International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 04 | Apr 2019 www.irjet.net p-ISSN: 2395-0072

CapSearch - " An Image Caption Generation based search"

Shaunak Baradkar[1], Aditya Bhatia[2], Prasad Gujar[3], and A. Prof. Vinita Mishra[4]

[1,2,3]*Student, Department of Information Technology, Vivekanand Education Society's Institute of Technology*

[4]*Assistant Professor, Dept. of Information Technology, VESIT, Mumbai, Maharashtra, India*

Abstract - *Due to the increasing use of social media, more and more data is collected in the form of images which will be useful only if we can determine what it represents. Searching for images is not as easy as searching text data. We intend to make it easy and accurate by searching for images based on their captions. We have implemented a Flask web app with a machine learning model at its core. The model generates captions using VGG16 and Convolutional Neural Network along with RNN. The model is trained on Flickr-8k dataset. Web app retrieves images based on a user's search query.*

Keywords: Convolution Neural network, Feature extraction, Object recognition, Image Caption Generation

1.INTRODUCTION

Searching for images is not as easy as searching text data. Images can be searched only if there are titles and are classified by humans. Since there are billions of images it is not possible to entitle it by hand every time.

Its purpose is to mimic the human ability to comprehend and process huge amounts of visual information into a descriptive language and store it for future retrieval. Search by example relies solely on the contents of the image. The image is analyzed, quantified, and stored in a database to retrieve similar images.

2. Related Work

Earlier image captioning methods relied on templates instead of a probabilistic generative model for generating the caption in natural language. Present applications use the internet to send/upload image data and then return indexed captions. This requires high bandwidth and data usage which we tend to reduce and optimize for better and fast user experience.

Our model is inspired by Show and Tell which uses GoogLeNet CNN to extract image features and generate captions using Long Short Term Memory cells. We differ from their implementation to optimize for real-time scenarios. Show, Attend and Tell makes use of new developments in machine translation and object detection to introduce an attention-based model that takes into account several "spots" on the image while generating the captions. They extract features from lower convolutional layer instead of extracting from the penultimate layers, resulting in a feature vector of length 14 × 14 × 512 for every image.

2.1 Methodology

Beam Search

Beam Search is an algorithm that explores the most promising predictions that would accompany the image. It takes the top N predictions and sorts them according to the probabilities. Thus it always returns the top N predictions. We take the one with the highest probability and go through it till we encounter <end> or reach the maximum caption length.

CNN for image embedding

The feature extraction of the image is done with the help of CNN's. (Convolutional Neural Networks).We use the VGG16 model for feature extraction.

Generally, CNN is used for image classification. In our case, we use CNN as an encoder for image feature extraction and use its last hidden layer as an input to the RNN decoder(LSTM) for caption(sentence) generation.

RNN

This is the next part of the image caption generation. a recurrent neural network (RNN) is typically viewed as the main generation component. The image features are injected into the RNN.

LSTM

Long Short-Term Memory (LSTM) networks are considered as an extension of recurrent neural networks, which basically extends the memory of RNNs.

In our case, LSTM is predicting the next word in a sentence(caption). Given the initial embedding (feature vector) of the image, the LSTM is trained to predict the most probable next value of the sequence.

2.2 Training the Model

The Flickr-8K contains 8,000 images that are each paired with five different captions which provide clear descriptions.
We used a pre-trained model to interpret the content of the photos. We used the Oxford Visual Geometry Group's VGG16. as part of a broader image captioning

2.3 Developing the Model

Photo Feature Extractor: This is a 16-layer VGG model pre-trained on the ImageNet dataset. We have pre-processed the photos with the VGG model (without the output layer) and will use the extracted features predicted by this model as input.

Sequence Processor: This is a word embedding layer for handling the text input, followed by a Long Short-Term Memory (LSTM) recurrent neural network layer.

Decoder: Both the feature extractor and sequence processor output a fixed-length vector. These are merged together and processed by a Dense layer to make a final prediction.

3. WORKFLOW

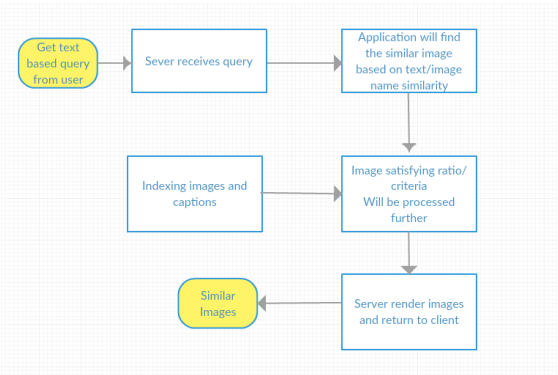

Diagram -1: Block Diagram of Workflow

Step 0: Generating and Indexing of Images and Captions.

We encode the data path in our local system and its caption and send them to the server so that it indexes it and becomes available for search.

We used VGG16 and CNN to extract the features from the image which is passed as input to the RNN—an LSTM generator. Finally, the caption for the image is generated using these words

Step 1: Take the search query as an input.

The input is sent as an HTTP request to the server goes for further processing.

Step 2: Searching of captions using similarity techniques.

The input text would be compared with the different captions and based on a similarity measure a score would be calculated for each result.

Only those results would be displayed who have a score within the pre-decided threshold which would vary based on performance.

Step 3: Displaying of results

The matched images will be displayed with the most relevant results appearing first.

4. IMPLEMENTATION

The implementation is carried out using Python installed on a system with 16 GB memory and i7 processor and 2 GB NVIDIA GTX GeForce 750Ti. We use Keras that contains inbuilt VGG16 implementation for the image feature extraction.
The python Sequencematcher - a python library is used for calculating the similarity score.
In this work, for the preliminary implementation and testing, we have used the Flickr-8k dataset.

5. RESULTS

The captions are generated when the images are given. Some of it includes are shown.

Processing Image Caption...
a group of people are posing for a picture
The Generated caption for this image is a group of people are posing for a picture .

Processing Image Caption...
a man in a red jacket is snowboarding

6. FUTURE SCOPE

1. Extend the functionality to accommodate custom datasets.
2. Make a highly scalable REST API which accepts the image and returns the caption of the image.
3. Use the following as a service over the web
4. Further development may also include working on improvising with more accurate predictions and search results

7. CONCLUSION

Thus we have presented a way to automate the process of organizing images by developing an Image captioning model for image searching to simplify the task of photo segregation for novice and professional users. Apart from this, CapSearch also aims to solve the problem of the requirement of the internet by facilitating offline search within the application.

8. REFERENCES

[1] Pranay Mathur∗, Aman Gill†, Aayush Yadav, Anurag Mishra, and Nand Kumar Bansode: Camera2Caption: A Real-Time Image Caption Generator

[2] Image Captioning-Based Image Search Engine: An Alternative to Retrieval by Metadata: SocProS 2017, Volume 2

[3] Oriol Vinyals , Alexander Toshev , Samy Bengio, Dumitru Erhan:Google Show and Tell

[4] Rohini K. Srihari: Automatic indexing and content-based retrieval of captioned images