Image Retrieval using Automatic Image Captioning

Aghil Karadathodi Prasad s1036614
Nolan Cardozo s1034065
Puja Prakash s1039329

I. Abstract

Due to the huge amount of images being shared, stored and viewed online on a daily basis, image retrieval applications have been gaining quite a lot of popularity in recent times. In this paper, we propose a two stage approach to this problem. We first use image captioning to describe the contents of an image followed by information retrieval algorithms to retrieve images relevant to the users query. Word centroid distance algorithm performed the best compared to the BM25 and the neural based Doc2Vec algorithms with mean precision of 72 %.

II. Introduction

In recent years, due to the convenience and affordability of digital products, there has been a huge shift towards image or video applications. Moreover, the boom in social media applications has led to millions of images being uploaded, shared and viewed on a daily basis. Furthermore, this is not just constrained to social media, as several other domains such as media and healthcare store and mine tons of images on a regular basis. This massive growth of image content eventually brings us to the need for better solutions for storing digital content and retrieval techniques.

In this project, we discuss the need for content-based information retrieval and propose advancements in deep learning specifically computer vision and language modeling as a solution to this task. At a higher level, we can think of this solution as retrieving matching images based on entering a few words describing the image in the system. In more detail, the input to the system will be a text query, for example, a boy playing and the output would be several images ranked based on the similarity to the text query.

This is a complex problem as we need to understand the components and semantics of the image to correctly classify objects, convert these objects and their semantic relations into text and retrieve efficiently. In this project, we first proceed by briefly describing some related work followed by a concise explanation of all the methods we used in this paper. Further, we discuss about the dataset and the implementation of both the image captioning module as well as the information retrieval algorithms - both traditional and neural embedding based. Lastly, we compare and contrast these methods briefly and conclude by suggesting some ideas on future work.

III. Related Work

In recent years, Content based image retrieval has received quite a lot of popularity. Most of the methods used earlier were to learn some features from an image and find related images based on some similarity measure. Although, recent advancements in computer vision and language modelling has led to a shift in the way the problem is solved. For example, Socher et al. (2014) uses an RNN to capture compositional semantics.

Another category of methods is to use image captioning to convert images into text which describes the image and then use a text retrieval system to extract relevant images. One such paper, is the state of the art for image captioning, (Peter Anderson, et al 2018) where top-down visual attention mechanisms are used to caption images. Moreover, several neural language embedding methods have shot to prominence lately for example, BERT and ELMO and are current state of the art for context based language modelling.

IV. Method

In this section, we describe the methods that we use for the image captioning model as well as the algorithms we use for image retrieval.

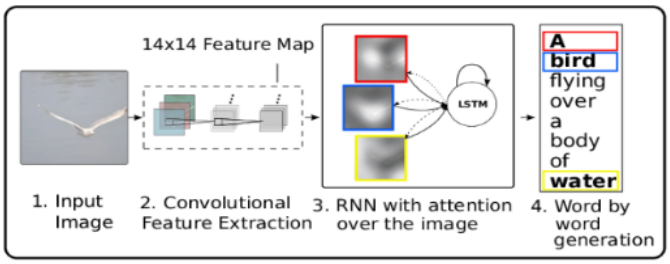

Fig. 1. Encoder-decoder architecture of the image captioning model

Fig. 2. Sample results of Automatic Image Captioning Model

A. Automatic Image Captioning

In this module, we implemented the Show, Attend, and Tell paper by xu et al. 2015 [1] which uses an attention mechanism to focus on that most important part of the image most relevant to the word it is going to utter next. To accomplish this, we use an encoder-decoder neural network architecture as indicated in the Figure 1.

1) Encoder: The encoder is a convolutional neural network that extracts a set of convolutional feature vectors from the image. These features are extracted from a lower convolutional layer instead of the fully connected layer. This helps the decoder selectively focusing on relevant parts of the image by choosing a subset of the feature vectors. The encoder need not be trained from scratch for this task. We use the concept of transfer learning in which we use a pre-trained Resnet 101 model which has been trained on the ImageNet classification task. We removed the softmax layer in the end which is used for classification and fine-tuned the model to better fit our encoding task.

2) Decoder: The task of the decoder is to use the generated feature vectors and generate a caption word-by-word. We use a Long short memory network (LSTM) network that generates a word every time step based on the feature vectors, the previous hidden state, and the previously generated word. Here, it is important to note that we use the attention mechanism. Without attention, the decoder would simply average the encoded image across all pixels and generate a word by word sequences. Alternatively, in a setting with attention, the decoder views relevant parts of the image at relevant points in the sequence

B. Image retrieval algorithms

We used 3 image retrieval algorithms to retrieve images whose caption was found similar to the users query given a dataset of captioned images. The image retrieval system consists of 2 components - matching and similarity scoring. At first, we use the matching operation to filter the relevant images. For example, the output can contain all the images with captions containing at least one of the users query terms. The scoring step is done to rank the images. The score denotes the relevance of a specific image to the query. The scores are used to create a ranked ordering of the matched images.

1) Doc2Vec[2]: In Doc2vec algorithm each document (image caption) is represented as feature vector in an N dimensional space where similar captions are placed closer forming a cluster. The users query is compared with the documents in the vector space and all the documents falling in the same cluster are extracted. These similar documents are ranked and top few images corresponding to these documents are retrieved.

2) Word centroid Distance model[3]: The model represents each document (image caption) as the centroid of its respective word vectors. The word vectors carry semantic information of the words, thus the centroid of the word vectors within a document represents its meaning to some extent. At the query time, the centroid of the querys word vector is calculated and the relevant image captions are found based on its cosine similarity to the centroids of the matching documents. Finally, the

matching images are finally retrieved based on the extracted image captions

3) BM25[4]: BM25 is a ranking algorithm used to rank matching documents based on the probability of their relevance to the given users query. The documents are ranked in decreasing order of their probabilities based on the users query terms appearing in the document. The top few images corresponding to these documents are retrieved.

V. Experimental setup

In this section, we discuss the dataset used for the image captioning model, the metrics used to evaluate the model and the results from this task. Further, we contrast and compare the three information retrieval algorithms implemented to retrieve the images.

A. Dataset

Neural network based methods produce optimal results when trained on a large amount of data. For this task, we used the COCO (Common Objects in Context) dataset (Lin et al., 2014). This dataset contains 123K images and every image is annotated with 5 different captions produced by different 5 different people, hence every caption is slightly (sometimes greatly) different from the other captions for the same image. This takes care of the annotation biases involved when annotated by just one individual. We used the split specified in Karpathy & Fei-Fei (2015): 113, 287 for training, 5, 000 for validation and test respectively.

B. Evaluation

1) Image Captioning: Results for the attention based image captioning model are reported using Bilingual Evaluation Understudy (BLUE 1-4) which is the standard evaluation metric discussed in most image captioning tasks. Though, there are several problems with BLUE and therefore we also used Metric for Evaluation of Translation with Explicit Ordering (METEOR). Sample results of image captioning model is captured in Fig 2

BLUE Scores				METEOR
BLUE-1	BLUE-2	BLUE-3	BLUE-4	
72.6	51.2	37.3	26.76	22.8

Fig. 3. The results of the Image captioning model

2) Image Retrieval: The Mean average precision (mAP) metric was used to evaluate and compare the image retrieval algorithms. We queried the system with 50 different text queries of varying lengths (ranging from one word to a sentence) and manually calculated the average precision based on the image results that the querying user thought were relevant. Further, the mean of the average precision for each of the algorithms was calculated.

The word centroid distance algorithm gave the best mean average precision results among the 3 different retrieval algorithm we tried. The average length of captions generated from the images using the image captioning model were mostly small to medium sized ones. The word centroid distance represents each document as the centroid of its respective word vectors. With lesser amount of document terms, the document centroid would explain the meaning of the document better. Since image captions were short and have less number of words, the algorithm gave good results. The performance of Doc2Vec algorithm was the least accurate because by modeling documents as a low-dimensional vector, a lot of information is lost. But then for single word queries, Doc2Vec gave better results comparable to both word centroid as well as BM25. BM25 also gave meaningful search results outperforming others in some search context by ranking the results better.

	Doc2Vec	Word Centroid	BM25
mAP	48%	72%	64%

Fig. 4. The results from the image retrieval algorithms . For a set of 50 different queries, we manually evaluated the relevance of the images returned and then calculated the mAP score

C. Web application

Finally, the image captioning and the image retrieval modules were brought together using a web application. This application was built using the Flask API in Python. The application consists of search page(Fig 5 Top part) where the user can type in a text query and the application navigates to a results page(Fig 5 Bottom part) that lists the relevant images that are most similar to the query. The application is called Image Fetch and is available at https://github.com/nolancardozo13/Fetch_Image.

Fig. 5. Image Fetch web application developed by our group. At the top is the search page and below is the results page using Word Centroid Distance model

VI. Future work

In this paper, our focus was to build a fairly good image captioning model followed by the comparison of several retrieval algorithms based on the caption generated. Though, the algorithms worked seamlessly in most cases, it failed at some complex tasks. For instance when the user entered the query dog playing under hot sun , the results had a picture of a ”hotdog”. This error was due to the system failing to understand the context in which the word was used. In future, we would like to try out more recent and state-of-the-art neural based embeddings such as ELMo and BERT [5] that perform well in identifying the context it appears in. For instance, if we have a word with multiple meanings, like ”bat”, ELMo and BERT tackle this problem by taking an entire sentence as input, and produce context-dependent embeddings of each word.

References

[1] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, “Show, attend and tell: neural image caption generation with visual attention,” In ICML’15 Proceedings of the 32nd International Conference on International Conference on Machine Learning, 2015.

[2] T. Thongtan and T. Phienthrakul, “Sentiment classification using document embeddings trained with cosine similarity,” In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[3] G.-I. Brokos, P. Malakasiotis, and I. Androutsopoulos, “Using centroids of word embeddings and word mover’s distance for biomedical document retrieval in question answering,” 2016.

[4] J. Whissell and C. Clarke, “Improving document clustering using okapi bm25 feature,” https://link.springer.com/article/10.1007/s10791-011-9163-y, 2011.

[5] Dai, Zhuyun, Callan, and Jamie, “Deeper text understanding for ir with contextual neural language modeling,” *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR19*, 2019. [Online]. Available: http://dx.doi.org/10.1145/3331184.3331303