Dynamic Modality Interaction Modeling for Image-Text Retrieval

Leigang Qu[1], Meng Liu[2*], Jianlong Wu[1], Zan Gao[3], Liqiang Nie[1*]
[1]Shandong University, Shandong, China, [2]Shandong Jianzhu University, Shandong, China,
[3]Shandong Artificial Intelligence Institute, Shandong, China
{leigangqu,mengliu.sdu,zangaonsh4522,nieliqiang}@gmail.com,jlwu1992@sdu.edu.cn

ABSTRACT

Image-text retrieval is a fundamental and crucial branch in information retrieval. Although much progress has been made in bridging vision and language, it remains challenging because of the difficult *intra-modal reasoning* and *cross-modal alignment*. Existing modality interaction methods have achieved impressive results on public datasets. However, they heavily rely on expert experience and empirical feedback towards the design of interaction patterns, therefore, lacking flexibility. To address these issues, we develop a novel modality interaction modeling network based upon the routing mechanism, which is the first unified and dynamic multimodal interaction framework towards image-text retrieval. In particular, we first design four types of cells as basic units to explore different levels of modality interactions, and then connect them in a dense strategy to construct a routing space. To endow the model with the capability of path decision, we integrate a dynamic router in each cell for pattern exploration. As the routers are conditioned on inputs, our model can dynamically learn different activated paths for different data. Extensive experiments on two benchmark datasets, *i.e.*, Flickr30K and MS-COCO, verify the superiority of our model compared with several state-of-the-art baselines.

CCS CONCEPTS

• **Information systems** → **Novelty in information retrieval**; **Multimedia and multimodal retrieval**.

KEYWORDS

Image-Text Matching; Cross-Modal Retrieval; Dynamic Routing

ACM Reference Format:
Leigang Qu, Meng Liu, Jianlong Wu, Zan Gao, Liqiang Nie. 2021. Dynamic Modality Interaction Modeling for Image-Text Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3404835.3462829

* Meng Liu (mengliu.sdu@gmail.com) and Liqiang Nie (nieliqiang@gmail.com) are the corresponding authors.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462829

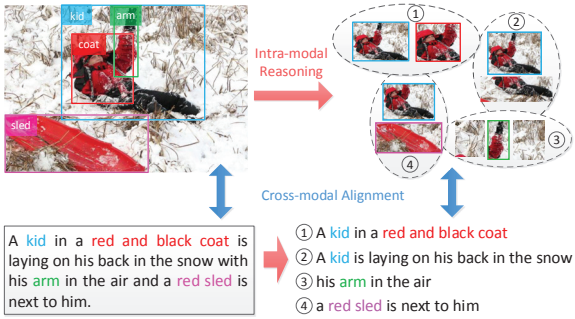


Figure 1: Illustration of two main challenges in image-text retrieval, *i.e.*, intra-modal reasoning and cross-modal alignment. The aligned visual regions and textual entities are highlighted in the same color, and the matched relations are marked with the same number.

1 INTRODUCTION

Visual media and natural language are the two most prevalent modalities exhibiting information in our daily life. It is essential for computers to understand, match, and transform such cross-modal data. Image-text retrieval, a fundamental and crucial problem in information retrieval, has attracted extensive attention in recent years [2, 8, 26]. It benefits a variety of applications, ranging from cross-modal retrieval [11, 12, 23, 41] to multimedia recommendation [36]. However, it is still a challenging task due to the requirement of the accurate reasoning of intra-modal relations and the precise alignment of cross-modal information. Specifically, the former requires recognizing and comprehending various relations within the visual or textual modality, such as the intermediate relation "a red sled is next to him" in the textual modality and the visual relation outlined in the dotted circle, as illustrated in Figure 1. The latter links items from different modalities to match with each other at different semantic levels. For instance, as shown in Figure 1, the visual region and the textual concept annotated with the same color, and the high-level relations marked with the same number should be well aligned.

Great efforts have been dedicated to tackling the above challenges via modeling various modality interactions over the past few years. According to used interaction patterns, they can be roughly divided into three categories: 1) **Intra-modal Interaction**. Towards the challenge of intra-modal reasoning, this pattern merely carries out interaction modeling independently for different modalities to explore relations among entities within the specific modality, as illustrated in Figure 2(a). Particularly, existing related work commonly utilizes graph convolutional network [16] or self-attention mechanism [26, 37] as intra-model interaction modules to derive comprehensive single-modal representations. 2) **Cross-modal Interaction**. Studies in this category focus on aligning cross-modal

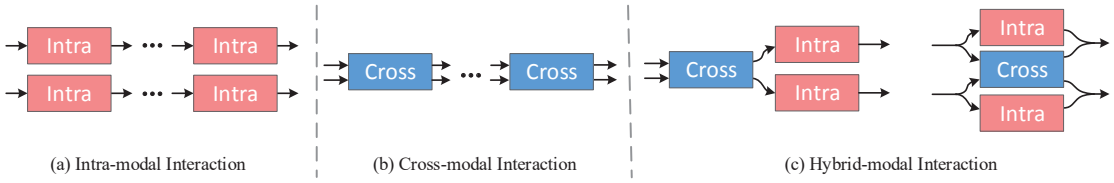


Figure 2: Illustration of existing modality interaction patterns. The red and the blue boxes represent the intra-modal interaction and the cross-modal interaction module, respectively.

entities, for example, aligning the visual region related to the "kid" and the word "kid" in Figure 1. This is accomplished by different cross-modal interaction operations, as shown in Figure 2(b). For example, the stacked cross-attention network [15] and an adaptive message passing method [33] are designed to capture cross-modal pairwise interactions. Moreover, to further delve into high-order correspondences, Chen *et al.* [3] proposed an iterative framework by stacking multiple cross-modal interaction modules in depth. And 3) **Hybrid-modal Interaction**. To further tackle the two aforementioned challenges, recent methods combining intra- and inter-modal interactions are developed, such as the serial pattern [22] and the parallel pattern [35, 40] displayed in Figure 2(c). Despite the significance and value of the methods in the above three categories, they still suffer from two critical shortcomings: *1) Their modality interaction patterns are hand-crafted, depending heavily on expert knowledge and empirical feedback, which may make some optimal interaction patterns still untapped. And 2) existing models are static, namely, all samples go through the same fixed computation flow. This may induce that even simple image-text pairs would be processed by some very complex interaction patterns.*

To tackle these downsides, we present a novel **D**ynam**I**c **M**odality int**E**raction modeling network (**DIME**), which is the first unified image-text retrieval framework with dynamic modality interaction pattern learning. As shown in Figure 3, we first tailor four types of cells to accomplish different interaction operations. Concretely, the rectified identity cell offers identical and non-linear transformation abilities, and the intra-modal reasoning cell is designed to capture context information and intra-modal relations. To enhance the visual-textual alignment, the global-local guidance cell and cross-modal refinement cell are designed with different granularities. Afterwards, we stack these cells in width and depth to construct a complete path space, such that a variety of unexplored interaction patterns can be considered. Meanwhile, we configure a dynamic router for each cell to generate data-dependent paths. Moreover, to drive similar images/texts to learn similar paths, we introduce a semantic consistency regularization. Extensive experimental results on two benchmark datasets, *i.e.*, Flickr30K [39] and MS-COCO [19], validate the effectiveness and superiority of our proposed method[1].

The main contributions of this work are three-fold:

- We present a dynamic modality interaction modeling framework towards image-text retrieval, which could cover existing interaction patterns and automatically learn other unexplored ones. To the best of our knowledge, it is the first work to dynamically explore different modality interaction patterns for varied data.
- We design four basic cells to model modality interactions with different granularities, settling both the intra-modal reasoning and the inter-modal alignment matters.
- To dynamically learn interaction patterns, we integrate a soft router in each cell. Furthermore, to constrain these dynamic routers for powerful path decision, we introduce a semantic consistency regularization term.

[1]Our codes and settings have been released at *https://sigir21.wixsite.com/dime*, to facilitate the future research.

2 RELATED WORK

2.1 Image-Text Retrieval

According to the granularity of semantic alignment, we roughly divide existing studies into two groups: global embedding based and local inference based methods. The former accomplishes semantic matching via mapping holistic images and sentences into a common modality agnostic embedding space, in which the visual-textual similarity is calculated [6, 7, 24, 42]. To be specific, DeViSE [7], the pioneering work of this line, adopts CNN and Skip-Gram [24] to map images and texts into a joint space for semantic alignment. Furthermore, to take full advantage of informative pairs, Faghri *et al.* [6] integrated the hard negative mining technology into the ranking loss, contributing to a significant improvement. Recently, Zheng *et al.* [42] proposed to discriminatively embed images and texts into a shared semantic space with an instance loss. Although these methods have achieved promising performance, they fail to explore fine-grained relations among visual regions and words.

The latter branch targets at cross-modal semantic alignments via exploiting fine-grained modality interactions [3, 4, 14–16, 22, 26, 34, 35, 40]. Particularly, Karpathy *et al.* [14] first detected visual regions with R-CNN, and then aggregated similarities between fragments for image-text matching. Inspired by the success of bottom-up attention [1], Lee *et al.* [15] proposed a stacked cross attention model for similarity prediction by considering the dense pairwise cross-modal interaction. To enhance the comprehensive understanding towards images and texts, some intra-modal interaction based methods have been proposed. For instance, Li *et al.* [16] performed local and global reasoning by building up connections between image regions, and Chen et al. [4] processed images and sentences symmetrically by reordering image objects according to the corresponding description. Qu *et al.* [26] designed a gating self-attention mechanism for context modeling and a multi-view summarization module for asymmetry matching. Thereafter, plentiful intra- and inter-modal interaction patterns have been explored and applied to the image-text retrieval task. To be more specific, Chen *et al.* [3] introduced an iterative matching model using cross-modal interaction with multiple steps. In contrast, Wei *et al.* [35] and Zhang et al. [40] exploited intra-modal and inter-modal correlations by designing

a parallel modality interaction framework. Liu *et al.* [22] serially performed inter- and intra-modal interaction for node-level and structure-level matching in a graph structured model.

Although significant progress has been made by considering different fine-grained interaction patterns, they are essentially handcrafted, heavily depending on expert experience. In addition, they overlook the difference in comprehension between data samples. To this end, designing a flexible and adaptable framework, which could provide a broad space of interaction patterns and automatically learn execution paths, becomes a critical challenge.

2.2 Dynamic Network

Different from neural architecture search [20] with static inference, dynamic networks [17, 18, 28, 38] generate execution paths on the fly conditioned on input samples. In particular, early dynamic methods aim at network compression by channel pruning [18] or layer skipping [31, 38]. For example, Wu *et al.* [38] designed a policy network to determine whether skip or execute convolutional blocks under a reinforcement learning setting. Recently, based on the dynamic mechanism, some researchers devote to tackling the intrinsic challenge of scale variation in the computer vision community. To be specific, Li *et al.* [17] proposed a dynamic routing network with soft conditional gate to search data-dependent scale transformation paths for semantic segmentation. Song *et al.* [28] designed a dynamic head with efficient fine-grained representation capability for object detection.

To the best of our knowledge, the dynamic mechanism has never been studied in the field of image-text retrieval. Different with current related work, our model is the first one to introduce the dynamic mechanism to learn modality interaction patterns for cross-modal semantic alignment.

3 METHODOLOGY

In this section, we elaborate each component of our model, as illustrated in Figure 3. Concretely, we first introduce the feature extraction process in Section 3.1 and four types of interaction cells in Section 3.2. Afterwards, we connect these cells to construct the routing space and present soft routers to perform routing process in Section 3.3. We ultimately detail the objective function utilized to optimize the network for image-text retrieval in Section 3.4.

3.1 Feature Representation

3.1.1 Visual Representation. Given an image I, we first extract region features with the bottom-up attention model[2] [27], and then select top-R ROIs according to the class confidence scores. The average pooling is applied to the feature maps of these regions to obtain their feature vectors, represented as $\mathbf{F} = [\mathbf{f}_1; \ldots; \mathbf{f}_R] \in \mathbb{R}^{R \times D_v}$, where D_v is the dimension of the extracted region features. Afterwards, we transform these feature vectors into a D-dimensional space via a fully-connected (FC) linear projection. The output visual region representation is denoted as $\mathbf{V} = [\mathbf{v}_1; \ldots; \mathbf{v}_R] \in \mathbb{R}^{R \times D}$. Meanwhile, we acquire the global representation $\bar{\mathbf{v}} \in \mathbb{R}^D$ of the given image I by adopting the average-pooling.

3.1.2 Textual Representation. For a given sentence T, we first utilize pre-trained BERT [5] as the textual encoder to extract word embeddings $\mathbf{E} = [\mathbf{e}_1; \ldots; \mathbf{e}_K] \in \mathbb{R}^{K \times D_t}$, where K denotes the number of words and D_t represents the dimension of word embeddings. We then adopt a bag of parallel 1-D convolution kernels with different sizes to capture phrase-level semantics. Afterwards, we concatenate the feature maps of these kernels, and then pass the result into a FC layer to obtain D-dimensional word features, denoted as $\mathbf{W} = [\mathbf{w}_1; \ldots; \mathbf{w}_K] \in \mathbb{R}^{K \times D}$. In this work, we adopt the max-pooling to obtain the global sentence feature $\bar{\mathbf{w}} \in \mathbb{R}^D$.

3.2 Modality Interaction Cells

To address the intra-modal reasoning and cross-modal alignment challenges, we tailor four[3] types of cells including the Intra-Modal Reasoning Cell (IMRC) for intra-modal reasoning challenge, the Global-Local Guidance Cell (GLGC) and the Cross-Modal Refinement Cell (CMRC) for cross-modal alignment challenge, and the Rectified Identity Cell (RIC) for discriminative clue retention, as shown in Figure 3. These cells are able to characterize modality interactions of different levels, endowing our model with excellent semantic representation and reasoning capability. Operations in these cells can be formally summarized as,

$$\mathbf{O}_i^{(l)} = \begin{cases} \mathcal{F}_i^{(l)}(\mathbf{X}_i^{(l)}), & i = 0 \text{ or } 1, \\ \mathcal{F}_i^{(l)}(\mathbf{X}_i^{(l)}, \bar{\mathbf{y}}), & i = 2, \\ \mathcal{F}_i^{(l)}(\mathbf{X}_i^{(l)}, \mathbf{Y}), & i = 3, \end{cases} \tag{1}$$

where $\mathcal{F}_i^{(l)}$ represents the interaction function of the i-th cell in the l-th layer, $\mathbf{X}_i^{(l)}$ denotes the input of the i-th cell in the l-th layer (introduced in Section 3.3.2) and $\mathbf{O}_i^{(l)} \in \mathbb{R}^{M \times D}$ denotes the corresponding output feature matrix. Each row of $\mathbf{O}_i^{(l)}$ represents the D-dimensional feature vector of a fragment. Due to the bidirectional nature of cross-modal retrieval, we separately utilize $\mathbf{X} \in \mathbb{R}^{M \times D}$ and $\bar{\mathbf{x}} \in \mathbb{R}^D$ to denote the local and global features of the query, respectively. Likewise, the local and global features of the gallery are represented as $\mathbf{Y} \in \mathbb{R}^{N \times D}$ and $\bar{\mathbf{y}} \in \mathbb{R}^D$, respectively.

In this work, we implement two single symmetrical versions of model. We hence set $\mathbf{X} := \mathbf{V}$ ($M := R$) and $\mathbf{Y} := \mathbf{W}$ ($N := K$) for the image-text (i-t) version, and $\mathbf{X} := \mathbf{W}$ ($M := K$) and $\mathbf{Y} := \mathbf{V}$ ($N := R$) for the text-image (t-i) version. Note that the same type cells in different layers (*e.g.*, $\mathcal{F}_i^{(l-1)}$ and $\mathcal{F}_i^{(l)}$) perform the same operation without sharing parameters. In what follows, we will omit the superscript layer index (l) for simplicity.

3.2.1 Rectified Identity Cell. Human can make sense of a simple image (or a short sentence) at a glance. We hence argue that complicated interaction operations may not always be essential, especially for simple images or sentences. Motivated by this, we intend to design a simple interaction cell that could skip unnecessary operations and retain discriminating clues. Besides, to alleviate the gradient vanishing problem, we present the rectified identity cell, formulated as $\mathcal{F}_0(\mathbf{X}) = \text{ReLU}(\mathbf{X})$.

[2] It is implemented by the Faster R-CNN [27] with ResNet-101 [10] as the backbone.

[3] In fact, we can also customize different numbers of cells according to diverse requirements of performance-efficiency trade-off.

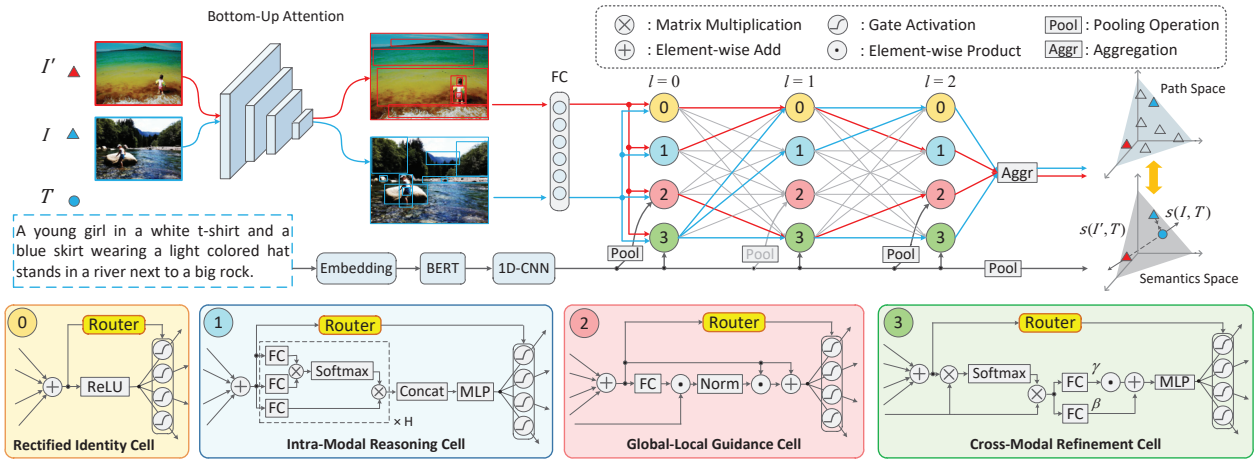


Figure 3: Schematic illustration of the proposed DIME framework (image-text version) with three layers. Four different cells are devised in blocks with different colors and stacked in a densely connected way to form a routing space. Given two different images denoted by triangles with different colors, their activated paths are shown with corresponding colors.

3.2.2 Intra-Modal Reasoning Cell. To capture semantic dependencies between local fragments (*i.e.*, words or visual regions), an intra-modal reasoning cell is designed. Concretely, we adopt the multi-head self-attention mechanism [30] to capture the intra-modal dependencies from different subspaces as,

$$MultiHead(\mathbf{X}) = Concat(\mathbf{h}_1, ..., \mathbf{h}_H) + \mathbf{X}, \tag{2}$$

where $Concat(\cdot)$ represents the concatenation operation across the feature dimension, H denotes the number of heads, and $\mathbf{h}_i = Att(\mathbf{X}\mathbf{W}_i^Q, \mathbf{X}\mathbf{W}_i^K, \mathbf{X}\mathbf{W}_i^V)$. In this work, Att refers to the scaled-dot product attention formulated as follows,

$$Att(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V}, \tag{3}$$

where Softmax is operated on each row and d_k is the channel number of $\mathbf{Q}$ and $\mathbf{K}$. Thereafter, a fully connected feed-forward network is executed to combine attention results from different heads.

Based on the above processes, we summarize our intra-modal reasoning cell as,

$$\mathcal{F}_1(\mathbf{X}) = FFN(MultiHead(\mathbf{X})), \tag{4}$$

where FFN denotes the feed forward network implemented by a two-layer multi-layer perceptron (MLP) with the ReLU activation function in between.

3.2.3 Global-Local Guidance Cell. Although local representations can encode abundant clues, global features condense contextual information and high-level semantics. Inspired by this, we adopt the global information of one modality as guidance to regulate the local fragment of another modality, which is formulated as,

$$\begin{cases} \mathbf{d}_r = FC(\mathbf{x}_r) \odot \bar{\mathbf{y}}, \\ \mathbf{x}'_r = (1 + Norm(\mathbf{d}_r)) \odot \mathbf{x}_r, \end{cases} \tag{5}$$

where $\mathbf{d}_r$ represents the guidance direction for the r-th local fragment, and $Norm(\cdot)$ denotes L2-normalization across fragment dimension. The above global-local guidance process is hence summarized as $\mathcal{F}_2(\mathbf{X}, \bar{\mathbf{y}}) = [\mathbf{x}'_1; ...; \mathbf{x}'_M]$.

3.2.4 Cross-Modal Refinement Cell. To further bridge the semantic gap and enrich representations, we refine fragment features by exploring local-local cross-modal interactions. Specially, we first calculate attention weights between fragments of divergent modalities as follows,

$$\alpha_{rk} = \frac{\exp(\lambda s_{rk})}{\sum_{k=1}^{N} \exp(\lambda s_{rk})}, \tag{6}$$

where λ is the inversed temperature factor and $s_{rk} = \cos(\mathbf{x}_r, \mathbf{y}_k)$. We can then obtain the context vector $\mathbf{c}_r = \sum_{k=1}^{N} \alpha_{rk}\mathbf{y}_k$.

Based on the cross-modal context information, we propose a conditional modulation strategy for refinement, in which the local features $\mathbf{x}_r$ can be enhanced semantically. To be more specific, the context vector $\mathbf{c}_r$ is first mapped to generate the scaling vector γ_r and the shifting vector β_r as follows,

$$\begin{cases} \gamma_r = \text{Tanh}(FC_\gamma(\mathbf{c}_r)), \\ \beta_r = FC_\beta(\mathbf{c}_r). \end{cases} \tag{7}$$

Afterwards, the refined local feature $\tilde{\mathbf{x}}_r$ is calculated by affine transformation followed by a MLP and shortcut connection, which is formulated as,

$$\tilde{\mathbf{x}}_r = MLP(\mathbf{x}_r \odot \gamma_r + \beta_r) + \mathbf{x}_r. \tag{8}$$

Combining the above steps, our cross-modal refinement cell is represented as $\mathcal{F}_3(\mathbf{X}, \mathbf{Y}) = [\tilde{\mathbf{x}}_1; ...; \tilde{\mathbf{x}}_M]$.

3.3 Soft Router

3.3.1 Routing Space. As shown in Figure 3, to give full play to the respective advantages of four cells, they are configured in parallel per layer. Besides, we connect them between the adjacent layers in a dense way. Based on this, each cell has the chance to receive all signals from the cells belonging to the last layer. More importantly, this dense connection scheme ensures the abundance and flexibility of the routing space, where many potential interaction patterns can be explored, including the ones shown in Figure 2.

3.3.2 Routing Process. After constructing the densely connected routing space, the routing process is conducted by our proposed soft router, which can be viewed as a procedure of path decision. Formally, the input of the i-th cell in the l-th layer is obtained by the following aggregation operation,

$$\mathbf{X}_i^{(l)} = \begin{cases} \mathbf{X}, & l = 0, \\ \sum_{j=0}^{C-1} g_{j,i}^{(l-1)} \mathbf{O}_j^{(l-1)}, & l > 0, \end{cases} \tag{9}$$

where $\mathbf{X} \in \mathbb{R}^{M \times D}$ denotes the features of local fragments (*e.g.*, $\mathbf{V}$ or $\mathbf{W}$), C indicates the total number of cells in each layer[4], and $\mathbf{O}_j^{(l-1)} \in \mathbb{R}^{M \times D}$ represents the output of j-th cell in the $(l-1)$-th layer (refer to Eqn. (1)). In this work, $g_{j,i}^{(l-1)} \in [0, 1]$ denotes the path probability from the j-th cell in the $(l-1)$-th layer to the i-th cell in the l-th layer. It is calculated as $\mathbf{g}_i^{(l)} = \mathcal{G}_i^{(l)}(\mathbf{X}_i^{(l)}) \in \mathbb{R}^C$, where $\mathcal{G}_i^{(l)}(\cdot)$ denotes the routing function of the i-th cell in the l-th layer. More concretely, this function is implemented by average pooling followed by a *MLP* and two activation functions as,

$$\mathcal{G}_i^{(l)}(\mathbf{X}_i^{(l)}) = \text{ReLU}\{\text{Tanh}[MLP(\frac{1}{M}\sum_{r=1}^{M} \mathbf{x}_{i,r}^{(l)})]\}, \tag{10}$$

where $\mathbf{x}_{i,r}^{(l)}$ is the r-th row vector of $\mathbf{X}_i^{(l)}$. Different from the hard gate used in [18, 38], we consider a soft version via generating continuous values as path probabilities, making direct gradient propagation available.

When the routing process ends, we can derive the final refined feature matrix $\mathbf{X}^* = \mathbf{X}_0^{(L)}$ through Eqn. (9) from the last layer, which only has one cell. We ultimately aggregate these local embeddeings (*i.e.*, the row vectors of $\mathbf{X}^*$) by pooling operation[5] to obtain the refined global representation $\mathbf{x}^*$. It will be used for similarity calculation with the global representation of another modality $\bar{\mathbf{y}}$.

3.4 Objective Function

3.4.1 Alignment Objective. To achieve semantic alignment of a given positive image-text pair (I, T), we utilize the hinge-based bidirectional triplet loss for optimization, which is defined as,

$$L_A = [\alpha - s(I,T) + s(I,\hat{T})]_+ + [\alpha - s(I,T) + s(\hat{I},T)]_+, \tag{11}$$

where α represents a margin factor, $[x]_+ = \max(x, 0)$, and $s(I, T)$ denotes the cosine similarity between the global representations of I and T. Specially, $s(I,T) = \cos(\mathbf{v}^*, \bar{\mathbf{w}})$ for the model of i-t version, and $s(I,T) = \cos(\bar{\mathbf{v}}, \mathbf{w}^*)$ for the t-i version. $\hat{T} = \text{argmax}_{j \neq T} s(I, j)$ and $\hat{I} = \text{argmax}_{i \neq I} s(i, T)$ are the hardest negatives in a mini-batch.

3.4.2 Path Regularization. Besides the complexity of input samples considered in Section 3.2.1, high-level semantics may also affect the learning of interaction patterns. In general, samples with similar semantics should learn similar routing paths. In other words, the routing distribution is expected to be consistent with the semantic distribution. To this end, we introduce a path regularization term by considering semantic similarities between samples.

Considering that the original BERT embeddings incorporate abundant semantic information, we leverage them to guide the routing learning. In particular, given an instance x (x is an image if we optimize the model of i-t version, otherwise it is a sentence.), we first extract word embeddings $\mathbf{E}_x$ by BERT[6], and then adopt average-pooling to obtain the semantic representation $\bar{\mathbf{e}}_x$ of x. Afterwards, we collect and concatenate the gate values from all routers, obtaining the path vector[7] $\mathbf{g}_x \in \mathbb{R}^{C^2(L-1)+C}$. To achieve semantic-path consistency, we formulate the regularization as,

$$L_P = \sum_{y \in \mathcal{B}} [\cos(\mathbf{g}_x, \mathbf{g}_y) - \cos(\bar{\mathbf{e}}_x, \bar{\mathbf{e}}_y)]^2, \tag{12}$$

where $\mathcal{B}$ is a collection of instances with the same modality as x.

Finally, we combine the above triplet loss and the path regularization term to obtain the total objective function as,

$$L = L_A + \lambda_P L_P, \tag{13}$$

where λ_P serves as the balance factor.

4 EXPERIMENTS

To justify the effectiveness of our DIME model, we carried out experiments under the bidirectional retrieval scenario involving 1) Image-to-Text (I2T) retrieval, i.e., retrieving sentences that can well depict the content of a given image; and 2) Text-to-Image (T2I) retrieval, i.e., retrieving images that are semantically consistent with a given text query.

4.1 Datasets

In this paper, we conducted experiments on the two widely-used benchmark datasets: Flickr30K and MS-COCO, to evaluate our proposed model and several state-of-the-art baselines.

Flickr30K [39]. This dataset consists of 31,783 images, where each image is described by 5 different sentences. Following the settings in previous work [4, 15, 26], this dataset is split into 29,783 training images, 1,000 validation images, and 1,000 testing images.

MS-COCO [19]. It is a large-scale dataset including 123,287 images, where each image is associated with 5 annotated sentences. Similarly, we followed the split of [4, 15, 26], *i.e.*, 113,287 images for training, 5,000 images for validation, and 5,000 images for testing. Meanwhile, two evaluation settings are considered in this paper: 1) **MS-COCO 1K**, the final result is calculated by averaging the results over 5-folds of 1K testing images; and 2) **MS-COCO 5K**, the evaluation result is directly calculated on the full 5K testing images.

4.2 Experimental Settings

4.2.1 Evaluation Protocols. Following the existing baselines [15, 16, 26], we adopted Recall at K, R@K (K=1, 5, and 10) for short, as the evaluation metrics, which are commonly utilized in the information retrieval community. To be specific, R@K is defined as the percentage of ground truth being retrieved at top-K results. The higher R@K indicates the better performance.

4.2.2 Implementation Details. We optimized our proposed model on 1 GeForce RTX 2080 Ti GPU using PyTorch library. The Adam optimizer is employed with a mini-batch size 64 and 30 epochs. The

[4]Since we have designed four types of cells, $C = 4$ in this paper.

[5]In our work, the final refined global representation is obtained by adding the average pooling result and the max pooling result.

[6]If x is an image instance, we apply BERT to the sentence corresponding to it.

[7]In a mini-batch with size B, B sentences/images would be interact with the given image/sentence x, hence we can get B path vectors for x. We then utilize average-pooling to obtain $\mathbf{g}_x$.

Table 1: Performance comparison between our proposed DIME and several state-of-the-art baselines on the Flickr30K and MS-COCO datasets. And statistical significance over R@1 between DIME* and the best baseline (i.e., CAMERA*) is determined by a t-test (△ denotes p-value < 0.01). The symbol '*' refers to the ensemble result. The best performance is highlighted in bold.

Method	Flickr30K Dataset						MSCOCO (1K) Dataset						MSCOCO (5K) Dataset					
	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
SCAN* [15]	67.4	90.3	95.8	48.6	77.7	85.2	72.7	94.8	98.4	58.8	88.4	94.8	50.4	82.2	90.0	38.6	69.3	80.4
CAMP [33]	68.1	89.7	95.2	51.5	77.1	85.3	72.3	94.8	98.3	58.5	87.9	95.0	50.1	82.1	89.7	39.0	68.9	80.2
BFAN* [21]	68.1	91.4	-	50.8	78.4	-	74.9	95.2	-	59.4	88.4	-	-	-	-	-	-	-
SAEM [37]	69.1	91.0	95.1	52.4	81.1	88.1	71.2	94.1	97.7	57.8	88.6	94.9	-	-	-	-	-	-
CAAN [40]	70.1	91.6	97.2	52.8	79.0	87.9	75.5	95.4	98.5	61.3	89.7	95.2	52.5	83.3	90.9	41.2	70.3	82.9
DP-RNN [4]	70.2	91.6	95.8	55.5	81.3	88.2	75.3	95.8	98.6	62.5	89.7	95.1	-	-	-	-	-	-
VSRN* [16]	71.3	90.6	96.0	54.7	81.8	88.2	76.2	94.8	98.2	62.8	89.7	95.1	53.0	81.1	89.4	40.5	70.6	81.1
SGM [32]	71.8	91.7	95.5	53.5	79.6	86.5	73.4	93.8	97.8	57.5	87.3	94.3	50.0	79.3	87.9	35.3	64.9	76.5
IMRAM [3]	74.1	93.0	96.6	53.9	79.4	87.2	76.7	95.6	98.5	61.7	89.1	95.0	53.7	83.2	91.0	39.6	69.1	79.8
MMCA [35]	74.2	92.8	96.4	54.8	81.4	87.8	74.8	95.6	97.7	61.6	89.8	95.2	54.0	82.5	90.7	38.7	69.7	80.8
GSMN* [22]	76.4	94.3	97.3	57.4	82.3	89.0	78.4	**96.4**	98.6	63.3	90.1	95.7	-	-	-	-	-	-
ADAPT* [34]	76.6	95.4	97.6	60.7	86.6	92.0	76.5	95.6	**98.9**	62.2	90.5	96.0	-	-	-	-	-	-
CAMERA* [26]	78.0	95.1	97.9	60.3	85.9	91.7	77.5	96.3	98.8	63.4	90.9	95.8	55.1	82.9	91.2	40.5	71.7	82.5
DIME (i-t)	77.4	95.0	97.4	60.1	85.5	91.8	77.9	95.9	98.3	63.0	90.5	96.2	56.1	83.2	91.1	40.2	70.7	81.4
DIME (t-i)	77,5	93.5	97.5	59.1	85.5	91.0	77.2	95.5	98.5	62.3	90.2	95.8	55.3	82.4	90.2	39.7	70.3	81.0
DIME*	**81.0**△	**95.9**	**98.4**	**63.6**△	**88.1**	**93.0**	**78.8**△	96.3	98.7	**64.8**△	**91.5**	**96.5**	**59.3**△	**85.4**	**91.9**	**43.1**△	**73.0**	**83.1**

learning rate is set as 0.0002 with decaying 10% of every 15 epochs. The snapshot with the highest sum of the recalls on the validation set is selected for testing. The dimension of visual features D_v is 2,048 and the number of visual regions R is 36. The basic version of the pre-trained BERT [5] is leveraged, equipped with 12 layers, 12 heads, 768 hidden units, and 110M parameters in total, to obtain the original word embeddings with dimension $D_t = 768$. The dimension of joint embedding space D is set to 256. As for the inverted temperature factor λ in Eqn. (6), we set it to 4 and 9 on the I2T and T2I tasks, respectively. In addition, the number of routing layers L, the trade-off parameter λ_P in Eqn. (13), and the head number of intra-modal reasoning cell H are set to 3, 0.5, and 16, respectively.

4.3 Performance Comparison

To justify the effectiveness of our proposed DIME model, we compared it with the following state-of-the-art baselines in the task of image-text retrieval.

- Methods that focus on exploring intra-modal interactions, namely, SAEM [37], VSRN [16], and CAMERA [26].
- Methods that aims to design different cross-modal interaction modules, namely, SCAN [15], CAMP [33], BFAN [21], IMRAM [3], and ADAPT [34].
- Methods that synchronously model intra- and cross-modal interactions, namely, CAAN [40], DP-RNN [4], SGM [32], MMCA [35], and GSMN [22].

Note that we directly quoted the results of these baselines from their original papers, except for CAMERA[8]. Bedsides, in addition to the single model DIME (i-t) and DIME (t-i), we provided an ensemble model DIME* for fair comparison, *i.e.*, averaging similarity scores of two single models.

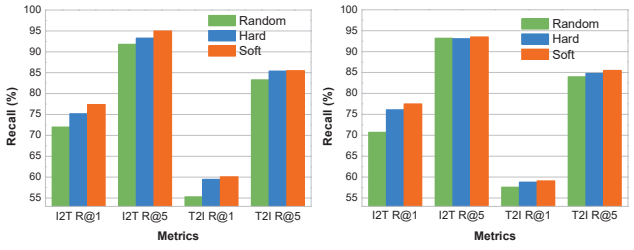
The comparison results are summarized in Table 1. By analyzing this table, we gained the following observations:

- Among intra-modal interaction pattern based methods, CAMERA* [26] surpasses VSRN* [16] by a large margin on two datasets. Although VSRN* builds up connections between image regions, CAMERA* designs more gorgeous internal operations. To be specific, it builds a gating self-attention mechanism for context modeling and a multi-view summarization module for asymmetry matching. This fact indicates that elaborately establishing interaction modules is extremely essential for image-text retrieval.
- Regarding cross-modal interaction pattern based approaches, IMRAM [3] outperforms SCAN* [15] on all criteria of two datasets. Because it could capture high-order correspondences via an iterative cross-attention framework, verifying the importance of aggregating high-order interactions and integrating the iteration strategy.
- Hybrid interaction pattern based methods (*e.g.*, MMCA [35] and GSMN* [22]) are superior to those based on the cross-modal interaction pattern (*e.g.*, SCAN* [15]) and the intra-modal interaction pattern (*e.g.*, VSRN* [16]). This reveals that jointly modeling the intra- and inter-modal interactions plays a significant role in image-text retrieval, contributing to more powerful representation and enhancing alignment.
- Our proposed model DIME* outperforms the compared baselines regarding R@K with different depth on Flickr30K and MS-COCO. Compared with CAMERA*, our approach obtains relative R@1 gains with 3.8% at I2T retrieval on Flickr30K, and it achieves improvement with nearly 1.7% and 7.6% R@1 gain on the MS-COCO (1K) and MS-COCO (5K) test set, respectively. Likewise, for T2I retrieval, our model produces the best results over previous methods on all metrics. The improvement indicates the feasibility and importance of dynamically exploring modality interaction patterns.
- The results of our single model DIME (i-t) and DIME (t-i) are competitive to some state-of-the-art methods, especially the prior ensemble models (*e.g.*, VSRN* [16] and GSMN* [22]). This further demonstrates the effectiveness and robustness of our proposed model and indicates the remarkable ability of our interaction modules. Moreover, it is can be observed

[8]We reproduced CAMERA (https://acmmmcamera.wixsite.com/camera) for the following significance test.

Table 2: The ablation study on Flickr30K to investigate the effect of different modality interaction cells. The best results are highlighted in bold.

Model	DIME (i-t)				DIME (t-i)			
	Image-to-Text		Text-to-Image		Image-to-Text		Text-to-Image	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
w/o RIC	76.4	93.2	59.8	85.4	75.7	94.1	58.8	83.9
w/o IMRC	75.6	93.5	58.1	84.8	73.3	91.9	58.0	84.1
w/o GLGC	75.9	94.3	59.9	85.4	75.0	**94.5**	**59.2**	85.1
w/o CMRC	68.9	90.1	51.2	80.7	61.2	86.2	47.7	77.4
Full	**77.4**	**95.0**	**60.1**	**85.5**	**77.5**	93.5	59.1	**85.5**



(a) DIME (i-t) Model (b) DIME (t-i) Model

Figure 4: The ablation study on Flickr30K to justify the effect of the soft router.

that the performance of DIME (t-i) is slightly inferior to that of DIME (i-t). This may be because that sentences are more abstract and subjective.

In addition, we also conducted the significance test over R@1 between our model and the most competitive baseline CAMERA*. To be specific, on Flickr30K, the p-values of I2T and T2I retrieval are 2.0E-3 and 9.2E-8, respectively. On MS-COCO (1K), the p-values are separately 9.4E-5 and 3.6E-7. Moreover, the corresponding p-values on MS-COCO (5K) are 4.9E-7 and 1.3E-8, respectively. It can be seen that these p-values are observably smaller than 0.01, indicating the statistically significant advantage of our model DIME.

4.4 Module Analysis

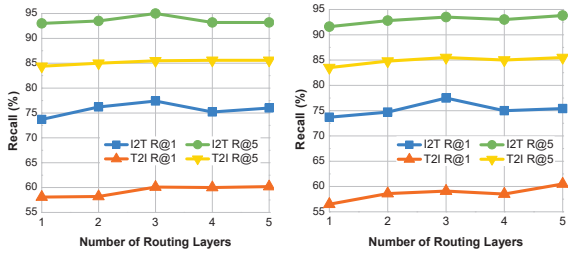
In this section, we carried out several experiments on Flickr30K using the single model (*i.e.*, DIME (i-t) and DIME (t-i)) to further analyze the effectiveness of our model. Specifically, we first explored how each component of our framework affects the image-text retrieval results, including four modality interaction cells, the router, and the path regularization term. We then displayed how the number of routing layers influences the retrieval performance.

4.4.1 Modality Interaction Cells. To gain the insights into our four interaction cells, we conducted ablation studies incrementally. To be more specific, we compared our model DIME with the following variants: 1) **w/o RIC**, removing the rectified identity cell; 2) **w/o IMRC**, eliminating the intra-modal reasoning cell; 3) **w/o GLGC**, without the global-local guidance cell; and 4) **w/o CMRC**, excluding the cross-modal refinement cell.

As reported in Table 2, compared with our model, the performance of **w/o CMRC** degrades dramatically. Particularly, It drops absolutely by 8.5% and 8.9% on R@1 of I2T and T2I for DIME (i-t), respectively. This demonstrates the vital importance of local-local cross-modal interaction as it can capture local discriminative clues. Besides, our model achieves better results than **w/o IMRC**,

Table 3: Performance comparison on Flickr30K with different trade-off values for path regularization. The best results are highlighted in bold.

λ_P	DIME (i-t)				DIME (t-i)			
	Image-to-Text		Text-to-Image		Image-to-Text		Text-to-Image	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
0	76.0	93.8	58.7	85.2	76.0	93.7	59.9	85.2
0.1	76.4	94.1	59.4	85.3	76.3	**94.7**	59.7	85.2
0.5	**77.4**	**95.0**	**60.1**	85.5	**77.5**	93.5	59.1	**85.5**
1.0	76.0	93.0	59.5	85.2	76.1	94.3	**60.0**	84.6
2.0	75.7	93.7	59.7	**85.8**	73.0	92.9	58.2	84.6



(a) DIME (i-t) Model (b) DIME (t-i) Model

Figure 5: Results comparison on Flickr30K regarding different number of routing layers L.

revealing that the intra-modal reasoning cell can enhance the representations of local fragments and boost the model performance. Moreover, the performance drop of **w/o GLGC** can be observed, indicating that it is important to consider the global information from another modality as the guidance to enhance the local representations of current modality. In general, our proposed model largely exceeds all variants on I2T and T2I retrieval, verifying the effectiveness and complementarity of four interaction cells.

4.4.2 The Router. To validate the impact of our proposed soft router, we conducted a series of experiments by introducing two variants: 1) **Random**, deriving the path probability of each cell from a uniform distribution (*i.e.*, $g \sim U[0, 1]$) independently; and 2) **Hard**, adopting the hard router. In other words, on the basis of [31], we introduced the gumbel-softmax [13] trick to discretize path values (*i.e.*, $g \in \{0, 1\}$), which enables the network to be optimized end-to-end by back propagation algorithm.

From the Figure 4, we observed that our proposed soft router achieves the best performance across all metrics consistently. Although the hard version does not perform as well as the soft version, it still obtains better results than the random variant. These facts demonstrate that the routers are capable of learning appropriate paths for different inputs automatically, enabling the model to explore more possible optimal modality interaction patterns.

4.4.3 Path Regularization. To justify the effectiveness of the routing regularization term discussed in Section 3.4.2, we designed a group of experiments by setting different λ_P in Eqn. (13). The results are summarized in Table 3.

From Table 3, we could see that the path regularization improves the performance as compared with the results in the first row (*i.e*, without regularization). This strongly illustrates the validity of the path regularization. Moreover, the performance first improves before reaching the saturation point (*i.e.*, λ_P=0.5), and then begins to decline slightly. The main reason may be that overly encouraging

Figure 6: Visualization of modality interaction path using T-SNE on Flickr30K. Each point denotes an image-level path vector learned via DIME (i-t). Different colors indicate different cluster labels assigned by spectral clustering [25].

the diversity of paths lead to a serious over-fitting phenomenon. Besides, the performance of our model changes within small ranges nearby the optimal setting. This justifies that our model is non-sensitive and robust to the parameter around its optimal setting.

4.4.4 Parameter Analysis. To explore the impact of the parameter L (*i.e.*, the number of routing layers), we conducted experiments by increasing it from 1 to 5. The results are shown in Figure 5.

From the comparison results, we could find that increasing the number of layers in an appropriate range (*i.e.*, from 1 to 3) can improve the retrieval performance by enhancing the representation ability of the model. This can be attributed to that more layers offer broader path space, thus increasing the probability of searching more unexplored superior patterns. However, when L is greater than 3, the performance begins to drop. The reason may be that the path space becomes very uneven, limiting the model optimization and further hindering the path learning.

4.5 Path Visualization

Apart from achieving the superior performance, the key advantage of DIME over other methods is that its dynamic interaction modeling strategy is able to adaptively assign different modality interaction paths for different inputs. To this end, we showed some images and visualized their path vectors learned by DIME (i-t). To be specific, we first obtained the path vector for each image, and then used t-SNE [29] to map the path vector into the two-dimensional space. Afterwards, we clustered these 2D vectors into five groups, where each group marked in one color, as shown in Figure 6.

From Figure 6, we could see that the images related to human (shown in black, blue, and green points) and the ones related to animals (shown in red and yellow points) can be well distinguished, according to the distribution of learned path vectors. For instance, there is a large margin between the red points (related to dogs) and the black ones (associated with basketball players), as they are clearly distinct. Although both red and yellow points are related to animals, they are still wide apart. Because there is much difference in their fine-grained semantics. Our proposed semantic regularization can transfer this knowledge to the routers, thus our model can learn entirely different paths for them. Likewise, diverse activity scenes are also discriminated by path vectors, such as the public events with lots of crowds (top green points), sports events (black points), and music activities (bottom blue points). These results demonstrate that 1) our model can intelligently learn specific semantic-aware paths for different inputs, therefore the distribution of learned paths is consistent with that of semantics to some extent. And 2) our proposed soft router is capable of perceiving, understanding, and reasoning multimodal data dynamically.

To gain the deep insights into our proposed dynamic modality interaction modeling scheme, we illustrated several results with different paths. Concretely, we employed 0.7 as the threshold to discretize the learned paths (*i.e.*, only showed paths of which the probability values greater than the threshold) for improving the intuitiveness. The results obtained from DIME (i-t) and DIME (t-i) are displayed in the first and second row of Figure 7, respectively. We gained the following observations. 1) More complicated inputs require more complex interaction paths, which is consistent with the perception of the human brain. For instance, the first image in the first row simply depicts a man in a horse, and it merely activates a few interaction cells. 2) The main difference between the second and the third columns lies in the higher layers of the routing space. Because samples with different semantics may share similar low-level patterns. And 3) samples with more detailed clues require

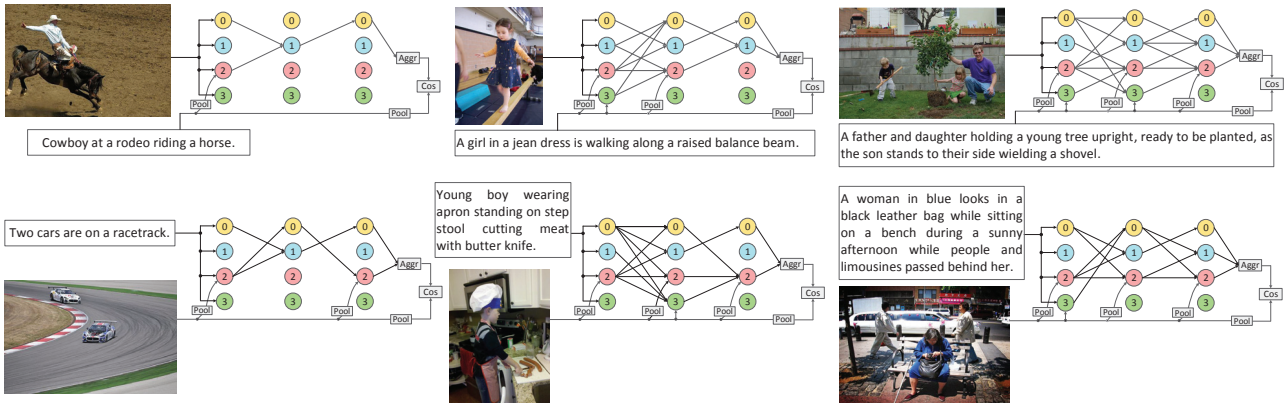


Figure 7: Illustrations of modality interaction patterns. Three examples in the first row are output by DIME (i-t), and the ones in the second row are generated by DIME (t-i).

Figure 8: Top-3 image-to-text retrieval results on Flickr30K. The ground-truth texts are marked with green checks, and the wrong results are indicated by cross marks.

Figure 9: Top-3 text-to-image retrieval results on Flickr30K. The matched images are annotated in green boxes, and the false ones are in red.

more cross-modal refinement cells since they need more cross-modal local-to-local interaction information. These observations reveal that our model can learn different interaction patterns for different samples dynamically. More importantly, these patterns obtained by automatic routing learning may provide some valuable insights for more efficient pattern design in the future.

4.6 Qualitative Results

To qualitatively validate the effectiveness of DIME, we displayed several typical examples on I2T retrieval and T2I retrieval in Figure 8 and Figure 9, respectively. Based on these retrieval results, we could see that our model could comprehend abstract short or complex long sentences accurately. Meanwhile, it is robust for simple or complex images, which is mainly attributed to the dynamic interaction modeling capability of our proposed model. Specially, the rank-3 sentence of the third image query in Figure 8 demonstrates that it is still challenging to count entities exactly due to the language prior problem [9] on existing datasets. Although the rank-1 image of the Query (b) in Figure 9 is not the ground-truth, it is still reasonable for semantic-level retrieval. Actually, it is even difficult for human to distinguish these top-3 results.

5 CONCLUSION AND FUTURE WORK

In this paper, we present a unified modality interaction modeling framework towards image-text retrieval, which is the first work on exploring interaction patterns by dynamic routing learning. Concretely, we first design four types of cells to execute different internal interaction operations and dynamic routers for routing learning. We then introduce a semantic-path consistence regularization for path decision. Extensive experimental results on two benchmarks have demonstrated the effectiveness and superiority of our proposed method.

In the future, we plan to explore more applications of dynamic mechanism in information retrieval systems under the constraint of given computing resources, making it more flexible and extensible.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation (NSF) of China, No.:62006140 and No.:62006142; the Key R&D Program of Shandong (Major scientific and technological innovation projects), No.:2020CXGC010111; the NSF of Shandong Province, No.:ZR2020QF106; new AI project towards the integration of education and industry in QLUT; Young creative team in universities of Shandong Province, No.:2020KJN012.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6077–6086.

[2] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 2018. Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-Image Embeddings. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 35–44.

[3] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. IMRAM: Iterative Matching with Recurrent Attention Memory for Cross-Modal Image-Text Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 12655–12663.

[4] Tianlang Chen and Jiebo Luo. 2020. Expressing Objects just like Words: Recurrent Visual Embedding for Image-Text Matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 10583–10590.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. ACL, 4171–4186.

[6] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 1–13.

[7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 2121–2129.

[8] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. Fashionbert: Text and Image Matching with Adaptive Loss for Cross-Modal Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2251–2260.

[9] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yibing Liu, Yinglong Wang, and Mohan Kankanhalli. 2019. Quantifying and Alleviating the Language Prior Problem in Visual Question Answering. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 75–84.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[11] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. 2019. Scalable Deep Multimodal Learning for Cross-Modal Retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 635–644.

[12] Yupeng Hu, Meng Liu, Xiaobin Su, Zan Gao, and Liqiang Nie. 2021. Video Moment Localization via Deep Cross-modal Hashing. *IEEE Transactions on Image Processing* 30 (2021), 4667–4677.

[13] Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 1–13.

[14] Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3128–3137.

[15] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. In *Proceedings of the European Conference on Computer Vision*. Springer, 201–216.

[16] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual Semantic Reasoning for Image-Text Matching. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 4654–4662.

[17] Yanwei Li, Lin Song, Yukang Chen, Zeming Li, Xiangyu Zhang, Xingang Wang, and Jian Sun. 2020. Learning Dynamic Routing for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 8553–8562.

[18] Ji Lin, Yongming Rao, Jiwen Lu, and Jie Zhou. 2017. Runtime Neural Pruning. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 2181–2191.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*. Springer, 740–755.

[20] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. 2019. Auto-deeplab: Hierarchical Neural Architecture Search for Semantic Image Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 82–92.

[21] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus Your Attention: A Bidirectional Focal Attention Network for Image-Text Matching. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 3–11.

[22] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph Structured Network for Image-Text Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 10921–10930.

[23] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive Moment Retrieval in Videos. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 15–24.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 1–12.

[25] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. 2002. On Spectral Clustering: Analysis and an Algorithm. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 849–856.

[26] Leigang Qu, Meng Liu, Da Cao, Liqiang Nie, and Qi Tian. 2020. Context-Aware Multi-View Summarization Network for Image-Text Matching. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1047–1055.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 91–99.

[28] Lin Song, Yanwei Li, Zhengkai Jiang, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. 2020. Fine-Grained Dynamic Head for Object Detection. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 1–11.

[29] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 5998–6008.

[31] Andreas Veit and Serge Belongie. 2018. Convolutional Networks with Adaptive Inference Graphs. In *Proceedings of the European Conference on Computer Vision*. Springer, 3–18.

[32] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. 2020. Cross-Modal Scene Graph Matching for Relationship-Aware Image-Text Retrieval. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*. IEEE, 1508–1517.

[33] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 5764–5773.

[34] Jonatas Wehrmann, Camila Kolling, and Rodrigo C Barros. 2020. Adaptive Cross-Modal Embeddings for Image-Text Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 2020. AAAI Press, 7718–7726.

[35] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-Modality Cross Attention Network for Image and Sentence Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10941–10950.

[36] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-Modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 1437–1445.

[37] Yiling Wu, Shuhui Wang, Guoli Song, and Qingming Huang. 2019. Learning Fragment Self-Attention Embeddings for Image-Text Matching. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 2088–2096.

[38] Zuxuan Wu, Tushar Nagarajan, Abhishek Kumar, Steven Rennie, Larry S Davis, Kristen Grauman, and Rogerio Feris. 2018. Blockdrop: Dynamic Inference Paths in Residual Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 8817–8826.

[39] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[40] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. 2020. Context-Aware Attention Network for Image-Text Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3536–3545.

[41] Zhu Zhang, Zhijie Lin, Zhou Zhao, and Zhenxin Xiao. 2019. Cross-modal Interaction Networks for Query-based Moment Retrieval in Videos. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 655–664.

[42] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-Path Convolutional Image-Text Embeddings with Instance Loss. *ACM Transactions on Multimedia Computing, Communications, and Applications* 16, 2 (2020), 1–23.