IMAGE CAPTION GENERATOR

By

K JIVITESH NARAYAN - 18BCE1290

SAI MAHESH V - 18BCE1107

A Project Report Submitted to

Dr. BHARADWAJA KUMAR

SCHOOL OF COMPUTER SCIENCES AND ENGINEERING

In partial fulfilment of the requirements for the course of

CSE4022

NATURAL LANGUAGE PROCESSING

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

VIT UNIVERSITY, CHENNAI CAMPUS

Vandalur-Kelambakkam Road

Chennai-600127

May 2021

BONAFIDE CERTIFICATE

Certified that this project report entitled “Image Caption Generator” is a bonafide work of **Jivitesh Narayan K (18BCE1290)** and **Sai Mahesh V (18BCE1107)** who carried out the J-component under my supervision and guidance.

Dr. BHARADWAJA KUMAR

Professor

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING (SCSE)

VIT UNIVERSITY, CHENNAI CAMPUS
CHENNAI-600127

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr. Bharadwaja Kumar**, Professor, SCSE, for his consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We are extremely grateful to the Dean of the SCOPE, VIT Chennai, for extending the facilities of the school towards our project and for the unstinting support. We also take this opportunity to thank all the faculty of the school for their support and their wisdom imparted to us throughout the course.

ABSTRACT

Images are a very effective visual aid for any story. But sometimes, the image by itself does not convey the entire context. Some images may look out of place when observed at a glance, but when given a proper setting and a caption, they start to make more sense. For example, an image might depict two individuals talking, but the topic of the conversation or the conversation itself would be a mystery to anyone viewing the scene. But, with the introduction of embedded captions that show the conversation, the viewer would be able to gain a clearer picture of the events leading up to and happening during the time the image was captured or created. This also helps the viewer of the image to understand and enjoy the meaning behind the image. Another branch of images which would benefit a lot from captions added to them are ones which would convey a sense of humour to the user, as the caption or the image can be used as a punchline while the other provides the baseline. The project deals with images of this kind, and to provide captions that make linguistic sense, natural language processing is crucial. Hence we combined the power of NLP with deep learning to generate a model that can learn the general format of the captions for a given image and generate original captions for the same image while conforming to the linguistic rules set by the English language.

TABLE OF CONTENTS

KEYWORDS

I. **LSTM - Long short-term memory** - Long short-term memory is an artificial recurrent neural network (RNN) architecture used in the field of deep learning.

II. **Bi-Directional LSTM** - Bidirectional LSTM, or biLSTM, is a sequence processing model that consists of two LSTMs - one taking the input in a forward direction, and the other in a backwards direction. This increases the amount of information available to the network, which, in turn, improves the context available to the algorithm.

III. **Deep Learning -** Deep learning is a type of machine learning and artificial learning that imitates the way humans gain certain types of knowledge, and is an important element of data sciences such as statistics and predictive modeling.

IV. **GPT-2 -** GPT2, or generative pre-trained transformer 2, is an open-source artificial intelligence created by OpenAI to facilitate the prediction of words in internet text. It displays a broad set of capabilities, including the ability to generate conditional synthetic text samples of extremely high quality.

V. **Fine-Tuning -** Fine-tuning is the adjustment of the parameters of the GPT-2 model to suit the new, unseen task. It can be seen as further training the model beyond the training done by OpenAI themselves in order to improve the model’s performance in specific tasks.

INTRODUCTION

Computers have struggled to do tasks that require a human touch up until recently. With the development of multi-layer perceptron and Deep learning concepts, computers acquired the ability to do human tasks. One such task is image captioning. In order to caption an image one needs to know the objects in the image, relationship between the objects and context of the image. Meme creation is a subset of image captioning. In meme generation, a particular format contains a certain sentiment. So in order to caption a meme, all you need to do is under the sentiment and generate the caption based on the understood caption.

In this project, we developed a system to generate humorous captions for a given image/meme format. We generated a dataset by web scraping the internet for a particular meme format and extracted the caption for that meme. This system is made using fine-tuning the experimental release of the large-scale unsupervised language model released by OpenAI, GPT-2 (117M). This method yielded the best result out of other methods (LSTM and Bi-directional LSTM).

DATASET

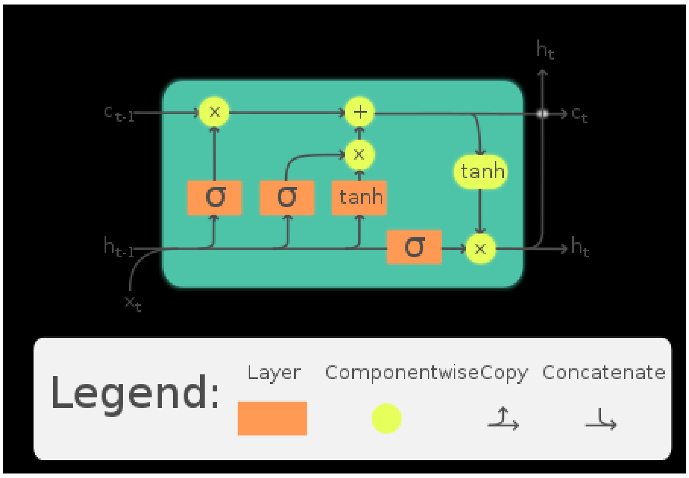
A custom dataset was generated for the use of this project. To achieve this, we scraped an online database of humorous captioned images, known as memes, through their webpage called “meme generator”. To ensure the proper working and focused performance of the model, each meme format was retrieved individually and stored in a separate file for the model’s usage. The size of the datasets for each of the formats ranged from 350 entries to 20000 entries, allowing the observation of the model’s reactions to widely varying dataset sizes.

METHODOLOGY

We used 3 methods to develop the text generation model for generation of the caption. They are 1) LSTM , 2) Bi-Directional LSTM and 3) GPT 2 (117M) fine-tuning.

LSTM

LSTM stands for long-short term memory. It is an RNN but overcomes its drawbacks (vanishing or exploding gradient problem). There are sigmoid and tanh activation functions used inside the layer. In the model built in the project, we use ReLU to connect the LSTM layer to other layers.



In this model, we performed tokenization, vocabulary for characters.

Modules: Keras, numpy and tensorflow

Level of prediction: character level

Model architecture:

Model: "sequential"

```
_________________________________________________________________
Layer (type)            Output Shape          Param #
=================================================================
lstm (LSTM)               (None, 100, 256)        264192

_________________________________________________________________
dropout (Dropout)           (None, 100, 256)        0

_________________________________________________________________
lstm_1 (LSTM)               (None, 256)             525312

_________________________________________________________________
dropout_1 (Dropout)           (None, 256)             0

_________________________________________________________________
dense (Dense)             (None, 46)              11822
=================================================================
```

Total params: 801,326

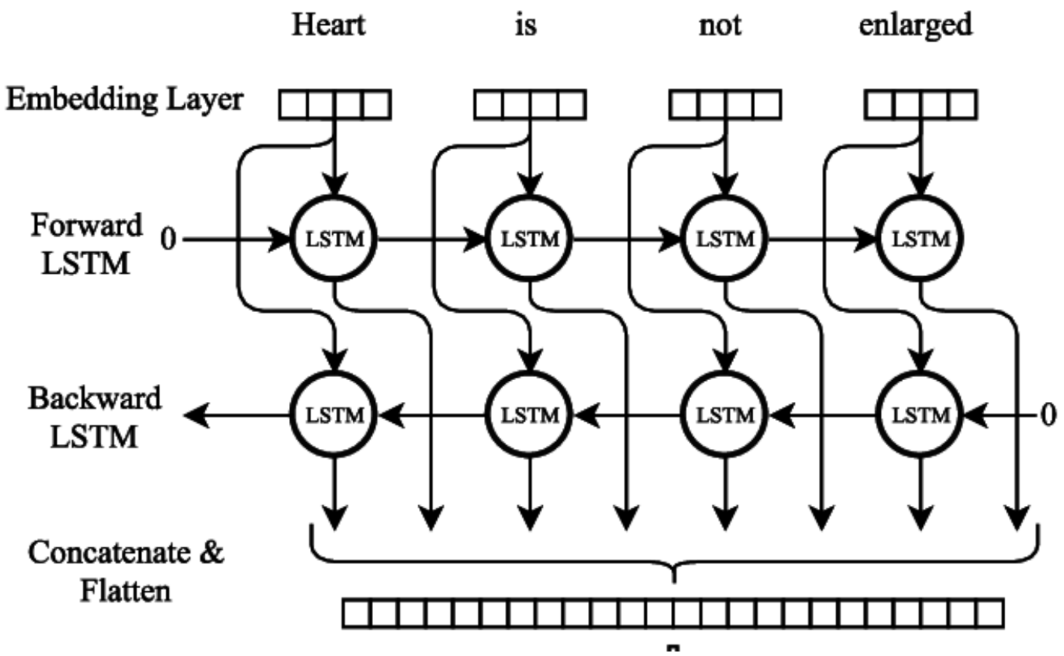
Trainable params: 801,326

Non-trainable params: 0

____

Bi-Directional LSTM

BiLSTM is just 2 LSTM layers but one in the forward direction and the other in the backward direction. This approach increases the sheer volume of data available to the model and increases the context availability



In this model, we performed tokenization, padding and vocabulary for words(word embedding).

Modules:keras, tensorflow, matplotlib, PIL

Level of prediction: word level

Model architecture:

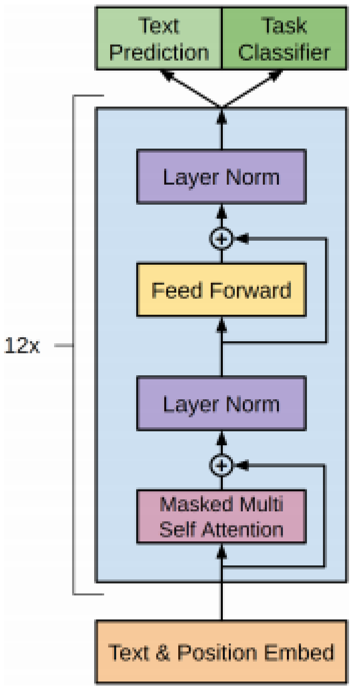
Model: "sequential"

```
________________________________________________________________
Layer (type)              Output Shape            Param #
=================================================================
embedding (Embedding)       (None, 49, 100)          295500
________________________________________________________________
```

```
bidirectional (Bidirectional (None, 49, 300)          301200

_________________________________________________________________

dropout (Dropout)           (None, 49, 300)          0

_________________________________________________________________

lstm_1 (LSTM)               (None, 100)              160400

_________________________________________________________________

dense (Dense)               (None, 1477)             149177

_________________________________________________________________

dense_1 (Dense)             (None, 2955)             4367490
=================================================================
Total params: 5,273,767
Trainable params: 5,273,767
Non-trainable params: 0

_________________________________________________________________
```

GPT 2 (117M) Fine-tuning

GPT 2 is a large-scale unsupervised language model which generates coherent sentences about arbitrary topics. This model can be fine-tuned to specific tasks. OpenAI developed GPT 2 with 40GB of text data and released a sample version as open source for researchers and academics to experiment. GPT 2 is a generative transformer model.



In this model, we used pos-tagging, fine-tuning

Modules used: gpt_2_simple, nltk, PIL

CODE

The notebooks are provided in the zip file.

LSTM

It will be titled as “LSTM.ipynb”

Bi-directional LSTM

It will be titled as “BiLSTM.ipynb”

GPT Fine-Tuning

It will be titled as “GPT.ipynb”

RESULTS

LSTM

Small model: less number of lstm layer

Large model: more number of lstm layers

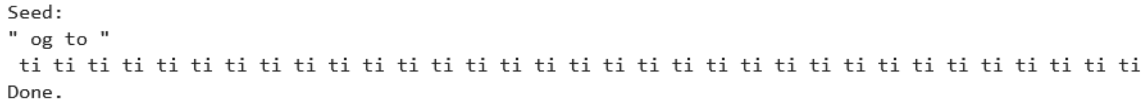
1st try:

No of characters generated: 100

Model: Small model

No. of epochs trained: 20 epochs.

Output:



```
Seed:
" og to "
 ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti ti
Done.
```

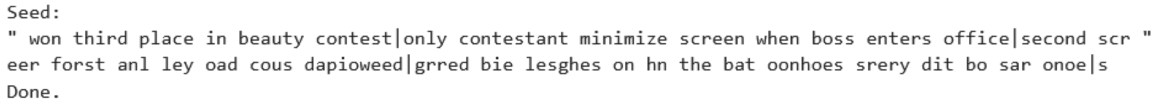
2nd try:

No of characters generated: 100

Model: Small

No of epochs trained: 50 epochs

output:



```
Seed:
" won third place in beauty contest|only contestant minimize screen when boss enters office|second scr "
eer forst anl ley oad cous dapioweed|grred bie lesghes on hn the bat oonhoes srery dit bo sar onoe|s
Done.
```

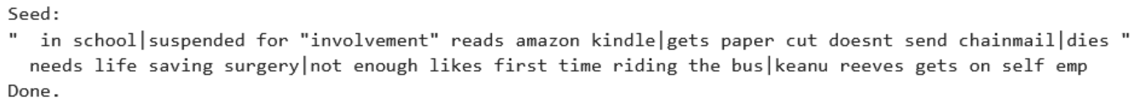
3rd try:

No of characters generated: 100

Model: Large

No of epochs: 100

Output:



```
Seed:
"  in school|suspended for "involvement" reads amazon kindle|gets paper cut doesnt send chainmail|dies "
  needs life saving surgery|not enough likes first time riding the bus|keanu reeves gets on self emp
Done.
```

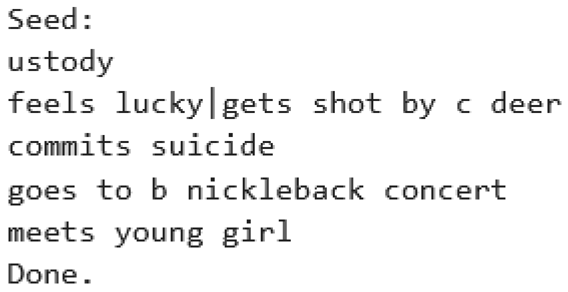
4th try:

No of characters generated: 100

Model: large

No of epochs: 100

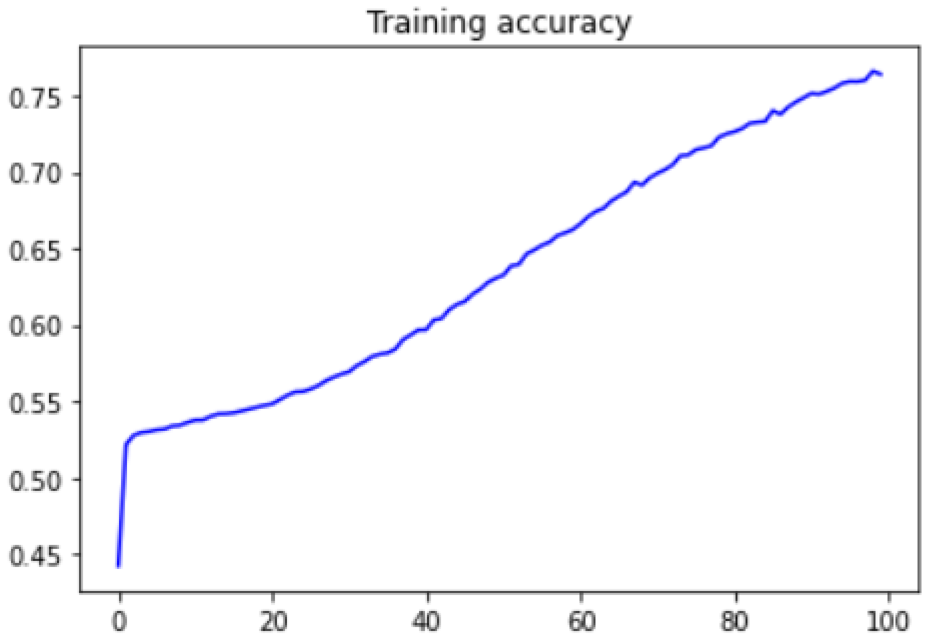
Output:

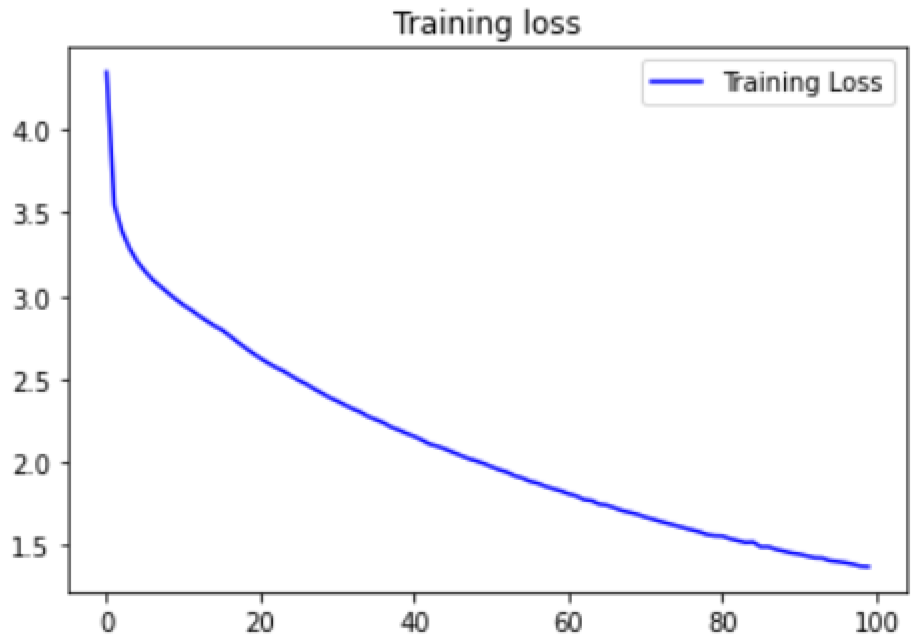


```
Seed:
ustody
feels lucky|gets shot by c deer
commits suicide
goes to b nickleback concert
meets young girl
Done.
```

Drawback: the model is overfitted as the generated text already exists in the dataset.

Bi-Directional LSTM





Output:

Sample from dataset: Textbooks y u no have CTRL-F

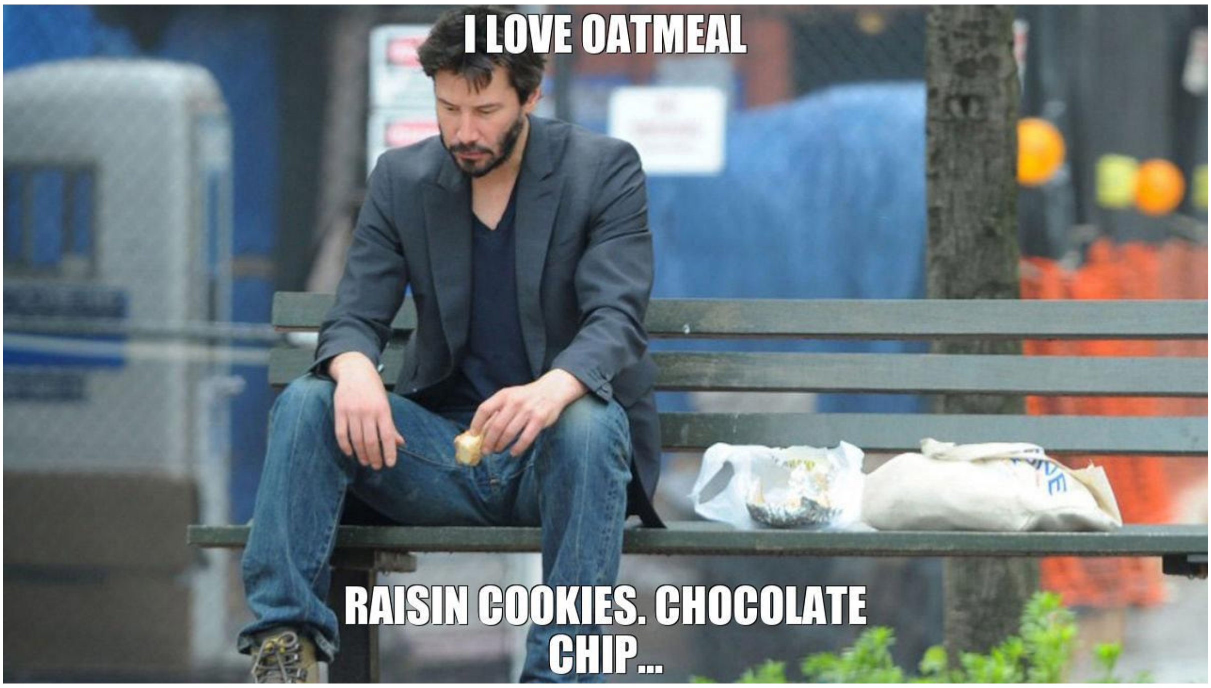
Seed = flavor

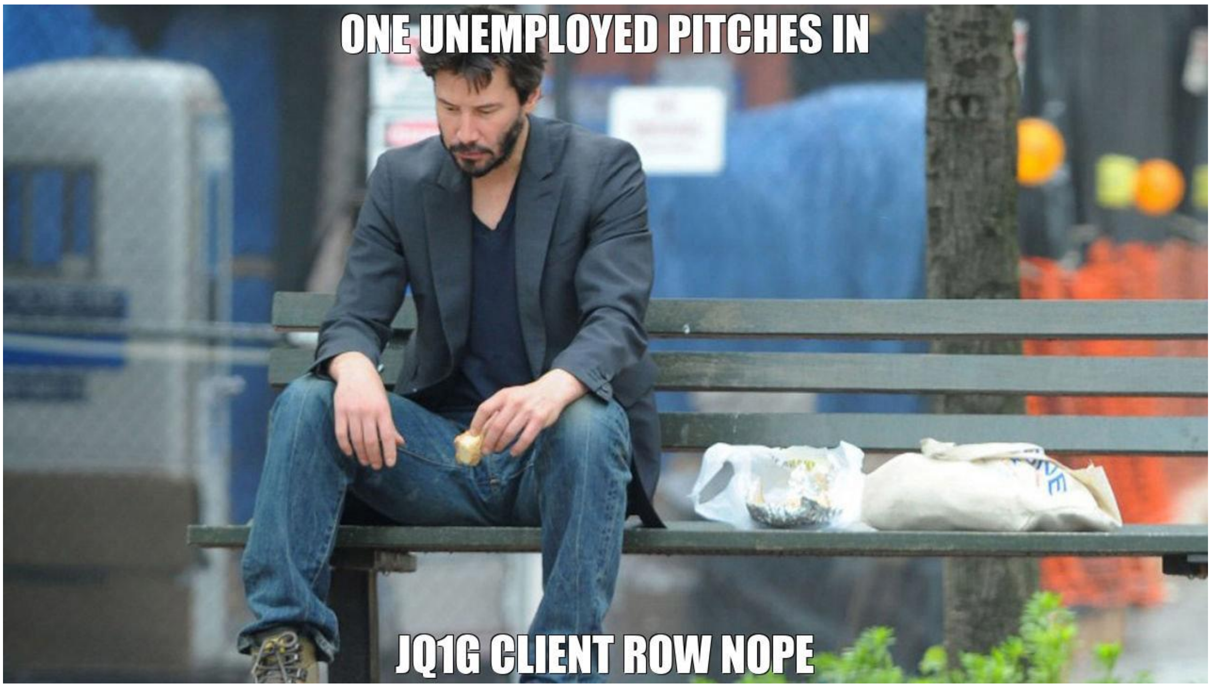
Generated text: flavor u no txt y u no talk 2 me no more first years of spot through morning shows truck morning

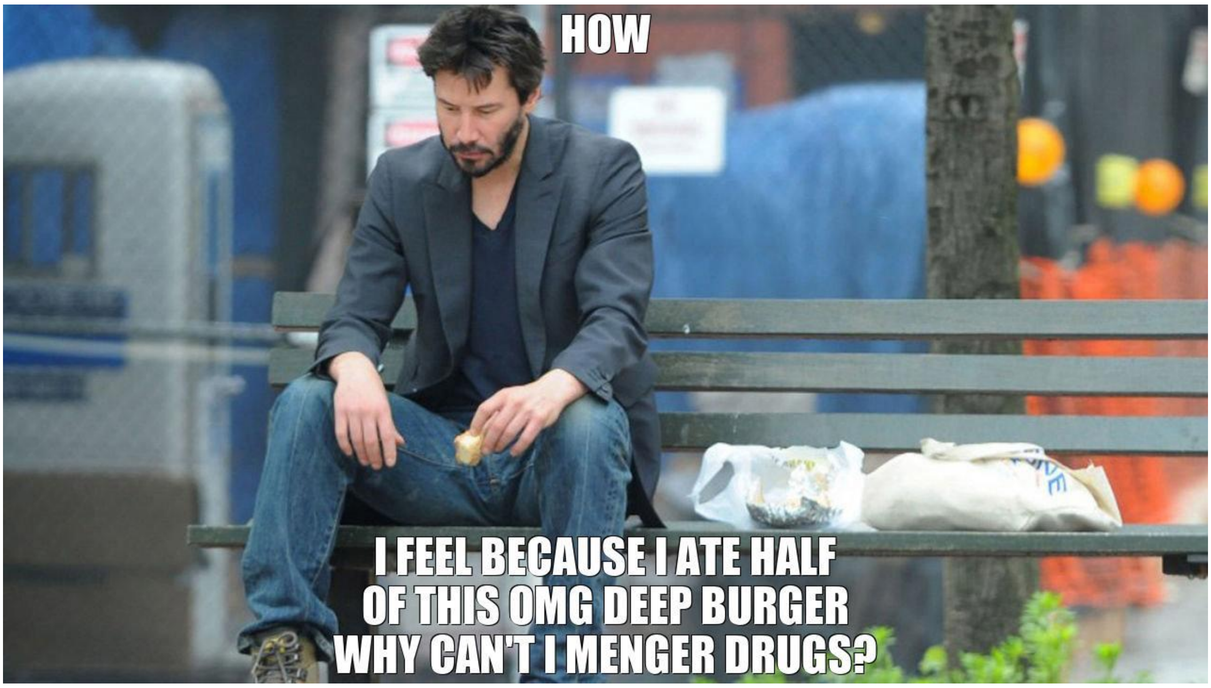
Drawback: the model is underfitted as the generated text shows little variance with different seeds.

GPT Fine-Tuning

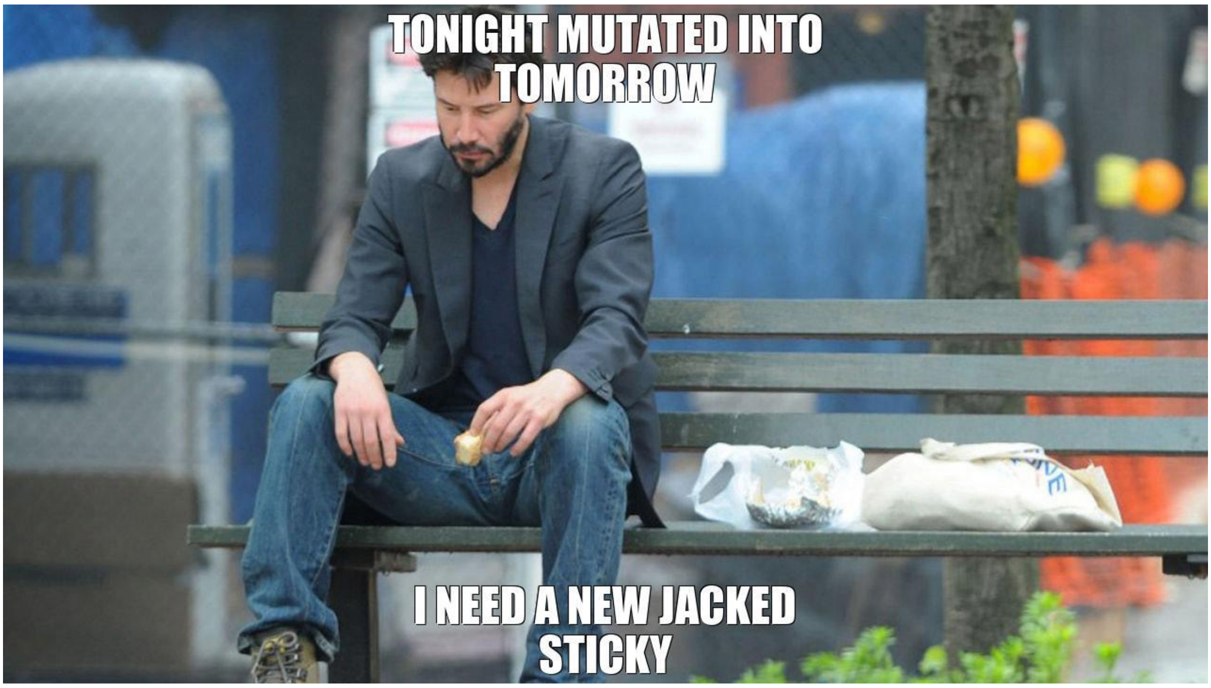
Output:






HOW
I FEEL BECAUSE I ATE HALF
OF THIS OMG DEEP BURGER
WHY CAN'T I MENGER DRUGS?

TODAY IS HOLI
BILL <-- BEEF



CONCLUSION

Adding captions to images allows the images to target a wider range of audiences, and can be used by multiple companies to use a single image to construct varying storylines. The project yields a conclusion that while the datasets can produce working results, the improvements in accuracy and quality of the output generated, such as the captions being more humorous and meaningful, when using datasets of increasing size denotes that the performance of the model is directly proportional to the size of the dataset to some extent.