Team 34: Enhancing Image Search with Fusion of Visual and Textual Information

Neshmia Hafeez 7048035
Babar Tameez 7047847
Maheen Saleh 7047619

Abstract

The task of image search typically relies on the extraction of visual aspects of an image and comparing them to find similar images, however, it may fail to capture the complete semantics of an image and hence not give the most relevant search results. We aim to bridge this gap by introducing semantic information of the image along with its visual information to create a more rich and useful representation of the image. This work investigates the impact of this fusion-based image search by fusing the image's visual and semantic embedding and performing qualitative and quantitative analysis.

1. Introduction

Image search which is based on the use of only the visual aspects of image mostly employs the use of low level image features like color histograms, edges etc. These features cannot address the underlying high level contextual information in the image which can be accessed through the textual description or the caption of the image.

The main idea of this work lies in the recognition that while visual features excel in capturing the surface attributes of images, they often fall short in encapsulating their rich semantics. In contrast, textual descriptions and captions offer a unique perspective, supplying contextual details that complement the visual domain. By fusing these two modalities together, a holistic representation of images can be made, encompassing both their visual and semantic description of the image.

2. Related Work

Fusion methods have been used for a wide array of tasks. One such example was the use of cross modal embeddings to map text (cooking recipes) and images (food images) to a common space [7] Thus, when a query image (an image of a dish) was input, its similarity was calculated with the existing dataset, and then the recipe existing close to the similar image in the common space was output. Hence, an image retrieval system is made using text and image embeddings, however, as opposed to our system, it works on image-to-text search instead of on image-to-image search.

Another use of fusion methods to enhance image retrieval was in [4], where the SDCD and the PHOW MSDSIFT histograms were fused together to compute a more comprehensive image search. Although this work uses a fusion method to run a better image search, it is a uni-modal fusion method as only two different kinds of histograms are fused together.

Fusion models have also been used in other use-cases as well. Fusion models containing Auxiliary Language Models (AuxLM) and Masked Language Models (MLM) have been used for the task of improved visual captioning [5]. Similarly, latent semantic fusion models have also been used for automatic image annotation [6]. Cross-modal fusion methods have also been used for image-sentence matching [9]. However, there has been little to no work in the use of fusion methods for image-to-image search. Thus, we propose a unique way of conducting image retrieval using a fusion method that uses semantic as well as visual information to perform a much more comprehensive image search.

3. Methodology

The methodology of this visual and semantic fusion-based image search is based on first generating the required embedding from the image using multiple pre-trained models, fusing them through different approaches to create a database of embeddings that be compared to the fused embeddings of the query image to search the relevant similar images through a similarity function.

The dataset used is MS COCO data set due to the large dataset size and its diverse images. The variety of classes and images in the COCO dataset help in better analysis for our experimentation. We have taken 10K images from the COCO dataset which includes all 80 classes of the dataset.

Table 1 shows the combinations of pretrained models

that have been employed to generate two different sets of embeddings of the dataset images.

No.	Visual Embedding	Captioning	Semantic Embedding
1	ResNet50	Visual Encoder Decoder (ViT and GPT2)	Sentence-BERT
2	EfficientNet	Visual Encoder Decoder (ViT and GPT2)	Universal Sentence Encoder

Table 1. Model Combinations for Visual and Semantic Embedding

The reason for generating two different sets of fused embeddings based on different model architectures is to observe and compare the consistency of the multiple fusion approaches employed to merge the visual and semantic embeddings.

3.1. Visual Embedding

For visual embeddings, two state-of-the-art pretrained CNN models have been used i.e. ResNet50 and EfficientNet.

ResNet50 [2] is chosen for generating visual embeddings because of its proven state-of-the-art performance in various tasks and to capture intricate features across different levels of abstraction.

EfficientNet's [8] capacity to deliver strong performance within efficient computational constraints makes it a valuable choice for generating image embedding. It has demonstrated robust performance across a wide range of image-related tasks, including image classification, object detection, segmentation, and more.

3.2. Image Captioning

Since we aim to generate image captions within the image search pipeline, therefore our dataset does not contain any captions. To generate the image captions, we used Visual Encoder-Decoder Model [3] i.e. ViT as Encoder and language model GPT 2 as decoder. The reason of not using a dataset with pre-existing captions is to address the fact that the query image will not have any caption and it would be generated at the time of image search, hence to add uniformity in the semantic representation of the image, we generated captions of the dataset images in the same manner as would be generated during a search query.

3.3. Semantic Embedding

The language models used to extract embeddings from the generated captions of images are the pre-trained SentenceBert Model (all-MiniLM-L6-v2) and Universal Sentence Encoder.

Sentence-BERT `https://www.sbert.net/docs/pretrained_models.html` employs transformer-based models like BERT to capture deep contextual and semantic understanding of sentences. This enables the model to encode not only the words but also the relationships between them.

USE [1] leverages transformer-based architectures to capture rich semantic information from sentences. This allows it to create embeddings that encapsulate the underlying meaning and context of sentences, making them valuable for a wide range of natural language understanding tasks.

3.4. Fusion Method

For fusing the visual and semantic embedding, we implemented the following three different approaches:

- Concatenate the visual and semantic embeddings without any normalization.
- Normalize the visual and semantic embeddings separately and later concatenate the normalized embeddings.
- Concatenate the visual and semantic embeddings and then normalize the resulting fused embedding.

3.5. Search Query Method

To perform an image search query on the dataset of fused embeddings, we first generate the fused embedding of the query image using the same models as used for the dataset embeddings. Cosine similarity has been used to compute the similarity of query embeddings with dataset embeddings and return the top matches.

4. Experimental Results and Analyses

Qualitative and quantitative analysis has been done for both combinations of the embedding models using all three fusion methods. For further comparison, we have also performed image search using only the visual embeddings and caption embeddings separately.

4.1. Quantitative Analysis

Micro precision and recall have been calculated to evaluate the image search quantitatively. To calculate the metrics, the test images were processed such that they contain only one object per image hence this evaluation is based on single-label images. Thresholding is done on the similarity score of the cosine function to declare images as matched or not matched.

The precision-recall curve for both combinations of models is shown in the figure 2 according to which the two

best performing methods are caption-based search and pre-normalized fusion (first normalize, then concatenate) based search.

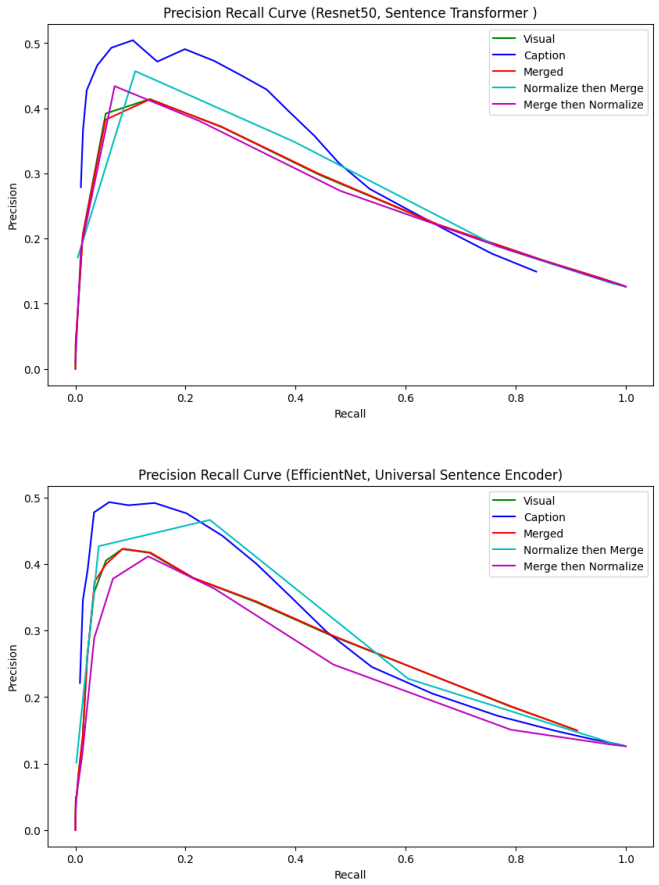


Figure 1. Precision-Recall Curve

Method	Precision	Recall
Visual Search	0.28	0.02
Semantic Search	0.46	0.03
Fusion Search	0.28	0.02
Pre-Normalized Fusion Search	0.28	0.54
Post-Normalized Fusion Search	0.41	0.12

Table 2. Precision and Recall Values for Resnet50 and Sentence Bert, at threshold = 0.83

Table 2 and 3 shows the precision and recall at a specific threshold according to which precision is highest for image search based on semantic(caption) embeddings only however it has a very low recall. The highest recall is for the fusion method where embeddings are first normalized and then concatenated. These observations of different fusion methods are consistent for both combinations of the embedding generation models.

Model	Precision	Recall
Visual Search	0.05	0.002
Semantic Search	0.49	0.09
Fused Search	0.05	0.002
Pre-Normalized Fused Search	0.12	0.99
Post-Normalized Fused Search	0.38	0.07

Table 3. Precision and Recall Values for EfficientNet and Universal Sentence Encoder, at threshold = 0.7

4.2. Qualitative Analysis

The results of the qualitative analysis are similar to quantitative analysis where caption-based embedding and pre-normalized fused embeddings-based search give better results than others. In Fig 2 and 3, the first image of every row is the query image. It shows the result of top-4 matches using cosine similarity for all the different implemented configurations of image search.

(a) Visual Search

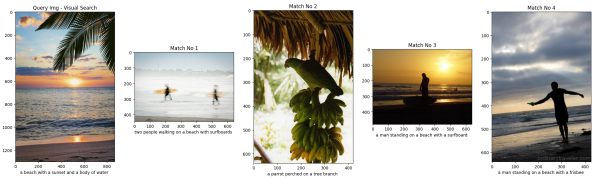
(b) Caption Search

(c) Fusion Search

(d) Pre Normalized Fusion Search

(e) Post Normalized Fusion Search

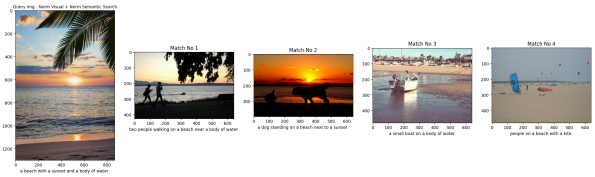
Figure 2. Results of image search with Resnet50 and Sentence-Bert, caption of query image: ”a beach with a sunset and a body of water”



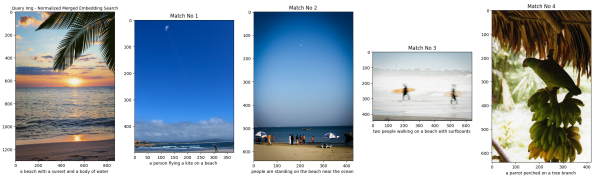
(a) Visual Search

(b) Caption Search

(c) Fusion Search



(d) Pre Normalized Fusion Search



(e) Post Normalized Fusion Search

Figure 3. Results of image search with EfficientNet and Universal Sentence Encoder, caption of query image: a beach with a sunset and a body of water

5. Conclusion

The above results strongly support the proposed idea that the fusion of semantic information along with the visual can leverage the image search task and that the best performance is obtained when the two embeddings are concatenated after they are normalized individually i.e. pre-normalized fusion of embeddings. The results remain consistent for both the combinations of pre-trained models used for generating captions and embeddings which adds to the validity of fusion results. In this work, only a few mathematical methods have been explored to fuse embeddings from the two modes, however, fusion through other approaches such as a neural network is expected to give better results. Moreover, other combinations and methods of generating embedding can be experimented with for further improvements.

References

[1] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Céspedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 169–180, 2018. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016. 2

[3] Ankur Kumar. The illustrated image captioning using transformers. 2022. 2

[4] Lili Nurliyana Abdullah Azreen Azman Mas Rina Mustaffa Leila Mansourian, Muhamad Taufik Abdullah. An effective fusion model for image retrieval. In *Multimedia Tools and Applications*, 2018. 1

[5] Marius Mosbach Dietrich Klakow Marimuthu Kalimuthu, Aditya Mogadala. Fusion models for improved visual captioning. 2020. 1

[6] Trong-Ton Pham, Nicolas Eric Maillot, Joo-Hwee Lim, and Jean-Pierre Chevallet. Latent semantic fusion model for image retrieval and annotation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 439–444, New York, NY, USA, 2007. Association for Computing Machinery. 1

[7] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3068–3076, 2017. 1

[8] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8567–8576, 2019. 2

[9] Xing Xu, Yifan Wang, Yixuan He, Yang Yang, Alan Hanjalic, and Heng Tao Shen. Cross-modal hybrid feature fusion for image-sentence matching. *ACM Trans. Multimedia Comput. Commun. Appl.*, 17(4), nov 2021. 1