GraDual: Graph-based Dual-modal Representation for Image-Text Matching

Siqu Long[*,1,3], Soyeon Caren Han[*,†,1,3], Xiaojun Wan[2,4], and Josiah Poon[1,3]

[1]School of Computer Science, The University of Sydney
[2]Wangxuan Institute of Computer Technology, Peking University
[3]{slon6753,caren.han,josiah.poon}@sydney.edu.au, [4]wanxiaojun@pku.edu.cn

Abstract

*Image-text retrieval task is a challenging task. It aims to measure the visual-semantic correspondence between an image and a text caption. This is tough mainly because the image lacks semantic context information as in its corresponding text caption, and the text representation is very limited to fully describe the details of an image. In this paper, we introduce **Gra**ph-based **Dual**-modal Representations (**GraDual**), including Vision-Integrated Text Embedding (VITE) and Context-Integrated Visual Embedding (CIVE), for image-text retrieval. The GraDual improves the coverage of each modality by exploiting textual context semantics for the image representation, and using visual features as a guidance for the text representation. To be specific, we design: 1) a dual-modal graph representation mechanism to solve the lack of coverage issue for each modality. 2) an intermediate graph embedding integration strategy to enhance the important pattern across other modality global features. 3) a dual-modal driven cross-modal matching network to generate a filtered representation of another modality. Extensive experiments on two benchmark datasets, MS-COCO and Flickr30K, demonstrates the superiority of the proposed GraDual in comparison to state-of-the-art methods.*

1. Introduction

Image-text matching is one of the fundamental tasks in the field of vision and language cross-modal research, and mainly focuses on two types of tasks, 1) an image retrieval for given sentence-based image description, and 2) a sentence retrieval from image queries. However, it is very challenging to accurately measure the visual-semantic similarity between a textual sentence and an image, due to the different nature of two modalities. For example, text representation is very limited to fully describe the visually realistic appearance like images, and image representation does not include the semantic contexts like text.

Existing image-text matching approaches focus on extracting high-level features from both images and sentences. The extracted features are jointly projected onto the same shared embedding space on top of full sentence and whole image representations [6, 26]. Those approaches have three sub-components: 1) visual feature processing, 2) language feature processing, and 3) multi-modal integration using the shared space. First, for visual feature processing, almost all approaches use pretrained CNNs/ResNet-101 models to extract visual features [15, 31]. For language feature processing, most models extract semantic information using pre-trained word representations, Word2Vec or Glove, [2, 31] with syntactic information from part-of-speech tagging or dependency parsing [19]. Then, it mainly focuses on the joint integration of extracted visual and language features, and measures the similarity by learning global or local region-word correspondence. The global correspondence learning methods aim to jointly project the whole image and text into a common latent space [22, 17, 27], where corresponding image and text can be unified into similar representations. The local region-word correspondence can be learned between salient regions and keywords [15, 29].

However, existing approaches do not consider the different nature and the lack of interchangeability of image and text representation, which would be very crucial for measuring the similarity of those two modalities in the sharing space. Their image representation still lacks semantic information of objects and relations to identify its corresponding textual caption, and the text/sentence representation covers quite limited information to fully describe the visual details of an image. For example, the given text caption in Fig-

[*]Equal contribution †Corresponding author

ure 1, *A man on a skateboard with a brown dog*, does not include the required visual information to identify the similar image; what objects are in the image (*aspect*)? where are those objects (*position*)? how to represent the relations between objects (*relations*)? The rich visual semantics should be included with the text descriptions to accurately retrieve the most similar photo-realistic visual output (image). Likewise, the plentiful contextual semantics should be integrated with the input image representations in order to find the most appropriate textual description.

Some graph-based approaches [19, 28, 11, 33] generate the graph representation to identify objects and relations but those by focusing on simply converting a single text or image information into an individual graph structure respectively at the graph generation stage. No cross-modal integration occurs until attention or final prediction calculation.

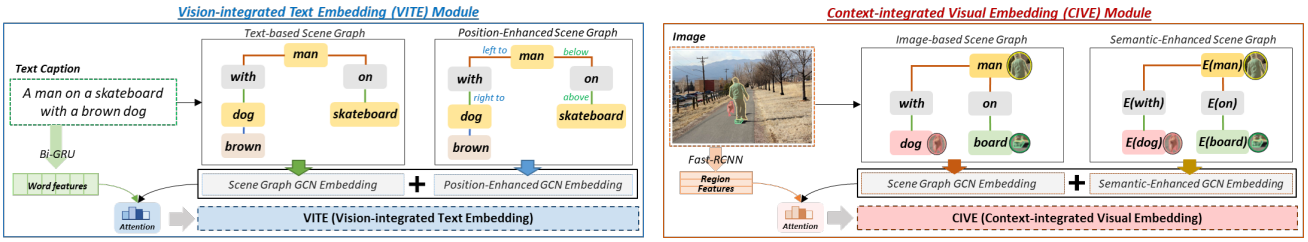
In this paper, we propose a Graph-based Dual-modal Representation (GraDual), the first model that focuses on producing the cross-modal graph representation by integrating the contextual semantic information from the two modalities into each other, using both 1) Vision-Integrated Text Embedding (VITE) that integrates the rich visual semantic information to the text representation and 2) Context-Integrated Visual Embedding (CIVE) that includes the plentiful contextual semantic information to the image representation, in the initial graph generation stage. Then, the GraDual utilises the general and global contextual semantic information learned based on graph structures from both modalities and integrate them into each other for better visual-textual aligned representation in the early stage. Finally, a dual-modal driven cross-modal matching network generates a filtered representation of another modality and retrieves the corresponding text/image.

The contributions of the paper can be summarised as: **1)** To the best of our knowledge, we propose the first Image-based Contextual Visual graph representation and Text-based Visual Semantic graph representation for the visual-language cross-modality research, especially image-text matching task. **2)** Our GraDual enhances the interchangeability of vision and language representation by integrating cross-modal information for visual-language matching tasks. **3)** We conduct experiments on two widely-used image-text matching dataset, Flickr30K and MSCOCO, showing our superiority over state-of-the-arts.

2. Related Work

Recent studies in text-image retrieval use a shared visual-semantic space that maps the represented visual and textual vectors and enforces the distance-measured similarity using ranking loss. Frome *et.al* [7] focused on aligning images and text at only global level. Karpathy *et.al* [14] advocated the needs of finer level matching considering local-level similarity by applying Regional Convolutional Neural Network (RCNN) and dependency tree relations. Similarly, Niu *et.al* [23] parsed the text into a constituency tree and jointly learns the text and image embeddings via multi-level matching loss. Karpathy *et.al* [13] then further build on the work in [14] by replacing the dependency tree relations with Bi-RNN-based textual representation, and the approach has been very popular among later research [8, 15, 16, 3]. Convolutional Neural Network (CNN) was applied for multi-level granularity in order to exploit the inter-modal correspondence at different levels such as character, word, phrase, sentence together [21, 30, 35]. Attention [25] has been applied to filter the important text phrases and image regions continuously at multiple steps [8, 22, 2]. Lee *et.al* [15] invoked object-word pair-wise attention to calculate visual-attended word representation to match with the image region (or vice-versa). Wehrmann *et.al* [31] proposed channel-wise attention for reconstruction of image/text adapted by factors learned from the other modality. More recently, many studies focus on building Vision-Language (VL) pretraining models and achieve promising performance on downstream VL tasks including text-image retrieval [20, 4, 18, 24, 10]. Drawing on the benefits of large-scale pretraining corpus, they tend to make less efforts on the design of initial representation and model architecture. Comparatively, in order to better utilize finite data resource, some studies incorporate graph structured information for initial representation learning [16, 32], global similarity calculation [19], global and local alignment reasoning [5]. All those approaches only use individual graph-based representations of either image or text, without considering the dual-modality representation to incorporate the information learned from the other modality at the initial representation stage. However, our Gradual integrates the cross-modal information for each image and text so it can enhance the interchangeability of both image and text representations at the early representation stage.

There are some dual-graph approaches [19, 28, 32, 11, 33] in VL tasks, which generate the graph representation by focusing on simply converting a single text or image information into an individual graph structure respectively at the graph generation stage. Hence, no cross-modal integration occurs until the cross-modal reasoning (attention-guided) [19, 11, 33] or final prediction calculation [28, 32]. However, our Gradual is the first model, which focuses on producing the cross-modal graph representation by integrating the contextual semantic information from the two modalities into each other for both 1) textual graph (with visual information) and 2) visual graph (with textual information) in the initial graph generation stage. Hence, the GraDual utilises the general and global contextual semantic information learned based on graph structures from both modalities and integrate them into each other for better visual-textual aligned representation in the early stage.



(a) VITE (Vision-Integrated Text Embedding)

(b) CIVE (Context-Integrated Visual Embedding)

Figure 1: Illustration of generating the Gradual(Graph-based Dual modal) representation: (a) Vision-Integrated Text Embedding (VITE) for textual sentence representation. (b) Context-Integrated Visual Embedding (CIVE) for image representation.

3. Methodology

We propose a **GraDual**: **Gra**ph-based **Dual**-modal Representation to enhance the interchangeability of single modal (vision or language) representation for improving image-text alignment. The overview of our proposed framework is illustrated in Figure 1.

3.1. GraDual

We propose two variants of GraDual representation: (1) Vision-Integrated Text Embedding (VITE) for sentence representation and (2) Context-Integrated Visual Embedding (CIVE) for image representation. Details of these two representations are provided as follows.

3.1.1 Vision-Integrated Text Embedding (VITE)

Vision-Integrated Text Embedding (VITE) aims to bridge textual representation with visual information, including attributes (*how objects look like*), position (*where those objects are located in*), and relations (*how the relations between multiple objects*). It has 5 steps: (1) Textual Scene Graph Generation, (2) Text-based Scene Graph Embedding, (3) Position-enhanced Scene Graph Embedding, (4) Multi-Graph Embedding Aggregation, (5) VITE Representation.

Textual Scene Graph Generation Given the raw text caption T for an image I, we parse T into a graph-based semantic representation called scene graph [12], which explicitly represents a scene using objects, their attributes and the relation between objects. The idea of scene graph would help extracting and representing the visual information of the objects, attributes, relations that are expressed by the raw text captions.

First, we apply the Stanford enhanced dependency parser [1] to recognise the syntactic structure of the textual description. The output of the syntactic parser is not enough to represent the image scene, especially the number of objects in the scene. Hence, we build quantity checker for detecting its quantifier expression (the number of objects), and duplicating object nodes for the scene graph. For example, if the textual caption contains "two men" or "three buildings", the scene graph should include two 'man' nodes, and three 'building' nodes. The attributes and relations of the duplicated objects are also copied accordingly. For our scene graph, all nouns are extracted and classified into objects classes, and all adjectives are defined as attributes of the paired objects for pairwise classification. The relation is then detected if the word in between is the predicate or preposition of the two object instances. The example of a text-based scene graph can be found in the Figure 1(a).

For each text T, a textual scene graph $G_T = (O, R, A)$ is generated, in which $O = \{o_1, o_2, ..., o_n\}$, $R = \{r_1, r_2, ..., r_m\}$ and $A = \{a_1, a_2, ...a_k\}$ are sets of objects, relations and attributes. For each object $o_i \in O$, we further assign a specific hypernym $p_t \in P_T = \{p_1, p_2, ..., p_t\}$.

Text-based Scene Graph Embedding We then encode the scene graph into vectorised feature representation that conveys the visual semantic cues from the image scene structure for each object, attribute and relation node. We first combine all the scene graphs as our textual scene graph G_{T_s} and apply GCNs to model the relative closeness of nodes and edges in the graph. Specifically, for $G_{T_s} = (O, R, A)$, the connection between a relation $r_m \in R$ and its two connected objects $o_i \in O$ are represented as edges $e_{o_i \to r_m}$ and $e_{r_m \to o_i}$ while the connection between an object o_i and its linked attribute $a_k \in A$ is represented as edge $e_{o_i \to a_k}$. The edge weights W_e is defined using equation 1, where $N(e)$ calculates the total count of edge e in the whole graph. The node's self-connection weight is set to 1. All the edge weights are complied into an adjacency matrix M_A and fed into a 2-layer GCN together with the graph degree matrix M_D, which is then trained by mapping each object to its assigned hypernym p_t. After training, we take the node embedding for object, attribute and relation as our textual scene graph embedding $TS_o, TS_a, TS_r \in \mathbb{R}^{D_{TS}}$.

$$W_e = \begin{cases} W_{e_{o_i \to r_m}} = \frac{N(e_{o_i \to r_m})}{N(e_{o_i \to R})} \\ W_{e_{r_m \to o_i}} = \frac{N(e_{r_m \to o_i})}{N(e_{r_m \to O})} \\ W_{e_{o_i \to a_k}} = \frac{N(e_{a_k \to o_i})}{N(e_{a_k \to O})} \end{cases} \quad (1)$$

Position-Enhanced Scene Graph Embedding The produced textual scene graph embedding represents the useful attribute and relation features of images but it mainly captures the semantic relations between objects, such as predicates (e.g. ride, eat). It provides the lingual semantics of objects and relations but it is not enough to fully align with geographical position of objects or relative position between objects. We construct an additional position-enhanced graph $G_{T_p} = (O, R)$ for representing the geographical information from visual content in the image that can be aligned with the positional semantics in the text. The object and relation nodes are directly from G_{T_s} while the edges between nodes are now decided by the geometric relation of bounding boxes of the connected objects, which is then mapped to a set of geometric relation types $g \in \{left\_to, right\_to, above, below, inside, surrounding\}$. Since the geometric relation between two objects may simultaneously satisfies more than one relation type of g, we instead construct 6 positional graph $G^g_{T_p}$ for the 6 relation types respectively. The weight calculation is the same as in equation 1 and node's self-connection weight is set to 1. We apply GCNs to train the 6 graphs separately based on the nodes and connections by classifying each object node into its assigned hypernym p_t. The derived node embeddings for object nodes and relation nodes from the 6 graphs are then concatenated at object level and node level, resulting in our textual positional graph embedding for object and relation as $TP_o, TP_r \in \mathbb{R}^{D_{TP}}$.

Multi-Graph Embedding Aggregation Since textual scene graph embedding and positional graph embedding convey complementary visual semantics, we apply min-pooling to aggregate the two graphs at node level for both object and relation to produce our integrated visual-semantic graph embedding $TV \in \mathbb{R}^{D_T}$, as is shown in equation 2. We also explored other aggregation mechanisms and the test results are provided in Section 5.3.

$$TV = \begin{cases} TV_o = min\_pooling(TS_o, TP_o) \\ TV_r = min\_pooling(TS_r, TP_r) \\ TV_a = TS_a \end{cases} \tag{2}$$

VITE Representation Our ultimate goal is to bridge the text representation using the visual semantic information from the learned graph embedding. In order to do this, for each image I with text T, we first concatenate its graph object embedding with its attribute and connected relation embedding to get the object-based visual-semantic graph embedding $TV_{obj} \in \mathbb{R}^{N \times 3D_T}$, in which N is the number of objects in the scene graph for T. Then, we filter TV_{obj} with referring to the text representation T_e for T via attention mechanism [25], as is formulated in equation 3. Here T_e is the encoded vectors for each word in the text T (See Section 3.4 for details). The final attended graph representation is the Vision-Integrated Text Embedding $VITE \in \mathbb{R}^{L \times 3D_T}$, in which L refers to the number of words in T.

$$\begin{aligned} &Attention\left(Q(T_e), K(TV_{obj}), V(TV_{obj})\right) \\ &= softmax\left((W^T T_e) TV_{obj}^T\right) TV_{obj} \end{aligned} \tag{3}$$

3.1.2 Context-Integrated Visual Embedding

Similarly, $CIVE$ aims to bridge visual representation with textual contextual semantics, which is constructed through the following 5 steps.

Visual Scene Graph Generation We extract a scene graph $G_I = (O, R)$ containing objects and associated relations from each image I using Motif-Net [34] and assign a specific hypernym for each object.

Image-based Scene Graph Embedding We combine all the scene graphs as our visual scene graph G_{I_s} and apply GCN training same as for G_{T_s} to learn the node embedding for object and relation as our visual scene graph embedding $VS_o, VS_r \in \mathbb{R}^{D_{VS}}$.

Semantic-Enhanced Scene Graph Embedding We then construct a complementary contextual graph $G_{I_c} = (O, R)$ preserving the graph structure of G_{I_s}. The only difference is to use the word vectors obtained by pretraining on the large text corpus as the initial node feature of objects and relations for GCN training, which incorporates the textual contextual semantics learned from the large text corpus into the graph embedding. We take the derived node embedding for object and relation as our visual contextual graph embedding $VC_o, VC_r \in \mathbb{R}^{D_{VC}}$.

Multi-Graph Embedding Aggregation Same to the aggregation for $VITE$ in equation 2, we min pool over the visual scene graph and contextual graph at node level for object and relation respectively to produce the integrated contextual-semantic graph embedding $VC \in \mathbb{R}^{D_V}$.

CIVE Representation To bridge the image representation using the textual contextual semantics from the learned graph embedding, we first concatenate the graph object embedding with its connected relation embedding for each image I to get the object-based contextual-semantic graph embedding $VT_{obj} \in \mathbb{R}^{N \times 2D_V}$. Then, we filter VT_{obj} using the image representation V_e for I via attention mechanism same to equation 3. Here V_e is the representation for regions in the image (See Section 3.4 for details). The final attended graph representation is the Context-Integrated Visual Embedding $CIVE \in \mathbb{R}^{K \times 2D_V}$, where K denotes the number of regions in I.

3.2. GraDual-based Cross-modal Matching

The essence of our matching mechanism is to utilize the information from one modality instance to generate a filtered representation of another modality for cross-modal matching, using our GraDual representation ($VITE$/$CIVE$) as a bridge, as depicted in Figure 2. We define two complementary formulations of GraDual-based cross-modal matching: Text-to-Image(GraDual-T2I) and Image-to-Text(GraDual-I2T). The generic formulation of

GraDual is illustrated as follows. Assume we have two modalities c_a and c_b. First, we summarize the initial modality representation $e_{c_a} \in \mathbb{R}^{N_{c_a} \times d}$, $e_{c_b} \in \mathbb{R}^{N_{c_b} \times d}$ and the $GraDual$ representation (either $VITE$ or $CIVE$) into global vectors using global $pooling$ and project them to a set of cross-modal transition vectors $V_{transition} \in \mathbb{R}^d$, as shown in equation 4.

$$V_{transition} = \begin{cases} v^{\alpha}_{c_a} = g(pooling(e_{c_a}), \theta_g) \\ v^{\beta}_{c_a} = p(pooling(e_{c_a}), \theta_p) \\ v^{G} = q(pooling(GraDual), \theta_q) \end{cases} \tag{4}$$

where g, p, q are linear projections. Then, we use $V_{transition}$ to guide the re-formulation of all the vectors $e_{c_{bi}}$ from modality c_b based on the scaling and shifting operation[1] in equation 5.

$$\bar{e}_{c_{bi}} = e_{c_{bi}} \odot v^{\alpha}_{c_a} + v^{\beta}_{c_a} + v^{G} \tag{5}$$

in which $\odot$ refers to point-wise vector multiplication and $\bar{e}_{c_b}$ denotes the representation matrix of modality c_b reformulated with the guidance of modality c_a and GraDual. Further, we adopt the $fovea$ module proposed in [31], which conducts a per-dimension λ-smoothed $softmax$ across all the N_{c_b} features in $\bar{e}_{c_b}$ and generates a channel-wise attention-like mask M as in equation 6. This mask is used for self-reconstruction to derive the final filtered representation $\bar{e}^{ftd}_{c_b}$ of modality c_b (equation 7).

$$M_{ij} = \left(\frac{e^{(\bar{e}_{c_{bij}})}}{\sum_{i=1}^{N_{c_b}} e^{(\bar{e}_{c_{bij}})}} \lambda \right) \tag{6}$$

$$\bar{e}^{ftd}_{c_b} = \frac{1}{N_{c_b}} \sum_{i=1}^{N_{c_b}} (\bar{e}_{c_b} \odot M)_i \tag{7}$$

The $\bar{e}^{ftd}_{c_b}$ is then normalized to have unit euclidean norm and the final cross-modal matching will be based on the inner product between $\bar{e}_{c_a}$ (global pooled e_{c_a}) and $\bar{e}^{ftd}_{c_b}$. To align with our GraDual-T2I and GraDual-I2T formulations, we can have c_a as text modality T and c_b as image modality I or the other way round. The modality representation e_{c_a} and e_{c_b} then can be either text representation T_e or image representation V_e accordingly while the $GraDual$ can be either $VITE$ or $CIVE$.

3.3. Objective Function

We adopt the hinge loss function that exponentially increases the relevance of the hard contrastives over time as formulated in equation 8.

$$L = \beta(\epsilon) \cdot L_m + (1 - \beta(\epsilon)) \cdot L_s, \beta = 1 - \eta^{\epsilon} \tag{8}$$

in which the β is the trade-off weight decided by the number of iterations ϵ and the exponential growth rate η. The

[1]We also tried to use GraDual v^G for both scaling and shifting but found that only shifting works better

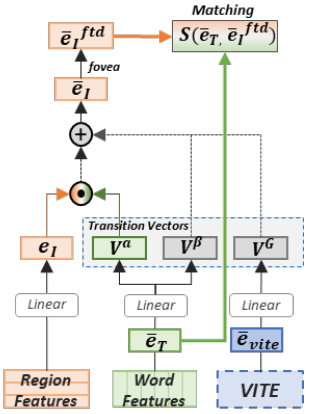


Figure 2: Illustration of the GraDual-based Cross Matching. The example is performed with the $GraDual$ representation ($VITE$). Note that the $\bar{e}_T$ represents the pooled word features and the $\bar{e}_{vite}$ is the pooled $VITE$ representation.

sum of hinges L_s and max of hinges L_m for cross-modal matching are formulated in equation 9 and 10, where r_a and r_b in function S represent the representation of similarity calculation for cross-modal matching. For GraDual-T2I, $r_a = \bar{e}_T$ and $r_b = \bar{e}^{ftd}_I$. For GraDual-I2T, $r_a = \bar{e}_I$ and $r_b = \bar{e}^{ftd}_T$.

$$\begin{aligned} L_s(r_a, r_b) = &\sum_{r_b'} [\alpha - s(r_a, r_b) + s(r_a, r_b')] \\ &+ \sum_{r_a'} [\alpha - s(r_b, r_a) + s(r_b, r_a')] \end{aligned} \tag{9}$$

$$\begin{aligned} L_m(r_a, r_b) = &\max_{r_b'} [\alpha - s(r_a, r_b) + s(r_a, r_b')]_+ \\ &+ \max_{r_a'} [\alpha - s(r_b, r_a) + s(r_b, r_a')]_+ \end{aligned} \tag{10}$$

3.4. Feature Representation

Text representation For each text T with L words w_i, we use Bidirectional Gated Recurrent Unit (Bi-GRU) to encode the bi-directional sequential context in text T and take the average of forward hidden states $\overrightarrow{h_i}$ and backward hidden states $\overleftarrow{h_i}$ as e_i to represent each word at position i, which results in our text representation $T_e \in \mathbb{R}^{L \times d}$. Here d is the dimension of cross-modal shared semantic space.

Image representation For each image I, we use the visual feature (2048d) of 36 salient object regions detected by the object detector from image scene graph parsing to represent the image. Then we project the region features onto the shared semantic space with text via a linear transformation, resulting in our image representation $V_e \in \mathbb{R}^{K \times d}$.

4. Experiment Setup

4.1. Datasets

We evaluate our proposed GraDual on the two widely used benchmark datasets for cross-modal (Text-Image) re-

No.	Models	Flickr30k							MS-COCO						
		Text retrieval			Image Retrieval			rSum	Text retrieval			Image Retrieval			rSum
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
1	SM-LSTM	42.4	67.5	79.9	29.7	60.1	72.1	351.0	52.4	81.7	90.8	38.6	73.4	94.6	421.5
2	VSE++	52.9	-	87.2	39.6	-	79.5	-	64.6	-	95.7	52.0	-	92.0	-
3	DPC	55.6	81.9	89.5	39.1	69.2	80.9	416.2	65.6	89.8	95.5	47.1	79.9	90.0	467.9
4	SCO	55.5	82.0	89.3	41.1	70.5	80.1	418.5	69.9	92.9	97.5	56.7	87.5	94.8	499.3
5	VRSN	71.3	90.6	96.0	54.7	81.8	88.2	482.6	76.2	94.8	98.2	62.8	89.7	95.1	516.8
6	SCAN(T2I+I2T)†	67.4	90.3	95.8	48.6	77.7	85.2	465.0	72.7	94.8	98.4	58.8	88.4	94.8	507.9
7	ADAPT(T2I+I2T)†	76.6	95.4	97.6	**60.7**	86.6	**92.0**	508.9	76.5	95.6	**98.9**	62.2	90.5	96.0	519.8
8	GSMN(sparse+dense)†	76.4	94.3	97.3	57.4	82.3	89.0	496.8	78.4	**96.4**	98.6	63.3	90.1	95.7	522.5
9	SGRAF†	77.8	94.1	97.4	58.5	83.0	88.8	499.6	**79.6**	96.2	98.5	63.2	90.7	96.1	524.3
10	Our GraDual-VITE,I2T	68.6	92.3	96.4	53.4	82.0	88.6	481.3	71.3	94.8	97.8	59.0	89.2	95.0	507.1
11	Our GraDual-VITE,T2I	76.1	94.7	97.7	57.7	84.1	90.5	500.8	76.8	95.9	98.3	63.7	90.8	95.6	521.1
12	Our GraDual-VITE,T2I+I2T†	**78.3**	**96.0**	**98.0**	60.4	**86.7**	**92.0**	**511.4**	77.0	**96.4**	98.6	**65.3**	**91.9**	**96.4**	**525.6**

Table 1: Cross-modal retrieval performance on Flickr30k test set and MS-COCO 1k test set. The reference of baseline models can be found in Section 4.3 The best result is **bolded** and † refers to ensemble models.

trieval task: MS-COCO[2] and Flickr30k[3]. MS-COCO provides 123,287 images with 5 manually written textual descriptions per image. Flickr30k contains around 31,000 images collected from the Flickr website with 5 crowed-sourced corresponding captions per image. We use the same splits as in the state-of-the-art approaches[15, 31]. MS-COCO is split into 113,287 images for training, 5,000 images for validation and 1,000 images for testing. For Flickr30k, models are trained on 29,000 images, validated on 1,000 images and tested on another 1,000 images.

4.2. Evaluation Metrics

The evaluation results are quantitatively measured using metrics same to the state-of-the-art studies [15, 31]. (1) $R@K$ (Recall at K), K=1,5,10 measures the percentage of queries which successfully retrieve the ground-truth as one of the first K results. (2) $rSum$ (Sum of $R@K$) calculates the total value of $R@K$ for both text and image retrieval, as is formulated in equation below, which provides general perspective for the overall retrieval performance. For both metrics, higher value means better performance.

$$rSum = \underbrace{R@1 + R@5 + R@10}_{\text{Text retrieval}} + \underbrace{R@1 + R@5 + R@10}_{\text{Image retrieval}}$$

4.3. Baselines

We compare our test result with the various streams of baseline models that have achieved state-of-the-art performance on MS-COCO 1k and/or Flickr30k: (1) **SM-LSTM [8]**, **SCAN [15]** and **ADAPT [31]** that try cross modal matching using attention or adaptation mechanism over raw image and text features, (2) **DPC [35]** and **VSE++ [6]** that focus on optimizing learning objective, (3) **SCO [9]** which re-design specific networks for learning region-word correspondence, (4) **VRSN [16]** that applies graph-based learning to enhance visual representation via visual positional information, (5) **GSMN [19]** and **SGRAF[5]** that incorporate graph structure for cross-modal similarity reasoning. Especially, compared to other single-variance models, SCAN, ADAPT, GSMN and SGRAF formulate complementary model variances and get the best performance through their ensemble.

[2] https://cocodataset.org/#home

[3] http://shannon.cs.illinois.edu/DenotationGraph/

	SG	Text retrieval			Image Retrieval			rSum
		R@1	R@5	R@10	R@1	R@5	R@10	
I2T	none	70.2	90.8	95.8	55.5	82.7	89.8	484.8
	T-sg	68.4	91.4	96.8	53.1	81.1	88.7	479.6
	I-sg	64.7	88.0	93.5	42.3	70.4	79.5	438.4
	VITE	68.6	92.3	96.4	53.4	82.0	88.6	481.3
	CIVE	64.4	88.7	95.3	49.5	78.8	86.9	463.6
	V+C	66.1	89.3	94.7	41.6	69.4	78.7	439.8
T2I	none	73.6	93.7	96.7	57.0	83.6	90.3	494.9
	T-sg	75.4	**95.0**	97.4	**57.8**	**84.1**	90.2	499.9
	I-sg	73.2	94.4	97.4	56.8	83.7	90.0	495.5
	VITE	**76.1**	94.7	**97.7**	57.7	**84.1**	**90.5**	**500.8**
	CIVE	74.9	94	97.5	56.5	83.7	89.7	496.4
	V+C	72.9	94.3	96.9	57.2	83.9	89.6	494.8

Table 2: Impact of different scene graph representations: cross-modal retrieval on Flickr30k. The best result for I2T/T2I is underlined and the overall best result is **bolded**.

4.4. Implementation Details

We use the pre-computed text and image regions for both MS-COCO and Flickr30k provided by the official github of ADAPT paper [4]. For image graph construction, we di-

[4] https://github.com/jwehrmann/retrieval.pytorch

rectly use the image scene graph from Sub-GC[5] and align the objects with the pre-computed image regions. For GCN hidden layers, we set $D_{TS} = D_{TP} = D_T = 300$ for text graph embeddings and $D_{VS} = D_{VC} = D_V = 200$ for image graph embeddings. The initial representation for image scene graph training uses glove of 300d pre-trained on Wikipedia[6]. We train all the graph embeddings using 2-layer GCNs with $batch\_size = 32$, $learning\_rate = 0.02$ and $epoch = 30$ (with early stopping). All the models are trained using 16 Intel(R) Core(TM) i9-9900X CPU @ 3.50GHz and NVIDIA Titan RTX 24GB, which takes around 20-30 mins for each epoch. The total number of trainable parameter for GraDual-VITE,T2I/I2T is 46,798,394 and 36,522,344 on MS-COCO and Flickr30k.

5. Evaluation Results

5.1. Comparisons with state-of-the-arts

Table 1 shows the test performance on Flickr30k and MS-COCO. We compare the test performance of our best GraDual variances and their ensembles with baseline models[7]. Like those baselines [15, 31, 19, 5], we ensemble our models by averaging their similarity of text-image pairs before retrieving. Firstly, it can be seen that GraDual-VITE,T2I (No.11) significantly outperforms all the non-ensemble baseline models (No.1-5) by a large margin in all metrics and even surpass several ensemble models in most metrics (e.g.SCAN, GSMN, SGRAF on Flickr30k and SCAN, ADAPT on MS-COCO), achieving rSum of 500.8 on Flickr30k and 521.1 on MS-COCO. Comparatively, GraDual-VITE,I2T (No.10) achieves inferior performance. This may be attributed to the image scene graph quality (See discussions in Section 5.2). In addition, comparing with the ensemble models (No.6-9), GraDual-VITE,T2I+I2T (No.12) achieves the top rSum on both datasets (511.4 on Flickr30k & 525.6 on MS-COCO), surpassing the best retrieval model by 2.5% on Flickr30k and the best retrieval model by 1.3% on MS-COCO. More specifically, significant improvements are found in both text and image retrieval tasks on Flickr30k and image retrieval on MS-COCO. Moreover, by comparing with ADAPT ensemble which uses $fovea$ module as GraDual, our GraDual-VITE ensemble improves R@1 by 1.7% for text retrieval on Flickr30k and 0.5%/3.1% for text/image retrieval on MS-COCO. This validates the effectiveness of cross-modal enhanced modality (GraDual) representation that increases the coverage between text/image at initial feature representation stage.

[5] https://github.com/YiwuZhong/Sub-GC/tree/master/data

[6] https://nlp.stanford.edu/projects/glove/

[7] GraDual variance analysis can be found in Table 2

	Text retrieval			Image Retrieval			rSum
	R@1	R@5	R@10	R@1	R@5	R@10	
concat	75.1	94.5	96.9	57.8	83.5	90.0	497.8
max pool.	74.6	94.2	**97.9**	57.3	83.7	90.0	497.6
mean pool.	**76.2**	94.3	97.4	**57.9**	**84.4**	90.4	500.5
min pool.	76.1	**94.7**	97.7	57.7	84.1	**90.5**	**500.8**

Table 3: Impact of different scene graph integration mechanism on GraDual-T2I, VITE: cross-modal retrieval on Flickr30k. The best result is **bolded**.

5.2. Impact of different scene graph representations

To explore the impact of different textual and visual graph representations, we conduct ablation studies for both GraDual-T2I and GraDual-I2T. Specifically, we compare our proposed VITE and CIVE representation with other three ablation variances: (1) **none** for no use of graph structured scene graph representation, (2) **T-sg** and (3) **I-sg** for using only textual or visual scene graph embedding (i.e TS or VS) respectively. (4) **VITE** and (5) **CIVE** for using only VITE or CIVE for GraDual-based Cross-model Matching, (6) **V+C**, which uses both VITE and CIVE (i.e. in the same way of using only VITE or CIVE in equation 4 and 5). The test result is provided in Table 2. Overall, T2I models perform better than their I2T counterparts, which aligns with the trend found from other bidirectional (T2I & I2T) retrieval approaches such as SCAN and ADAPT. This is expected because text is a subset of semantics contained in an image, which may make using text as query source to attend the visual content for formulating filtered visual representation (T2I) more efficient than the opposite (I2T). In addition, the low performance of object detection model may limit searching the relevance between image representation and its corresponding text. Thus, using text as reference to attend or filter on the image (T2I) would lead to more relevant attended or filtered content. We also found several observations from the T2I models: (1) when adding basic text (T-sg) or image (I-sg) scene graph representations compared to using no scene graph embedding (none), the overall rSum improves. Especially, T-sg improves significantly more compared to I-sg. (2) incorporating cross-modal semantics via VITE or CIVE further improves the overall performance compared to T-sg or I-sg respectively. We originally expected the similar patterns from I2T models as well. However, the test result shows that adding I-sg or CIVE with I2T network structure fails to generate better result than using none scene graph representation. This might be attributed to the quality of image scene graph.[8] However, it can be still found that adding textual graph representa-

[8] Originally both detected objects and relations contain many 'background' instances, we cleared only those in the objects. In addition, no attributes are included in the scene graphs due to unavailability.

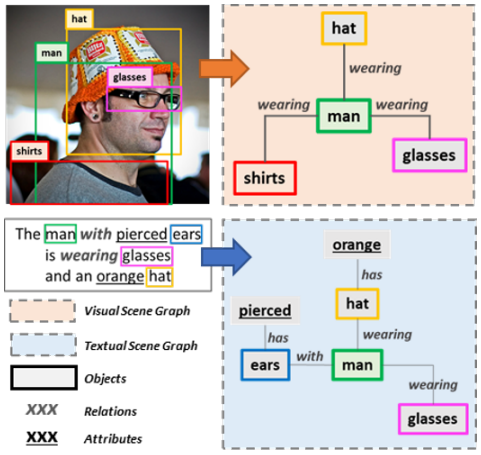


Figure 3: Visualization of visual and textual scene graph on Flickr30k. Upper row shows the image (left) with objects and the visual scene graph (right). Bottom row illustrates the caption (left) and the textual scene graph (right).

tion (both T-sg and VITE) benefit the text retrieval for I2T variance in terms of R@5 and R@10. Overall, incorporating cross-modal semantics via VITE or CIVE improves the overall performance of I2T models compared to T-sg or I-sg respectively. Moreover, when we included VITE and CIVE together (V+C), the overall performance decreases dramatically in both T2I and I2T variances. This may be attributed to the confusion caused by adding two sources of textual or visual information (from the original textual/visual feature and VITE/CIVE) for the modality reformulation.

5.3. Impact of different scene graph aggregation

Our proposed GraDual-VITE applies min pooling for aggregating a textual scene graph and a positional graph. We conduct experiments with different graph aggregation mechanisms in Table 3. Regarding rSum, we find that concatenation results in almost the same result compared to max pooling (497.8, 497.6) while mean and min pooling perform better (500.5, 500.8). More specifically, mean pooling achieves the best R@1 for text retrieval as well as R@1 and R@5 for image retrieval. However, drawing on the improvement on R@5 and R@10 by a larger margin, min pooling outperforms mean pooling and achieves the best result overall (rSum).

5.4. Qualitative Analysis

In Figure 3, we visualise the relevance between image (visual scene graph) and caption (textual scene graph). It can be seen that the major objects man, hat and glasses in the text and image can be clearly represented in both the textual and visual scene graphs together with the corresponding relations between those objects such as wearing between man and hat/glasses, which builds the key elements for the cross-modal alignment. The attribute orange for the man's hat and the positional relations of man-hat, man-glasses are further integrated via $VITE$ and thus generating the visual-enhanced text representation that can align better with the corresponding relevant image content based on the visual clues such as the color of the objects and the geometrical relation between the objects.

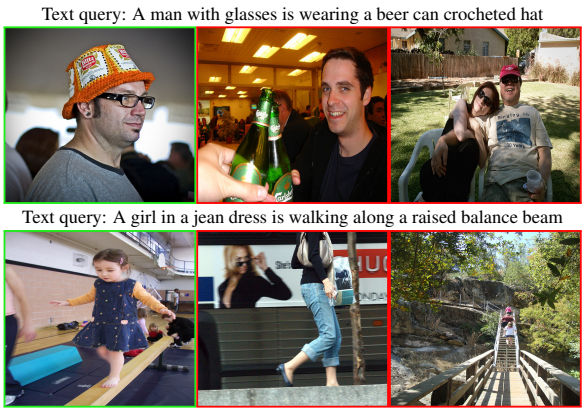


Figure 4: Visualization of image retrieval result. The top 3 images are retrieved for each text. Our approach always retrieves the ground-truth in Top 1 rank.

6. Case Study

We further visualize image retrieval results using our proposed GraDual in Figure 4: We illustrate the same image retrieval samples used in GSMN [19]. It can be observed that our GraDual model can identify the correct image from the similar sets for the two given queries via the matching process illustrated in Section 5.4, which successfully recognize the one man wearing glasses and hat as well as the one girl in a jean dress above the beam with considering the visual geometrical relations and match them between the text query and the image content. We also visualize text retrieval results and provide the analysis in Appendix A.

7. Conclusion

We proposed **Gra**ph-based **Dual**-modal Representations (**GraDual**) for text and image retrieval, which includes Vision-Integrated Text Embeddings (VITE) and Context-Integrated Visual Embedding (CIVE). It improves the coverage of textual and visual modalities by incorporating rich contextual semantics from one modality to enhance the initial representation of the other modality via graph-based learning. We demonstrated promising results of our GraDual-based cross-modal retrieval model by outperforming numerous state-of-the-art counterparts in most of R@K and the overall rSum on both Flickr30k and MS-COCO.

References

[1] Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014.

[2] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12655–12663, 2020.

[3] Tianlang Chen and Jiebo Luo. Expressing objects just like words: Recurrent visual embedding for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10583–10590, 2020.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120, 2020.

[5] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, 2021.

[6] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. 2017.

[7] Andrea Frome, Greg S Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: a deep visual-semantic embedding model. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 2121–2129, 2013.

[8] Yan Huang, Wei Wang, and Liang Wang. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318, 2017.

[9] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6163–6171, 2018.

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. 2021.

[11] Chenchen Jing, Yuwei Wu, Mingtao Pei, Yao Hu, Yunde Jia, and Qi Wu. Visual-semantic graph matching for visual grounding. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4041–4050, 2020.

[12] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[14] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2*, pages 1889–1897, 2014.

[15] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018.

[16] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4654–4662, 2019.

[17] Shuang Li, Tong Xiao, Hongsheng Li, Wei Yang, and Xiaogang Wang. Identity-aware textual-visual matching with latent co-attention. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1890–1899, 2017.

[18] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. 2020.

[19] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10921–10930, 2020.

[20] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32:13–23, 2019.

[21] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, pages 2623–2631, 2015.

[22] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 299–307, 2017.

[23] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 1881–1889, 2017.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. 2021.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.

[26] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. 2015.

[27] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018.

[28] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1508–1517, 2020.

[29] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 3792–3798. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[30] Jônatas Wehrmann and Rodrigo C Barros. Bidirectional retrieval made simple. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7718–7726, 2018.

[31] Jonatas Wehrmann, Camila Kolling, and Rodrigo C Barros. Adaptive cross-modal embeddings for image-text alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12313–12320, 2020.

[32] Keyu Wen, Xiaodong Gu, and Qingrong Cheng. Learning dual semantic relations with graph attention for image-text matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[33] Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9952–9961, 2020.

[34] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018.

[35] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–23, 2020.