

ViSiX - A Deep Learning Model for Automatic Image Captioning using InceptionV3 and Long Short-Term Memory Recurrent Neural Networks with Visual Attention

By

Vignesh Baalaji S | Sandhya Sridhar | Sadhana B

20MAI1002 | 20MAI1012 | 20MAI1007


School of Computer Science and Engineering

Vellore Institute of Technology

Chennai, India.



Abstract

- Image Captioning is a method for enabling the machine to generate a description of the image fed to it.
 - In other words, we can say that it enables the machine to obtain a perception of the image.
 - In this project, we propose a deep transfer learning model, titled. ViSiX, for image captioning using Inception V3 as the feature extractor and Bidirectional LSTMs for generation of captions based on the extracted feature vectors.
 - ViSiX is trained and tested using the Flickr 8K dataset and MS COCO Dataset separately.
- 

Literature Survey

S. No	Reference	Methodology	Results
1	DeepCap: A Deep Learning Model to Caption Black and White Images	This approach uses pre trained model Inception V3 and LSTM for captioning black and white images.	Accuracy = 45.77%
2	Image Caption Generation using Deep Learning Technique	This proposed model consist of Convolutional Neural Network for feature extraction from images and Recurrent Neural Network for caption generation.	BLEU = 53.35%
3	Deep Learning for Military Image Captioning	This generative model is based on a deep recurrent architecture combined with pre trained image-to-vector model Inception V3 via CNN and word2vec model via skip-gram model.	BLEU = 48.5%
4	Show and Tell: A Neural Image Caption Generator	This model is based end-to-end on a neural network consisting of a vision CNN followed by a language generating RNN. The model is trained to maximize the likelihood of the target description sentence given the training image	BLEU = 66%
5	Image Captioning using Deep Neural Architectures	This approach uses pre trained model Inception V3 ,word embedding given to series of LSTM .	BLEU = 65%

Dataset Description

About Flickr 8K Dataset

- A new benchmark collection for sentence-based image description and search, consisting of 8,091 images
- Each image is paired with five different captions which provide clear descriptions of the salient entities and events, giving a total of 40,000 captions.
- The images were chosen from six different Flickr groups, and tend not to contain any well-known people or locations, but were manually selected to depict a variety of scenes and situations
- The dataset consists of train and test images separately, and the model is trained and tested on them.

About Microsoft COCO Dataset

(Common Objects in Context)

- The dataset contains over 82,000 images, each of which has at least 5 different caption annotations with over 80 object categories.
- This amounts to a total of 4,10,000 captions in total.
- Once again, the dataset is split following an 80:20 ratio for training, and validation, respectively.

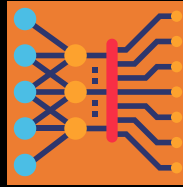
$$\alpha^0 = 1 [a0]$$

Proposed Methodology

$$\arcsin(z)$$

$$x_{n+1} =$$

Modules



Extraction of Feature Vectors

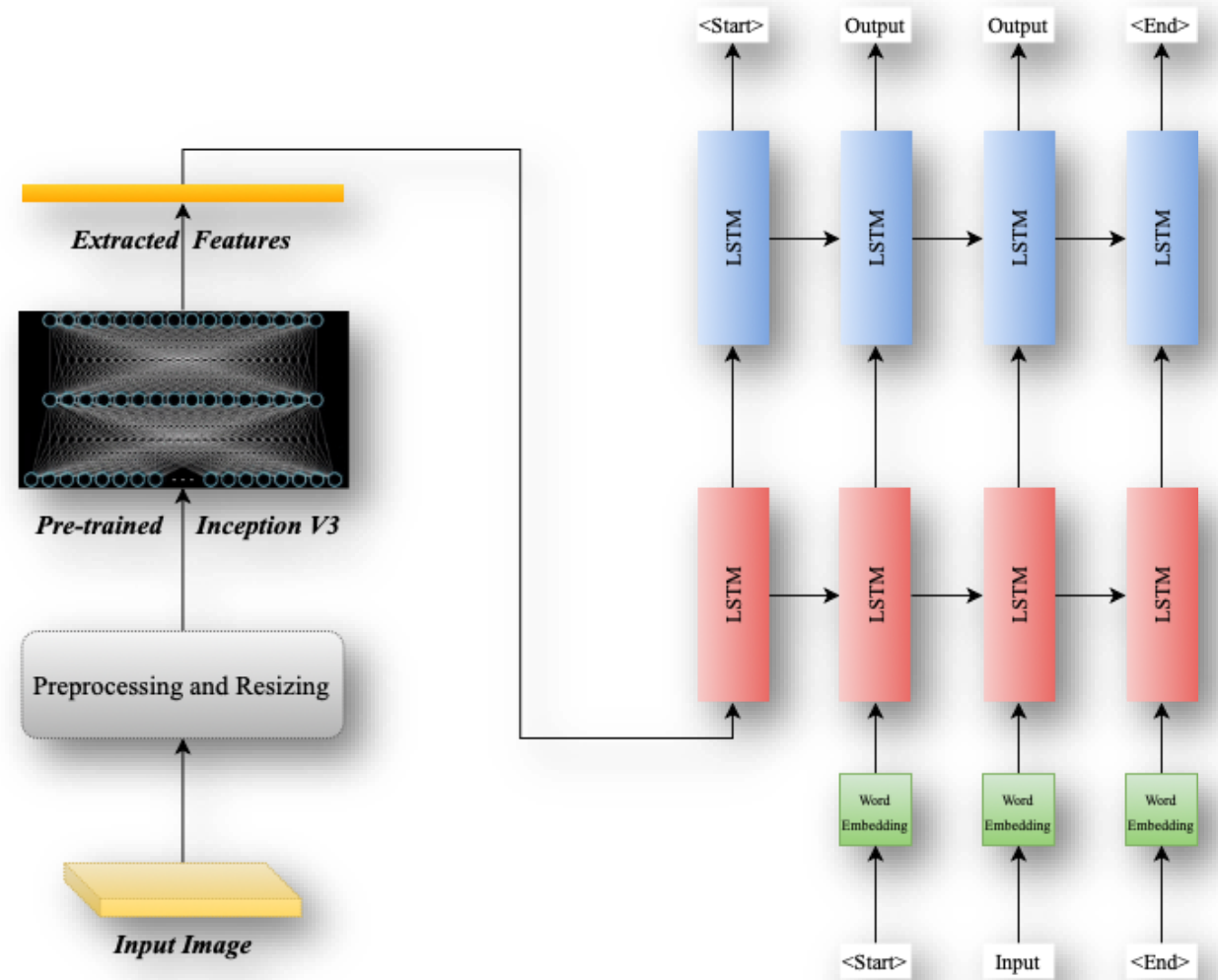


Sequence Modelling



Streamlit Web Application

Model Architecture



Implementation With MS COCO Dataset

A sample of 30,000 images with corresponding captions from the MSCOCO Dataset is used for training our model.

Images are preprocessed by resizing and normalizing them and the inputs are given to the pre trained Inception V3 model for feature extraction. Similarly, the captions are also preprocessed to convert sequences of same length.

After splitting the data into train and validation sets, the data is trained using attention mechanism, which consist of CNN encoder and RNN decoder. After the training, the model is tested with the validation sets.

Additionally, Image Captioning is performed with visual attention called Bahdanau's attention which enables us to see what parts of the image the model focuses on as it generates a caption.

Implementation With Flickr8k Dataset

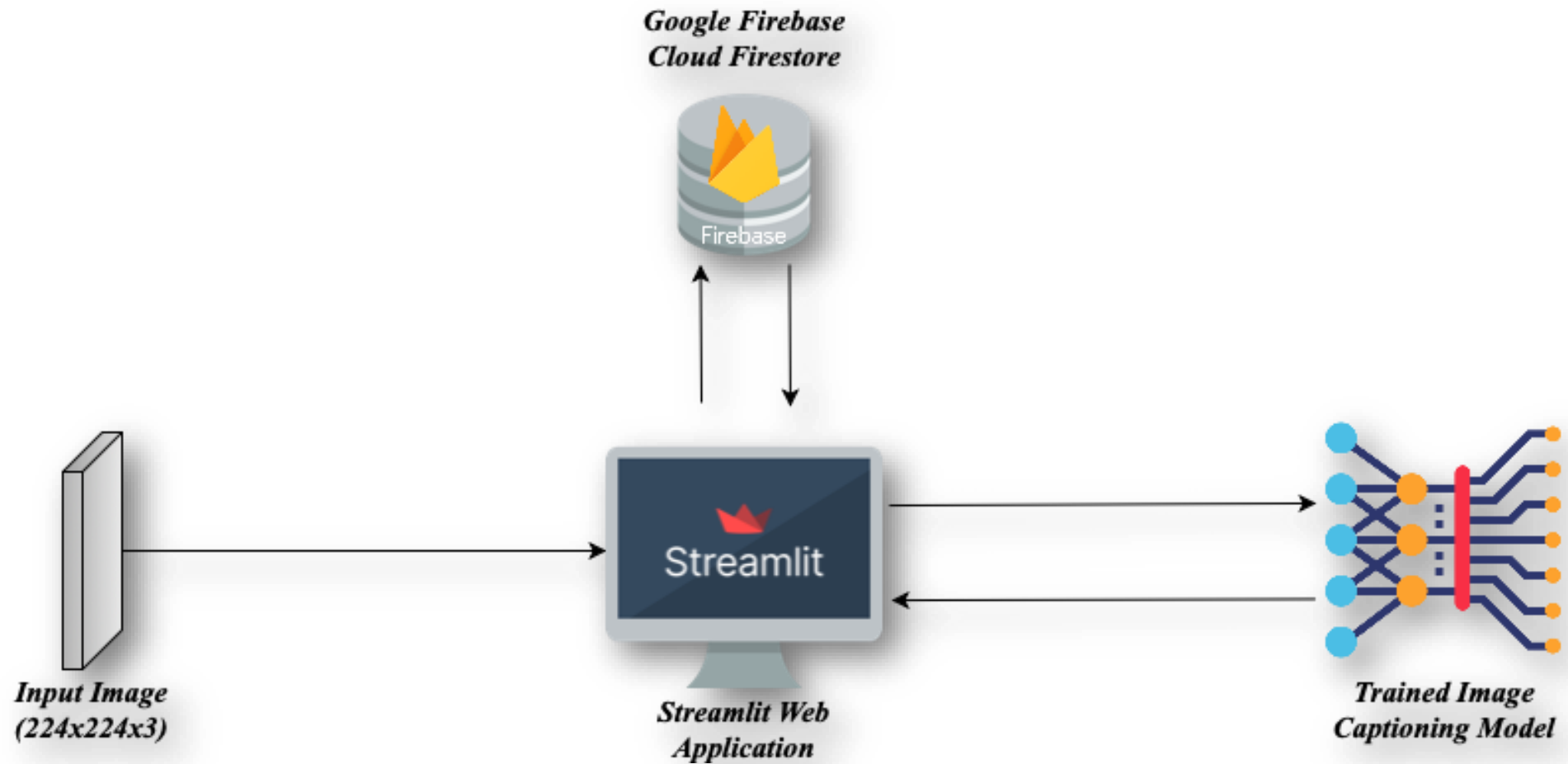
The 8091 images in the Flickr 8k dataset are already split into training and test set.

The transfer learning is done using one of the pre-trained models, Inception v3, and feature extraction is performed.

Loading and processing of captions is performed, followed by standard processes of text data like tokenizing, converting to lower case, removing punctuation etc are applied.

The pre-trained GloVe embedding is downloaded, and GloVe vectors are loaded.

Lastly, the LSTM model for Image caption generation is defined and trained.

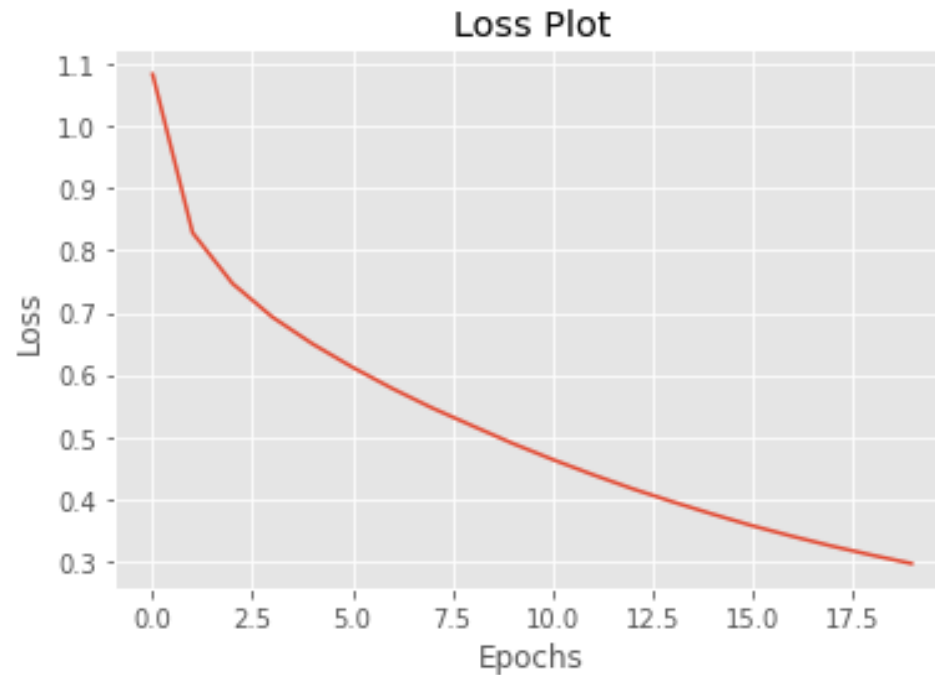


Deployment Architecture



Performance Analysis

Results and Discussion

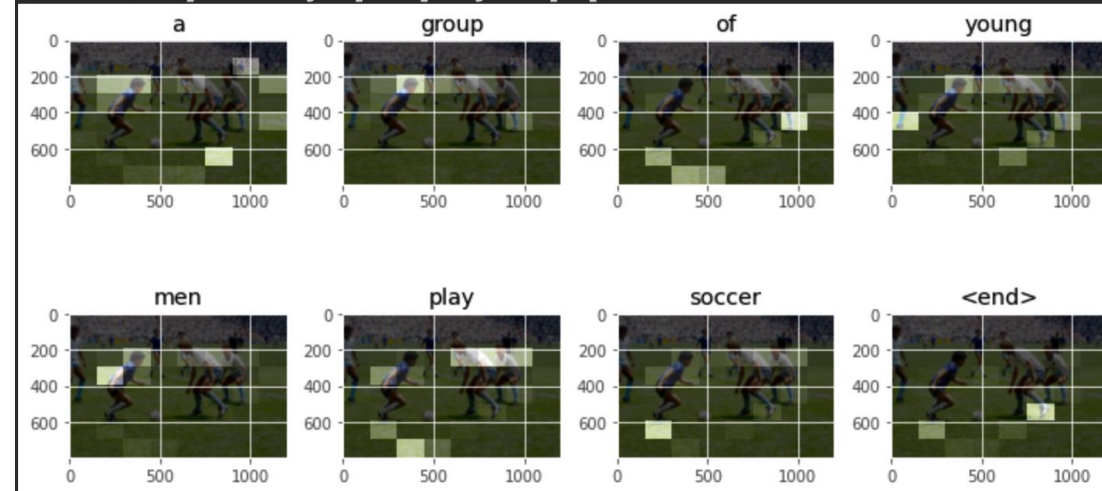


- The model was trained for 20 epochs with Adam as its optimizer and Sparse Categorical Cross Entropy as its loss function.
- The minimum loss achieved after 20 epochs is 0.29.
- A steady decrease in the loss function indicates a perfect convergence towards the minima.
- The performance of the model will be further analyzed in terms of evaluation metrics specific to Natural Language Processing, such as BLEU, which expands as Bilingual Evaluation Understudy

Output of the MS COCO Dataset Model



Prediction Caption: a group of young men play soccer <end>



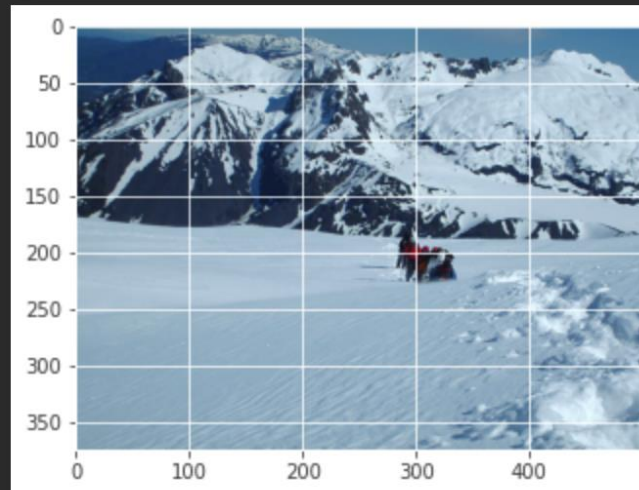
GENERATED CAPTION: startseq man in red and red outfit leans into sharp turn endseq

▶ -0:00



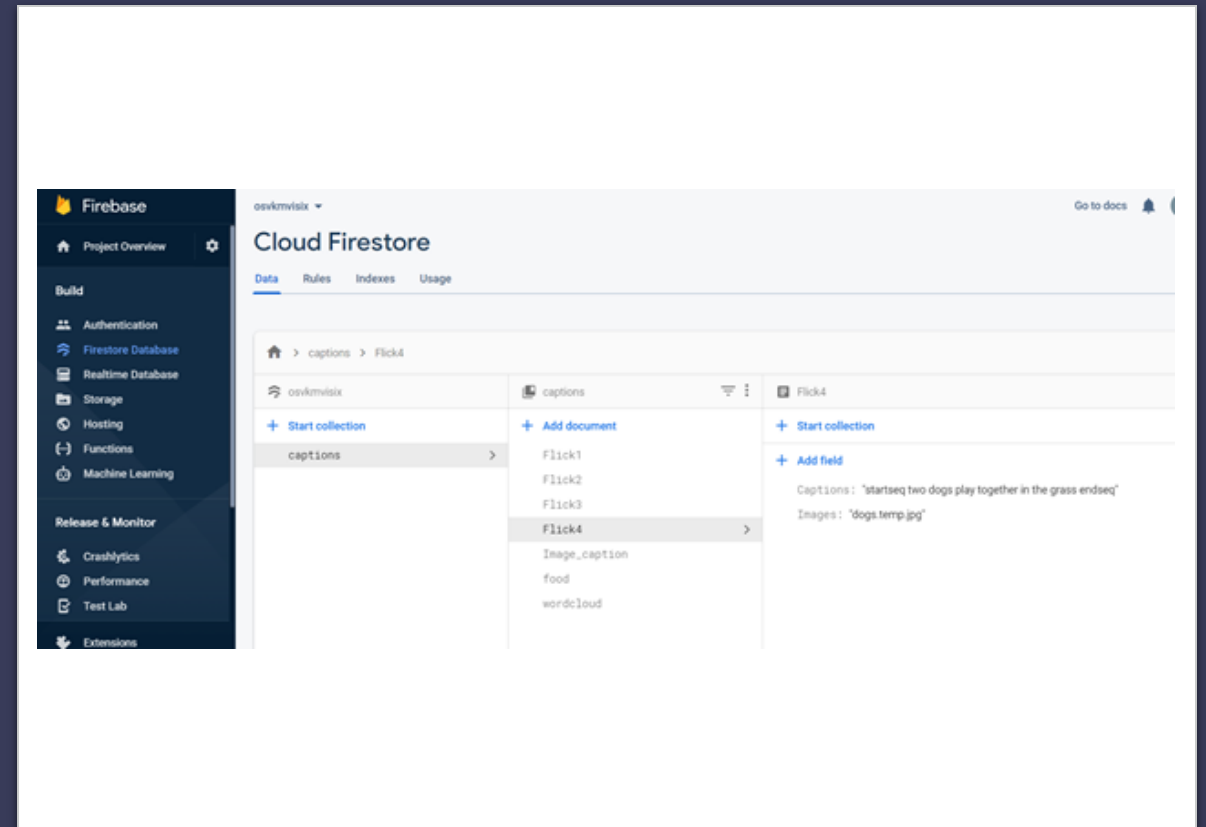
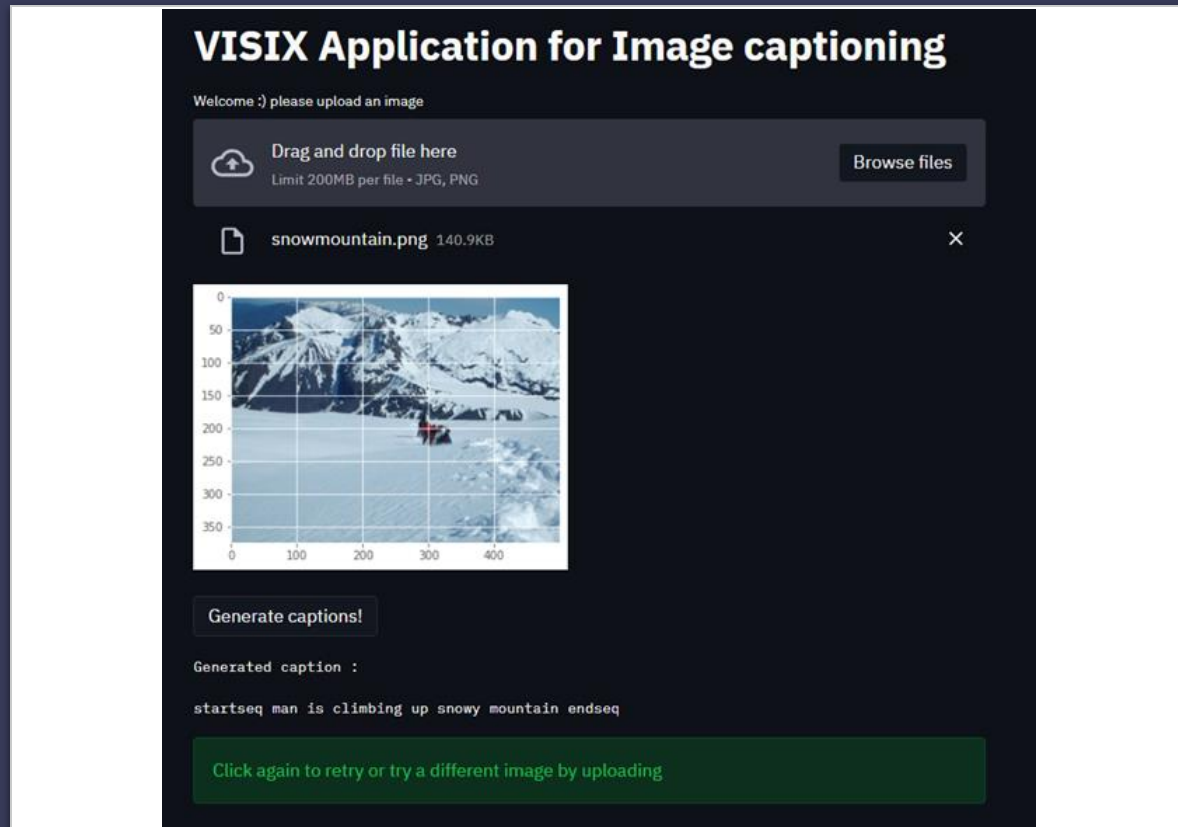
GENERATED CAPTION: startseq man is climbing up snowy mountain endseq

▶ -0:00



Output of the Flickr8K Dataset Model

Streamlit Web Application Screenshot



Project Demo

Let's have a sneak peek into the real-time working of the idea.



Tools and Libraries



Thank You

"Machine Intelligence is the last invention that humanity will ever need to make"

— Nick Bostrom —