Enhancing Cross-Modal Retrieval Based on Modality-Specific and Embedding Spaces

RINTARO YANAGI1, (Graduate Student Member, IEEE), REN TOGO2, (Member, IEEE), TAKAHIRO OGAWA3, (Senior Member, IEEE), AND MIKI HASEYAMA3, (Senior Member, IEEE)

1Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan
2Education and Research Center for Mathematical and Data Science, Hokkaido University, Sapporo 060-0812, Japan
3Faculty of Information Science and Technology, Division of Media and Network Technologies, Hokkaido University, Sapporo 060-0814, Japan

Corresponding author: Rintaro Yanagi (yanagi@lmd.ist.hokudai.ac.jp)

This work was supported in part by the MIC/SCOPE under Grant #181601001.

ABSTRACT A new approach that drastically improves cross-modal retrieval performance in vision and language (hereinafter referred to as “vision and language retrieval”) is proposed in this paper. Vision and language retrieval takes data of one modality as a query to retrieve relevant data of another modality, and it enables flexible retrieval across different modalities. Most of the existing methods learn optimal embeddings of visual and lingual information to a single common representation space. However, we argue that the forced embedding optimization results in loss of key information for sentences and images. In this paper, we propose an effective utilization of representation spaces in a simple but robust vision and language retrieval method. The proposed method makes use of multiple individual representation spaces through text-to-image and image-to-text models. Experimental results showed that the proposed approach enhances the performance of existing methods that embed visual and lingual information to a single common representation space.

INDEX TERMS Multimedia information retrieval, cross-modal retrieval, vision and language, text-to-image model, image-to-text model.

I. INTRODUCTION

Single-modal retrieval such as document retrieval from key-

word queries [1] and image retrieval from an image query [2] has been traditionally conducted. However, these approaches only perform retrieval on the same modality. Web services now contain not only lingual descriptions but also images and videos [3]. Therefore, a vision and language retrieval method that retrieves relevant data of one modality by uti-

lizing a query of another modality [4]–[10] is needed for many practical applications such as hot topic detection, personalized recommendation and multimodal retrieval [11], [12]. We can easily associate vision with language and vice versa; however, such an association is still difficult in computer vision because of the large gap between vision and language [13]–[15]. Actually, different modalities have diverse representations and distributions, and these heteroge-

neous characteristics make it difficult to directly measure the similarities of vision and language [16].

Popular vision and language retrieval methods are embed-

ding approaches [4]–[15]. They aim to find a joint mapping of instances from visual and lingual representation spaces to a common representation space so that related instances from the source representation space are mapped to nearby places in the target space. Specifically, the goal of these approaches is to jointly learn two mapping functions \( f: \mathcal{L} \rightarrow \mathcal{E} \) and \( g: \mathcal{V} \rightarrow \mathcal{E} \), where \( \mathcal{L} \) and \( \mathcal{V} \) are lingual and visual representation spaces, respectively, and \( \mathcal{E} \) is a common representation space. For example, in a text-to-image retrieval scenario, a lingual feature from a query in space \( \mathcal{L} \) and visual features from candidate images in space \( \mathcal{V} \) are projected into a learned common representation space \( \mathcal{E} \) that can compare the two different modalities. This embedding approach is currently one of the most popular approaches.

Although progress has been made in embedding approaches, there is still room for improvement, especially in the representation spaces. Most of the embedding approaches try to embed sentence and image information into a single representation space, \( \mathcal{E} \); however, sentence and image information might not be expressed as effectively in space \( \mathcal{E} \) as in
spaces $\mathcal{L}$ and $\mathcal{V}$ due to the forced embedding optimization. For example, it can be considered that word relationship information in a sentence (e.g., verbs and prepositions) and detailed texture information in an image (e.g., object surface and scenery) can be effectively expressed in spaces $\mathcal{L}$ and $\mathcal{V}$, respectively [17]–[19]. If we can make use of spaces $\mathcal{L}$ and $\mathcal{V}$ in addition to the learned common representation space $\mathcal{E}$, more robust vision and language retrieval would become feasible.

In this paper, we evaluate the effectiveness of representation spaces $\mathcal{L}$ and $\mathcal{V}$ and show that the retrieval performance of conventional embedding approaches can be enhanced by effectively utilizing spaces $\mathcal{L}$ and $\mathcal{V}$. As shown in Fig. 1, we further make use of spaces $\mathcal{L}$ and $\mathcal{V}$ to enhance the retrieval performance of embedding approaches that utilize only space $\mathcal{E}$. How do we utilize visual representation space and lingual representation space? We simply adopt two types of generative models: an image-to-text model $\mathcal{M}_{\mathcal{V} \rightarrow \mathcal{L}}$ [20]–[24] and a text-to-image model $\mathcal{M}_{\mathcal{L} \rightarrow \mathcal{V}}$ [25]–[28]. Generative models can generate one type of multimedia content from a different type of multimedia content by simulating the observed distribution. An image-to-text model learns the mapping of space $\mathcal{V}$ to space $\mathcal{L}$, and a text-to-image model learns the mapping of space $\mathcal{L}$ to space $\mathcal{V}$. In the proposed approach, candidate images and a query sentence are respectively translated into synthesized sentences and a synthesized image through $\mathcal{M}_{\mathcal{V} \rightarrow \mathcal{L}}$ and $\mathcal{M}_{\mathcal{L} \rightarrow \mathcal{V}}$. Then we calculate similarities between the query sentence and synthesized sentences in space $\mathcal{L}$ and similarities between the synthesized image and candidate images in space $\mathcal{V}$. By effectively using spaces $\mathcal{L}$ and $\mathcal{V}$ in addition to space $\mathcal{E}$, the retrieval performance of conventional embedding methods that only utilize space $\mathcal{E}$ can be enhanced.

The contributions of this paper are as follows. First, we verify the effectiveness of spaces $\mathcal{L}$ and $\mathcal{V}$ based on image-to-text and text-to-image models. Then we enhance the retrieval performance of conventional embedding methods that only utilize space $\mathcal{E}$ utilizing the text-to-image and image-to-text models.

We review related works in Sec. II. Then we describe the proposed vision and language retrieval method in Sec. III. Finally, we evaluate the effectiveness of each representation space in Sec. IV and the proposed approach in Sec. V.

II. RELATED WORKS

We briefly review some of the most relevant works on vision and language retrieval and explain the differences between these methods and our approach. Also, we describe studies on generative models that are utilized in our approach.

A. VISION AND LANGUAGE RETRIEVAL

The retrieval performance of vision and language retrieval methods is greatly dependent on the lingual and visual feature representation. Traditional methods based on Canonical Correlation Analysis (CCA) [29] or Topic-regression Multi-modal LDA (Tr-mm LDA) [30] utilized handcrafted feature extraction techniques such as Scale Invariant Feature Transform (SIFT). They can retrieve desired images by considering various factors such as rotation and scaling. However, such handcrafted features cannot represent the high-level semantics of multimedia data [31]. Feature extraction techniques that can represent high-level semantics have therefore been studied [32]. Recently, it has been shown that deep neural networks can learn the joint lingual and visual semantic feature representation and achieve better performance [18], [33], [34].

To embed such lingual and visual features into a common representation space, most of the vision and language methods aim to learn a common embedding representation space $\mathcal{E}$ in which the similarity between samples from different modalities can be measured directly. The methods learn a proper embedding function from different modalities utilizing similar or dissimilar cross-modal data pairs. Kiros et al. [13] utilized a Gated Recurrent Unit (GRU) [35] and a Convolutional Neural Network (CNN). In their method, the distances between matched text-image pairs are smaller than the distances between mismatched pairs. Faghri et al. [36] extended the method of [13] with the idea of hard negative mining. Their method focuses on maximizing violating negative pairs, and they reported improvement of convergence rates. Also, as an extension of the method of [13], Zhang and Lu [9] proposed an embedding method that calculates a loss function with KL divergence. By utilizing KL divergence, they realized discriminative text-image embeddings. The above-described methods can improve vision and language retrieval performance by focusing on the effective utilization of the common representation space $\mathcal{E}$. However, space $\mathcal{E}$ might not express the same sentence and image information as that expressed in spaces $\mathcal{L}$ and $\mathcal{V}$. The proposed approach further makes use of multiple representation spaces ($\mathcal{L}$ and $\mathcal{V}$) to enhance their retrieval performance.

On the other hand, there are methods that utilize multiple representation spaces for cross-modal retrieval. To utilize multiple representation spaces, Eisenschitz and Wolf [37] adopted bi-directional neural architectures, and Wei et al. [38] assumed...
jointly optimized a correlation space (between text and images) and a linear regression space (from one modal space, i.e., text or image, to the semantic space). They improve cross-modal retrieval performance by utilizing multiple representation spaces. Different from their approach, we propose an approach that additionally introduces multiple representation spaces (\(\mathcal{L}\) and \(\mathcal{V}\)) from generative models to several conventional embedding methods that only utilize space \(\mathcal{E}\) and enhances their retrieval performance.

**B. TEXT-TO-IMAGE AND IMAGE-TO-TEXT MODELS**

Image-to-text models [20]–[24] can generate a description or sentence representing information of an input image. Image-to-text generation tasks have been widely studied in the field of computer vision. Vinyals et al. [20] used a deep recurrent architecture to generate a sentence representing an input image. Xu et al. [21] extended this method based on an attention mechanism that can focus on relevant parts of the image and improve the performance of image-to-text generation.

Text-to-image generation has been one of the most attractive research topics in recent years. Text-to-image Generative Adversarial Networks (GANs) [25]–[28], [39]–[43] are the most basic and popular approaches. They are deep-learning methods and can generate an image representing information of an input sentence by alternately training their generators and discriminators. By conditioning on an input sentence, generators generate images, and discriminators discriminate whether input images are from a real data distribution or generators generate images, and discriminators discriminate. By continuously predicting the...0th words and an input image \(I\) is a training sample pair \((T_{\text{train}}, I_{\text{train}})\) in a training phase, to generate a sentence that expresses an input image, we train the RNN to maximize the following objective function:

\[
\sum_{(T_{\text{train}}, I_{\text{train}})} \log p(T_{\text{train}} | I_{\text{train}}),
\]

where \(T_{\text{train}}\) and \(I_{\text{train}}\) is a training sample pair \((T_{\text{train}} and I_{\text{train}} being a sentence and an image, respectively) in a training dataset and \(T_x\) \((x = 0, 1, \ldots, X; X being the number of words in a sentence \(T_{\text{train}}\)) is the \(x\)th word of sentence \(T_{\text{train}}\). By continuously predicting the \(x\)th word from \(x − 1, x − 2 \ldots 0\)th words and an input image \(I_{\text{train}}\) in the training phase, they can generate a sentence that contains information on word relationships.

Based on the trained image-to-text model \(M_{\text{TV}} \rightarrow \mathcal{L}\), we calculate image-to-text-based similarity in space \(\mathcal{L}\). We generate sentences \(T_n^{DB}\) that represent information on candidate images \(I_n^{DB}\). From an input sentence \(T^Q\) and generated sentences \(T_n^{DB}\), we calculate lingual features \(f_n^{LQ} \in \mathbb{R}^{D_L}\) and \(f_n^{DB} \in \mathbb{R}^{D_L}\), where \(D_L\) represents the dimension of the extracted lingual features. Then we calculate the cosine similarities \(s_n^{LQ}\) between \(f_n^{LQ}\) and \(f_n^{DB}\), where \(s_n^{LQ}\) indicate the similarities in the lingual representation space between the input sentence \(T^Q\) and the candidate images \(I_n^{DB}\).

**III. OUR VISION AND LANGUAGE RETRIEVAL APPROACH**

We present a simple but effective vision and language retrieval approach that utilizes spaces \(\mathcal{L}\) and \(\mathcal{V}\) in addition to space \(\mathcal{E}\) with focus on a text-to-image retrieval scenario. The goal of a text-to-image retrieval scenario is retrieval of a relevant image from only an input sentence as a query. This cross-modal retrieval setting does not have descriptions of candidate images. Our retrieval approach consists of two phases: similarity calculation and retrieval. In the first phase, we calculate three types of similarities between an input sentence and candidate images. Our approach calculates similarity \(s_n^{LQ}(n = 1, 2, \ldots, N; N being the number of candidate images)\) in lingual space \(\mathcal{L}\) through an image-to-text model \(M_{\text{TV}} \rightarrow \mathcal{L}\) and \(s_n^{DB}\) in visual space \(\mathcal{V}\) through a text-to-image model \(M_{\text{TV}} \rightarrow \mathcal{V}\) in addition to similarity \(s_n^{LQ}\) in space \(\mathcal{E}\). In the second phase, we simply integrate the three types of similarities as follows:

\[
s_n = \alpha s_n^{LQ} + \beta s_n^{DB} + \gamma s_n^{V},
\]

\[
\alpha + \beta + \gamma = 1,
\]

where \(\alpha\), \(\beta\), and \(\gamma\) are parameters that balance the percentages of similarity combinations. Finally, we rank candidate images in descending order of \(s_n\). By using spaces \(\mathcal{L}\) and \(\mathcal{V}\) with space \(\mathcal{E}\), the retrieval performance of vision and language retrieval could be enhanced. The details of each similarity calculation are shown below.

**A. SIMILARITY CALCULATION IN SPACE \(\mathcal{L}\)**

Theoretically, since any type of image-to-text model can be applied as \(M_{\text{TV}} \rightarrow \mathcal{L}\) in our approach, an explanation based on a basic image-to-text model [20] is presented. A basic image-to-text model consists of a single Recurrent Neural Network (RNN) [45]. In the training phase, to generate a sentence that expresses an input image, we train the RNN to maximize the following objective function:

\[
\sum_{(T_{\text{train}}, I_{\text{train}})} \log p(T_{\text{train}} | I_{\text{train}}),
\]

\[
\log p(T_{\text{train}} | I_{\text{train}}) = \sum_{x=0}^{X} \log p(T_{\text{train}}, I_{\text{train}}), T_0, \ldots, T_{x-1},
\]

where \(T_{\text{train}}\) and \(I_{\text{train}}\) is a training sample pair \((T_{\text{train}} and I_{\text{train}} being a sentence and an image, respectively) in a training dataset and \(T_x\) \((x = 0, 1, \ldots, X; X being the number of words in a sentence \(T_{\text{train}}\)) is the \(x\)th word of sentence \(T_{\text{train}}\). By continuously predicting the \(x\)th word from \(x − 1, x − 2 \ldots 0\)th words and an input image \(I_{\text{train}}\) in the training phase, they can generate a sentence that contains information on word relationships.

Based on the trained image-to-text model \(M_{\text{TV}} \rightarrow \mathcal{L}\), we calculate image-to-text-based similarity in space \(\mathcal{L}\). We generate sentences \(T_n^{DB}\) that represent information on candidate images \(I_n^{DB}\). From an input sentence \(T^Q\) and generated sentences \(T_n^{DB}\), we calculate lingual features \(f_n^{LQ} \in \mathbb{R}^{D_L}\) and \(f_n^{DB} \in \mathbb{R}^{D_L}\), where \(D_L\) represents the dimension of the extracted lingual features. Then we calculate the cosine similarities \(s_n^{LQ}\) between \(f_n^{LQ}\) and \(f_n^{DB}\), where \(s_n^{LQ}\) indicate the similarities in the lingual representation space between the input sentence \(T^Q\) and the candidate images \(I_n^{DB}\).
identify networks (discriminators) based on the loss function as follows:

\[
\min_G \max_D L_G - L_D, \tag{5}
\]

\[
L_G = -\frac{1}{2} \mathbb{E}_{n \sim p_G} [\log(D(I^n))] - \frac{1}{2} \mathbb{E}_{n \sim p_G} [\log(D(I^n, T^n))], \tag{6}
\]

\[
L_D = -\frac{1}{2} \mathbb{E}_{n \sim p_R} [\log(D(I^n))] + \frac{1}{2} \mathbb{E}_{n \sim p_G} [\log(1 - D(I^n))] - \frac{1}{2} \mathbb{E}_{n \sim p_R} [\log(D(I^n, T^n))] + \frac{1}{2} \mathbb{E}_{n \sim p_G} [\log(1 - D(I^n, T^n))]. \tag{7}
\]

where \(I^n\) is a generated image from the generated data distribution \(p_G\), \(\hat{I}^R\) is an image from real data distribution \(p_R\), and \(T^n\) is an input sentence. In Eq. (6), the first term determines whether \(I^n\) is from \(p_R\) or \(p_G\), while the second term determines whether \(I^n\) expresses \(T^n\) or not. Also, in Eq. (7), the first and second terms respectively determine whether \(I^n\) and \(I^n\) are from \(p_R\) or \(p_G\), while the third and fourth terms determine whether \(I^n\) and \(I^n\) express \(T^n\) or not, respectively. Namely, by conditioning on \(T^n\), generators learn these parameters to generate \(I^n\) that discriminators cannot discriminate from \(\hat{I}^R\), while discriminators learn these parameters to discriminate \(\hat{I}^R\) and \(I^n\). This adversarial training enables the model \(M_{E \rightarrow V}\) to generate a real image representing visual information (such as color and texture) of an input sentence.

Based on the trained text-to-image model \(M_{E \rightarrow V}\), we generate an image \(I^Q\) that represents information of an input sentence \(T^Q\). Next, we calculate visual features \(f^Q \in \mathbb{R}^{D_V}\) and \(f^DB_n \in \mathbb{R}^{D_V}\) from \(I^Q\) and \(I^DB_n\), where \(D_V\) represents the dimension of the extracted visual features. Then we simply calculate the cosine similarities \(s_n^V\) between \(f^Q\) and \(f^DB_n\). These values indicate the similarity in a visual representation space between the input sentence \(T^Q\) and the candidate images \(I^DB_n\).

\section*{C. SIMILARITY CALCULATION IN SPACE \(\mathcal{E}\)}

As stated before, the proposed retrieval method corresponds to an enhancing method that improves the performance of arbitrary conventional embedding methods. Therefore, the selection of these embedding methods is not the main contribution of this paper. We can use any type of embedding method and we only explain the calculation of similarity in the embedding space \(\mathcal{E}\). We will verify the effectiveness of the proposed approach by using several embedding methods in Sec. \(\mathcal{V}\).

To calculate similarities \(s_n^E\), firstly, we calculate a lingual feature \(f^LE_0\) from a query sentence \(T^Q\) and visual features \(f^VDB_n\) from the candidate images \(I^DB_n\). Then we embed \(f^LE_0\) and \(f^VDB_n\) into a common representation space based on the trained embedding models. We define embedded lingual features as \(f^L_{E_0} \in \mathbb{R}^{D_E}\) and embedded visual features as \(f^V_{DB_n} \in \mathbb{R}^{D_E}\), where \(D_E\) represents the dimension of the embedded features. Then we simply calculate the cosine similarities \(s_n^E\) between \(f^L_{E_0}\) and \(f^V_{DB_n}\), where \(s_n^E\) indicates the similarities in the common representation space between the input sentence \(T^Q\) and the candidate images \(I^DB_n\).

\section*{IV. VERIFYING THE EFFECTIVENESS OF LINGUAL AND VISUAL SPACES FOR RETRIEVAL}

In this section, we first show experimental results to verify the effectiveness of using each of the spaces \(\mathcal{E}\) and \(\mathcal{V}\). More robust vision and language retrieval that further focuses on sentence and image information will be expected by verifying the effectiveness of using spaces \(\mathcal{E}\) and \(\mathcal{V}\).

\section*{A. EXPERIMENTAL SETUP}

We used the Microsoft Common Objects in Context (MSCOCO) [50] dataset that consists of daily scene images and their descriptions in the experiments following the conventional vision and language methods shown in Table 1. The MSCOCO dataset contains 82,783 training images and 40,504 validation images, each of which is associated with five descriptions. Based on the data split provided by [13], we utilized 5,000 captions as queries, and for each caption, we retrieved a relevant image from the 5,000 test images.

As the text-to-image model and image-to-text model, we utilized the state-of-the-art methods Dynamic Memory GAN [28] and Show attend and tell method [21], respectively. Also, output of the BERT model provided by Reimers and Gurevych [18] was used as features of the query and generated sentences (\(f^E_0\) and \(f^V_{DB_n}\)), and output of the DenseNet-121 model [34] was used as features of the generated and candidate images (\(f^V_{DB_n}\) and \(f^V_{DB_n}\)). As embedding methods, we adopted state-of-the-art methods shown in Table 1. It should be noted that all of the methods are implemented on the basis of open source codes provided by each author.

\section*{B. VERIFYING THE EFFECTIVENESS OF LINGUAL SPACE}

The relationships between words provide vital information for discriminating similar images. We can describe these relationships with language well, and lingual feature extractors are trained to consider sentence structures [18]. We assume that space \(\mathcal{E}\) focuses on representing word relationships in a sentence and thus can express this information in a sentence better than conventional embedding methods using space \(\mathcal{E}\) can. If this assumption is satisfied, the retrieval performance becomes better. Namely, by cooperatively using space \(\mathcal{E}\) with the conventional embedding methods using space \(\mathcal{E}\), the performance of conventional embedding methods can be enhanced.

At first, we removed the verbs and prepositions of the input sentences and randomly shuffled words in the sentences in which they had been removed. We defined these sentences as changed sentences. Then we inputted these...
TABLE 1. Methods used in preliminary experiments described in Sec. IV and in retrieval performance evaluation described in Sec. V. “Description” shows an overview of each method.

| Method   | Description                                                                 |
|----------|-----------------------------------------------------------------------------|
| BL₇      | Using only similarity on a lingual representation space 7                    |
| BLᵥ      | Using only similarity on a visual representation space v                    |
| Kiros’14 [13] | Using LSTM and CNN to compare visual and sentence features by embedding them into a common semantic space E   |
| Vendrov’16 [14] | Taking the order relationship between words into consideration in addition to the mechanism of [13] |
| Faghri’17 [36] | Reducing the number of negative samples between a query and an objective sample in addition to the mechanism of [13] |
| Liu’17 [10] | Adopting residual and recurrent modules to unify images and sentences into a more discriminative space     |
| Zheng’17 [46] | Constructing a dual path CNN that can extract lingual and visual features directly from data |
| Zhang’18 [9] | Adding two loss functions that introduce the cross-modal feature projection operation for learning discriminative text-image embeddings |
| Gu’18 [47] | Incorporating generative processes into the cross-modal feature embedding to learn the abstract and local features |
| Huang’18 [48] | Using multi-regional multi-label CNN that can simultaneously predict multiple concepts to learn the semantic concepts of images |
| Song’19 [8] | Considering ambiguous instances with multiple meanings and partial cross-modal associations |
| Ji’19 [49] | Constructing a saliency-guided attention network that collaboratively performs visual and textual attentions to model the fine-grained interplay between both modalities |

---

FIGURE 2. Mean and median ranks obtained by methods based on normal sentences (red bars) and methods based on sentences in which verbs and prepositions had been removed and words had been shuffled (blue bars). In mean and median ranks, lower values indicate better performance.

---

changed sentences into various methods. Finally, we compared retrieval performances of methods utilizing normal sentences with retrieval performance of methods utilizing changed sentences.

Figure 2 shows the mean and median ranks of the retrieval results for the MSCOCO dataset. Red bars show the results based on normal sentences, and blue bars show the results based on sentences in which verbs and prepositions had been removed and words had been shuffled. While the values of blue bars are similar to those of red bars in BLᵥ and other conventional methods using space E, the value of the blue bar is smaller than that of the red bar in BL₇. These results imply that BLᵥ and other conventional methods using space E mainly focus on the objectives in sentences to retrieve desired images. On the other hand, the results indicate that verbs, prepositions and word orders are important for space BL₇. The results suggest that the performance of conventional embedding methods using space E is enhanced by cooperatively using space L since the conventional embedding methods further focus on word relationship information.

C. VERIFYING THE EFFECTIVENESS OF VISUAL SPACE

Texture information is vital information for retrieving a target image. Recent papers [19] have shown that CNN-based visual feature extractors trained on ImageNet focus on texture information. We assume that space V focuses on representing texture information and thus can express texture information better than conventional embedding methods using space E.
can. We compared the effect when we removed texture information of the images in the conventional embedding methods based on the experiments in [19]. If space $V$ can express texture information better than other embedding methods using space $E$, the performance of conventional embedding methods using space $E$ could be enhanced by cooperatively using space $V$.

At first, we converted candidate images $C_n$ to an edge-based representation using a canny edge extractor following a conventional method [19] and defined them as changed candidate images. Then we retrieved changed candidate images using the methods shown in Table 1. Finally, we compared the retrieval performance of methods using normal candidate images with the retrieval performance of methods using changed candidate images.

Figure 3 shows the retrieval results of mean and median ranks for the MSCOCO dataset. The yellow bars show the results based on normal candidate images, and the green bars show the results based on candidate images converted to an edge-based representation. A comparison of the differences between the values of the yellow bars and green bars shows that the difference for BL$_V$ using space $V$ (1,696 in mean rank and 1,498 in median rank) is considerably larger than the differences for other methods using space $E$ (in other methods, average values of 828 in mean rank and 538 in median rank). Images converted to an edge-based representation have little or no texture information [19]. We therefore verified that texture information is more important for space $V$ than other representation spaces. This implies that the performance of conventional embedding methods can be enhanced by cooperatively using space $V$ since the conventional embedding methods further focus on texture information.

V. RETRIEVAL PERFORMANCE EVALUATION

In this section, we evaluate the effectiveness of the proposed approach by comparing it with some state-of-the-art retrieval methods.

A. QUANTITATIVE EVALUATION ON MSCOCO DATASET

Following evaluations of the conventional embedding methods, we report results using mean rank, median rank and Recall@$k$ as follows:

$$Recall@k = \frac{r_k}{N} \quad (k = 1, 10, 100),$$

where $r_k$ is the number of correctly retrieved images in the top-$k$ retrieval results, and $N (= 5,000)$ is the number of input sentences. Furthermore, in our framework, we simply set $\alpha = \beta = \gamma$ in Eq. (1) as $\alpha = \beta = \gamma = \frac{1}{3}$. Note that we utilized the same experimental setup as that described in Subsec. IV-A. In addition, we used the comparative methods in Table 1 that are described in Subsec. IV-A.

Tables 2 shows Recall@$k$, mean and median ranks for the MSCOCO dataset, and Fig. 4 shows “examples of retrieval results by the proposed approach with the method of Ji ’19” and “examples of retrieval results by the method of Ji ’19”. Note that the proposed approach can be applied to all of the existing embedding methods since its proposition is the use of a combination of the embedding space and the original visual and lingual spaces as shown in Table 2. In Fig. 4, we show the results obtained by the method of Ji ’19 since this method shows the best results in Table 2.

As shown in Table 2, the proposed approach using spaces $E$ and $V$ in addition to space $\hat{E}$ outperforms the conventional embedding methods. Specifically, in each method, we can see that the proposed approach drastically enhances the mean and median ranks compared to the state-of-the-art methods.
This means that the proposed approach is effective for various conventional embedding methods to improve the retrieval performance of vision and language retrieval. Furthermore, in each of the conventional methods, we can see that retrieval performance is better when both space $L$ and space $V$ are used than when only space $L$ or space $V$ is used. It can therefore be assumed that both spaces contribute to the performance enhancement. Also, from Fig. 4, we qualitatively confirmed that “the proposed approach with the method of Ji ‘19” can realize retrieval that considers objects in images and these relationships comparing with the “the method of Ji ‘19”.

C. PARAMETER SEARCH AND ANALYSIS OF RANKS IN EACH SPACE
In our proposed approach, we set parameters $\alpha$, $\beta$ and $\gamma$ that balance the percentages of similarity combinations.
We simply set each parameter as $\alpha = \beta = \gamma = \frac{1}{3}$ in the above experiments; however, there may be better parameter settings. We can further boost the performance of the proposed approach by adjusting each parameter. By searching for the best parameters, we can observe the importance of the similarities $s^L_n, s^E_n$, and $s^V_n$. Thus, we show the results of a parameter search.

In this experiment, we simply conducted a grid search in the MSCOCO dataset. We changed each parameter by 0.1 and observed the mean and median ranks at each of the parameter settings. Here, we utilized the method of Ji '19 to calculate $s^E_n$ since it outputs the best results as shown in Subsec. V-B.

Figures 5 and 6 show the results of mean and median ranks, respectively. As shown in Fig. 5, the best mean rank is obtained around the settings $\alpha = 0.3, \beta = 0.4$ and $\gamma = 0.3$. Also, as shown in Fig. 6, the best median rank is obtained around the settings $\alpha = 0.3, \beta = 0.5$ and $\gamma = 0.2$. These results mean that the similarity $s^V_n$ is the most important information; however, the similarities $s^E_n$ and $s^L_n$ are also important information, and we can enhance the performance for vision and language retrieval by using spaces $L, E$, and $V$ simultaneously. Also, we observed the rank of each sample utilizing only $L$ for the MSCOCO dataset, $E$ or $V$ in Fig. 7. Each axis shows the rank when using only each space. From Fig. 7, we can see that samples that can be retrieved in the lower rank are different in each space. This result implies that the proposed approach improves the retrieval performance of vision and language retrieval by

\[ \text{TABLE 2. Recall@k (R@k), mean and median ranks on the MSCOCO dataset. Higher values indicate better performance in Recall@k, and lower values indicate better performance in mean and median ranks.} \]

| Method | R@1 | R@10 | R@100 | Mean | Median |
|--------|-----|------|-------|------|-------|
| BL_L   | 0.11| 0.27 | 0.67  | 154.36| 29   |
| BL_V   | 0.12| 0.28 | 0.56  | 165.17| 28   |
| BL_L+BL_V | 0.16| 0.39 | 0.70  | 109.49| 27   |
| Kiro* 14+BL_L | 0.12| 0.34 | 0.83  | 166.86| 30   |
| Kiro* 14+BL_V | 0.17| 0.41 | 0.82  | 126.46| 18   |
| Kiro* 14+BL_L+BL_V (Ours) | 0.22| 0.48 | 0.73  | 77.56 | 16   |
| Vendrov 15 | 0.13| 0.34 | 0.75  | 155.86| 28   |
| Vendrov 16+BL_L | 0.17| 0.35 | 0.79  | 155.33| 24   |
| Vendrov 16+BL_V | 0.13| 0.32 | 0.77  | 110.09| 20   |
| Vendrov 16+BL_L+BL_V (Ours) | 0.23| 0.49 | 0.85  | 75.43 | 13   |
| Fuglir '17 | 0.15| 0.38 | 0.80  | 151.95| 25   |
| Fuglir '17+BL_L | 0.16| 0.36 | 0.80  | 146.77| 21   |
| Fuglir '17+BL_V | 0.15| 0.36 | 0.80  | 140.95| 19   |
| Fuglir '17+BL_L+BL_V (Ours) | 0.19| 0.45 | 0.89  | 79.44 | 15   |
| Liu '17 | 0.16| 0.40 | 0.78  | 153.34| 25   |
| Liu '17+BL_L | 0.18| 0.40 | 0.78  | 148.55| 24   |
| Liu '17+BL_V | 0.17| 0.40 | 0.81  | 100.04| 23   |
| Liu '17+BL_L+BL_V (Ours) | 0.23| 0.48 | 0.87  | 69.66 | 13   |
| Zheng '17 | 0.18| 0.42 | 0.77  | 147.41| 24   |
| Zheng '17+BL_L | 0.21| 0.45 | 0.79  | 132.36| 21   |
| Zheng '17+BL_V | 0.18| 0.43 | 0.80  | 127.24| 21   |
| Zheng '17+BL_L+BL_V (Ours) | 0.22| 0.50 | 0.87  | 89.55 | 12   |
| Zhang '18 | 0.19| 0.41 | 0.80  | 142.84| 25   |
| Zhang '18+BL_L | 0.21| 0.44 | 0.82  | 122.32| 24   |
| Zhang '18+BL_V | 0.18| 0.51 | 0.86  | 109.98| 12   |
| Zhang '18+BL_L+BL_V (Ours) | 0.25| 0.51 | 0.86  | 87.33 | 12   |
| Gu '18 | 0.20| 0.43 | 0.77  | 140.55| 22   |
| Gu '18+BL_L | 0.21| 0.45 | 0.79  | 135.32| 21   |
| Gu '18+BL_V | 0.21| 0.44 | 0.80  | 130.96| 21   |
| Gu '18+BL_L+BL_V (Ours) | 0.23| 0.47 | 0.80  | 120.53| 19   |
| Huang '18 | 0.21| 0.46 | 0.85  | 126.67| 14   |
| Huang '18+BL_L | 0.22| 0.46 | 0.87  | 100.37| 13   |
| Huang '18+BL_V | 0.23| 0.54 | 0.86  | 111.79| 13   |
| Huang '18+BL_L+BL_V (Ours) | 0.23| 0.48 | 0.91  | 79.66 | 12   |
| Song '19 | 0.21| 0.51 | 0.85  | 112.02| 10   |
| Song '19+BL_L | 0.26| 0.52 | 0.89  | 70.55 | 10   |
| Song '19+BL_V | 0.21| 0.49 | 0.87  | 55.60 | 13   |
| Song '19+BL_L+BL_V (Ours) | 0.26| 0.56 | 0.94  | 49.38 | 9    |
| Ji '19 | 0.30| 0.52 | 0.88  | 90.99 | 8    |
| Ji '19+BL_L | 0.31| 0.54 | 0.88  | 67.55 | 8    |
| Ji '19+BL_V | 0.31| 0.57 | 0.92  | 63.33 | 7    |
| Ji '19+BL_L+BL_V (Ours) | 0.33| 0.57 | 0.94  | 44.54 | 7    |

\[ \text{TABLE 3. Recall@k (R@k), mean and median ranks on Flickr30k dataset. Higher values indicate better performance in Recall@k, and lower values indicate better performance in mean and median ranks.} \]

| Method | R@1 | R@10 | R@100 | Mean | Median |
|--------|-----|------|-------|------|-------|
| BL_L   | 0.21| 0.37 | 0.76  | 123.43| 9    |
| BL_V   | 0.22| 0.60 | 0.98  | 122.67| 8    |
| BL_L+BL_V | 0.27| 0.62 | 0.99  | 107.79| 9    |
| Kiro* 14+BL_L | 0.22| 0.59 | 0.99  | 122.33| 8    |
| Kiro* 14+BL_V | 0.25| 0.66 | 1.00  | 99.67 | 7    |
| Kiro* 14+BL_L+BL_V (Ours) | 0.26| 0.64 | 0.99  | 109.66| 7    |
| Vendrov 16 | 0.23| 0.63 | 1.00  | 98.72 | 7    |
| Vendrov 16+BL_L | 0.32| 0.67 | 1.00  | 90.11 | 6    |
| Vendrov 16+BL_V | 0.31| 0.70 | 1.00  | 86.49 | 7    |
| Vendrov 16+BL_L+BL_V (Ours) | 0.34| 0.79 | 1.00  | 78.78| 6    |

\[ \text{FIGURE 5. Mean ranks (lower values indicate better performance) obtained by changing the parameters } \alpha, \beta \text{ and } \gamma \text{ for the MSCOCO dataset. The vertical line shows parameter } \beta, \text{ and the horizontal line shows parameter } \gamma. \text{ Parameter } \gamma \text{ is calculated by } \gamma = 1 - \alpha + \beta. \]
we will attempt to consider adaptive fusion of similarities. In a future work, we will examine this in our future work.

VI. CONCLUSION

In this paper, we have proposed an approach that utilizes multiple representation spaces from generative models for vision and language retrieval. Experimental results verified that the proposed approach enhances the performance of conventional embedding methods that only utilize space $E$. Furthermore, by analyzing each component, the advantage of each representation space was shown. In a future work, we will attempt to consider adaptive fusion of similarities.

REFERENCES

[1] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt, “Survey of temporal information retrieval and related applications,” ACM Comput. Surveys, vol. 47, no. 2, pp. 1–15, 2015.

[2] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, “A survey of content-based image retrieval with high-level semantics,” Pattern Recognit., vol. 40, no. 1, pp. 282–282, 2007.

[3] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” Mobile Netw. Appl., vol. 19, no. 2, pp. 171–209, Apr. 2014.
T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, Y. Huang, Q. Wu, C. Song, and L. Wang, “Learning semantic concepts,” in *Encyclopedia of Statistics in Behavioral Science*. Hoboken, NJ, USA: Wiley, 2005.

D. Puthullydihy, H. T. Attias, and S. S. Nagarajan, “Topic regression multi-modal latent Dirichlet allocation for image annotation,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3408–3415.

L. Zheng, Y. Yang, and Q. Tian, “SIFT meets CNN: A decade survey of instance retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1224–1244, May 2018.

Y. Zhang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, “A multimedia retrieval framework based on semi-supervised ranking and relevance feedback,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 723–742, Apr. 2012.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. IEEE Int. Conf. Learn. Represent.*, Apr. 2015, pp. 1–9.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), Jul. 2017, pp. 4700–4708.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” 2014, arXiv:1412.3555 [Online]. Available: http://arxiv.org/abs/1412.3555

F. Faghri, D. J. Fleet, J. Ryan Kiros, and S. Fidler, “VSE++: Improving visual-semantic embeddings with hard negatives,” 2017, arXiv:1707.05612 [Online]. Available: http://arxiv.org/abs/1707.05612

A. Eisenschitz and L. Wolf, “Linking image and text with 2-way nets,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), Jul. 2017, pp. 4601–4611.

Y. Wei, Y. Zhao, Z. Zhu, S. Wei, Y. Xiao, J. Feng, and S. Yan, “Modality-dependent cross-media retrieval,” *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, pp. 1–13, Jul. 2016.

N. Bodla, G. Hua, and R. Chellappa, “Semi-supervised FusedGAN for conditional image generation,” 2018, arXiv:1801.05551 [Online]. Available: http://arxiv.org/abs/1801.05551

S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, “Learning what and where to draw,” in *Proc. Adv. Neur. Inf. Process. Syst.*, 2016, pp. 217–225.

Z. Zhang, Y. Xie, and L. Yang, “Photographic text-to-image synthesis with a hierarchically-structured adversarial network,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6199–6208.

A. Dash, J. Cristian Borges Gamboa, S. Ahmed, M. Liwicki, and M. Zeshan Afzal, “TAC–GAN—text conditioned auxiliary classifier generative adversarial network,” 2017, arXiv:1703.06412 [Online]. Available: http://arxiv.org/abs/1703.06412

H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, “StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vis.* (ICCV), Oct. 2017, pp. 5907–5915.

T. Qiao, J. Zhang, D. Xu, and D. Tao, “MirrorGAN: Learning text-to-image generation by redescription,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (CVPR), Jun. 2019, pp. 1505–1514.

S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

Z. Zheng, L. Zheng, M. Garrett, Y. Yang, and Y.-D. Shen, “Dual-path convolutional image-text embeddings with instance loss,” *ACM Trans. Multimedia Comput., Commun., Appl.*, 2020, doi: 10.1145/3383184.

J. Gu, J. Cai, S. Joty, L. Niu, and G. Wang, “Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7191–7199.

Y. Huang, Q. Wu, C. Song, and L. Wang, “Learning semantic concepts and order for image and sentence matching,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6163–6171.

Z. Ji, H. Wang, J. Han, and Y. Pang, “Saliency-guided attention network for image-sentence matching,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (ICCV), Oct. 2019, pp. 5754–5763.

T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in *Proc. IEEE Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 740–755.

B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models,” in *Proc. IEEE Int. Conf. Comput. Vis.* (ICCV), Dec. 2015, pp. 2641–2649.

A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* (CVPR), Jun. 2015, pp. 3128–3137.