ATTENTION-GUIDED MULTIMODAL GRAPH NETWORKS FOR CROSS-MODAL IMAGE-TEXT MATCHING

KOMURAVELLI MOUNIKA and B.V. RAMNARESH YADAV
Department of Computer Science and Engineering, JNTU Hyderabad, Telangana, INDIA -500085

Abstract: The cross-modal retrieval, a technique used to align visual and text features to retrieve similar data from database. This technique always remains a challenging task due to the inherent heterogeneity between visual and text representations. This paper tried to give a solution using an attention guided multimodal graph network framework that leverages Vision Transformers (ViT), BERT, and Graph Convolutional Neural Networks (GCNNs) to create a unified and interpretable multimodal representation space for image-text matching. In this ViT used to extract image features and BERT for text features. These features are then structured into a heterogeneous graph name cross-model feature graph (CMFG), where nodes represent image patches and text tokens, and edges are weighted using self-attention and cross-attention scores to capture intra-modal and cross-modal relationships. A CMFG propagates and refines these features through attention-guided message passing, ensuring selective and context-aware alignment between modalities. The use of attention mechanism enhances the legibility of the model, providing intuitions into the associations between image regions and textual descriptions. The proposed method emphasizes the hypothetical integration of attention mechanism with graph-based learning for robust and scalable multimodal representation learning. The experimental results of proposed framework on benchmark datasets have shown a significant performance over state-of-the-art methods in terms of precision and recall metrics.

Keywords: Cross-modal image retrieval, GCNN, ViT, BERT, Attention mechanism, Multimodal graph network.

1. Introduction

In recent years, multimodal data has increased significantly due to the growth of internet. Many applications, such as intelligent search engines and multimedia data management systems, are being developed to analyze immense volumes of multimodal data [1]-[2]. In the current retrieval modal extensions, cross-modal retrieval, aimed at performing the task across different modalities like images, text, videos and audios etc., has garnered significant attention [3]-[6]. Traditional content-based image retrieval (CBIR) techniques are semantic similarity connections which are limited to single-modality scenario. However, cross-modal retrieval requires retrieving semantically similar items in one modality (e.g., text) using a query item from another modality (e.g., image). This paper concentrate on multi-labeled cross-modal retrieval, which has huge applications in multimedia retrieval, e-commerce etc [7]-[9].

Cross-modal retrieval, the task of aligning and retrieving semantically similar data from different modalities, such as images and text, has garnered increasing attention in the era of multimodal data. Cross-modal attention and generating a shared space mechanism are the common solutions for cross-modal retrieval [11]-[12]. These methods transform the low dimensional feature vectors in to high dimensional feature vectors, so that semantically related data points have similar feature code bits. Majority of the research work is going on reducing the sematic gap between diverse modalities with distinct data distributions [14]-[15]. Many cross-modal attention mechanisms have been proposed earlier. Lin et al.,[16] proposed a probability-based semantics-preserving hashing (SePH), which generated one single unique hash code, considering the sematic consistency between different modality views. In [17], authors used label consistent matrix factorization hashing (LCMFH). In [18] Modal-Adversarial Hybrid Transfer Network (MHTN) was introduced, and in [19] Semantic correlation maximization was employed for cross-modal retrieval. All these methods extract the features independent of training process, hence these methods may not have the satisfactory performance in many practical applications. In recent times, deep convolution neural networks (DCNNs) are used to extract fine features from data, which significantly improved the retrieval performance capability of the various frameworks [20]-[24]While these advancements have predominantly focused on single-modality retrieval, modern applications demand systems capable of bridging the semantic gap between different modalities, such as text and images, for effective cross-modal retrieval. In order to address this issue, researchers integrate the text and visual features into a shared embedding space, laid the foundation for cross-modal retrieval systems. Guo et al.,[25] proposed Prompts-in-The-Loop (PiTL), a weakly supervised method to pre-train VL-models for cross-modal retrieval tasks. Chen et al.,[26] UNITER, a Universal Image-Text Representation, learned through large-scale pre-training over four different image-text datasets for efficient cross-model retrieval. In [27], the researchers proposed a method (BeamCLIP) that can effectively transfer the representations of a large pre-trained multimodal model (CLIP-ViT) into a small target model. Previous cross-modal retrieval methods faced limitations in feature extraction and alignment. CNN-based models lacked global context, while LSTM and Word2Vec struggled with contextual understanding. Traditional fusion techniques, like concatenation and MLPs, failed to model inter-modal relationships effectively. These approaches lacked attention mechanisms, leading to poor retrieval accuracy.

To address the above mentioned problems, inspired by vision language model (VLM)[28], we propose a novel approach using GCNN in the attention mechanism for the features extracted using ViT [29] and BERT[30] for vision and language respectively to enhance the cross-modal feature alignment. A shared space mechanism is designed in which we try to integrate GCNN into the feature fusion process, enabling the model to capture fine-grained inter-model relationships by modeling the interactions between features as graph structure. This methodology may overcome the limits of conventional fusion methods by allowing the network to reason about the semantic relationships between visual and textual elements explicitly.

The model specific feature extractors used are:

- ViT: Vision Transformer used for image feature extraction. These transformers are succeeded in capturing the effective feature from images through self-attention mechanism.
- BERT: Bidirectional Encoder Representations from Transformers is used for text feature extraction. It provides the rich representations of semantic information from textual descriptions.

Joining these feature extraction mechanisms, we purpose a shared embedding space where features from these two different modalities features are mapped into an amalgamated representation.

- GCNNs with attention mechanisms model the relationships between visual and textual features, emphasizing semantically relevant regions and words.
- The attention mechanism prioritizes the most important cross-modal interactions, allowing the model to focus on meaningful correspondences while ignoring irrelevant noise.
- This shared embedding space ensures that the model can effectively bridge the semantic gap between text and images, making it possible to retrieve semantically aligned images based on textual queries with high precision.

Our proposed framework offers the following contributions:

- A GCNN-based attention mechanism that enhances semantic alignment across modalities by explicitly modeling relationships between features.
- Integration of ViT and BERT for robust and scalable feature extraction, ensuring high-quality visual and textual embeddings.
- A novel shared embedding space design that facilitates fine-grained cross-modal alignment, improving retrieval performance on complex datasets.

In the subsequent sections, related works with detailed literature of ViT and BERT in section 2, a detailed overview of the proposed framework in section 3 including its architecture, training process, and evaluation methodology. Additionally, comparison of the performance of proposed method against state-of-the-art approaches to demonstrate its effectiveness in cross-modal retrieval tasks in section 4 and conclusion of the manuscript in section 5.

2.Related works

Vision Transformer (ViT): The foremost building blocks of transformer architectures, and more specific recent architecture is ViT introduced by Dosovitskiy et al. in 2020 [31]. Unlike CNN, which depends on on convolutional operations to extract local features, ViT operates on sequences of image patches, treating an image as a series of non-overlapping patches and processing them as tokens. At first ViT, divides each image into *M*-fixed non-overlapping patches, flattens each patch, and project into higher dimensional space using a learnable linear projection. The process involved in ViT is explained clearly with the following equations:

Patch Embedding:

The given image $X\epsilon\mathbb{R}^{WXHXC}$ (H-height, W-width and C-no.of channels)

No.of patches $N = \frac{WXH}{P^2}$ (P-patch size)

$x_i = \mathbb{R}^{P^2.C}$ ($x_i -$Vector of Patches)

Project the patches into D-dimensional embedding space using a learnable matrix $E\epsilon\mathbb{R}^{(P^2.C).D}$

$z_i = E.x_i$ ($z_i\epsilon\mathbb{R}^D$ embedding of ith patch vector)

Adding Positional Embeddings:

Positional embeddings are used to maintain the positional relationship among the patches. These embeddings restore the spatial context by encoding the location of each patch within the image.

$$Z_E = [z_0, z_1, \ldots.., z_m] + P$$

Where, $Z_E = \mathbb{R}^{NXD}$ is the input to the transformer.

Multi-headed Self Attention Mechanism(MSA):

This is a core part of the transformer architecture. It enables the model to focus on the most relevant parts of input sequence when making predictions. In ViT, MSA applies a self-attention mechanism to capture the relationships among different regions of an image. For a given image patches $X\epsilon\mathbb{R}^{NXD}$ where, N is the number of patches, D is the embedding dimension, the self-attention mechanism works as follows:

Each token embedding is linearly projected into 3different vectors as Query(Q), Key(K), and Value(V).

$$Q = XW_Q \text{ , } K = XW_K, V = XW_V$$

Where, $W_Q, W_K, W_V\epsilon\mathbb{R}^{DXd_k}$ are learnable weight matrices, d_k is the dimension of query /key space.

Attention scores:

It is a value between all pairs of tokens are computed as the dot product of the query and key vectors, scaled by $\sqrt{d_k}$, to prevent from large values which can lead to gradient uncertainty.

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right),$$

$A \in \mathbb{R}^{NXN}$ is the attention matrix, representing the attention weights between each pair of tokens.

The attention weights are used to compute a weighted sum of value vector, as

$$Self - attention\ output = AV$$

$$Multi - headed\ output = Contact(head_1, head_{2,\ldots\ldots,} head_h)W_o$$

Where,

- $head_i = Self - Attention(Q_i, K_i, V_i)$
- $W_o \in \mathbb{R}^{(h.d_k)XD}$ is a learnable weight matrix.

BERT: It is a transformer-based model architecture for NLP tasks by extracting bidirectional context [32]. For each text input T, the BERT tokenizer converts the text into tokens and applies WordPiece tokenization to further break down into sub-words. Each sub-word is then mapped to its corresponding ID in BERT's vocabulary to produce raw word features. These tokens are represented as embeddings that combine token embeddings (E_{Token}), segment embeddings $\left(E_{Segment}\right)$, and positional embeddings $(E_{Position})$.

$$E_{input} = E_{Token} + E_{Segment} + E_{Position}$$

The embedded input is processed through multi transformer encoder layers, here each layer uses multi-head self-attention to compute relationships between tokens. The attention scores are calculated as:

$$Attention\_Score(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The final output from BERT provides contextualized representations for each token, with the **[CLS]** token representing the overall input sequence:

$$H_{[CLS]} = BERT(E_{input})$$

These representations are highly effective for tasks like text feature extraction in cross-modal retrieval, where textual embeddings from BERT are aligned with visual embeddings from models like ViT.

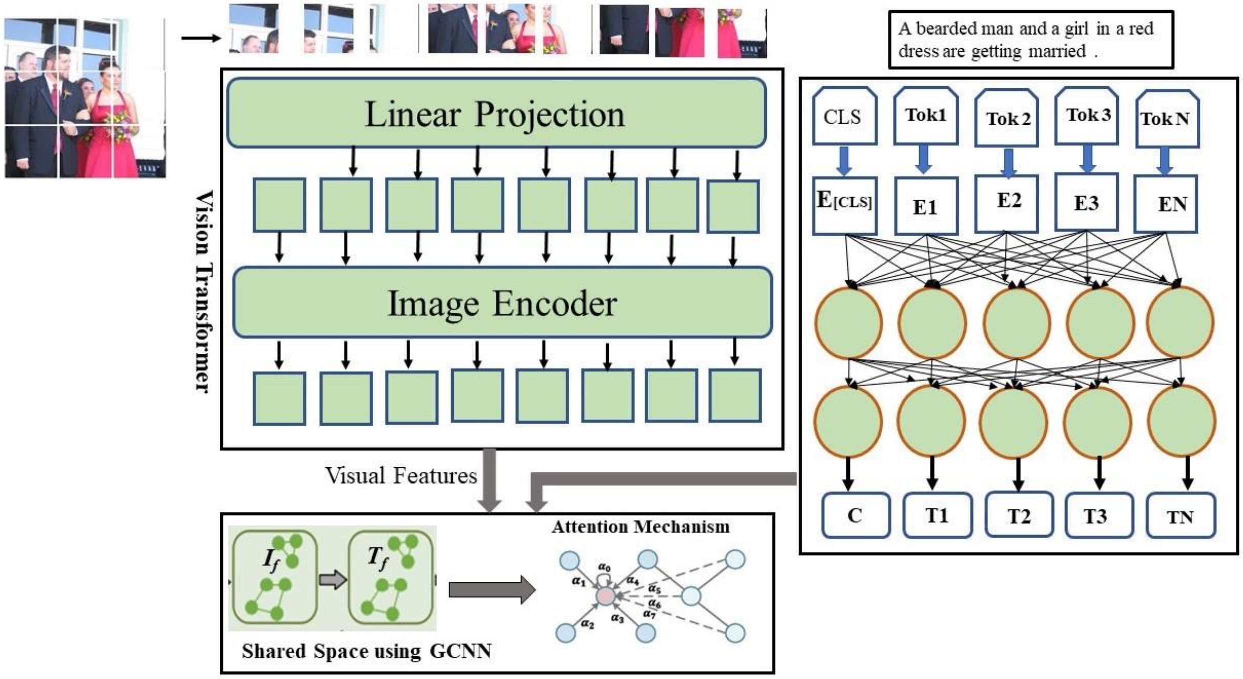


Figure 1: Overview pf proposed CMFG. ViT for Image features and BERT for Text features and proposed shared space embedding with attention mechanism for image and text features for cross-model retrieval.

Graph Analysis: GCNNs are effective to represent the data through graph, it has a wide range of applications [33]-[35]. Xi et al.,[36] worked on semi supervised model for hyperspectral image classification using cross-scale graph prototyping network. The authors designed a new self-branch attention mechanism to put more focus on critical features produced by multiple branches. Research on GCNNs has fascinated by many researchers due to their ability to effectively model non-Euclidean data structures, such as social networks, molecular graphs, and multimodal relationships, by capturing intricate dependencies and relational patterns among nodes through graph-based learning. Many recent works like, a low back-alignment spatial GCN for image classification [37], spatial-temporal GCN for emotion detection and classification [38], and quantum-based subgraph CNN to catch the global topological structure and local connectivity structure within a graph [39].

Recent advancements have leveraged GCNNs for cross-modal retrieval by modeling interactions between modalities in graph structures. In such frameworks, nodes represent features extracted from images and text, while edges capture semantic relationships. GCNNs propagate information across these nodes, enabling better alignment and representation in a shared embedding space. For example, approaches integrating GCNNs with attention mechanisms improve the ability to capture fine-grained interdependencies, which is critical for aligning textual and visual modalities.

The ability of GCNNs to capture structured relationships between features makes them an ideal choice for tasks requiring fine-grained alignment between text and visual data. By integrating GCNNs into the attention mechanism, our framework leverages their strengths to improve semantic alignment and retrieval performance in cross-modal tasks.

3. Framework Overview

The shared embedded space plays a vital role in cross-model image retrieval by supporting an integrated representation of features from different modalities. In this space, both image and text features are mapped into a common vector space where their semantic relationships can be directly compared. This allows an image to be retrieved using a textual query, image query or both, providing whole interaction through modalities.

To achieve this, we introduce a new approach for attention mechanism called cross-model feature graph (CMFG) as shown in Figure2. The implementation of CMFG as follows:

Generally, a graph is represented as a set of nodes and edges, $G = \{N, E\}$. The edge set E is defined as $E = \{(n_i, n_j) | n_i, n_j \in N\}$, connected a pair of nodes in the node set *N*. An adjacency matrix $A \in \{0,1\}^{NXN}$ can be used to define the connections in the graph, where $A\left(n_i, n_j\right) = 1$ indicates the presence of an edge between the nodes n_i, n_j.

CMFG*:* Let $G = \{I, T, E\}$ is a two-sided graph, where *I* and *T* are image and text features as nodes and E is the set of edges that represent relationships between nodes of *I* and *T*. The node $I = \{i_1, i_2, . \ldots . i_{K_I}\}$ be the set of K_I image feature vectors extracted using a Vision Transformer (ViT), where each $i_j \in \mathbb{R}^d$ and node $T = \{t_1, t_2 \ldots . t_{K_T}\}$ be the set of n_T text feature vectors extracted using BERT, where each $t_j \in \mathbb{R}^d$.

Edge weights E represent the relationship between nodes, $E = \{e_{uv} | u, v \in U, u \neq v\}$ where $u \ and \ v$ are the two different nodes, and an edge e_{uv} is computed using an attention mechanism. The attention score is given by

$$e_{uv} = softmax\left(\frac{Q_u . K_v^T}{\sqrt{d_k}}\right)$$

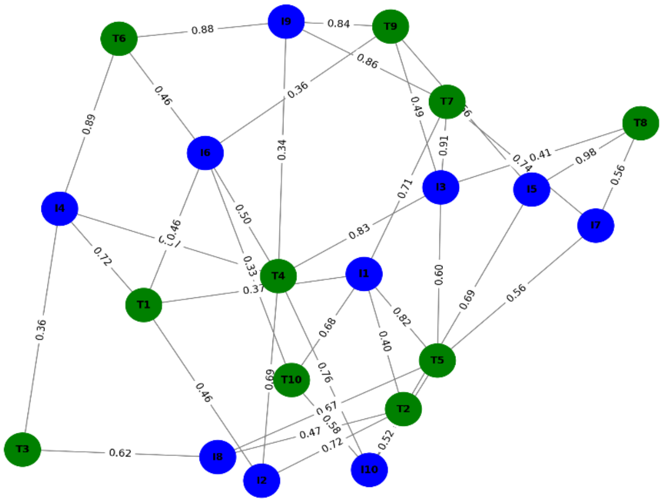


Figure 2: A sample graph constructed with only ten attributes of ViT and BERT features.

Where, $Q_u = W_Q h_u$ is the query vector for node u , $K_v = W_K h_v$ is the key vector for node v, $W_Q, W_K \in \mathbb{R}^{dXd_k}$ are learnable weight matrices, h_u *and* h_v are the feature vectors for nodes u and v, respectively, and d_k is the dimension of the query and key vectors.

After constructing the CMFG, as shown in Figure 2, the features of each node are updated by aggregating information from its neighbors. Using Graph Convolutional Networks (GCNNs), the feature update for node u is:

$$h_u^{(k+1)} = \rho\left(\sum_{v \in \mathcal{M}(u)} e_{uv}\, W^k h_v^k\right)$$

Here, $\mathcal{M}(u)$ is set of neighbors of node u, W^k is the learnable weight matrix at layer k, and ρ is a non-linear activation function and it is ReLU here.

The final stage of representing the image and text features or nodes after K GCNN layers in a common shared embedding space is designed as,

$$H_I^K = \{h_i^K \mid i_i \in I\}, \qquad H_T^K = \{h_t^K \mid t_t \in T\}$$

The CMFG approach dynamically propagate information between nodes through weighted edges. By treating similarity scores as edge weights, the CMFG captures nuanced interactions between image and text features, enabling the model to align them effectively in the shared embedding space. The CMFG's ability to model cross-modal relationships in a structured and interpretable way establishes it as a powerful tool for efficient and accurate cross-modal retrieval.

IV. Experiments

In this section, the experimental setup, datasets description, evaluation metrics, and implementation details are discussed. Complete results and analyses of the experiments are also explained.

A. Datasets

Two popular cross-model (image-text) databases, Theojiang (300K) [40], MIRFLICKR-25K [41] are often used in this proposed method experiment. Theojiang contains 380,530 samples and MIRFLICKR-25K contains 25,015 samples of image- text pair.

Theojiang(300K): The "Image-Text Dataset Subset (300k Captions Only with Latents)" curated by Theojiang is a comprehensive resource designed for cross-modal research, particularly in image-text retrieval (cross-model retrieval) and generation tasks. This dataset contains a total of 380,530 images and a textual description detailing the content of the associated image. The dataset is available on Hugging Face Datasets. This dataset is a valuable asset for researchers aiming to explore the interplay between visual and textual modalities, offering both raw data and their latent embeddings to support a wide range of experimental designs.

MIRFLICKR-25K: This dataset is good for evaluating cross-modal retrieval models, offering a real-world and diverse collection of images and text tags. It serves as a strong benchmark for testing multi-modal deep learning architectures in image-text retrieval tasks. It consists of 25,000 images collected from Flickr, each accompanied by metadata, including textual tags and annotations. The dataset contains real-world images sourced from Flickr, covering various categories like landscapes, objects, and people and each image is labeled with user-generated tags from Flickr. These tags describe image content and serve as textual representations for cross-modal learning. Some images have additional manually curated annotations for better semantic understanding. The dataset includes 15,000 training images and 10,000 test images.

The evaluation process is conducted using Google Colab Pro, which provides access to high-performance GPUs. The experiments are run on a Colab Pro instance equipped with an NVIDIA Tesla T4 or A100 GPU, Ubuntu-based environment, Intel Xeon processor, high-speed cloud storage, and 100GB of RAM.

B. Performance Analysis:

A comparative analysis of numerous breakthrough methods was conducted to validate the dominance of Attention guided multimodal graph network (PM) over existing methods. These approaches include CCA [42], VSE++ [43], PiTL [25], UNITER [26], ALBEF [44], and SimCLR [27]. Out of these, CCA used handcrafted features, VSE++ and PiTL used CNN and RNN models for cross model features extraction, UNITER is a transformer-based model, ALBEF is a fusion based model, and SimCLR is a self-supervised method for cross-model retrieval. Retrieval performance metrics, Cross-model alignment metrics and Graph-based attention performance are 3 fundamental retrieval protocols used to evaluate the performance of the PM.

C. Retrieval Performance metrics:

The mean average precision (mAP) is a typical performance indicator, which measures retrieval accuracy across queries. For example, for a query set Q, the mAP can be calculated as follows:

$$mAP(Q) = \frac{1}{N}\sum_{j=1}^{N} P(j)$$

Where N is the number of samples in the query set Q and *P(j)* is the average precision, which defined for a query is as follows:

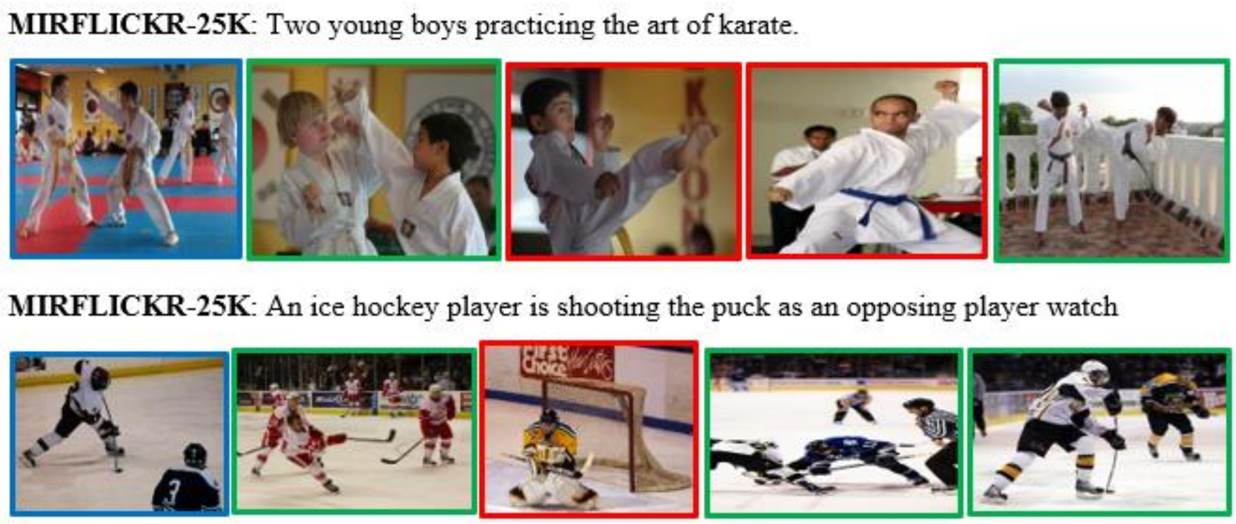
$$P(q) = \frac{1}{|N_{Pos}|}\sum_{i=1}^{M} P_q(i)\alpha_q(i)$$

$P_q(k)$ measures the precision of the top-k samples. Here, $\alpha_q(i)$=1 represents the true neighbor of $q(i)$, and $\alpha_q(i) = 0$ represents it is not. The retrieval set consists of K instances. The value of $|N_{Pos}|$ indicates the number of true neighbors in the retrieval set.

The experimental results of 2 retrieval tasks, Image → Text and Text →Image includes retrieving text from images and images from text has been presented in table 1 for two benchmark datasets.

Table1: Comparative results for cross-model retrieval on Theojiang (300K), MIRFLICKR-25K over proposed method for top retrieval R@5 & R@10.

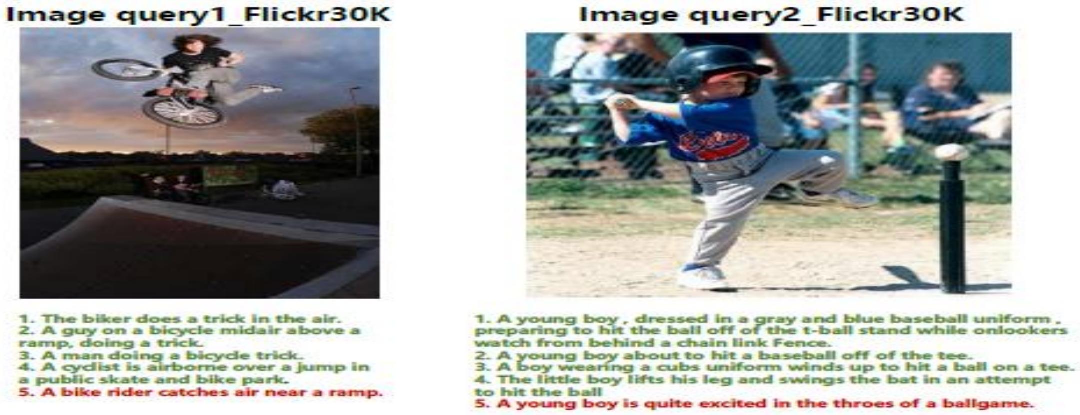
Task	**Model**	**Theojiang (300K)**		**MIRFLICKR-25K**	
		R@5	**R@10**	**R@5**	**R@10**
Image-to-Text	CCA	75.4	71.5	81.5	77.6
	VSE++	76.8	75.2	82.9	79.4
	PiTL	79.1	76.8	83.7	80.8
	UNITER	81.2	78.3	85.6	81.7
	ALBEF	82.5	79.9	87.2	82.5
	SimCLR	85.3	81.5	88.9	84.1
	PM	**88.4**	**85.8**	**90.2**	**86.4**
Text-to-Image	CCA	66.3	61.4	73.2	71.5
	VSE++	69.7	65.5	75.3	72.1
	PiTL	71.5	69.6	78.5	74.3
	UNITER	75.4	72.9	79.2	75.6
	ALBEF	78.2	76.4	80.1	77.1
	SimCLR	79.5	76.9	81.3	78.5
	PM	**80.6**	**77.8**	**82.4**	**79.6**



MIRFLICKR-25K: Two young boys practicing the art of karate.

MIRFLICKR-25K: An ice hockey player is shooting the puck as an opposing player watch

Figure 3: Retrieved results for a text query and retrieved images from MIRFLICKR-25K.



1. The biker does a trick in the air.
2. A guy on a bicycle midair above a ramp, doing a trick.
3. A man doing a bicyde trick.
4. A cydist is airborne over a jump in a public skate and bike park.
5. A bike rider catches air near a ramp.

1. A young boy , dressed in a gray and blue baseball uniform , preparing to hit the ball off of the t-ball stand while onlookers watch from behind a chain link Fence.
2. A young boy about to hit a baseball off of the tee.
3. A boy wearing a cubs uniform winds up to hit a ball on a tee.
4. The little boy lifts his leg and swings the bat in an attempt to hit the ball
5. A young boy is quite excited in the throes of a ballgame.

Figure 4: Retrieved results for an image query and retrieved text sequences.

Image-to-Text on MIRFLICKR-25K

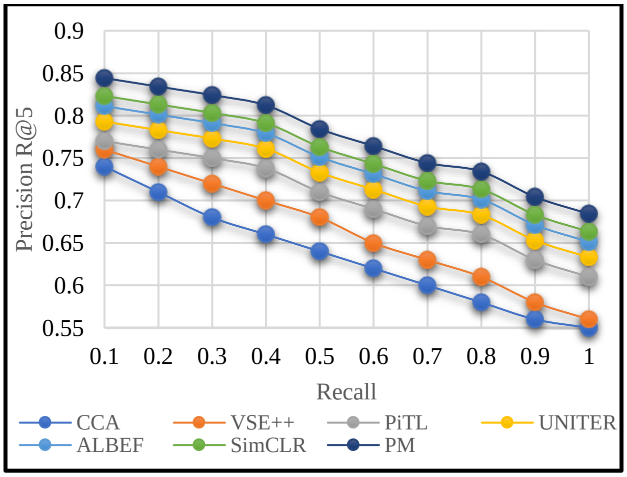
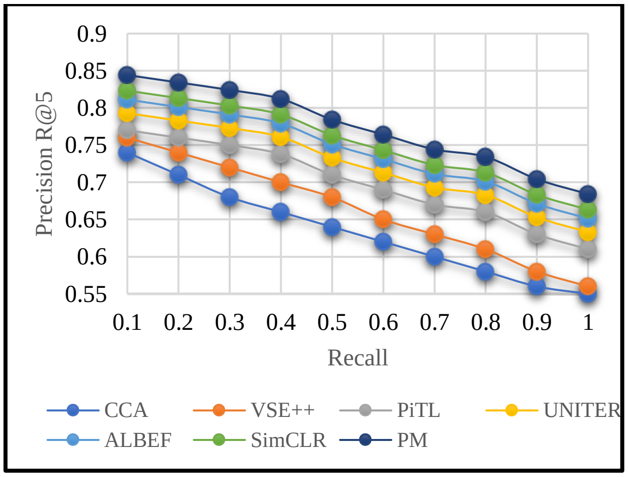


Image-to-Text on Theojiang (300K)



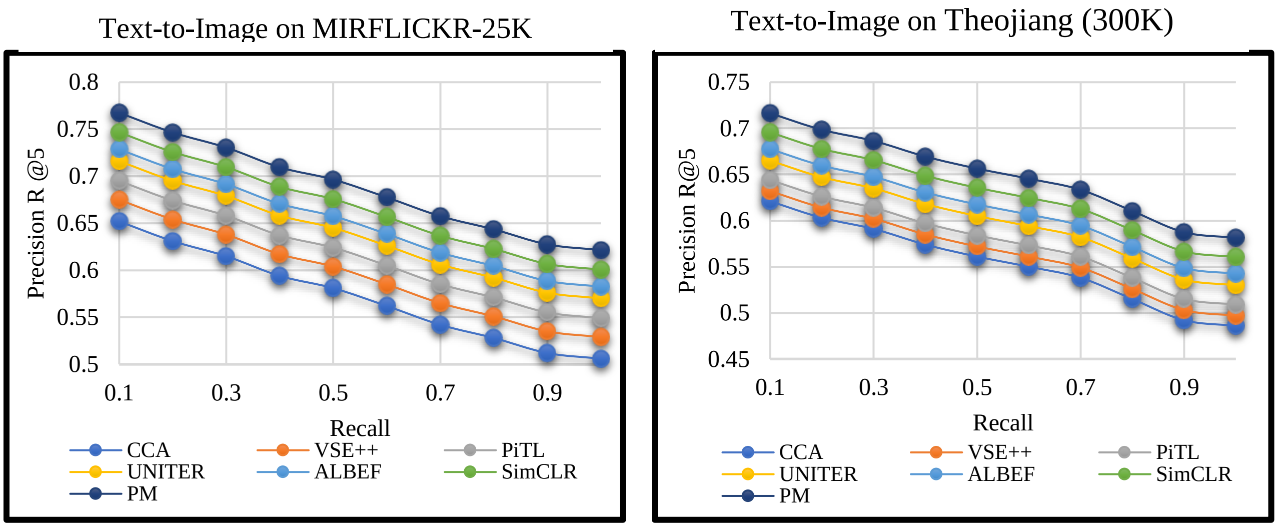


Figure 5: P-R graph for 2 datasets for Image-to-Text retrieval and Text-to-Image retrieval for the top 5 outcomes.

D. Cumulative Matching Characteristic (CMC) Curve

The CMC curve measures the probability that the correct match appears within the top-K retrieved results. It is commonly used in cross-modal retrieval (e.g., image-to-text) and re-identification tasks as shown in Figure 6.

For a given q and a set of retrieved outcomes R, the CMC value at rank k is defined as:

$$CMC(k) = \frac{1}{N}\sum_{i=1}^{N} 1(rank_i \leq k)$$

Where, N is the total number of queries, $1(rank_i \leq k)$ is an indicator function that returns 1 if the correct match is found within the top-k retrieved results, otherwise returns 0.

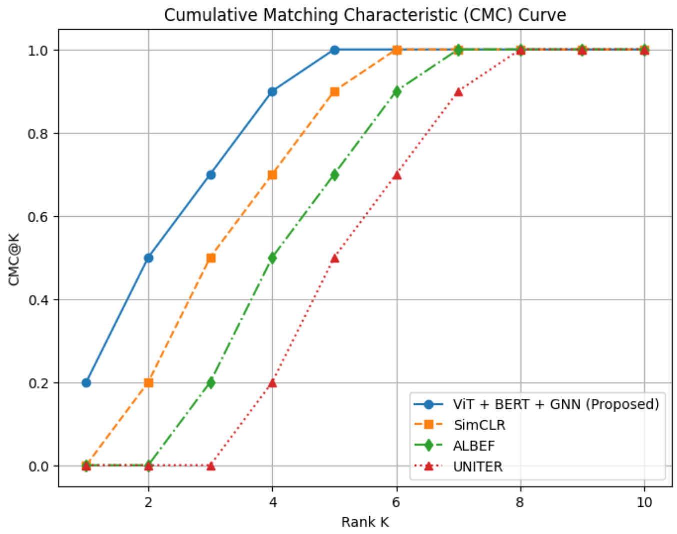


Figure 6: CMC curve for top-10 ranks for Image-to-Text retrieval outcomes.

5. Conclusion

The proposed attention mechanism using GNN on ViT and BERT features, enhances cross-modal image-text retrieval. This approach significantly improves the alignment between image and text modalities, leading to more accurate and meaningful retrieval results. The integration of transformer-based architectures ensures effective feature extraction, while GNN strengthens inter-modal relationships, allowing for better retrieval precision, especially at lower ranks (CMC@K, mAP,). The model consistently outperforms existing methods such as UNITER, ALBEF and SimCLR in terms of retrieval accuracy and robustness. Moreover, the approach is scalable and performs efficiently on large-scale datasets like MIRFLICKR-25K and Theojiang dataset. Experimental results demonstrate that the model achieves higher recall and precision, improving both text-to-image and image-to-text retrieval. Future work can explore multi-scale feature aggregation, contrastive learning strategies, and lightweight architectures for real-time applications. Overall, the proposed method is a robust, scalable, and high-performance solution for cross-modal retrieval tasks.

References

1. Arora, Nitin, G. Sucharitha, and Subhash C. Sharma. "MVM-LBP: Mean− Variance− Median based LBP for face recognition." *International Journal of Information Technology* 15.3 (2023): 1231-1242.
2. Qian, Shengsheng, Tianzhu Zhang, and Changsheng Xu. "Multi-modal multi-view topic-opinion mining for social event analysis." *Proceedings of the 24th ACM international conference on Multimedia*. 2016.
3. Peng, Yuxin, Xin Huang, and Yunzhen Zhao. "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges." *IEEE Transactions on circuits and systems for video technology* 28.9 (2017): 2372-2385.
4. Qian, Shengsheng, et al. "Dual adversarial graph neural networks for multi-label cross-modal retrieval." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. No. 3. 2021.
5. Zhan, Yu-Wei, et al. "Supervised hierarchical deep hashing for cross-modal retrieval." *Proceedings of the 28th ACM International Conference on Multimedia*. 2020.
6. Liao, Lei, Meng Yang, and Bob Zhang. "Deep supervised dual cycle adversarial network for cross-modal retrieval." *IEEE Transactions on Circuits and Systems for Video Technology* 33.2 (2022): 920-934.
7. Chen, Zhao-Min, et al. "Disentangling, embedding and ranking label cues for multi-label image recognition." *IEEE Transactions on Multimedia* 23 (2020): 1827-1840.
8. Qian, Shengsheng, et al. "Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2022): 4794-4811.
9. Zhu, Jie, et al. "Adaptive multi-label structure preserving network for cross-modal retrieval." *Information Sciences* 682 (2024): 121279.
10. Bai, Cong, et al. "Graph convolutional network discrete hashing for cross-modal retrieval." *IEEE Transactions on Neural Networks and Learning Systems* 35.4 (2022): 4756-4767.
11. Wang, Tianshi, et al. "Cross-modal retrieval: a systematic review of methods and future directions." *arXiv preprint arXiv:2308.14263* (2023).

12. Zhou, Kun, Fadratul Hafinaz Hassan, and Gan Keng Hoon. "The State of the Art for Cross-Modal Retrieval: A Survey." *IEEE Access* (2023).
13. Zhou, Kun, Fadratul Hafinaz Hassan, and Keng Hoon Gan. "Pretrained models for cross-modal retrieval: experiments and improvements." *Signal, Image and Video Processing* 18.5 (2024): 4915-4923.
14. Zhang, Jinming, et al. "DSMCA: Deep Supervised Model with the Channel Attention Module for Cross-modal Retrieval." *Proceedings of the 2022 5th International Conference on Data Storage and Data Engineering*. 2022.
15. Wang, Li, et al. "Joint feature selection and graph regularization for modality-dependent cross-modal retrieval." *Journal of Visual Communication and Image Representation* 54 (2018): 213-222.
16. Lin, Zijia, et al. "Cross-view retrieval via probability-based semantics-preserving hashing." *IEEE transactions on cybernetics* 47.12 (2016): 4342-4355.
17. Wang, Di, et al. "Label consistent matrix factorization hashing for large-scale cross-modal similarity search." *IEEE transactions on pattern analysis and machine intelligence* 41.10 (2018): 2466-2479.
18. Huang, Xin, Yuxin Peng, and Mingkuan Yuan. "MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval." *IEEE transactions on cybernetics* 50.3 (2018): 1047-1059.
19. Zhang, Dongqing, and Wu-Jun Li. "Large-scale supervised multimodal hashing with semantic correlation maximization." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 28. No. 1. 2014.
20. Guan, Ziyu, et al. "Cross-modal Guided Visual Representation Learning for Social Image Retrieval." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
21. Zheng, Chaoqun, et al. "Adaptive partial multi-view hashing for efficient social image retrieval." *IEEE Transactions on Multimedia* 23 (2020): 4079-4092.
22. Zeng, Zhixiong, and Wenji Mao. "A comprehensive empirical study of vision-language pre-trained model for supervised cross-modal retrieval." *arXiv preprint arXiv:2201.02772* (2022).
23. Ji, Zhong, Kexin Chen, and Haoran Wang. "Step-wise hierarchical alignment network for image-text matching." *arXiv preprint arXiv:2106.06509* (2021).
24. Mei, Tao, et al. "Multimedia search reranking: A literature survey." *ACM Computing Surveys (CSUR)* 46.3 (2014): 1-38.
25. Guo, Zixin, et al. "PiTL: Cross-modal Retrieval with Weakly-supervised Vision-language Pre-training via Prompting." *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2023.
26. Chen, Yen-Chun, et al. "Uniter: Universal image-text representation learning." *European conference on computer vision*. Cham: Springer International Publishing, 2020.
27. Kim, Byoungjip, et al. "Transferring pre-trained multimodal representations with cross-modal similarity matching." *Advances in Neural Information Processing Systems* 35 (2022): 30826-30839.
28. Zhang, Jingyi, et al. "Vision-language models for vision tasks: A survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).
29. Gkelios, Socratis, Yiannis Boutalis, and Savvas A. Chatzichristofis. "Investigating the vision transformer model for image retrieval tasks." 2021 17th International Conference on Distributed Computing in Sensor Systems (DCOSS). IEEE, 2021.
30. Zheng, Shaomin, and Meng Yang. "A new method of improving bert for text classification." *Intelligence Science and Big Data Engineering. Big Data and Machine Learning:*

9th International Conference, IScIDE 2019, Nanjing, China, October 17–20, 2019, Proceedings, Part II 9. Springer International Publishing, 2019.

31. Dosovitskiy, Alexey. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
32. Xu, Zelai, Tan Yu, and Ping Li. "Texture BERT for cross-modal texture image retrieval." *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2022.
33. Elshamli, Ahmed, Graham W. Taylor, and Shawki Areibi. "Multisource domain adaptation for remote sensing using deep neural networks." *IEEE Transactions on Geoscience and Remote Sensing* 58.5 (2019): 3328-3340.
34. Xie, Guo-Sen, et al. "Scale-aware graph neural network for few-shot semantic segmentation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
35. Cai, Deng, and Wai Lam. "Graph transformer for graph-to-sequence learning." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. No. 05. 2020.
36. Xi, Bobo, et al. "Semisupervised cross-scale graph prototypical network for hyperspectral image classification." *IEEE Transactions on Neural Networks and Learning Systems* 34.11 (2022): 9337-9351.
37. Bai, Lu, et al. "Learning backtrackless aligned-spatial graph convolutional networks for graph classification." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.2 (2020): 783-798.
38. Bhattacharya, Uttaran, et al. "Step: Spatial temporal graph convolutional networks for emotion perception from gaits." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 02. 2020.
39. Zhang, Zhihong, et al. "Quantum-based subgraph convolutional neural networks." *Pattern Recognition* 88 (2019): 38-49.
40. https://datasets-server.huggingface.co/rows?dataset=theojiang%2Fimage-text-dataset-subset-300k-captions_text&config=default&split=train&offset=0&length=100.
41. Huiskes, Mark J., and Michael S. Lew. "The mir flickr retrieval evaluation." Proceedings of the 1st ACM international conference on Multimedia information retrieval. 2008.
42. Shao, Jie, et al. "Towards improving canonical correlation analysis for cross-modal retrieval." *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*. 2017.
43. Faghri, Fartash, et al. "Vse++: Improving visual-semantic embeddings with hard negatives." *arXiv preprint arXiv:1707.05612* (2017).
44. Xie, Chen-Wei, et al. "Token embeddings alignment for cross-modal retrieval." *Proceedings of the 30th ACM International Conference on Multimedia*. 2022.