

# Deep Learning-Based Multimodal Image Retrieval Combining Image and Text

Md Imran Sarker

Department of Computer Science  
University of Arkansas at Little Rock  
Little Rock, Arkansas, USA  
[misarker@ualr.edu](mailto:misarker@ualr.edu)

Mariofanna Milanova

Department of Computer Science  
University of Arkansas at Little Rock  
Little Rock, Arkansas, USA  
[mgmilanova@ualr.edu](mailto:mgmilanova@ualr.edu)

**Abstract**— *Multimodal learning is omnipresent in our lives. Human absorbs features in different ways, whether through pictures or text. Combining these features in computational science, especially in Image retrieval problems, poses two significant challenges: how and when to fuse them. Most image retrieval systems use images or text data associated with the image. In this paper, we study the image retrieval task, where the input query is an image plus text sentence that describes the image. The system starts a query triggered by input image and text while taking the help of the Transformer model, which puts attention on both modalities and combines embedded features through the feature fusion technique. We proposed a feature fusion layer using modified Text Image Residual Gating in our work. We have used two methods based on the features extracted from the fusion layer. First, we trained K Nearest Neighbor (KNN) algorithm on the training data, and later we used test data to find a similar image. Second, we used the clustering technique and a support vector machine to compute the nearest neighbor points and cluster the center to see a similar image. We found that SVM (Support vector Machine) is more effective from the results, giving an overall accuracy of 92%.*

**Keywords**—*Image Retrieval, Multimodal, Computer Vision, Deep Learning*

## I. INTRODUCTION

Image retrieval problem has been an active area of research for the last twenty years. With different data sources available in the public domain, the researcher has started looking into the use case of image and text fusion for image retrieval problems. Humans recognize images based on high-level concepts. On the other hand, content-based image retrieval extracts visual low-level. As a result, a content-based image retrieval system needs to overcome that drawback and improve performance. Text-based image retrieval lacks in analyzing content. Because semantics are not always associated with image content. There exists an enormous gap between low-level visual features and high-level semantic information. To overcome this gap, researchers focus on

multimodal fusion techniques to make a robust image retrieval system. Deep Neural Networks have provided methods for similarity matching within text and images with the aid of embeddings gained from other Deep Learning methods, such as Convolutional Neural Networks (CNN) and Bidirectional Encoder Representations from Transformers (BERT)[1]. An interesting aspect where additional research could be done is retrieving images by providing a reference image and text feedback from the user. Finding similarity for matching purposes in multimodal data is also an interesting area, as it has practical applications. Research by Liwei Wang et al. [2] investigated two-branch neural networks for learning the similarity between two data modalities. They used the Flickr30k entities dataset for phrase localization. The MSCOCO dataset was used for bi-directional image sentence retrieval. The work mainly focused on neural architecture for a core problem underlying most image-text tasks-how to measure the semantic similarity between visual data, e.g, image, and text. Their work serves as an attractive alternative to the embedding network for region phase matching but doesn't work for image-sentence retrieval. A modified text concerning an input image might not precisely describe the user's intentions by a single image or text. This research effort is mainly related to text guided Image Retrieval approach [3]. Yanbei Chen et al [4] proposed a Visiolinguistic Attention Learning model which takes an input image and text to retrieve the image. The retrieved image shows a change in certain aspects of the given text. The text modifies the visual content of the reference image.

Other researchers have done similar work to improve multimodal image retrieval tasks. [5][6][7][8]. Fashion IQ and Fashion200k datasets have been used to validate model performance. Nam Vao [8] et al. extended the research of Multimodal Image Retrieval by proposing residual connection, which is a way of composing image and text. They have achieved (SOTA) results on this task. However, their approach doesn't perform well in the real world as the

model focuses more on image space and less on query text. Less importance on text happens when the model carries long and detailed sentences [8]. The above facts motivated us to explore the transformer model and fusion techniques in multimodal image retrieval. To the author's knowledge, the proposed work is the first to perform the feature of SIFT and BoVW along with the transformer model for image retrieval. The key contribution of this project is summarized as follows:

1. Multimodal Query (Image & Text) has been trained with a Transformer model to get embedded features.
2. The features have been combined with modified Text Image Residual Gating.
3. Unsupervised machine learning algorithm has been used to retrieve similar images.

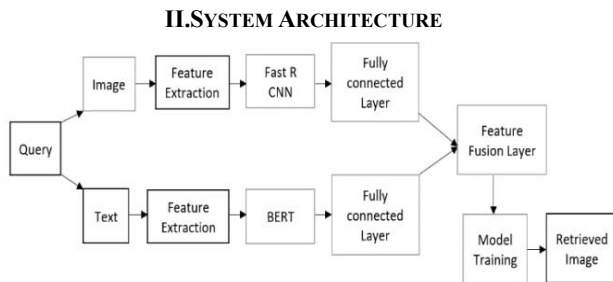


Figure 1: System Architecture Flowchart

### III.METHOD

#### A. Feature Extraction & Transformer Model

We aim to solve image retrieval using the multimodal task as discussed earlier. In our method, we extracted and embedded features from text & image query to find retrieval images. Our work has been inspired by VisualBert model [10] to carry out self-attention tasks in both modalities. VisualBert consists of a stack of Transformer layers that implicitly align elements of an input text and regions in an associated input image with self-attention. We used the BERT model [1] to work with the text query. We employ a pre-trained BERT model for extracting text features instead of LSTM. All the words in a sentence are mapped to a set of embeddings. Each embedding  $e \in E$  is computed as 1) a token of embedding, 2) a segment embedding 3) and a position embedding. The input embeddings are then passed into the attention layer, which produces a context representation of the sub-words. To work with Image query, we have used Scale Invariant Feature Transform and a Bag of Visual Words to model an image. Each embedding  $f \in F$  corresponds to the bounding region in the image. Detectron2 [11] has been used to derive an object detector from the image. The bounding region has been computed by the Fast R CNN model. The visual embeddings are then passed into a multi-layer transformer model.

#### B. Feature Fusion layer

To combine image and text features, we used text image residual getting, but we have updated this method because we don't want to modify text as our expected output. Nam Vo et al. [9]. In their work, they used an Eiffel tower image and

asked the system to find visually similar images but modified in small ways. But in our task, the output won't be different from the input image. Instead, it would be a similar image.

$$\phi_{xt} = \omega_g \text{fgate}(\phi_x, \phi_t) \quad (1)$$

In equation 1,  $\text{fgate}$  is gating features.  $\omega_g$  is the learning weights.  $\phi_x$  is the image, and  $\phi_t$  is the text. Inspired by [14, 15, 16], we propose combining image and text features using equation 2. In our work, we have only used  $\text{fgate}$ . The gating connection is computed by:

$$\text{fgate}(\phi_x, \phi_t) = \sigma(Wg_2 * \text{RELU}(Wg_1 * [\phi_x, \phi_t]) \odot \phi_x) \quad (2)$$

Where,

$\sigma$  is the sigmoid function

$\odot$  is element wise product

$*$  represents 2d convolution with batch normalization

$Wg_1$  and  $Wg_2$  are 3x3 convolution filters

To keep the image in feature map compatible  $\phi_x$ , we transmit  $\phi_t$  in such a way that height and width dimension falls under the shape.

#### C. Model Training

Feature decomposition keeps the image and text embedded into one single vector. The similarity between two data points can be attained using KNN (K nearest neighbors). KNN is a machine learning model that identifies the closest approximate neighbors to the input data. To retrieve a similar image, the comparison is done by converting the image into vectors. The neighbors are identified by comparing the fusion vector with the trained model in a multi-dimensional plane. The value of K determines how far we want to expand the search comparison results. We used Euclidean distance when comparing an input image vector with the trained data vector. We used exact search, which is linear search and space partitioning, to keep the quality of image retrieval. Suppose x and y are input and database feature vectors. Then the Euclidian distance between the two vectors is explained in equation 3.

$$d_{euclid} = \sqrt{\sum_{i=1}^n (xi - yi)^2} \quad (3)$$

The dataset we used has constraints. We have more than 2000 different captions in the NLVR (dev) dataset for 6000 images/captions pairs. So, we have roughly three samples per caption only. To improve the model, we used K-means clustering and trained those clusters with the Support Vector Machine algorithm [12] [13]. By taking fifteen clusters, the result got improved. Equation 4 was used to calculate accuracy.

$$\text{Accuracy} = \frac{TP+TN}{(TP+TN+FP+FN)} \quad (4)$$

Where  $TP$  = True Positives,  $TN$  = True Negatives,  $FP$  = False Positives, and  $FN$  = False Negatives







Query No.	Query Input (Text & Image)	Results
1.	 <p data-bbox="272 657 787 688">"Human Wearing Graduation gown and Cap"</p>	 <p data-bbox="824 289 1425 678">0 25 50 75 100 125</p>
2.	 <p data-bbox="272 1245 625 1276">"In an image at least one fluiet"</p>	 <p data-bbox="824 709 1425 1266">0 25 50 75 100 125</p>
3.	 <p data-bbox="272 1776 657 1808">"A school bus on the left in Image"</p>	 <p data-bbox="824 1304 1425 1791">0 25 50 75 100 125</p>

Figure 2: Result from Multimodal Image Retrieval combining image and text

## IV. EXPERIMENTAL RESULT

In this paper, we used NLVR Dataset. NLVR is a dataset for joint reasoning about natural language and images with a focus on semantic diversity, compositionality, and visual reasoning challenges. NLVR dataset is popular for visual question-answering problems and visual reasoning. Having a benchmark performance in visual reasoning, we choose the NLVR dataset for our experiment. Because of computational limitations, we have used only the development dataset, and this development dataset covers 6000 images and JSON test data. Using the KNN algorithm, we get 79 percent accuracy. The SVM algorithm showed 92 percent accuracy (Figure 3). We have visualized our result on page 4. In figure 2, we also have shared the result of our proposed method adding a few examples of query (text & image) and result.

Our main metric for image retrieval is accuracy. We computed the percentage of test queries where at least one target is the correct image within the top K retrieved images. Each experiment is repeated 10 times to obtain a stable retrieval performance. Two of the most well-known assessment measures are precision and Recall. In any case, the weakness of recall is that it is determined for the whole retrieved set and is unaffected by the rankings of the significant substances in the retrieved list.

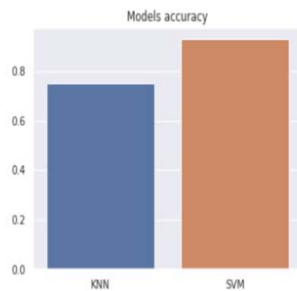


Figure 3: KNN & SVM Algorithm Accuracy

## CONCLUSION

In this paper, we explored Multimodal based image and text query for image retrieval. We experimentally evaluated the feature fusion layer modifying image text residual gating. In the future, we would like to try much bigger and benchmark datasets like Fashion200k and Fashion IQ. To make our model robust, we hope to compare it with state of the arts multimodal fusion architectures.

## ACKNOWLEDGMENT

The research is funded by the NSF I-Corps #21552-National Innovation. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

## References

- [1] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019, May 24). Bert: Pre-training of deep bidirectional Transformers for language understanding. <https://arxiv.org/abs/1810.04805>
- [2] Wang, L., Li, Y., Huang, J., & Lazebnik, S. (2018, May 01). Learning two-branch neural networks for image-text matching tasks. <https://arxiv.org/abs/1704.03470>
- [3] Liu, Y., De Nadai, M., Cai, D., Li, H., Alameda-Pineda, X., Sebe, N., & Lepri, B. (2020, August 10). Describe what to change: A text-guided unsupervised image-to-image translation approach. <https://arxiv.org/abs/2008.04200>
- [4] Chen, Y., Gong, S., & Bazzami, L. (n.d.). Image search with text feedback by Visiolinguistic attention learning. <https://ieeexplore.ieee.org/document/9157634>
- [5] Yu, Y., Lee, S., Choi, Y., & Kim, G. (2020, March 30). CurlingNet: Compositional learning between images and text for fashion IQ Data. <https://arxiv.org/abs/2003.12299>
- [6] Lee, S., Kim, D., & Han, B. (n.d.). Cosmo: Content-style modulation for image retrieval with text feedback. <https://ieeexplore.ieee.org/document/9577437>
- [7] Shin, M., Cho, Y., Ko, B., & Gu, G. (2021, October 26). RTIC: Residual learning for text and image composition using graph Convolutional Network. <https://arxiv.org/abs/2104.03015>
- [8] Anwaar, M., Labintcev, E., & Kleinsteuber, M. (2021, May 31). Compositional learning of image-text query for Image retrieval. <https://arxiv.org/abs/2006.11149>
- [9] Vo, N., Jiang, L., Sun, C., & Murphy, K. (n.d.). Composing text and image for Image retrieval - an empirical odyssey.
- [10] Li, L., Yatskar, M., Yin, D., Hsieh, C., & Chang, K. (2019, August 09). VisualBERT: A simple and performant baseline for vision and language. <https://arxiv.org/abs/1908.03557>
- [11] Liu, X., Sarker, M., Milanova, M., & O’Gorman, L. (2021). Video-based monitoring and analytics of human gait for Companion Robot. [https://link.springer.com/chapter/10.1007/978-981-33-4676-5\\_2](https://link.springer.com/chapter/10.1007/978-981-33-4676-5_2)
- [12] <https://ai.facebook.com/tools/detectron2/>
- [13] O’Gorman, L., Liu, X., Sarker, M. I., & Milanova, M. (2021). Video analytics gait trend measurement for Fall Prevention and Health Monitoring. <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9413226>
- [14] K. He, X. Zhang, S. Ren, and J. Sun. (2016) Deep residual learning for image recognition. CVPR.
- [15] A. Micch, I. Laptev, and J. Sivic. (2017) Learnable pooling with context gating for video classification. arXiv preprint arXiv:1706.06905.
- [16] R. K. Srivastava, K. Greff, and J. Schmidhuber. (2015) Highway networks. arXiv preprint arXiv:1505.00387