Optimizing Reverse Image Search by Generating and Assigning Suitable Captions to Images

Dhvani Kansara, Aditya Shinde, Yashi Suba, Abhijit Joshi

[1,2,3] Dwarkadas J. Sanghvi College of Engineering, Mumbai, India
[4] Professor, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India

Abstract. Reverse image search is a content-based image retrieval (CBIR) query technique that takes a sample image as an input and search is performed based on it. Reverse image search is characterized by a lack of search terms. Using reverse image search, one can find the original source of images, find plagiarized photos, detect fake accounts on social media, etc. Reverse image search works by uploading an image by the user and searching of images is carried out by using the corresponding meta tags, html tags or color distributions of the image. The search engine currently functions by using this information and not the context of the image, thus resulting in inaccurate outcomes. Labelling the image using keywords relevant to the image is a very generic explanation to an image and can lead to ambiguous interpretations. This can be overcome by giving a more specific description to an image thereby prevents misinterpretations and provides more specific results. The proposed system aims at optimizing the search results by depicting more clearly the relationships between the objects in an image with the help of Image Captioning

Keywords: Captioning, VGGNet-16, CNN, LSTM.

1 Introduction

Reverse image search is a query technique that formulates a search query based on the sample image provided to it. It works on the principle of content-based image retrieval (CBIR) which allows users to obtain results related to the content of the sample image provided by the user. It also helps to find out other derivative or manipulated versions of the sample image. Thus reverse image search helps to find images that most closely resemble the input image.

The human mind remembers more of what it sees than it reads. Images help us visualize the information we are looking for. Thus, it is widely crowd-pleasing, and a large amount of population depend on image search for various activities which was concluded by the survey conducted by the authors. It showed that about 80% of people use Reverse Image Search frequently for different purposes. Out of which about 70% of people did not receive expected results. The reason is, the current system generates its output based on a single keyword search, color distribution or meta tags relating to the image. Such superficial information is not able to provide accurate results because it does not take into consideration the content of the image. This motivated us to improve and optimize the search engine. Visual knowledge fed to the computer through images can be interpreted with the help of descriptions or captions. Captions are the detailed description of images that help to distinctly locate an image from its pool. These descriptions are composed up of objects represented in the image and the relationship among those objects. Using captioning, one can find results that are more accurate and apt to the user's search query. Captioning also allows us to filter out any irrelevant and disturbing results that are widely published on the internet. Thus, developing the said model will help users to search images that are more in sync with their query and provide results efficiently.

Thus, the system presented in this paper will generate captions or descriptions for the images. This will help in optimizing the reverse image search technique by searching images based on the captions rather than the current criterion for search. This will provide optimized results because the interpretations are universal and scope for confusion is minimized. This will increase user engagement, popularity and improved Search Engine Optimization of the system.

2 Literature Review

In the literature review, various approaches for image captioning were discussed, and the drawbacks of the existing systems were identified. The details are given in the subsequent sections.

2.1 Existing System

Most of the systems for image captioning use the basic pipeline of image feature extraction followed by sentence generation. The captions generated by some of them are not grammatically correct. A few systems provide a complex but an accurate way for getting captions with correct captions and grammar.

The existing system for Reverse Image Search labels the entire image by a single keyword. For example, for an image of a beach, the existing system will refer to it as 'beach' and fetch results accordingly. This leads to inaccurate query results, because a broad range of images can refer to the word 'beach'.

(Elamri and Planque, 2016) generated image captions using deep learning techniques. This system was developed for visually impaired people, so that they can better understand their surroundings. One of the objective of the system is to automate caption generation of online images which will make the web a more inviting place for visually impaired surfers. In this system, Image features are extracted using Convolutional neural network. Then, the dimensions of this image feature vector are reduced using Principal Component Analysis (PCA). This resulting feature vector is then fed into the recurrent neural network model for sentence generation. The generated description of the image is in valid English. This system was capable to generate precise captions.

(Xinwei and et al., 2016) used the method of POS (Parts of Speech) tags. This is used to guide the training and testing process. POS tags are used to generate relative context for the sentences that are developed. An important feature presented in this is the use of a recursive method which manages a memory cell to keep a track of the complete sentence and thus uses the context from the entire sentence to append the next word while generating the caption. Also, POS tags ensure that the right part of speech appears at the right place ensuring grammatical soundness of captions generated.

(Zhongliang and et al., 2017) presented a novel method for caption generation using pre-trained vectors. The new scope covered by them is in terms of diversity of images. It uses a model that is already trained by oxford to learn input features in addition to the new features learnt from the user specific dataset. Thus, making the model more exhaustive with respect to object identification. The concept of 'attention mechanism' is used in this model. It presents a distinct approach wherein, attention is concentrated on features using probability distribution of objects based on their brightness and intensity in the overall image. For sentence generation, Long Short Term Memory (LSTM) model is used (an improvement to RNN) which uses memory to remember context of words.

(Zhou, and et al., 2017) presented a model which showed better performances than other state of art models developed until 2016. The authors talk about the use of Text-Conditional based attention; it aims at training the model for 'captioning' rather than mere feature extraction. The scope of this model is expanded to cover detailed and intelligent feature extraction where the models tries to find objects related to the object detected previously. This theory covers methods such that the model becomes intelligent enough to recognize relations between objects of an image.

(Asakawa and et al., 2017) trained the model in unsupervised environment by remembering long term details in memory. In order to incorporate a complete reference, a more robust system is developed using autoencoders. The method of using autoencoders and para-graph vectors was found to have better context in this methodology. This approach has an encoder and a decoder, both work recursively by using words as input and giving sentences. In this case, the intermediate sentences are generally ungrammatical and do not transit smoothly from one to other. In later iterations, it takes text from neighboring sentences rather than only the current sentence. Later, formation of sentences was recommended using a generative model which is based on a regularized version of standard autoencoder. Once the features are translated into probability distributions, the process of extracting words for sentences be-comes faster. Autoencoders yielded similar results as the standard recursive models but Improved Autoencoders can be increasingly used for more narrative descriptions and for generation of longer passages.

(Chen, and et al., 2017) replaced the CNN part with three state-of-the-art architectures (VGGNet, AlexNet, GoogleNet). It is found that VGGNet performs best according to the BLEU score. A simplified version of Gated Recurrent Unit (GRU) (a variant of LSTM) is used as a new recurrent network which is implemented using Caffe. Caffe provides a modifiable framework for the state-of-the-art deep learning algorithms. The simplified GRU achieves comparable result when it is compared with the LSTM method. But it has few parameters which saves memory and is faster in training.

In all works stated above, performance is measured using BLEU metric. It measures the difference and deviation be-tween captions generated by the computer and those given by humans.

2.2 Literature Related to Works

The information gained from reading all the existing work is that the captions are generated in two steps: first, feature extraction using object detection followed by sentence generation using language interpretation.

Feature Extraction: This is the process of extracting features and objects from images uploaded by the user. The main approach followed by all the existing systems is, initially the image is divided into various parts for identifying components in each section. This makes it easy for the model to focus on each part and retrieve maximum information for each section. This also ensures that any important information is not left out. Using the similar approach (Karpathy et al., 2015) presents a very accurate approach by detecting the minute details and even those which can be missed by any interpreter. (Zhoul and et al., 2016) used attention mechanism along with object detection through which the model tries to predict the object it has to focus. (Khurana, and Awasthi, 2013) used template matching for object detection. In this technique, templates of objects are stored and each input image object is compared with the stored templates to find out a match.

Sentence generation: In this process, captions are generated using language interpretation based on the features extracted in the previous step. In the existing systems, a recursive approach is used where sentences are generated based on Parts of Speech i.e. using grammatical context and relevance to previous words of sentences. A recursive method is used where each word of the sentence which is accepted by the model is followed by the next word, which is most suitable in context and grammar. Other approaches using autoencoders and probability distributions for generating sentences are also proposed by (Asakawa et al., 2017).

2.2 Observations made on the existing systems

The major drawbacks in the above systems and the approaches are:

- They fail to generate accurate and contextual captions for moderate-to-high complex images.
- They sometimes overestimate the relationships between objects by forcefully trying to focus on things which are not present in the image.
- The models using attention mechanism generates captions which ignores the background details due to reduced intensity of such objects but provides a concise description including all the important objects.

- Autoencoders are not feasible for generating concise captions, they are more suited for producing large paragraphs rather than captions.

In the proposed work we tried to overcome these issues.

4 The Proposed Approach

After a thorough literature survey, it was concluded that Convolutional Neural Network (CNN) is the most suitable method for the purpose of feature extraction, thus objects from the image will be extracted using CNN. In order to upgrade the object detection, a pre-trained model by Oxford called as 'VGGNet' is used. This will lead to improved accuracy as initial weights in the neural networks will not be random, rather will be based on a model which is already trained on many images. The system uses additional datasets on top of it, thus leading to improved accuracy.

In order to generate the sentences, Long Short-Term Memory (LSTM) is used. LSTM is a form of recurrent neural network that uses an internal memory to remember its inputs. Thus, it is appropriate for problems involving sequential data (same as in our problem). The internal working of LSTM uses three gates, the input, output and forget gate. The forget gate determines which information is important and which is not. Based upon the decision taken by the forget gate, the algorithm deletes or remembers certain information. The importance of information is assigned by the weights calculated during the training process. Thus, after receiving the input, the image is passed through the system in order to receive the most appropriate caption. Figure 1 shows the flow of information while generating the caption of an uploaded image.

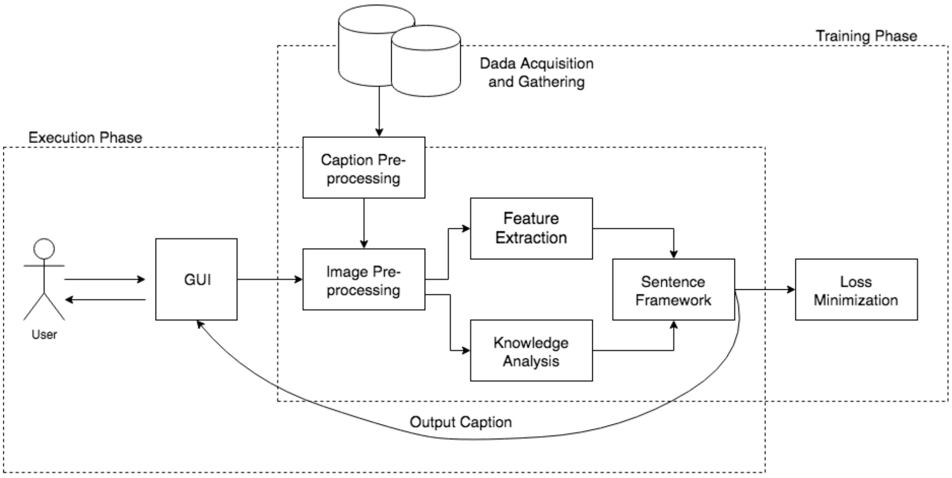


Fig. 1. System Architecture

The overall caption generation process is divided into two phases, namely the training phase and the execution phase.

In the training phase, the required data is acquired from various sources and then it is gathered at one place which is later used for the training purpose.

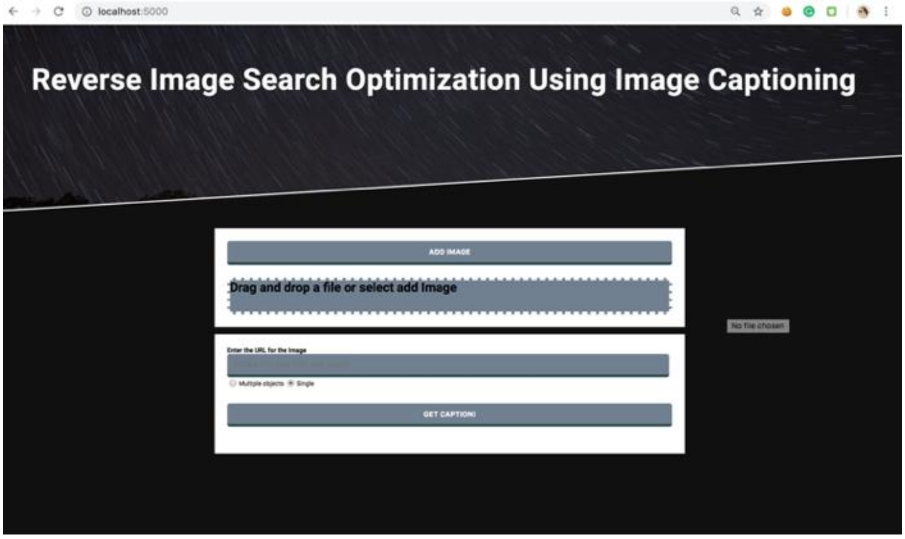
1. Acquisition and Gathering: Two datasets are used for training in order to obtain a good accuracy:
   Flickr8k: Consisting of 8,000 images and 5 different descriptions/captions for each image provided by different people based on their perspectives.
   Flickr30k: Similar to above with 30,000 images and respective captions. These images also contain celebrity images to detect renowned faces in an image.

   Thus, the proposed model uses a total of 38,000 images for training. All captions are labelled by humans based on their perspective about the image. The majority of captions consist of succinct descriptions with an average length of 10 words. Because of the volunteer nature of the project, language errors are relatively common (in a random sampling of 100 captions, five had either a spelling mistake or a significant grammar mistake). These datasets contain a text file with the image names and corresponding 5 captions. Another database contains all the images. The captions are extracted by reading the text file line wise and creating a list of each caption along with corresponding image name. Images are read from the database and stored in a matrix form. It is necessary to preprocess the captions and images from the dataset so that the model training is easier and more efficient.
2. Pre-processing: After the creation of the required dataset of images and captions, captions are preprocessed by splitting each of the five caption for an image and inserting <START> and <END> tags at the beginning and end of each caption. The words in each of the caption are tokenized based on their context. In the process of tokenization, vector values are assigned to each word in a multi-dimensional free space. For example, the apple vector and fruit vector will be nearby but apple vector and tiger vector will be mapped distantly in free space. Images are also preprocessed by converting them from RGB to BGR format, for faster processing during training. After preprocessing, the objects in the image are identified by feature extraction, which is the next stage.
3. Feature Extraction and Knowledge Analysis: Feature extraction algorithm, here Convolution Neural Network (CNN) is applied on the image to extract the objects contained in it. First it splits the image into segments and then extracts features/objects from each part. The convolution layer is applied on these segments and convolutes a kernel (feature vector) over the pixels of the image, which gives a feature map. In this way, the network first learns edges and curves (based on the feature vector) and then slowly understands complex and intricate shapes. Since we used a pre-trained model- VGGNet, this process is faster. The output of the convolution layer is fed into the pooling layer which reduces the spatial dimensions by averaging or taking the maximum value of the sub regions in the feature maps. This improves computation performance and also reduces chances of over fitting. The next layer is flatten layer, which converts 2D matrix into 1D vector. At the end, the output objects are extracted using an activation function. The objects that are extracted in this step are related using a sentence framework.

4. Sentence Framework: In order to determine the relationship between objects, the model is then trained to understand and interpret the captions in the training data, i.e. the model is taught to learn 'English' with the use of POS (Parts of Speech) tags so that it can generate sentences. The algorithm chosen is such that, internal memory is used to process the sequence of inputs, which is called Long Short Term Memory (LSTM). In this algorithm the information cycles through a loop. The input to the neural network is the current input as well as the output from the previous inputs, based on their importance and significance. For example, if the object extractor has fetched 'man', 'TV' etc. Then this step will associate them by 'man' is watching 'TV'. These sentences will be generated by the knowledge analysis performed in the previous step.
5. Thus, the system generates caption for the given input image. In the execution phase, this generated caption will be directly provided back to the user because the model is already trained. But during the training phase, in order to improve the accuracy, it is necessary to minimize the loss of information that might be produced in this step.
6. Loss Minimization: This step is performed only during the training phase to minimize the inaccuracy while the model is still in the learning phase. Here, the model learns from the wrongly generated captions so as to update the weights in the neural network using backpropagation. For this, the error is calculated by comparing the caption generated by our model and the caption already present in the training set. The difference between these vectors is calculated to check the error, and weights of the neural network are updated to minimize this error. After the above steps are performed, the model is tested to check the accuracy of the generated captions with the test dataset.
7. Thus, the system is ready to take user input and generate appropriate captions by generating output based on the trained and tested model, this is called the execution phase. During the execution, the user image passes through the feature extraction and sentence generation framework to provide the output to the user.

5 Results and Analysis

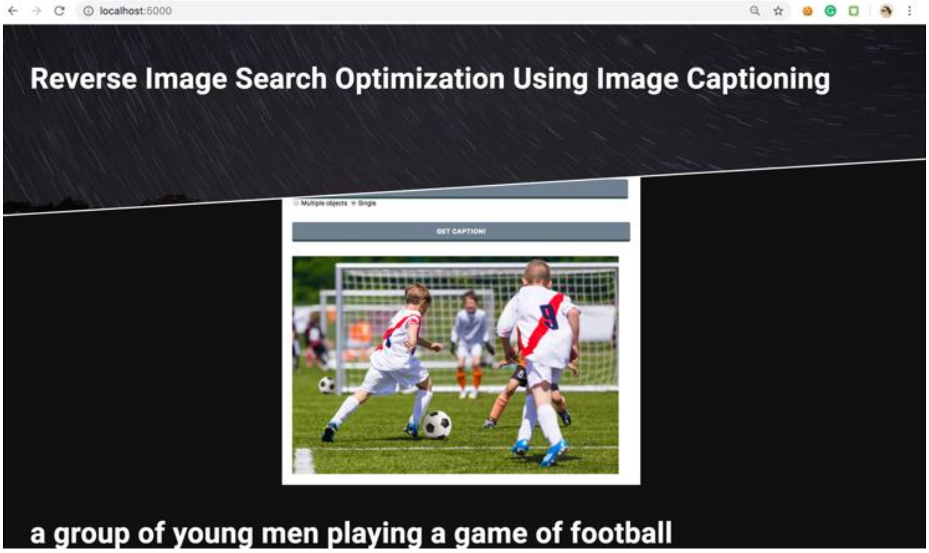
Figure 2 shows the home page of our system. The user can upload/drag an image from their directory or can also upload the URL of the image. An API is also developed which can be integrated with a search engine to improve the search results.



localhost:5000
Reverse Image Search Optimization Using Image Captioning
ADD IMAGE
Drag and drop a file or select add Image
No file chosen
Enter the URL for the Image
○ Multiple objects ● Single
GET CAPTION!

Fig. 2. GUI for client

Figure 3 represents a general output caption, “a group of young men playing a game of football” for the input image uploaded by the user. After performing Black Box Testing for multiple classes of images, it was found that the caption generator could label images in domains of sports, movies, environment, nature, animals, birds, etc. accurately.



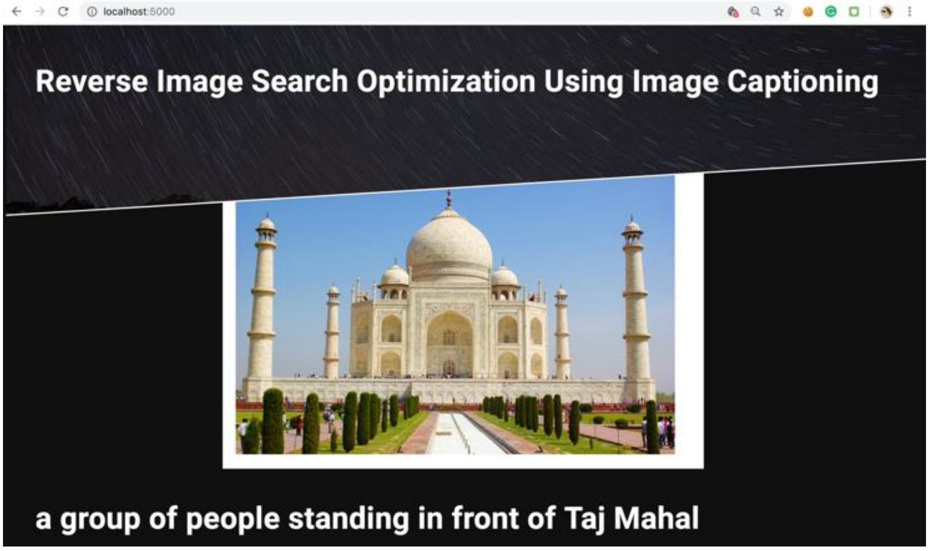
localhost:5000
Reverse Image Search Optimization Using Image Captioning
○ Multiple objects ● Single
GET CAPTION!

a group of young men playing a game of football

Fig. 3. Example output 1

Figure 4 shows that the model is able to recognize renowned monuments as well. For an image of Taj Mahal uploaded by the user, the system caption generated was, “a group of people standing in front of Taj Mahal”, rather than calling it just another building. In addition to this, it is also able to identify renowned celebrities, and other revered personalities as well.

Figure 4 shows that the model is able to recognize renowned monuments as well. For an image of Taj Mahal uploaded by the user, the system caption generated was, “a group of people standing in front of Taj Mahal”, rather than calling it just another building. In addition to this, it is also able to identify renowned celebrities, and other revered personalities as well.



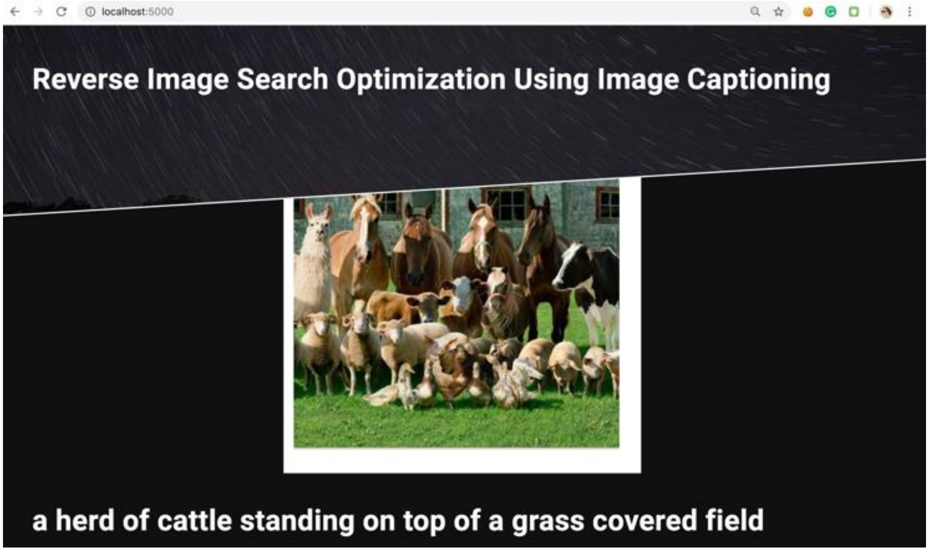
Reverse Image Search Optimization Using Image Captioning

a group of people standing in front of Taj Mahal

Fig. 4. Example output 2

When recognizing multiple objects in the image, the model gave output in a succinct form which could not list all the objects in the image rather provide a generalized caption,” a herd of cattle standing on top of a grass covered field” for the whole image as seen in figure 5.

In order to address this problem, the model contains a feature where the user can select a radio button of whether the user wants to display a succinct caption or a dense caption. A dense caption for figure 6 would look like one in Figure 6. Such a caption will enlist all the details included in the image. The mechanism used in figure 5 is called ‘attention mechanism’ where focus is given on the image as a whole whereas in figure 6, the focus is on each individual object present in the image.

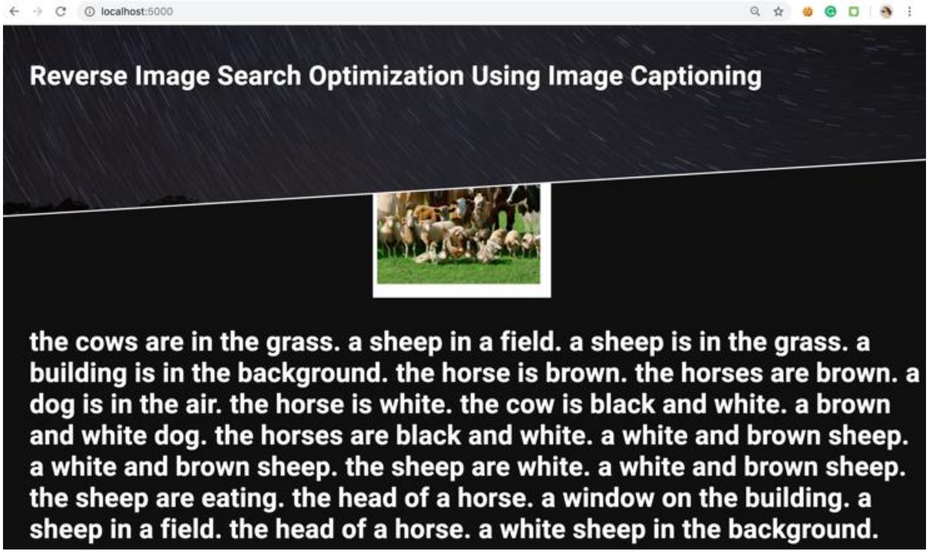


localhost:5000

Reverse Image Search Optimization Using Image Captioning

a herd of cattle standing on top of a grass covered field

Fig. 5. Example output 3



localhost:5000

Reverse Image Search Optimization Using Image Captioning

the cows are in the grass. a sheep in a field. a sheep is in the grass. a building is in the background. the horse is brown. the horses are brown. a dog is in the air. the horse is white. the cow is black and white. a brown and white dog. the horses are black and white. a white and brown sheep. a white and brown sheep. the sheep are white. a white and brown sheep. the sheep are eating. the head of a horse. a window on the building. a sheep in a field. the head of a horse. a white sheep in the background.

Fig. 6. Example output 4

6 Conclusion

In the proposed model of Image Captioning and thus Reverse image optimization, images are taken as the input from the user and efforts are made to formulate most specific and relevant description to the image. This description will be in the form of captions and given as the output to the users. These descriptions are the universal to prevent any misinterpretation. The search engine is directly queried using these descriptions so that the user gets the resultant images which most closely resembles the input image. The proposed system will help the users to find the relationships between objects in the image and information about the image.

References

1. Christopher Elamri, Teun de Planque (2016) Automated Neural Image Caption Generator for Visually Impaired People, Department of Computer Science, Stanford University. https://cs224d.stanford.edu/reports/mcelamri.pdf (Accessed 18 October 2018)
2. Xinwei He, Baoguang Shi, Xiang Bai, Gui-Song Xia, Zhaoxiang Zhang, Weisheng Dong (2016), 'Image Caption Generation with Part of Speech Guidance', IEEE Conference on Computer Vision and Pattern Recognition 2016, School of Electronic Information and Communications, Huazhong University.
3. Zhongliang Yang, Yu-Jin Zhang, Sadaqat ur Rehman, Yongfeng Huang (2016) *Image Caption with Object Detection and Localization*, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. https://arxiv.org/pdf/1706.02430.pdf
4. Luowei Zhou, Chenliang Xu, Parker Koch, and Jason J. Corse (2017) *Image Caption Generation with Text-Conditional Semantic Attention*, Robotics Institute, University of Michigan, Department of Computer Science, University of Rochester, Electrical and Computer Engineering, University of Michigan, 2017.
5. Asakawa, Shin & Ogata, Takashi. (2017) 'Comparison Between Variational Autoencoder and Encoder-Decoder Models for Short Conversation' Proceedings of International Conference on Artificial Life and Robotics. 22. 639-642. 10.5954/ICAROB.2017.OS1-4.
6. Jianhui Chen, Wenqiang Dong, Minchen Li (2017) *Image Caption Generator based on deep neural networks.* Department of Computer Science, University of British Columbia. https://www.cs.ubc.ca/~carenini/TEACHING/CPSC503-19/FINAL-PROJECTS-2016/image_caption_generator_final_report.pdf (Accessed 12 November 2018)
7. Khushboo Khurana, Reetu Awasthi, 'Techniques for Object Recognition in Images and Multi-Object Detection' International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Volume 2, Issue 4, April 2013.
8. Andrej Karpathy, Fei-Fei Li, 'Deep Visual-Semantic Alignments for Generating Image Descriptions', IEEE Conference on Computer Vision and Pattern Recognition 2015.