**Image Caption Generation**

-Mayurkumar M patel(online)

-Tirth Shah

**Introduction:**

Image captioning is one of the most often used technologies in the current day. Additionally, there are built-in tools that use machine translation to automatically create and deliver a caption for a specific image. Picture captioning is the process of creating a description for an image. Recognizing the significant things, their characteristics, and the connections between the objects in a picture are necessary. It produces statements that are both syntactically and semantically sound. We introduce a learning strategy that combines machine translation, text categorization and computer vision to characterize pictures and provide captions. This study seeks to identify the various things present in an image, understand their relationships, and produce captions. The key components for creation of picture captions are computer vision and natural language processing. The usage of image segmentation by Facebook and Google Photos and, more importantly, the extension of its use to video frames, may both be utilized by image caption generators. They'll be able to effortlessly automate the task of an image interpreter. Not to mention that it has a huge potential for assisting those who are blind.

1. The real-world significance of the project:

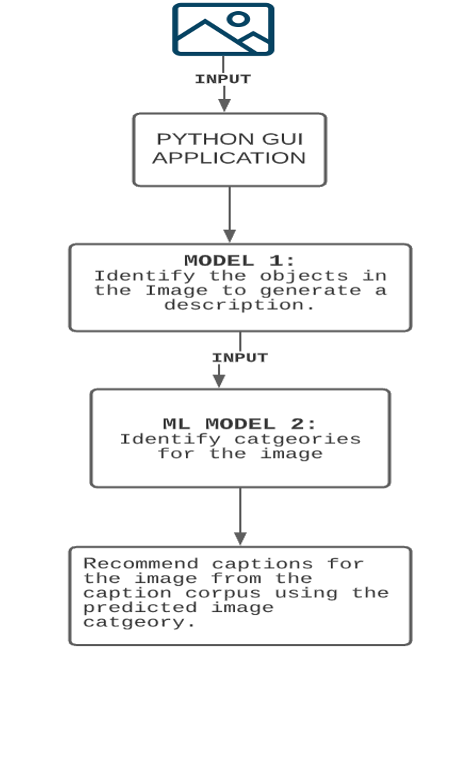
* Currently, there are several NLP apps available that can extract insights or summaries from a given text data set, such as an essay. People who would profit from automatic insights from photographs can get the same advantages.
* Explaining what occurs in a film, frame by frame, would undoubtedly be a little longer term use case.
* Social media Platforms like Facebook may instantly deduce your location (such as a seashore or bistro), what you're wearing (such as a specific color), and—more importantly—what you're doing from a photograph (in a way). To better understand, view an illustration.
* The flexibility to function as a "Image Caption Generator" for social networking sites and other image-based platforms. Social networking sites like Instagram, Facebook, Twitter, and Pinterest all rely heavily on how well-written captions and hashtags are used in the image descriptions. As users upload photographs into the environment, our algorithm evaluates the items in the image to provide categories that may be used as "hashtags" and further suggests descriptions to the user.

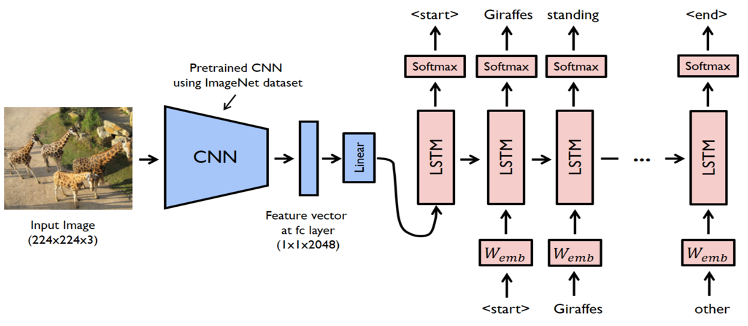
1. Methods and an overview of the tools used:

For first, we tried InceptionV3 and VGG16 model for image processing, both were a little heavier to run on an 8000-image dataset and hard to use for training as the accuracy generated was low. Later, we resided with Resnet 50, as it is a smaller and faster model as compared to the other two. Apart from that, the initial process was to use the dataset, so we used a Flicker 8k Dataset, which consists of 8K images paired with 5 captions each, which provides proper description of the image. For text-categorization, through which we are labeling the images, in categories like ‘Nature’, ‘Adventure’ etc., in that part we created a DNN model based on YouTube API data for text categorization which consists of 20K Titles (As shown in below in the image). For training this model, first we tokenize and vectorize the dataset and then use ‘Google News Vectors Negative 300’ for weights for Word2Vec and KeyedVectors and then pass these encodings to DNN for model’s training and predicting categories.

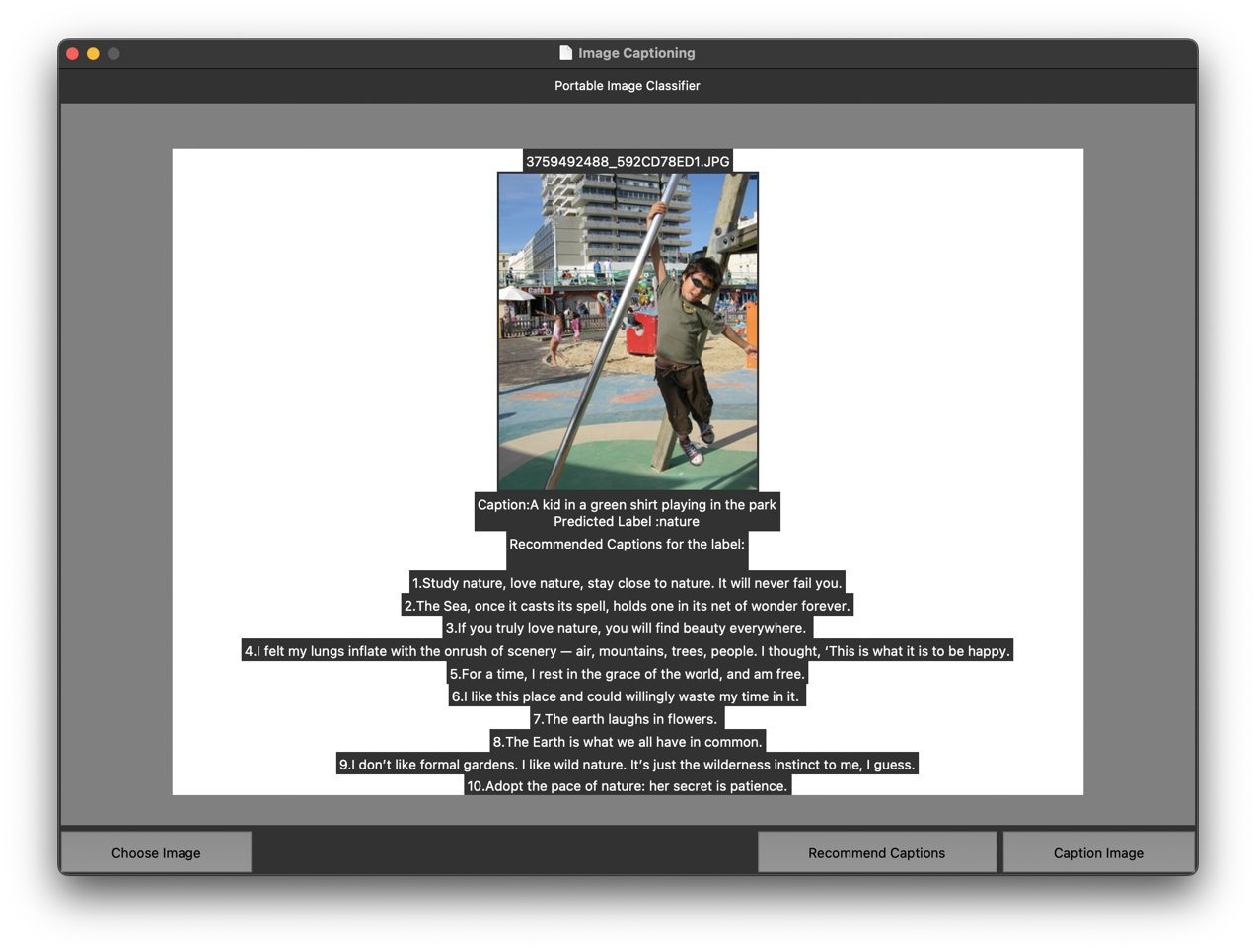


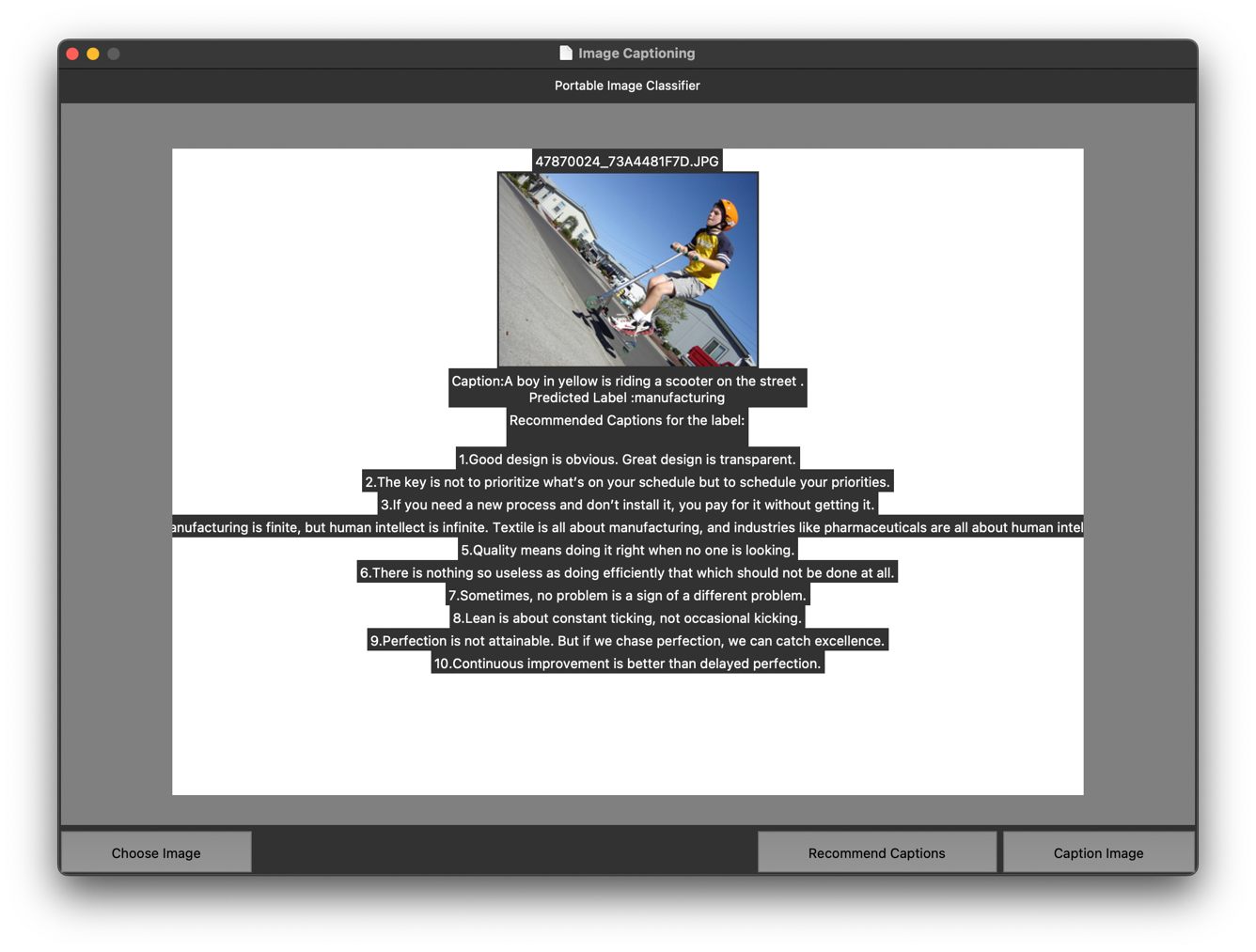
After that, we trained the Image Processing model i.e., based on Flicker 8K dataset, we used Resnet 50 and encoded the images to vectors and passed those image encodings to LSTM layers for generating captions based on it. And to test the same/ use the same, we created a front-end/GUI which is shown below, where in you can upload an image and generate a caption, see the category for the image and also get recommended captions, which are nothing, but a set of captions pre-defined for the labels used.





Sample Outputs:





1. Tools used:
2. Python (libraries used):

* NumPy
* Pandas
* Tkinter
* Scikitlearn
* NLTK
* TensorFlow
* Keras

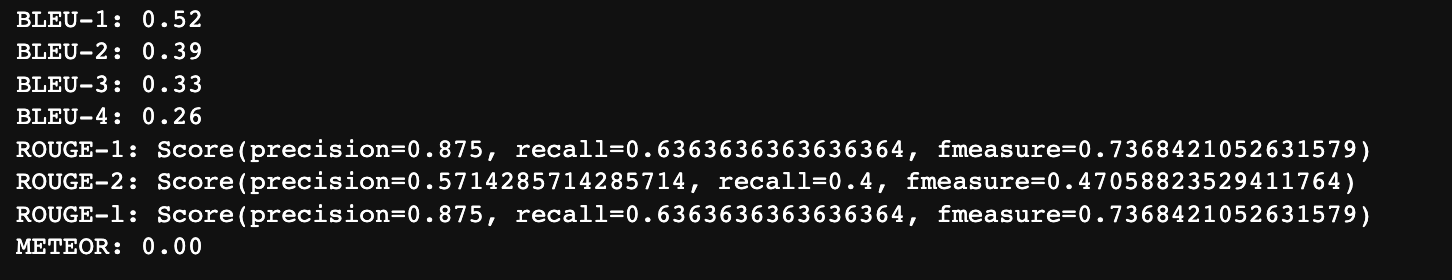
1. VS Code
2. Kaggle (For model training)
3. Dataset used:

We used ‘Flicker 8K dataset’ for Images which included captions/ description of the image which consists of 8000 images with 5 captions for each image. This dataset was created by University of Illinois at Urbana-Champaign. And for text-categorization, we used ‘YouTube API data for text categorization’, which is made available by Kaggle. It consists of 20K Titles, and is based on Video Descriptions, Titles, categories etc.

1. The results of your evaluation:

Based on the training and evaluation, we were able to rectify that the model is not that accurately identifying the objects in the image and often misclassifies some categories.

The evaluation metrics showed the same. We used BLEU (Bilingual Evaluation Understudy Scores) and Rouge (Recall-oriented Understudy) scores for evaluation.



The root cause to get low accuracy was the lack of availability of computing power and the ability to use a bigger dataset for model training without having a time constraint i.e., limited availability of labeled data.

1. Discussion of the significance of your results – what lessons can be gleaned:

Based on the results we achieved, one can realize that there are some more models that can be used to train the model, like the new ResNet101 which is a heavier and takes more time to train, but gives better accuracy. Apart from that, the text-categorization can be improved by taking a better dataset, as the YouTube data may not be grammatically that great (it may have spelling errors or incomplete sentences) which may have led to inconsistent classification of categories.

1. What would one can extend or do differently in the future:

One can use the same method over a better, powerful machine, and use a Flicker 30K dataset over 8K and also use ResNet101 for more accuracy, but instead of using ‘YouTube API data for text-categorization’, one should try a different way for getting the labels, as this dataset is a reason for misclassification of categories, the dataset has to be bigger than the one used currently. As, based on the previous versions of the same, the accuracy might be less for text-categorization but it is a concept that is done differently.

**Citation:**

1. [Wang, Haoran, Yue Zhang, and Xiaosheng Yu. "An overview of image caption generation methods." *Computational intelligence and neuroscience* 2020 (2020).](https://www.hindawi.com/journals/cin/2020/3062706/)
2. [Ding, Songtao, Shiru Qu, Yuling Xi, Arun Kumar Sangaiah, and Shaohua Wan. "Image caption generation with high-level image features." *Pattern Recognition Letters* 123 (2019): 89-95.](https://www.sciencedirect.com/science/article/abs/pii/S0167865519301047)
3. [Chen, Xinlei, and C. Lawrence Zitnick. "Learning a recurrent visual representation for image caption generation." *arXiv preprint arXiv:1411.5654* (2014).](https://arxiv.org/abs/1411.5654)
4. [P. Anderson, X. He, C. Buehler et al., “Bottom-up and top-down attention for image captioning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.](https://openaccess.thecvf.com/content_cvpr_2018/papers/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.pdf)
5. [Farhadi, A. et al. (2010). Every Picture Tells a Story: Generating Sentences from Images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds) Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6314. Springer, Berlin, Heidelberg](https://www.cs.cmu.edu/~afarhadi/papers/sentence.pdf)
6. [O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.](https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Vinyals_Show_and_Tell_2015_CVPR_paper.pdf)
7. [Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *International conference on machine learning*. PMLR, 2015.](https://arxiv.org/pdf/1502.03044.pdf)
8. [J. Aneja, A. Deshpande, and S. Alexander, “Convolutional image captioning,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.](https://openaccess.thecvf.com/content_cvpr_2018/papers/Aneja_Convolutional_Image_Captioning_CVPR_2018_paper.pdf)