Ensembling of IR and Deep Learning Models for Image Retrieval

Akshay Jain, Anand Kumar, Ashok Kumar, Manish Rathore
Information Technology Department, NIT Karnataka

***Abstract*— Information retrieval has become a crucial part of today's digital life and it is very much essential to develop some really good ways to store and retrieve information from the web efficiently. In this paper, we discuss about content-based image retrieval, which refers to the process of retrieving images based on their content (CBIR). A query (either a text or an image) would be entered by the user into the system, which would then compute weights based on how closely each object (a caption or picture) in the database fits the query and rank the images in accordance with this value. The user is then given the top-ranking pictures.**

Keywords: **Image Retrieval, Convolutional Neural Network, Deep Learning, Query Image, CBIR**

I. INTRODUCTION

In recent years, the Demands for multimedia applications on the Internet are increasing, and the importance of intelligent image retrieval has also increased. So we need a way to make sure that the Queries are processed in such a way that the Semantic Gap is as minimal as possible. The initial step in transforming human vision into numerical descriptions that computers can alter is feature extraction. The derived characteristics have a significant impact on the retrieval accuracy. However, the user's needs are what determine this decision. The performance of the model can be enhanced by feeding the retrieved characteristics to a supervised or unsupervised machine learning method. A recent trend in image retrieval research focuses on using deep learning to improve accuracy at the expense of longer run times. The high-dimensional features that are generally created while attempting to convert visual picture material into a numerical feature format are another problem that adversely affects model performance (i.e., memory consumption, scalability, speed, and accuracy). These sparse distributions of high-dimensional representations are referred to as the "curse of dimensionality." Dimensionality reduction provides a solution to this issue. Numerous thorough investigations on a variety of suggested dimensionality reduction techniques are found in the literature. Similarity measurements are yet another crucial factor that affects how well the model performs. The arrangement of the feature vector determines this metric, therefore picking the wrong one will lead to less uniform pictures being returned and less accurate model systems. In other words, by employing a suitable similarity measure, a high level of accuracy may be attained.

In our project, the image retrieval method that has been used is Content-Based Image Retrieval(CBIR). We have designed different models for extracting the images from the database which includes using the content features in the image through the VGG model and using the given captions for a text-based query for outputting the images. For the purpose of text-based query, we have implemented two different ways, the TF-IDF weighing scheme and BERT model separately to get the sentence embeddings of the captions. Based on the similarity measure we are taking the top 10 results as output to the query.

II. LITERATURE REVIEW

Many studies have been conducted on picture retrieval, some of which are covered below. Authors Jenni and Ali conducted some research in 2015 and 2020, and the research was based on image retrieval from an image database. The Content Based Picture Retrieval (CBIR) algorithm was used in those studies to search for an image from an image dataset and to get photos that had about the same sort of content as the sought image (Jenni et al., 2015). (Ali et al., 2020). These studies' methodology was entirely automated. Between the searched images and the retrieved images, however, there was a "semantic gap." Non-relevant picture retrieval was caused by the disconnect between the images' low-level features and their high-level ideas (Bai et al., 2018). Numerous investigations have been conducted over the last three decades to close this semantic gap (Shrivastava Tyagi, 2017). Many techniques were developed to convert abstract visual concepts into features, and these techniques formed the cornerstone of CBIR.

Two categories of features from the photos were identified (global features and local features). This classification was based on the manner in which the researchers extracted these traits. The photos' colours, textures, forms, and spatial information were contained in global features. This feature was assisting in giving the complete image a representation. Because of the classifications they performed, researchers got the benefit of quick feature extraction and similarity computation. On the other hand, they were unable to differentiate between the image's background and its objects. As a result, it was only appropriate for object categorization and detection (Halawani et al., 2006), not complicated scene search and object identification (Ghrabat et al., 2019). In contrast to global features, the previously described local characteristics were helpful for picture retrieval, matching, and recognition tasks (Halawani et al., 2006). Author Bansal conducted another another study in 2020 on "Object recognition, which entails identifying and categorising items in photographs" (Bansal et al., 2020). This study reduced object recognition to a subtask of categorization. Local features were referred to as prominent areas or components of an image, such as borders,

corners, and spots. They can withstand backdrop changes, scaling, rotation, translation, clutter, and partial occlusion (Halawani et al., 2006).

Sneha Choudhary, Haritha Guttikonda, Dibyendu Roy Chowdhury, and Gerard P. Learmonth released a research on document retrieval in 2020 utilising BERT and TF-IDF. To create an end-to-end document retrieval system, they added an ensemble model of BERT and TF-IDF techniques. This model outperformed conventional retrieval algorithms by a wide margin when evaluated on a portion of the MS MARCO dataset.

III. Dataset

We have used Flickr8K dataset for our proposed method. Flickr8K contains 8,000 images, each with 5 different captions, clearly describing important entities and events. The images selected from Flickr's six groups have been hand-picked to represent a variety of scenes and situations, rather than typically familiar people and places. Sample images of the dataset look like this :

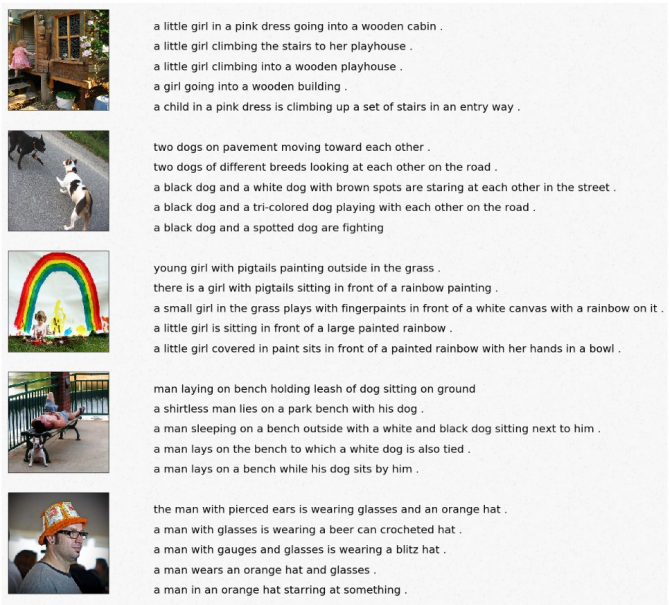

Fig. 1: Dataset Example

Because of the availability of the five captions for a single image we are able to perform a text-based search on these images easily.

IV. Methodology

In proposed approach, Feature extraction is being performed with the help of the VGG-16 Pre-Trained Model, we need to give our images to model and then it'll give us the desired relevant features of the Image which can be used in CBIR framework to carry out further procedures.

A. CBIR framework

Figure 2 illustrates the conventional CBIR framework, which comprises of some essential and some optional phases. A question image is submitted by the user as the first stage of CBIR. All operations performed on the query picture are carried out in the same order on all database images. These programmes, which are referred to as online processes, often execute on the query screen when a user submits a question. The same procedure, known as the offline method, can be used on the dataset's picture data before sending the query. The architecture of the framework might contain optional preprocessing stages like resizing, segmentation, denoising, rescaling, etc. The feature extraction phase follows this optional stage. The most crucial stage in the conversion of visual notions into numerical form is this one. Low-level features, such as colour, shape, texture, and spatial information, as well as local descriptors can be used to represent the retrieved features. Normalization or categorization comes after feature extraction in the preprocessing step. In order to find the most pertinent images, the final step compares all other images in the data set to the features retrieved from the query image. Another phase that could involve user involvement is relevance feedback, which evaluates the relevance and relevance of the photos that are returned. Many methods have been suggested to apply pertinent input to enhance CBIR performance. For our proposed method we have used VGG16 model to extract all such features.

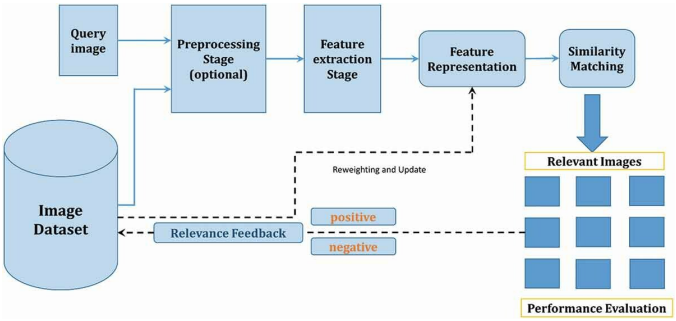

Fig. 2: CBIR Framework

The figure shows the entire framework of how a typical CBIR works and the same approach has been used in our proposed methodology for the feature extraction we have used the VGG-16 model which does the job of feature extraction quite efficiently.

B. Text-based framework(tf-idf)

Apart from searching for the Image by the Image as input itself, users can also type the text describing their needs and as in our dataset we have approximately five captions for each single image which describes the content in that particular image. Also each caption is attached to image with unique ID of the image. So we are having corpus of the captions of the images attached with IDs of the images. Now we are pre-processing the corpus like removing punctuation, removing white spaces, removing non alphanumeric values, stemming and all. Then after we are making text vector of each captions by finding their weights and storing it into a pickle file. So dataset is preprocessed and now we have text query. We are preprocessing this text query same as we preprocessed the captions. Then we can go for retrieving

the text query by finding similarity

$$\text{tf} - \text{idf}(i, j, D) = \text{tf}(i, j) \cdot \text{idf}(i, D)$$

Once caption will be predicted then we can easily find the ID attached with it then using ID we can print the image. So for doing this we have made use of simple vector space model which use TF-IDF(term-frequency and inverse document frequency) technique to match the query typed and the matching of the corresponding captions and based on the score that we are getting from the ranking function we are displaying the top-ranked images to the users.

C. BERT based framework

For getting the embedding of the captions we have used the BERT model which converts the sentences into the embedding vectors each of size 768. When a new query comes we generate a vector of it using BERT and then compute the nearest captions in our dataset and output the results.

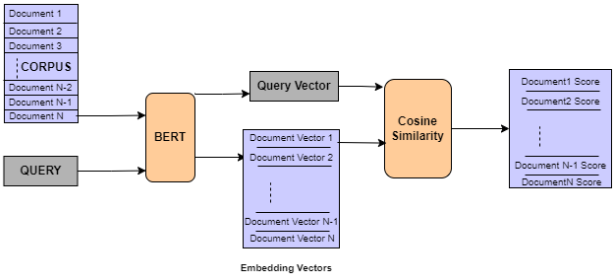

Fig. 3: BERT Model flowchart

V. Results and Analysis

Searched Image :-

Fig. 4: Searched Image

Results :-

Fig. 5: Retrieved Image Outputs

Fig. 4: Searched Image & Fig. 5 Image Outputs

We can see in the figures, we have used an image for searching and we got the results. In the searched image, a boy is playing a sport and as a result, we got images in which players are playing a different kind of sport.

Query : boy running with horse
Results :-

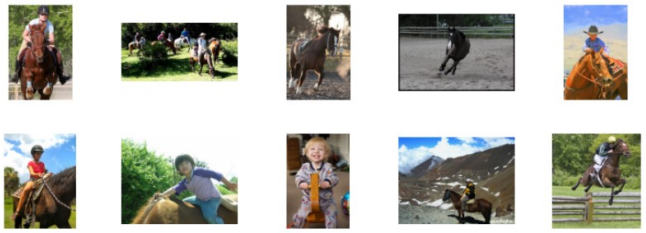

Fig. 6: Retrieved Results : boy running with horse

Query : man on a hill or mountain
Results :-

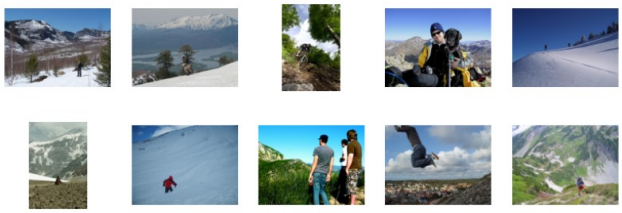

Fig. 7: Retrieved Results : man on a hill or mountain

Fig. 6: Text Query 1 & Fig. 7: Text Query 2

As we have seen in previous image-based retrieval we got good results that were similar to the searched images. Now in these images, we are retrieving the images via texts query. In figure 5, the query is "boy running with horse" and we got the images in results which are containing horses, boy, running boy, and a combination of these. So we can say that we got the expected results. Same in figure 6 also we searched the texts query "man on a hill or mountain" and we got the expected results.

S. No	Query	Precision@K (K=10)
	VGG-16 Results	
1	A woman shows off a newborn baby to a young boy and girl	0.8
2	A dog walks down the dirt road as a person follows	0.8
	Average =	**0.8**
	BERT Result	
3	boy running with horse	0.9
4	man on a hill or mountain	0.8
	Average =	**0.85**
	TF-IDF Results	
5	boy running with horse	0.8
6	man on a hill or mountain	0.9
	Average =	**0.85**
	Ensembled Model (TF-IDF + VGG Net)	
7	boy running with horse	0.9
8	man on a hill or mountain	0.9
	Average =	**0.9**

Fig. 8: Precisions of Retrieved Results

Fig. 8: Precisions of Retrieved Results

We have calculated the Precision@K of retrieved results for each method used in retrieving the images at K=10. We got 80% average precision for VGG-16, 85% for BERT, 85% TF-IDF and for Ensembled model of TF-IDF and VGG-Net, we got 90%. Here we can see that ensembled model is performing better than all others methods.

VI. CONCLUSIONS

In present time, the computational power is increasing and cost of storage is reducing. And the data we deal with is also increasing. But without having a way of retrieving the information and querying it, the collection of data will not be usefull. So the information retrieval systems comes into picture which helps us to retrieve the data from large collections. As of now we have seen the many retrieval systems based on texts. However, in this paper, we deal with photos and retrieve them depending on the content they contain.

Using the Flicker8K dataset, we successfully created a system for image retrieval based on the content present in it. We applied three different approaches to retrieve the images from our dataset namely VGG-based, TF-IDF based, and using the BERT model as well.

REFERENCES

[1] Liu Y, Zhang D, Lu G, and Ma W Y, 2007, A survey of content-based image retrieval with highlevel semantics, Pattern Recognition, 40(1), pp 262–282.

[2] Le Cun Y, Bengio Y, and Hinton G, 2015, Deep learning, Nature, 521(7553), pp. 436–444.

[3] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, and Wojna Z, 2016, Rethinking the inception architecture for computer vision, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818– 2826

[4] Babenko A, Slesarev A, Chigorin A, and Lempitsky V, 2014, Neural codes for image retrieval, European conference on computer vision, Springer, pp. 584–599

[5] Xia R, Pan Y, Lai H, Liu C, and Yan S, 2014, Supervised hashing for image retrieval via image representation learning." AAAI, 1(2), pp. 2-7.

[6] Chen J. C, and Liu C. F, 2015, Visual-based deep learning for clothing from large database, in Proceedings of the ASE Big Data Social Informatics. ACM, pp. 42-48.

[7] Jenni, K., Mandala, S., Sunar, M. S. (2015). Content based image retrieval using colour strings comparison. Procedia Computer Science, 50, 374–379. https://doi.org/10.1016/j.procs.2015.04.032

[8] Ali, F., and Hashem, A. (2020, June). Content Based Image Retrieval (CBIR) by statistical methods. Baghdad Science Journal, 17 (2(SI)), 694. https://doi.org/10.21123/bsj.2020.17.2(SI).0694

[9] Bai, C., Chen, J., Huang, L., Kpalma, K., Chen, S. (2018, January). Saliency-based multi-feature modeling for semantic image retrieval. Journal of Visual Communication and Image Representation, 50, 199–204. https://doi.org/10.1016/j.jvcir.2017.11.021

[10] Shrivastava, N., Tyagi, V. (2017, December). Corrigendum to "Content based image retrieval based on relative locations of multiple regions of interest using selective regions matching" [Information Sciences 259 (2014) 212–224]. Information Sciences, 421, 273. https://doi.org/10.1016/j.ins.2017.09.017

[11] Halawani, A. H., Teynor, A., Setia, L., Brunner, G., Retrieval, C. I. (2006, January). Fundamentals and Applications of Image

[12] Ghrabat, M. J. J., Ma, G., Maolood, I. Y., Alresheedi, S. S., Abduljabbar, Z. A. (2019, December). An effective image retrieval based on optimized genetic algorithm utilized a novel SVM-based convolutional neural network classifier. Human-centric Computing and Information Sciences, 9(1), 31. https://doi.org/10.1186/s13673-019-0191-8

[13] BERT For Measuring Text Similarity https://towardsdatascience.com/bert-for-measuring-text-similarity-eec91c6bf9e1

[14] Text Classification with TF-IDF, LSTM, BERT: a comparison of performance https://medium.com/@claude.feldges/text-classification-with-tf-idf-lstm-bert-a-quantitative-comparison-b8409b556cb3