
- Predictions

CAPTION GENERATION



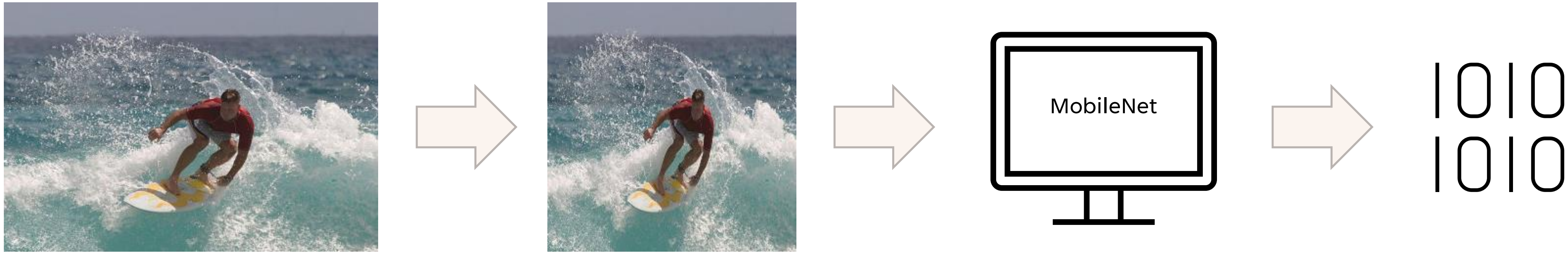
Input 1: Feature map (7, 7, 576)

Input 2: Caption tokens

TensorFlow Dataset

Caption Generator

INPUT 1: FEATURE MAP



MobileNet

Input:
Image

Resized
(224, 224, 3)

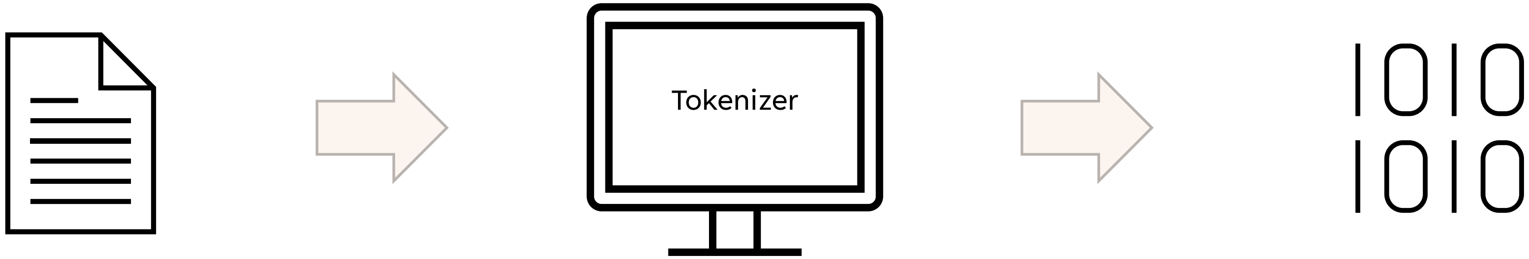
Passed into
image encoder

Output:
Feature map
(7, 7, 576)

IMAGE ENCODER – MOBILENETV3

- Image classification model without classification layer
- Developed by Google researchers
- Pre-trained on the ImageNet dataset
  - 1.4M images
  - 1,000 classes

INPUT 2: CAPTION TOKENS



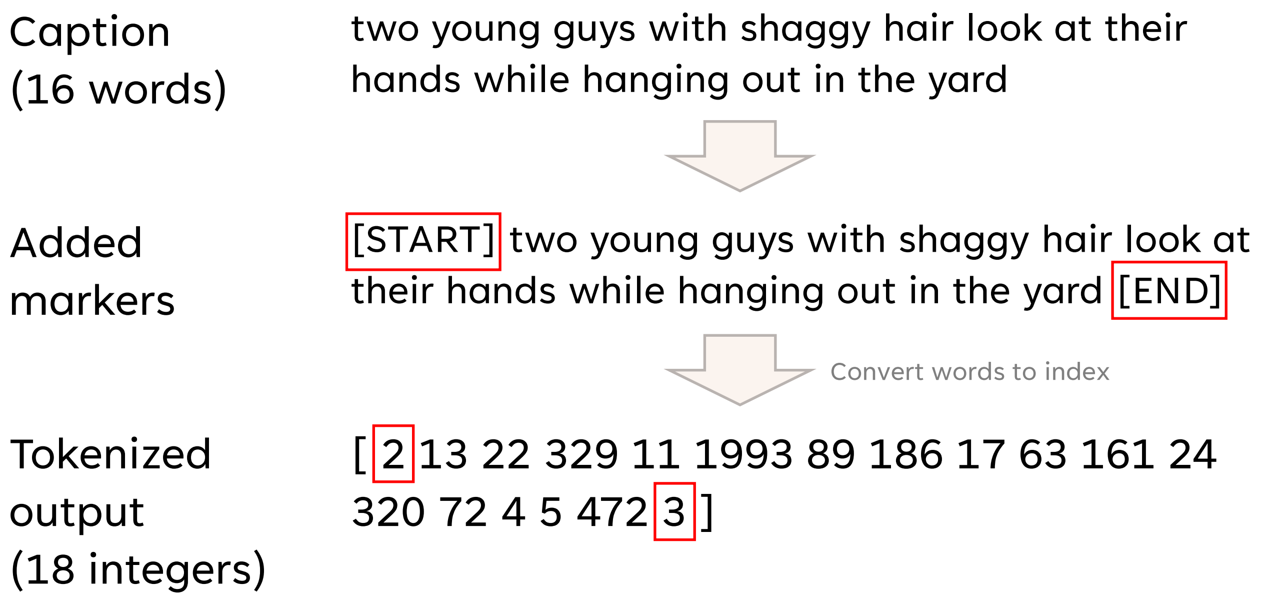
Input:
Caption

Passed into
TextVectorization
Layer

Output:
Integer sequence

TOKENIZER

- TextVectorization layer adapted on training captions to compute a vocabulary with 6,000 words
- Adds a [START] and [END] marker



Caption (16 words)

two young guys with shaggy hair look at their hands while hanging out in the yard

Added markers

[START] two young guys with shaggy hair look at their hands while hanging out in the yard [END]

Convert words to index

Tokenized output (18 integers)

[2 13 22 329 11 1993 89 186 17 63 161 24 320 72 4 5 472 3]

CAPTION GENERATOR

Token Output layer

- Generates token probabilities

Feed Forward layer

- Passes information through

Cross Attention layer

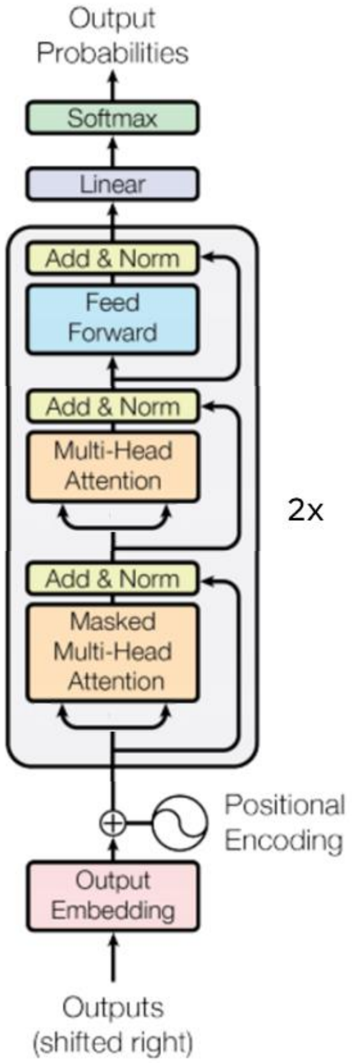
- Returns the attention_scores for plotting

Causal Self Attention layer

- Creates a causal mask to ensure output only depends on previous elements

Sequence Embedding layer

- Provides a sequential order to the tokens



Output

break

Tokenized output [13 22 329 11 1993 89 186 17 63 161 24 320 72 4 5 472 3]

Convert index to words

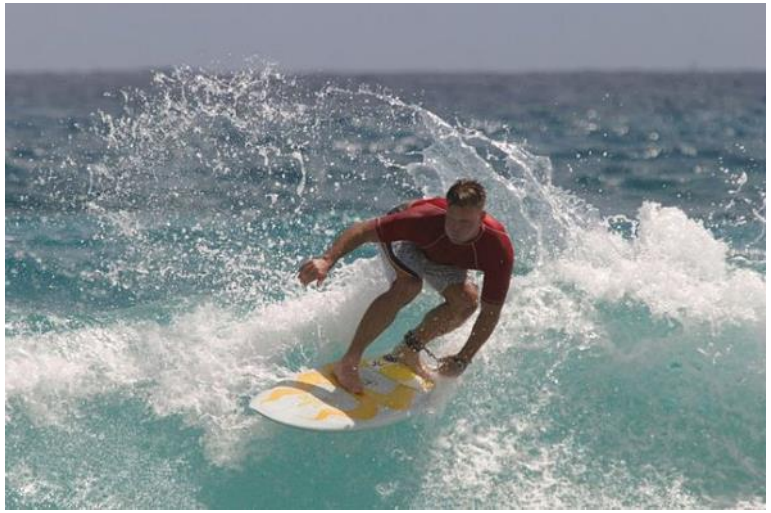
Caption two young guys with shaggy hair look at their hands while hanging out in the yard [END]

Image adapted from https://www.tensorflow.org/text/tutorials/transformer

TRAINING

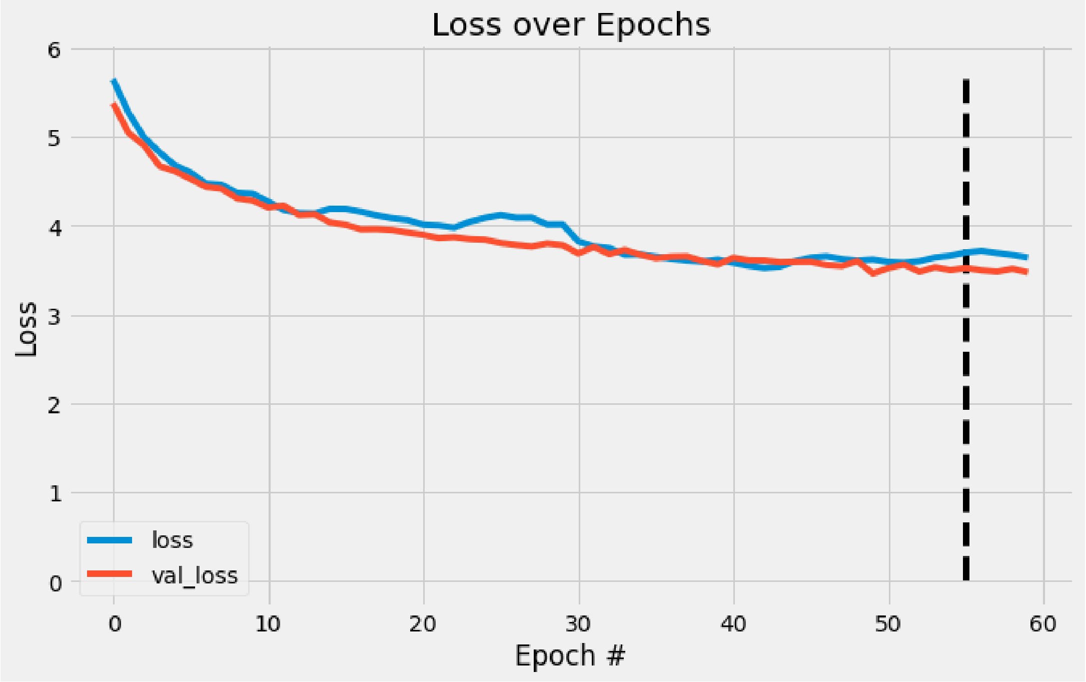
Custom ‘GenerateText’ callback to visualise training progress

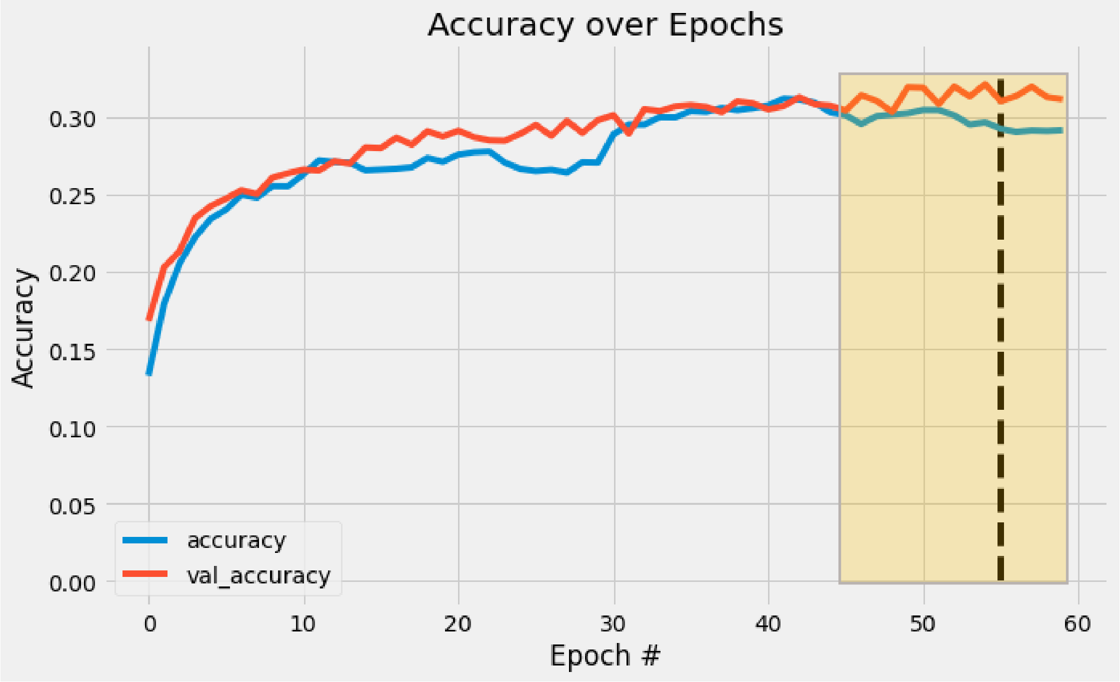
- Epoch 1
  “man in the the the the the the the the the the the the the the the”
- Epoch 5
  “two men are playing in the water”
- Epoch 10
  “man in red shirt is playing in the water”
- Epoch 20
  “man in red shirt and blue shirt is in the water”
- Epoch 40
  “surfer is surfing wave”
- Epoch 50
  “surfer in red wetsuit is surfing”
- Epoch 55
  “man in red shirt is surfing in the ocean”
- Epoch 60
  “surfer is surfing in the ocean”

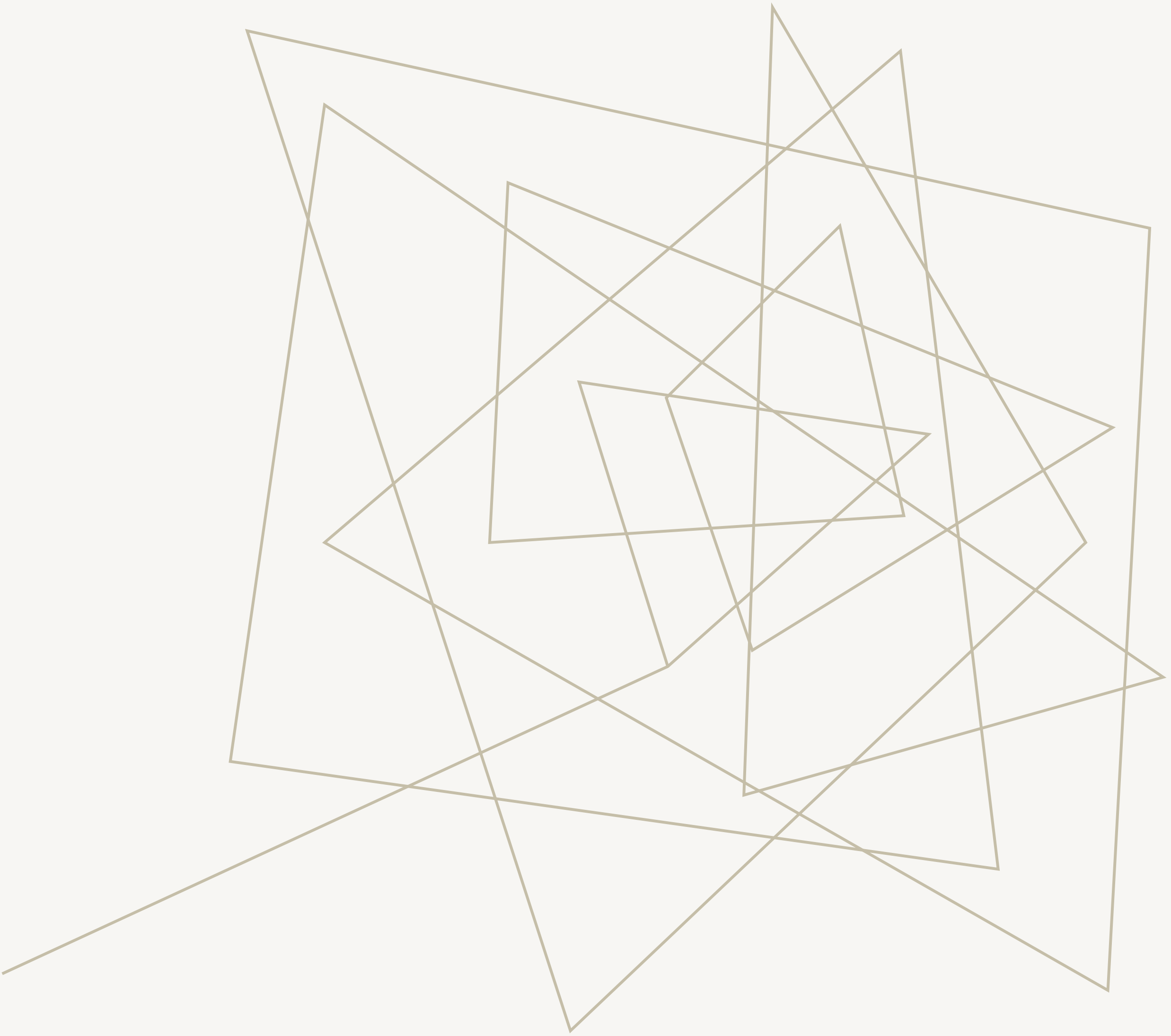


Achieved highest val_accuracy at Epoch 55

- ‘EarlyStop’ callback if no improvements after 5 Epochs
- Training stopped after Epoch 60
- Could have potentially stopped at Epoch 45, as there appears to be a divergence between train and validation scores, which may suggest overfitting







DOES IT WORK?

BLEU SCORE

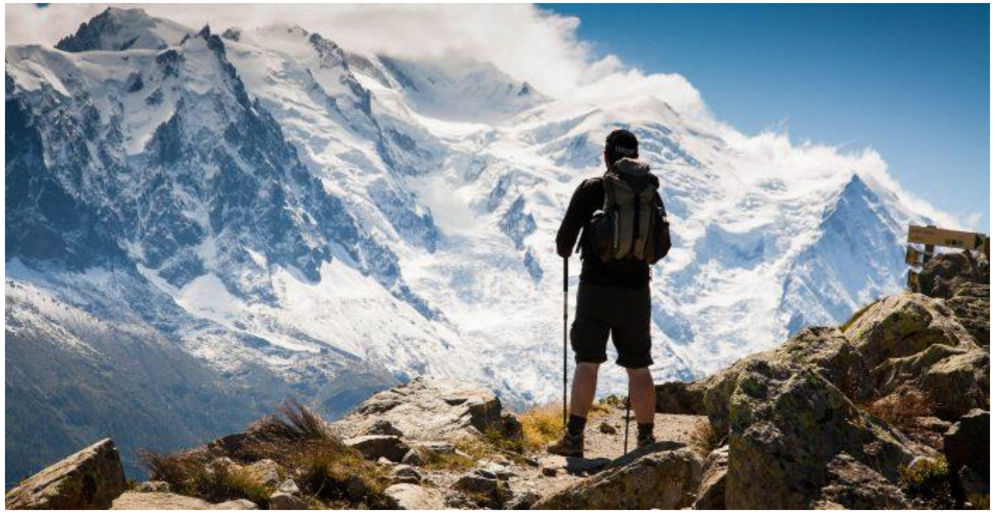
- Bilingual Evaluation Understudy (BLEU)
- Originally developed to compare performance of translation models
- Easy to compute and understand

Reference	“I like dogs very much”
Prediction	“I like dogs too”
1-gram	“I”, “like”, “dogs”, “very”, “much” “I”, “like”, “dogs”, “too”
2-gram	“I like”, “like dogs”, “dogs very”, “very much” “I like”, “like dogs”, “dogs too”
3-gram	“I like dogs”, “like dogs very”, “dogs very much” “I like dogs”, “like dogs too”
4-gram	“I like dogs very”, “like dogs very much” “I like dogs too”

BLEU SCORE

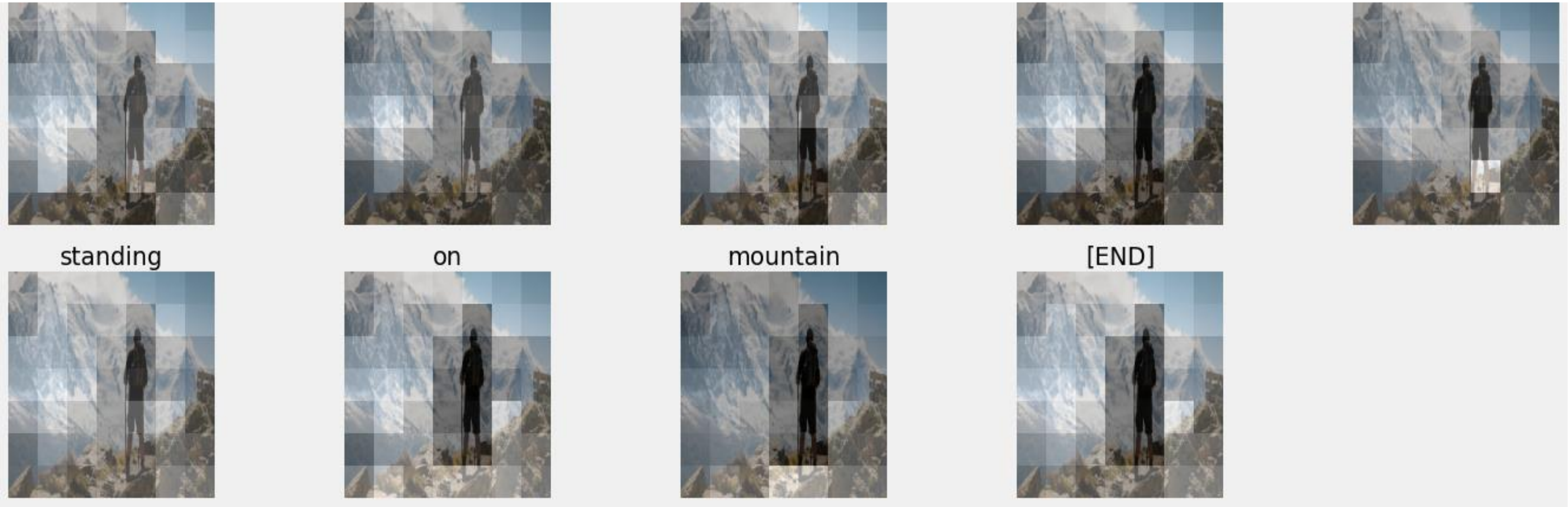
	Weightage					
	1-gram	2-gram	3-gram	4-gram	Benchmark[1]	**Model Score**
BLEU-1	1	0	0	0	0.60	**0.52**
BLEU-2	0.5	0.5	0	0	0.41	**0.32**
BLEU-3	0.33	0.33	0.33	0	0.27	**0.20**
BLEU-4	0.25	0.25	0.25	0.25	0.18	**0.11**

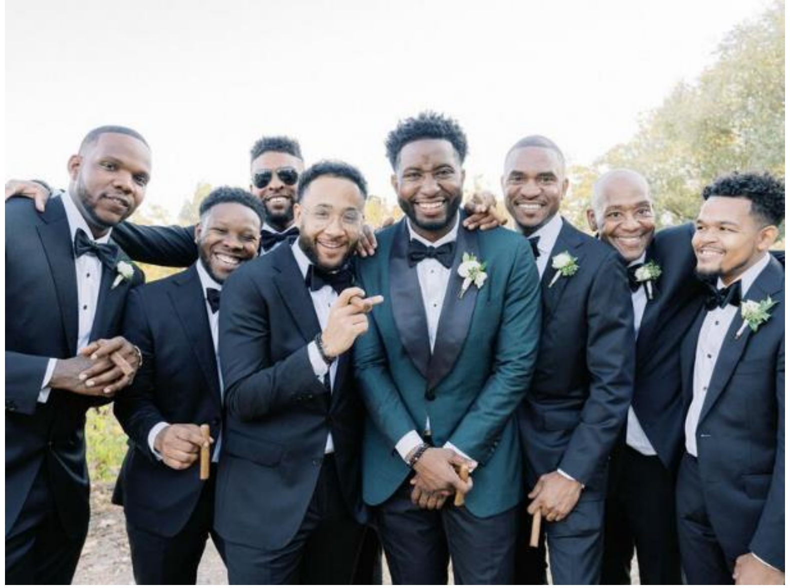
[1] Estimated benchmark based on a high performance model. Source: *Where to put the Image in an Image Caption Generator*



PREDICTED CAPTION

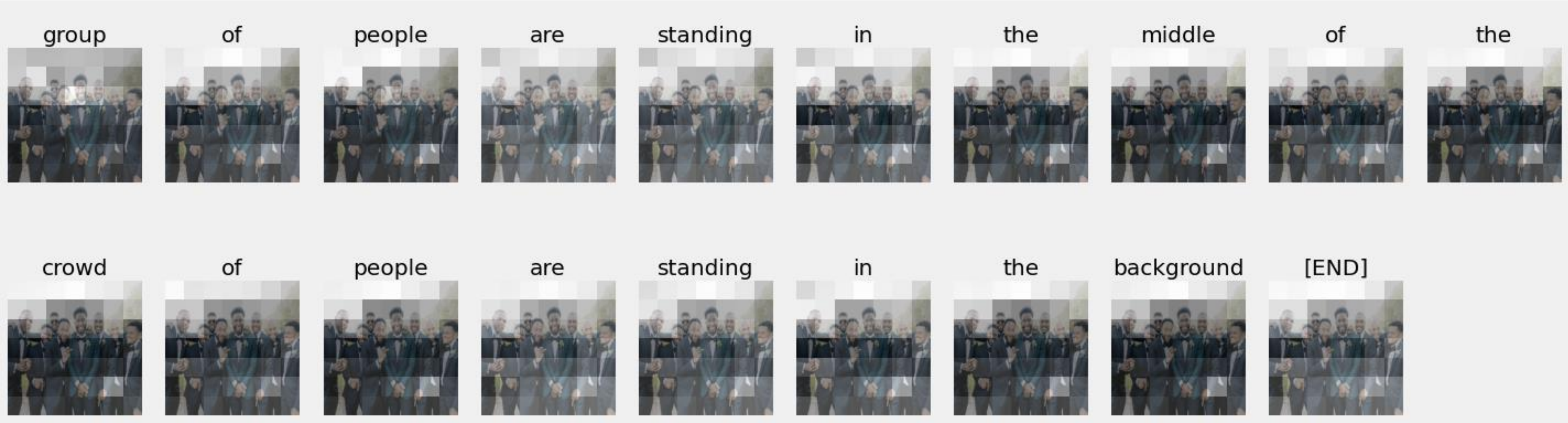
man in blue shirt is standing on mountain

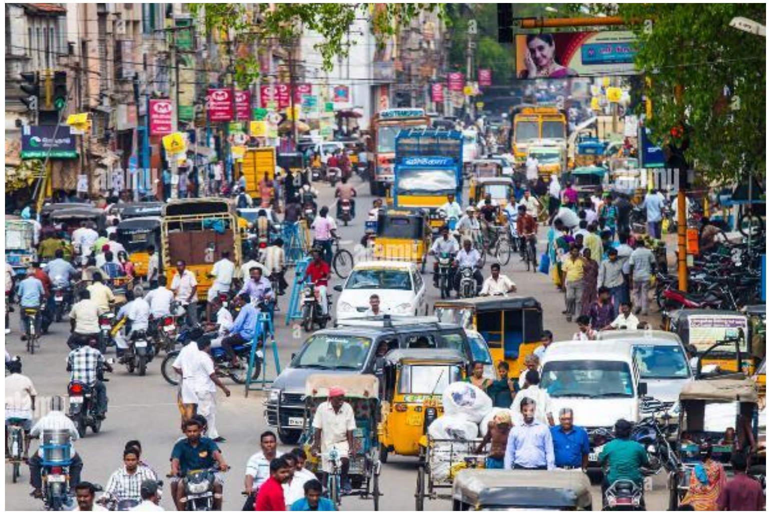




PREDICTED CAPTION

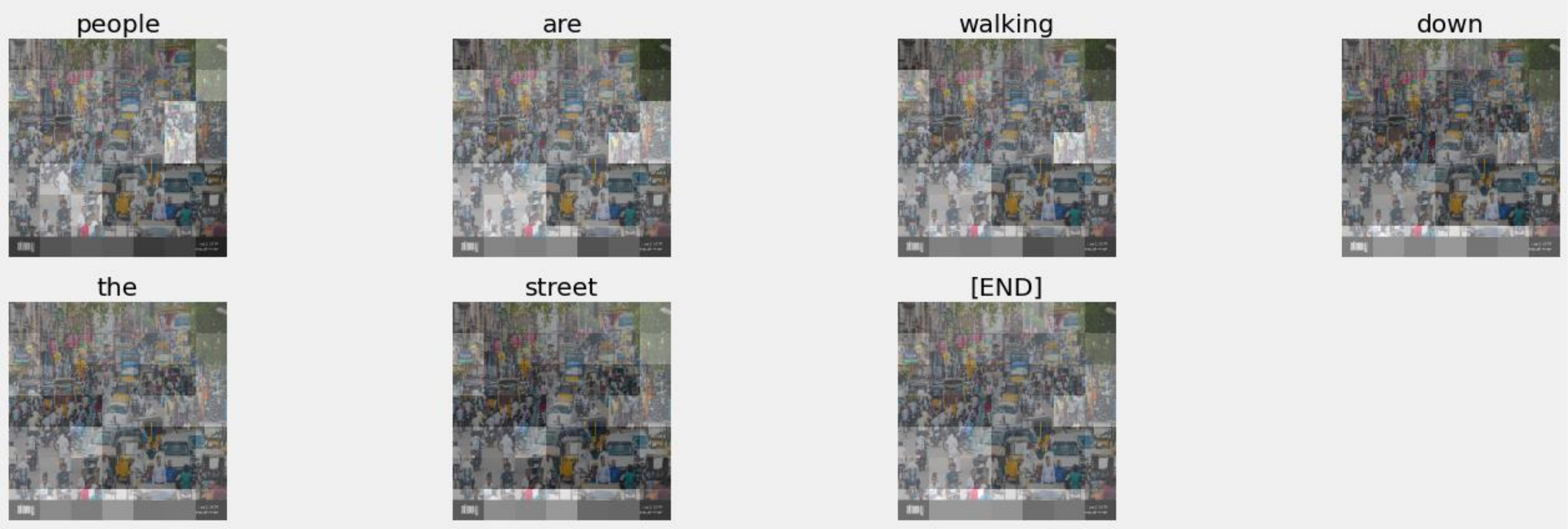
group of people are standing in the middle of the crowd of people are standing in the background

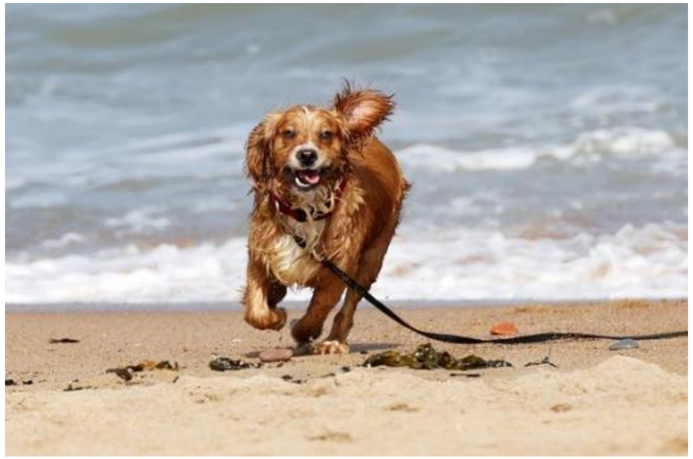




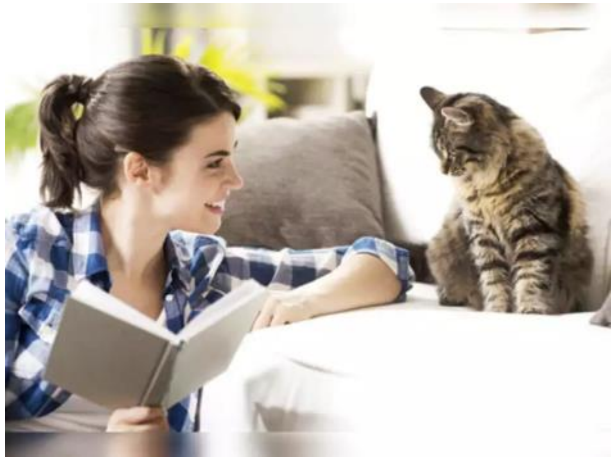
PREDICTED CAPTION

people are walking down the street

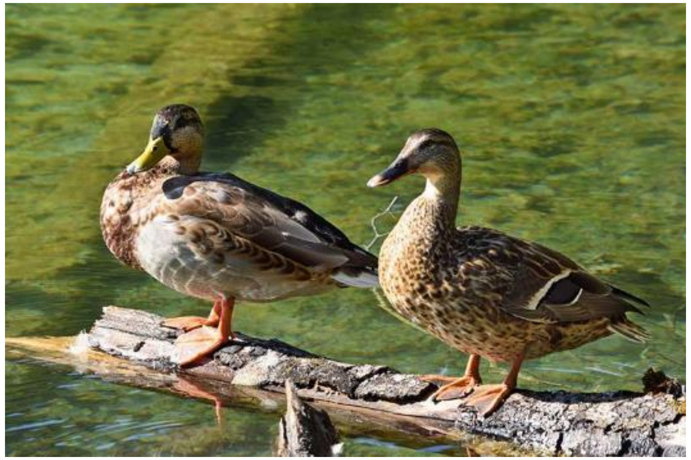




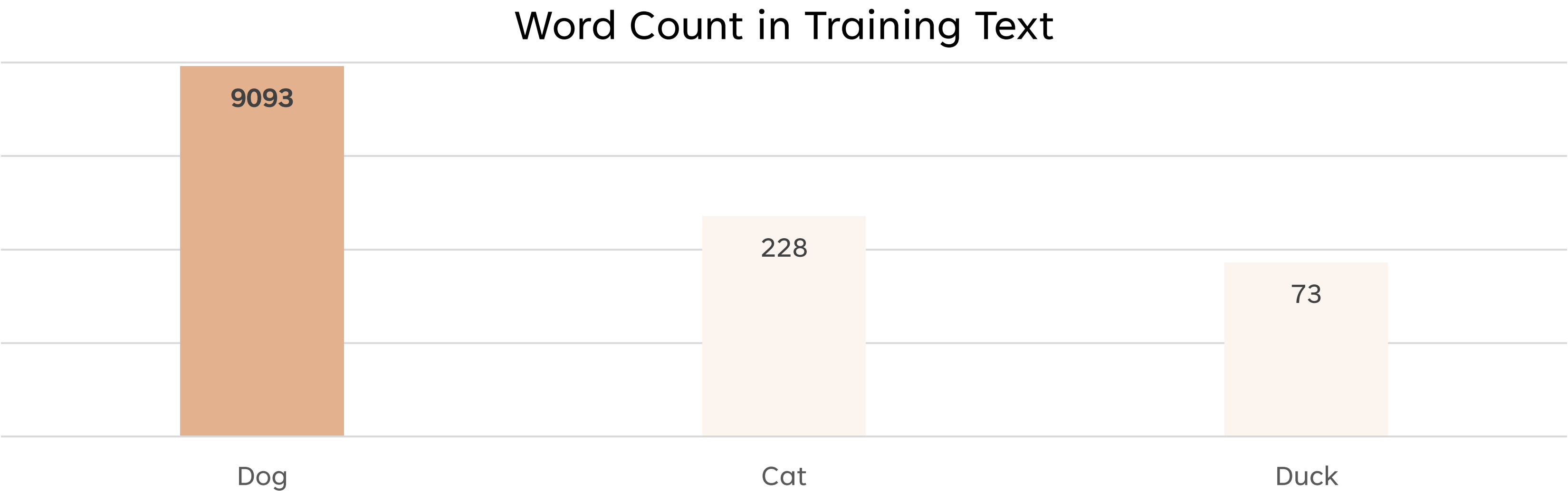
brown dog is running through the sand

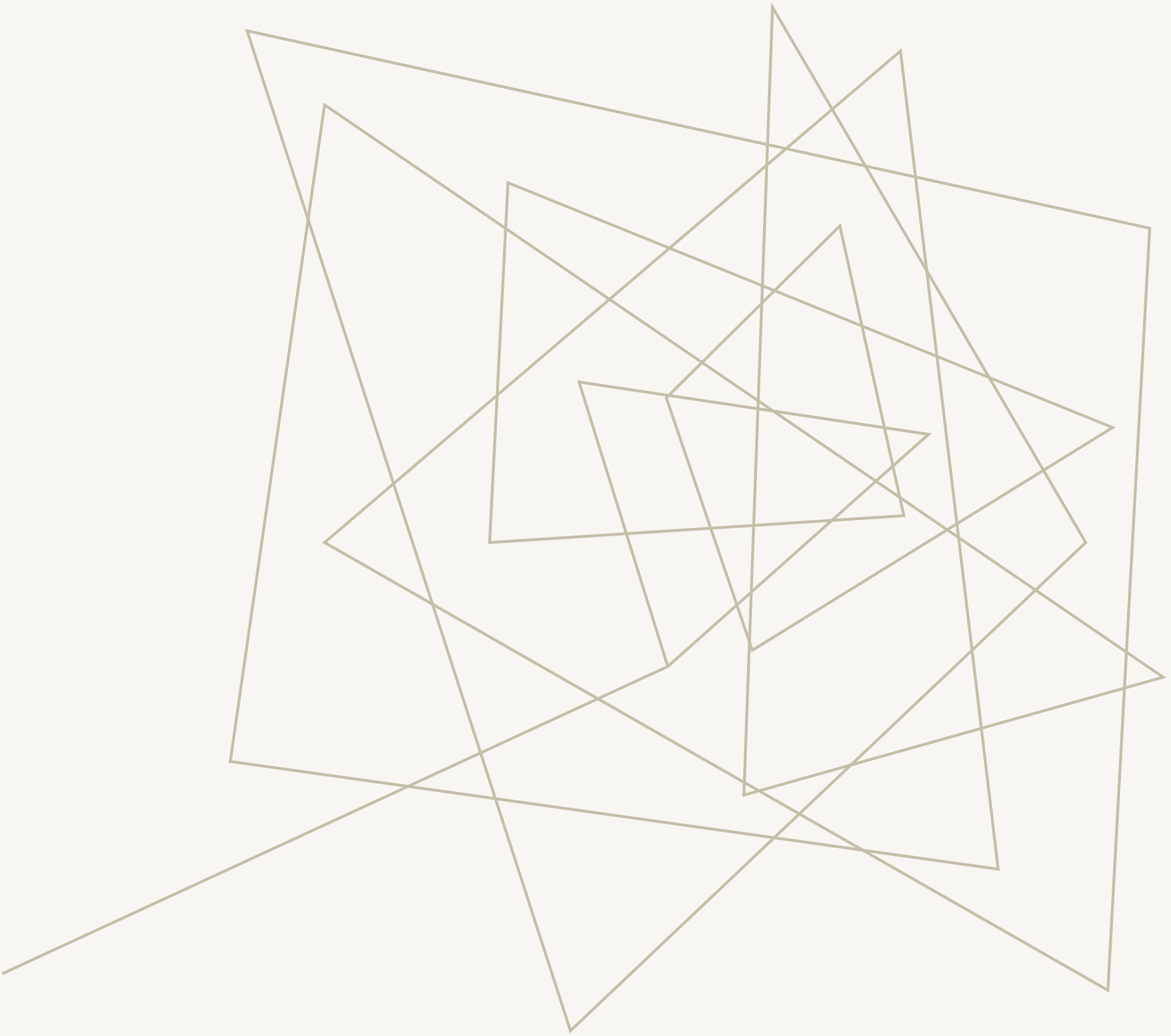


young girl is sitting on the floor



man is standing on the ground with his head





CONCLUSION

FUTURE WORK

CHOICE OF IMAGE ENCODER

Different model with higher prediction accuracy and perhaps more parameters, or utilize transfer learning to fine-tune the model to a specific style of photography

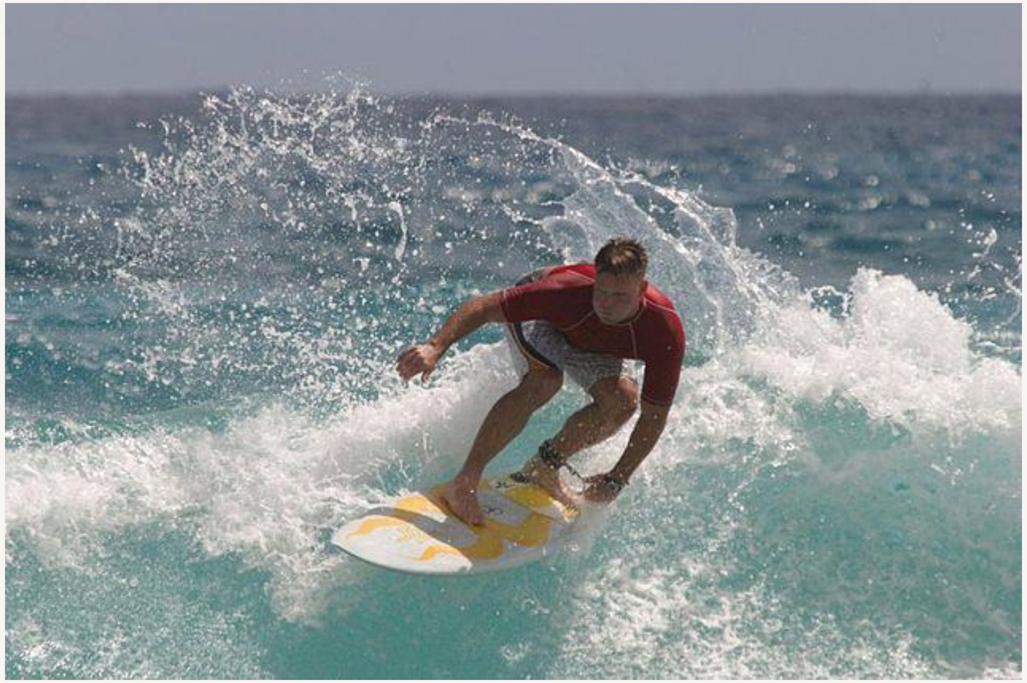
DATASET

Use of larger datasets such as Microsoft's COCO or Google's Conceptual Captions allows greater exposure

HYPERPARAMETER TUNING

Unfortunately I did not have time to perform hyperparameter tuning, which could have boosted the final performance of the model.

IMAGE CAPTION GENERATOR



STEP 1: Caption Generator

Man in red shirt is surfing in the ocean

Completed

STEP 2: Text Generation AI

Man in a red shirt is surfing in the ocean off Australia's Gold Coast, one of the world's top surfing destinations.

Next Step

SUMMARY

I was able to achieve the goal of training a caption generator with a **BLEU-1 score of 0.52** and produces captions that are intelligible.

Navigating my way through the TensorFlow API and reading up on the various aspects of deep learning was an eye opening experience, and perhaps a fitting project to mark the end of the Data Science Immersive and the start of a longer data science journey ahead.

THANK YOU

Sing Ee Shawn

✉ eeshawn11@gmail.com

https://github.com/eeshawn11/

in https://www.linkedin.com/in/shawn-sing/