Content/Context Based Image Retrieval

Parth Chhabra
parth19069@iiitd.ac.in

Moksh Aggarwal
moksh19177@iiitd.ac.in

Srijan Garg
srijan19448@iiitd.ac.in

Mohd. Siraj Ansari
siraj19176@iiitd.ac.in

1 INTRODUCTION

1.1 MOTIVATION

Image retrieval is a burgeoning area of research in the field of deep learning due to its diverse applications across a range of domains, including medical image analysis, content discovery and recommendation in fields such as fashion, social media, and e-commerce, and planning and monitoring in geographical information systems. It also has potential applications in image annotation, classification, and even image-based search engines. Despite the existence of several systems for image retrieval, they are often limited in their ability to handle either visual features or context, thereby necessitating the development of a more comprehensive and integrated approach. As such, the primary objective of our system is to construct a sophisticated image retrieval system that effectively incorporates both visual features and contextual information, thereby enabling the retrieval of images that are not only visually similar but also contextually relevant. Our proposed system has the potential to provide highly personalized and relevant search results, thereby improving the search process efficiency and enhancing the user experience. Moreover, the development of such a system will contribute to the advancement of the field of deep learning by providing a more comprehensive approach to image retrieval that takes into account both visual and contextual features, thereby improving the accuracy and relevance of the search results.

1.2 PROBLEM STATEMENT

We aim to tackle two main problems of searching images: given an input query image, find images similar to it. For example, if we give image of a cat as an input, we expect the information retrieval system to give images of other cats as output. We would also like to remove false positives i.e. we would want to avoid retrieving images of dogs. This kind of retrieval is known as content-based image retrieval.

The second kind of retrieval is called context-based image retrieval. Here, we give textual input to the information retrieval system and the system outputs images that are best described by this text. For example, if the textual input is "Dancing people", the system should return images of people dancing. It should avoid false positives e.g. images of animals since we didn't specify it in our textual query.

Note that since image data is usually very large, it is infeasible to compare the input image or text with all possible images in the dataset. To overcome this problem, we usually find/learn an embedding for each image which contains only the important information about the image. In a good embedding space, similar images are closer to each other whereas dissimilar images are farther apart.

Models like CLIP use contrastive loss which does exactly this: it rewards similar images having similar embeddings whereas it penalizes dissimilar images having similar embeddings.

2 LITERATURE REVIEW

[7] describes the CLIP model developed by OpenAI. It is a multimodal vision-language model which can be used for tasks like image-to-text retrieval and in our specific case, text-to-image retrieval. CLIP is also a backbone model for many advanced image retrieval models and is usually either used as a pre-trained model or the pre-trained model is fine-tuned to get better results on cross-domain tasks.

The paper [5] describes an autoencoder framework that generates a binary mask for an image. The binary mask is then used for image retrieval. The paper also describes the advantages of binary masks over real-valued representations. Binary masks are cheap to store and very fast to compare using bitwise operations. If the masks are small enough, they can be used for "semantic" addressing where similar images can be directly accessed by flipping some bits of the mask. The obvious trade-off of using a binary mask is loss of information since for a mask of size n, there are only 2^n possible masks.

The paper [8] describes a deep hashing framework, named Deep Incremental Hashing Network (DIHN) for learning binary hash codes of images in an incremental way i.e. as new images are added to the dataset, the codes of previous images remain unchanged and the codes of the new images are learned directly. This method is fast as the model does not need to be retrained including the new images. Since the codes for original images remain unchanged, they don't need to be fed into the model to get new codes. The method provides a considerable speedup over conventional CNN-based models while not compromising the state-of-the-art retrieval accuracy.

The authors of [2] propose a self-supervised method to train feature representations of the images in the database. Instead of just having one query image, their method is built to incorporate multiple query images and then by using attention-based architecture, they extract features from diverse image aspects like object aspects and scene aspects which will in turn take advantage from the feature representations learnt. Moreover, they use videos in their training data instead of random images, in which they make use of naturally occurring object transformations. This helps them to avoid unnecessary data augmentation.

Unsupervised disentangling refers to disentangling factors of variation without any information about the labels of the data. β-VAE [3] bounds the KL divergence term in a standard VAE which forces the latent distribution to be closer to the standard normal

and thus leads to disentangled representations. Since it is a modification of the standard VAE, it doesn't require labelled data. [4] introduces an unsupervised model that uses autoencoders to learn disentangled representations of images. It uses autoencoders to generate feature representations which are divided into chunks of equal size. Each chunk represents some discernible attribute of the image. The paper interprets disentanglement as invariance i.e. in disentangled representations, encoding of each image attribute into feature chunks should be invariant to changes in other image attributes and vice versa. To achieve this, they mix chunks from features of two different images and pass the mixed feature vector through the decoder. To avoid the shortcut problem, they introduce a classifier that forces each chunk to have some information about the image.

Semi-supervised disentangling models have given good results in recent times. [1] uses latent optimization to optimize the disentangled representations directly. [6] introduces a semi-supervised conditional generative model which divides the image into two sources of variation: an observed variable s (specified factor) and a continuous latent variable z (unspecified factor; characterizes the remaining variability). s is given by a vector of real numbers as opposed to a class ordinal or one-hot vector. z is given by a Gaussian distribution characterized by its mean and variance. Any stochasticity in s comes from the data distribution. It uses a discriminator to counter degenerate cases where all the image information is stored in the unspecified space and none is stored in the specified space.

Learning disentangled representations allows us to make conditional image retrieval models since the representations are now divided into chunks based on different factors of variation.

3 DATASETS

3.1 Flickr30

The Flickr30k dataset is a popular benchmark for sentence-based picture portrayal. The dataset has over 31,000 images. Each image in the dataset has five reference sentences provided by human annotators.

The dataset can be used to analyse the performance of models on text-to-image (context-based image retrieval) retrieval tasks.

Figure 1: Flikr30k example

3.2 GPR1200

GPR1200 is a benchmark dataset for standard content-based image retrieval.

It contains 1200 categories and 10 class examples. This dataset is a standard dataset used to find similar images given an input image.

Figure 2: GPR1200 example

4 METHODOLOGY

4.1 Content Based

We have extracted different types of visual features like Local Binary Patter(LBP), Hu Moments(shape-based features), Histogram of Oriented Gradients(HOG) and DL based features like ResNet50 feature vector(ResNet Features), ResNet50 Multi-classification(Tag features).

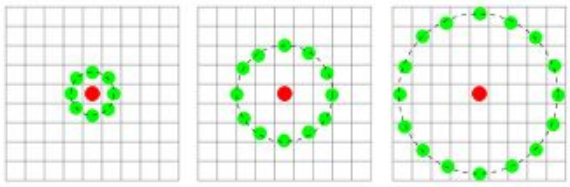

Figure 3: Local Binary Pattern

Figure 4: Hu Moments

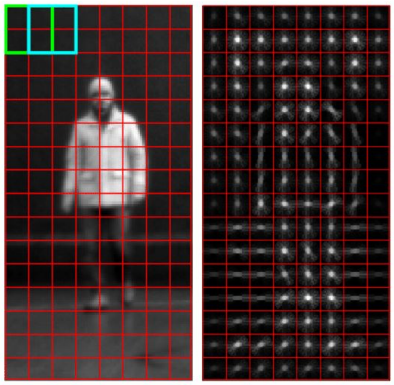

Figure 5: Histogram of Oriented Gradients

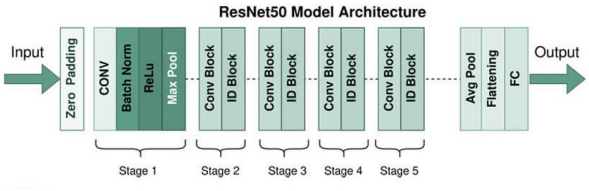

Figure 6: ResNet50 Model

Then, for scoring images based on the above features we have used cosine similarity for tag features and distanced based similarity using Euclidean distance for rest of the features. Also, we have scored images using a combination of all of the above features and their respective similarity measures. For all of the above scoring techniques, we have retrieved top 9 images based on scores.

$$\text{Cosine Similarity}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}$$

$$\text{Euclidean Distance}(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}$$

$$\text{Distance Based Similarity}(a, b) = \frac{1}{1 + \text{Euclidean Distance}(a, b)}$$

4.2 Context Based

For context-based image retrieval, we make use of the CLIP model. The CLIP (Contrastive Language-Image Pre-Training) model is a neural network that can understand both images and text. Its methodology involves pre-training on a large dataset of image and text pairs using a contrastive loss function, a transformer-based architecture, and fine-tuning for specific tasks. The model accepts multimodal inputs and learns a joint representation space that captures the relationships between images and text. This approach enables the model to perform tasks such as image captioning, visual question answering, and image retrieval.

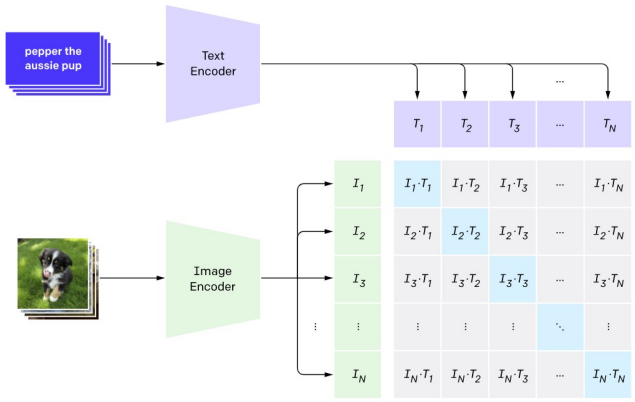

Figure 7: CLIP Model

5 RESULTS AND ANALYSIS

5.1 Content-based image retrieval

The results for content-based image retrieval is provided in Table 1. The performance using different sets of visual features is measured for this case using precision and recall as the metric.

Features	Precision	Recall
Tag Features[ResNet-50]	0.3911	0.3911
LBP Features	0.1836	0.1836
Shape-based Features	0.1259	0.1259
HOG Features	0.2095	0.2095
ResNet Features	0.2969	0.2969
Combined Features	**0.42169**	**0.42169**

Table 1: Content Based Image Retrieval: Performance

Also, the Figure 8 represents the images that we retrieve for a given query image in case of content-based retrieval.

From the results, we conclude that combining all the visual features generates the best results in case of content-based retrieval. The reason behind this can be explained by a variety of factors like :

- **Complementarity**: These feature extraction techniques each focus on a distinct visual element of an image. ResNet-50 is effective in identifying high-level semantic features, whereas LBP, HOG, and shape-based features are effective at capturing texture, shape, and structural information. Combining these elements makes it possible to capture a more comprehensive and diverse range of image features, improving the representation of the image's content.

Figure 8: Content-based Retrieval Results

- **Discriminative Power**: The discriminative capability of the feature set can be improved by combining features from several sources. As a result, it is simpler to discriminate between images that have a common theme, increasing the retrieval performance.
- **Robustness**: Combining multiple features increases the robustness of representations to variations in factors like scale, illumination, rotation etc. that can influence the appearance of images even with similar content

5.2 Context-based image retrieval

In the case of context-based image retrieval, the Figure 9 represents the results that we got for an input textual query: "people dancing". As from the result for the query, we can clearly observe that the retrieved images directly contain the people engaged in dancing (or similar) activities or are either quite close to the context in which dancing is a recurring theme.

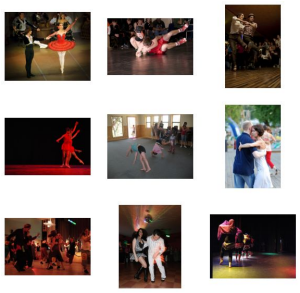

Figure 9: Context-based Retrieval Results

6 CONCLUSION

We test different methods of content-based image retrieval. The results are summarized in 1. From the results, it is evident that Tag Features [ResNet-50] performs the best. However, combining all the features provides us with the best results.

We also perform context-based image retrieval. We observe that the query "people dancing" gives us images of people dancing which shows us that the model captures the context of the textual input it receives and appropriately processes it to fetch relevant images. The CLIP-based model developed by OpenAI provides the backbone for the image retrieval model. It uses both text and image to learn a model that can perform the efficient text to image retrieval.

7 FUTURE WORK

Future work in the field of image retrieval could focus on improving feature extraction techniques, incorporating multi-modal data, implementing attention mechanisms, developing domain-specific retrieval models, exploring interactive retrieval systems, utilizing transfer learning, and creating privacy-preserving models. Current feature extraction methods can be further optimized to capture more diverse and robust visual features, while the incorporation of other types of data, such as audio or video, could improve accuracy. Attention mechanisms could enhance model performance by focusing on relevant regions of an image, and domain-specific retrieval models could leverage specialized knowledge and features. Interactive retrieval systems, transfer learning, and privacy-preserving models are also areas of interest, with the potential to improve accuracy, efficiency, and security. Further exploration in these areas could advance the field of image retrieval and expand its potential applications in various domains.

REFERENCES

[1] Aviv Gabbay and Yedid Hoshen. 2019. Demystifying inter-class disentanglement. *arXiv preprint arXiv:1906.11796* (2019).

[2] Hariprasath Govindarajan, Peter Lindskog, Dennis Lundström, Amanda Olmin, Jacob Roll, and Fredrik Lindsten. 2021. Self-Supervised Representation Learning for Content Based Image Retrieval of Complex Scenes. In *2021 IEEE Intelligent Vehicles Symposium Workshops (IV Workshops)*. IEEE, 249–256.

[3] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. (2016).

[4] Qiyang Hu, Attila Szabó, Tiziano Portenier, Paolo Favaro, and Matthias Zwicker. 2018. Disentangling Factors of Variation by Mixing Them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Alex Krizhevsky and Geoffrey E Hinton. 2011. Using very deep autoencoders for content-based image retrieval.. In *ESANN*, Vol. 1. Citeseer, 2.

[6] Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. 2016. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/ef0917ea498b1665ad6c701057155abe-Paper.pdf

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

[8] Dayan Wu, Qi Dai, Jing Liu, Bo Li, and Weiping Wang. 2019. Deep incremental hashing network for efficient image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9069–9077.