Grounded Vocabulary for Image Retrieval Using a Modified Multi-Generator Generative Adversarial Network

KUEKYENG KIM, CHANJUN PARK, JAEHYUNG SEO, AND HEUISEOK LIM, (Member, IEEE)
Department of Computer Science and Engineering, Korea University, Seoul 02841, South Korea
Corresponding author: Heuiseok Lim (limhseok@korea.ac.kr)

This work was supported in part by the Ministry of Science and ICT, South Korea, through the Information Technology Research Center Support Program supervised by the Institute for Information and Communications Technology Planning and Evaluation under Grant IITP-2018-0-01405; in part by the Korean Government (A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques) under Grant 2020-0-00368; and in part by the Ministry of Science and ICT, South Korea, through the ICT Creative Conرسل Program supervised by the Institute for Information and Communications Technology Planning and Evaluation under Grant IITP-2021-2020-0-01819.

ABSTRACT
With the recent increase in requirement of both natural-language and visual information, the demand for research on seamless multi-modal processing for effective retrieval of these types of information has increased. However, because of the unstructured nature of images, it is difficult to retrieve images that accurately represent the input text. In this study, we utilized an augmented version of a multi-generator generative adversarial network that uses BERT embeddings and attention maps as input to enable grounded vocabulary for visual representations. We compared the performance of our proposed model with those of other state-of-the-art text input-based image retrieval methods on the MSCOCO and Flikr30K datasets, and the results showed the potential of our proposed method. Even with limited vocabulary, our proposed model was comparable to other state-of-the-art performances on R@10 or even exceed them in R@1. Moreover, we revealed the unique properties of our method by demonstrating how it could perform successfully even when using more descriptive text or short sentences as input.

INDEX TERMS
Artificial intelligence, artificial neural network, computer vision, image processing, search methods.

I. INTRODUCTION
In recent times, the demand for seamless multi-modal processing has increased notably. The significance of multi-modal processing has risen at an astounding rate as recent advances in both natural-language and image processing have blurred the line between the two fields. For further advancement of artificial intelligence, the two fields must merge. However, one of the biggest obstacles to this is the inability to appropriately retrieve related information across the platforms. While text retrieval has advanced enough to rival that performed by humans, and even outperform in some cases, this is not the case for image retrieval. Image retrieval based on input images faces multiple obstacles owing to the unstructured nature of images and because the low-level features extracted directly from a given input image cannot be interpreted or used to determine what the user wants to search for, given just the input image [1]. Because keywords or text can be used to easily express semantic concepts or be interpreted to determine what the user wishes to search for, research on image retrieval has naturally steered toward that based on text input. While most early approaches simply tagged images with keywords or labels, more recent approaches have started to focus on directly accessing the content of the image with the given text input.

Currently, research on multi-modal processing has become more popular as both fields of processing have advanced to a point where they can rival even humans on certain tasks. However, when it comes to real-life tasks, it is not difficult to find instances where humans still outperform many applications [2]. Thus, for further advancement, recent research has started to use parallel inputs from different mediums to gather more information and context that cannot be gained...
by dealing with only one medium; hence, this is termed as “multi-modal processing” [3].

This is also the basis of the proposed model; i.e., it aims to parallelly align and compare natural-language and visual information from a single equalized perspective, allowing the capture and retrieval of images with greater accuracy based on subtle details through “seamless multi-modal processing.”

Herein, we introduce a new image-retrieval approach that utilizes multi-modal processing. By utilizing an augmented version of the multi-generator generative adversarial network (MGAN) [4] along with Bidirectional Encoder Representations from Transformers (BERT) embeddings [5] and attention maps as input, we could successfully construct a library of vocabularies that we applied for image retrieval. The model could be used to generate a visual representation that matches not only the object of the given input text but also the details of the object.

Through this research, we offer a new and unique image-retrieval approach. We aim to combine two types of information, namely text and image information, into a singular platform that can be used to retrieve images with high accuracy even when using descriptive text or short sentences as input. Additionally, because our model operates on learning and saving individual vocabularies, libraries of grounded vocabulary can be created, which can be applied not only to identify objects and details but also to a variety of other computer vision tasks.

In Section 2, we discuss the works related to our approach, including previous attempts in the fields. Next, Section 3 presents details of the methodology employed by our proposed modified MGAN model (MMG-GAN), its overall processes, and the activation and training of each generator. Section 4 presents the way the dataset was augmented and specifics of the training conditions. It also presents comparisons between the performances of our model and other recent image retrieval methods and an evaluation of the changes in performance of our model depending on the number of details that it considers. This is followed by an analysis about the evaluations and results, highlighting the proposed model’s strengths and some shortcomings. Finally, in Section 5, we present the conclusions of our study, propose methods to further improve our model, and suggest other directions to further our research.

II. RELATED WORK
A. IMAGE RETRIEVAL
Image retrieval using text and cross-modal retrieval have made significant progress since 2012. With the successful implementation of AlexNet [6], deep convolutional neural networks were introduced to the field. An increase in the demand for multi-modal tasks, such as visual question and answering, image or video captioning, phrase localization, knowledge transfer, and text-to-image generation tasks, the demand for cross-modal retrieval also increased. One of the most common ways of performing cross-modal retrieval is image retrieval through text input.

Recent image-text retrieval research focused on cross-modal processes based on natural language processing and vision leading to one of the most prominent state-of-the-art methods, SGRAF by Diao et al. [7]. By learning vector-based similarity, the local and global alignments were characterized, thus encoding image-text and region-word alignments, followed by the use of the SGR and SAF modules to capture meaningful relationships and alignments.

Another remarkable approach was TERAN by Messina et al. [8], where they undertook the task of cross-modal retrieval through image-sentence matching based on word-region alignments using supervision only at the global image-sentence level. Li et al. [9], with VSRN, also attempted to address cross-modal retrieval based on graph convolutional networks (GCNs) to enhance visual representations with local region relationship reasoning and global semantic reasoning. Lee et al. [10] introduced a stacked cross attention model with latent semantics, giving useful alignments between image regions and text words.

Even though these approaches deliver impressive performances, they lack the ability to capture specific details in many instances. Because these methods typically capture only the relationships between the objects in the image and texts and then compare those for image retrieval, they usually omit the detail or status of the object described within lengthy texts. In this study, we attempted to overcome this obstacle by combining representations of natural language vocabulary and visual data into a singular representation to enable seamless multi-modal processing, similar to that of grounded vocabulary. By unifying the two different types of information into a unified representation, swapping between the two types of information achieves significant improvements. Thus, we use a feature representation of images as the main form of representation, along with sentence attention maps, to construct structures within those representations. Consequently, details on the abstract natural language representations are reinforced, and the complexity of the unstructured visual representations is simplified to fairly balance the processing of both media.

B. GENERATIVE ADVERSARIAL NETWORK
In the field of deep learning, in general, the generative adversarial network (GAN) introduced by Goodfellow et al. [11] was a revolutionary model that successfully found its way into many types of research and applications. Designing a generator and a discriminator model to be locked on a zero-sum non-cooperative game allowed the two models to train as if it was supervised, despite the fact that it was actually unsupervised as the two models attempted to minmax each other out. While the generator model trained to generate convincing examples, the discriminator trained to distinguish real examples from the generated ones, allowing the two sub-modules to train. Approaching unsupervised training as a supervised one allowed both modules to be trained beyond the abilities of a fixed supervision model on several domains.
However, the approach itself was far from perfect, and since then, multiple types of variations have emerged. One of those variations was the MGAN by Szegedy et al. [4]. This particular variation attempted to overcome the mode-collapse problem using multiple generators to create multiple samples. Only one of the generated samples was randomly selected for determination by the discriminator, along with its classifier specifying which generator it came from, thereby preventing the generator to collapse to a single output at all times. This is a very simple yet effective approach to address the problem. This is the reason we chose this variation as the basis of our model. Combined with the reasoning behind conditional GANs proposed by Mirza et al. [12], we constructed individual generators to train on specific individual vocabularies related to objects and specific details. During this process, it was crucial to prevent vocabularies that converged from collapsing into each other. For example, both “Husband” and “Wife” are meaningful words that often go together; using a single generator to process these words can easily cause the two instances to collapse into each other during training. The possible entanglement of two vocabularies can cause the output to be warped. However, we overcame this problem using multiple generators to process individual vocabularies, thus avoiding the “vocabulary collapsing problem” and allowing each visual representation to act independently.

C. MSCOCO DATA

For the majority of our research, we chose to use the Microsoft Common Objects in Context (MSCOCO) dataset [13] to train and evaluate our proposed method. During this research, the model needed to be trained on a dataset containing a variety of objects in various contexts; it needed local data on each of the objects within each image while having paired text aligning with the situation. The MSCOCO dataset [13] is one of the few datasets that fits this criteria well because it is one of the largest datasets containing various types of images, captions, and precise 2D localization of objects. It is the best option when building an image retrieval system based on grounded vocabulary since the dataset itself contains 123,287 images, and each image is paired with five text annotations.

D. BERT

Equally important to our model was the way we fed the input texts into our module. Compared to visual information, linguistic information is fairly more structured and easier to interpret but can also be much more abstract. Therefore, the input must contain not only the given vocabulary itself but also contextual information about its surroundings. While there were numerous candidates to choose from, BERT by Hoang et al. [5] was chosen because of its outstanding ability to interpret contexts. On its initial release, BERT was revolutionary owing to its design to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right contexts in all layers. Consequently, this considerably improved the way language was represented in the model, which was well suited to the needs of our proposed method.

III. PROPOSED METHOD

The proposed model can be simply summarized as a modified version of the MG-GAN model. Unlike the original version from Hoang et al. [4], our proposed model does not use all the generators at once but a selected number based on the text input.

Because each generator, assigned to specific vocabularies, is trained separately, there are three major advantages. First, each generator only deals with its target vocabulary and other
closely related vocabularies, which reduces the level of stress on each generator as it only needs to be able to generate an accurate representation of one vocabulary concept. Second, since each generator works independently and is called from the library of generators only when an associated vocabulary is called, the generated representations can act individually. This reduces the chance of each representation being affected by a totally unrelated factor. Finally, our proposed model excels at finding specific images as it not only searches for the object but also determines the state of each object.

Moreover, we did not use a raw image to represent itself. Rather, we used a pretrained inception-v4 convolutional neural network (CNN) [14] to extract visual representations from the image. The reason for using visual extractions instead of actual images is similar to why specific vocabularies are assigned to individual generators: to reduce unnecessary complexity introduced by the low-level features of the image and increase coherency by recognizing semantic concepts from the image. In real life, the same text can generate numerous amounts of images; therefore, we utilized a format that is more consistent in recognizing and generating features. Inception-v4 CNN [14] has previously been used to generate text in image-captioning tasks [15]; it was applied here to reduce the complexity of the low-level features of the image while retaining the major features, making the search more consistent and easier.

For example, as seen in Fig. 1, given the input text “A Running Dog,” the model will instantly call upon the two generators assigned to “dog” and “run,” respectively. While the “dog” generator simply generates a representation of a “dog,” the generator assigned to “run” generates a series of features that can be added to an object representation to represent the detail of “running.” This is possible because the given input is not just text but BERT embeddings that consider the surrounding vocabulary within the input text. Therefore, the generated representation of “dog” is reinforced with the detail of “running.” The level and intensity of reinforcement are determined by the BERT attention map based on the given input text. The attention map also determines the detail that needs more focus. Each visual representation of the objects in the input text that is generated will be used by the discriminator to locate the best matching image. Consequently, the image that contains the most matching objects and also aligns best with the input text’s description will be ranked the highest and returned as the resulting image.

During the training phase, depicted in Fig. 2, the generators train themselves to generate accurate representations of their assigned counterparts, and the discriminator trains to locate and distinguish the objects and details in the generated visual representations from those in an actual image that aligns with the input text. Then, depending on the results, losses are returned to the discriminator and each associated generator to advance their training. This training is continued until the discriminator can no longer favor the objects and details from the actual image over the generated visual representations. Due to our unique structure, each individual generator is trained independently. Thus, this leads each generator to converge at a different rate. While the training continues until all generators converge, some generators that are called upon more often than others are bound to converge earlier.
than others. In order to prevent early converging generators from overfitting, we simply put each generator to stop any additional training if the training loss no longer changes for over 20 epochs.

Once the training is complete, a description can be fed into the model’s generators, which will generate visual representations. These generated visual representations are then fed into the image search module, which will utilize the discriminator to compare the generated text-to-image results with images processed through the inception-v4 CNN, and each image is ranked based on the size of loss from the discriminator. The smaller the loss, the higher the ranking.

Note that the initial computational cost to train our proposed module from ground up is quite high compared to other state-of-the-art models, as our model does need to establish a solid number of vocabulary it can recognize in order to work properly. But once a certain number of vocabulary are established, the strength in our proposed method comes from the fact that each generator is independent. For every instance, only generators that are associated to the instance are called from the entire library during each processing. Thus compared to the computational cost needed to train, the computational cost demanded to use the model itself is significantly lower.

A. BERT FEATURES AND ATTENTION MAPS

For our methodology to work, the given model must be able to structure a representation of an image based on the description in a given text. While objects themselves can be recognized through single words, details such as the status and location of the object and relations with other objects need deeper analysis of the given sentence. To structure the appropriate visual representation of a given text, we implemented BERT attention maps along with their features.

BERT is a transformer model introduced by Vaswani et al. [16], and it uses scaled dot-production attention. However, since the transformer has a multi-layer, multi-head attention mechanism, its attention map consists of multiple attention patterns for each layer and head. Incorporating the entirety of this map increases the complexity of the model to impractical proportions. In this study, we did not use the entire attention map but specific layers according to the amount of detail that we wanted to refer to during use.

For this, we incorporated Vig et al. [17] work on “neuron view” when choosing the layer of attention pattern to construct the number of added details when generating a visual representation of the target object. Using the equation from Vig et al. [17], the attention distribution at position \( i \) in a sequence \( x \) could be given by

\[
x_i = \text{softmax} \left( \frac{q_1 \cdot k_i}{\sqrt{d}}, \frac{q_2 \cdot k_i}{\sqrt{d}}, \ldots, \frac{q_N \cdot k_i}{\sqrt{d}} \right)
\]

where \( q_n \) and key vector \( k_i \) are replaced with their corresponding generators \( Q_n \) and \( K_i \), and each assigned BERT embedding is added as an input, with \( v_o \) being the embedding from the main object vocabulary and \( v_d \) the \( n \)th vocabulary that is regarded as a detail to the main object.

B. GENERATOR AND INPUT

Stemming from the use of multiple generators, the generators in this study had three major requirements: low computation, fast processing, and accurate training even when using a small dataset. To fulfill these criteria, we chose the U-net architecture by Ronneberger et al. [18] for our generator.

U-net was first introduced in 2015 as a CNN architecture for fast and precise segmentation of images for biomedical purposes. Since its introduction, it has become a popular architecture for use not only in biomedical applications but also in natural images.

For our research, the U-Net that we chose to use is based on the model variant introduced in U-GAN: GANs with U-net for retinal vessel segmentation by Wu et al. [19]. It is an encoder-decoder model used for image translation, where skip connections are used to connect layers in the encoder with corresponding layers in the decoder that have the same sized feature maps. Earning its name “U-net” based on the shape of the network, it was able to enhance its processing speed by reducing the ratio of overlapping compared to its predecessors by applying a skip architecture. While preceding models used a sliding window technique that re-evaluated patches resulting in wasted time, U-nets would skip patches that were already evaluated and focus on patches not yet evaluated, reducing the total time spent on evaluation.

The generator that utilizes this architecture is a basic encoder-decoder network with eight encoding layers and seven decoding layers. Note that the encoding layer has one layer more than the decoding layer. This is because the input for the generator is a text representation, and the extra layer is needed in order to match the dimensions of the U-net generator model.

For every generator \( G(T) \), the goal is to not only generate the correct object of the given vocabulary but also match the description given by the surrounding context. Thus, given an input text representation \( T \) with the designated vocabulary included, the generator is trained to maximize probability \( \theta \) of generating the correct image object representation \( R \) but
minimize the probability to contort representation \( \nu \) on the base vocabulary during the process. This is represented in the following Equation 3.

\[
\theta^*, \nu^* = \arg \min_{\nu} \max_{\theta} \sum_{T, R} \log p(R|T; (\theta, \nu, \text{loss})) 
\]

A similar equation for image captioning is used in Show and tell—a neural image caption generator by Vinyals et al. [15]; for every input text representation \( T \), the best way to maximize the probability of generating an accurate image representation \( R \) while retaining it is to acquire appropriate parameters \( \theta \) and \( \nu \). In our case, however, the size and length of both object representation \( R \) and input text representation \( T \) are fixed, albeit the two have different sizes. Note that a given input sentence can be segmented into multiple vocabularies and contexts owing to the text’s structured nature, but the resulting image cannot. Because the image itself cannot be segmented or masked in any way for comparison, the best way to increase the probability of the total sum of \( R \) to be correct is for each generated \( R_s \) to be correct. Thus, given a set of vocabularies \( N \), while each designated generator contortion parameter \( \nu \) is independent, the parameter passed on to maximize the probability in adopting the surrounding context itself is shared throughout the sentence.

\[
\log p(R|T) = \sum_{i=0}^{N} \log p(R_i|T_i; \theta) 
\]

Unlike Equation 3, which represents the process of a single generator, notice that \( \nu \) is not included during the formation of the entire image in Equation 4 since the independence of the vocabulary itself is not relevant when forming a correct representation of the entire image. In this stage of forming the image with individual objects, it is assumed that the generated objects will not be contorted beyond basic recognition. Therefore, the only matter of concern is maximizing the probability of creating an accurate representation of the entire description when put together.

Clearly, neither the entire representation nor the vocabulary description can appropriately train themselves without any feedback from the discriminator and more so for updating the \( \theta \) parameter as it is used extensively to generate image object representations during the generator phase. Thus, the following problem arises: How do the generators receive appropriate feedback to increase each of their parameters? This problem leads us to the discussion of our discriminator.

**C. DISCRIMINATOR AND LOSSES**

In contrast to the multiple generators, our proposed model has only one discriminator. Once the total sum of \( R \) is given by the multiple generators, the discriminator compares the resulting \( R \) to the representation of an actual image for comparison.

The discriminator itself is quite simple. It is a Markovian discriminator similar to the convolutional “PatchGAN” classifier [12] that attempts to compare two given image representations of an object and its details and find the one that is “fake.” The discriminator will convolutionally run across the given input, distinguish which of the two images are real, and give output \( D \). By incorporating our discriminator, the objective of the entire model can be shown in the following Equation 5.

\[
S_{cGAN}(G, D) = \mathbb{E}_{T,R} [\log D(T, R)] \\
+ \mathbb{E}_T [\log (1 - D(T, G(T, \theta, \nu)))] 
\]

To appropriately deceive the discriminator, the sum of the generator’s output should not only appear real but also be near to the ground truth output with \( L2 \) distance in mind [20]. Much like PatchGAN [12], the reconstruction “\( L2 \) loss” is included to optimize the generators as the \( L2 \) loss will guide the generators to minimize the Euclidean distance of the generated representation from the actual image representation.

\[
S_{L2}(G) = \mathbb{E}_{T,R} [(y - G(T, \theta, \nu))^2] 
\]

The reason behind the use of the traditional “\( L2 \) loss” as seen in Equation 6 is to force the output to be conditioned on the input. As the discriminator should not only be deceived that the given output from the generators are like that of actual images, but also be recognized as the objects from the paired image. Thus, the use of this loss to train not only the generator but also the discriminator is vital to achieve accurate representations from the given text description. The discriminator later acts as our search engine for the proposed model.

\[
G^* = \arg \min_G \max_D S_{cGAN}(G, D) + \lambda S_{L2}(G) \cdot I 
\]

After each loss is given, the losses are returned to each generator. Equation 7 represents the way a loss is returned to each generator; it can be seen that each will receive their loss differently depending on how much influence (I) it had on the actual representation when dealing with the object. This is done for two reasons: 1) to focus on representing and improving the nouns and verbs of an image, which are usually the most prominent vocabularies to specify an object or describe the details when searching; and 2) to prevent certain auxiliary verbs to be represented prominently and cause a mode collapse problem, subsequently over-fitting the discriminator model during training. The influence of each word is determined by the attention map ranging between 0 to 1 on a relative scale. An object within the image will be set as the main focus, and all other words will describe its details. Therefore, by setting the generator associated with the main object to receive the most feedbacks and determining the level of feedback to the other generators, which created surrounding details, based on their relationship with the main object, we were able to achieve ideal optimization.

As each object within the given text is added, we can observe the resulting output generating the detail of each object, allowing images to be searched with high accuracy and rich detail.
TABLE 1. Comparison of image retrieval results on the MSCOCO and Flikr30K data test sets.

| Methods          | MSCOCO data | Flikr30K data |
|------------------|-------------|---------------|
|                  | R@1 | R@5 | R@10 | R@5 | R@10 |
| CAMP (Wang et al., 2019c) | 58.5 | 87.9 | 95.0 | 51.5 | 77.1 | 85.3 |
| SCAN (Lee et al., 2018) | 58.8 | 88.4 | 94.8 | 48.6 | 77.7 | 85.2 |
| SGM (Wang et al., 2020) | 57.5 | 87.3 | 94.3 | 53.5 | 79.6 | 86.5 |
| RDAN (Hu et al., 2019) | 61.6 | 89.2 | 94.7 | 54.1 | 80.9 | 87.2 |
| MMCA (Wei et al., 2020) | 61.6 | 89.8 | 95.2 | 54.8 | 81.4 | 87.8 |
| BFAN (Liu et al., 2019) | 59.4 | 88.4 | - | 50.8 | 78.4 | - |
| CAAN (Zhang et al., 2020) | 61.3 | 89.7 | 95.2 | 52.8 | 79.0 | 87.9 |
| DPRNN (Chen & Luo, 2020) | 62.5 | 89.7 | 95.1 | 55.5 | 81.3 | 88.2 |
| PFAN (Wang et al., 2019b) | 61.6 | 89.6 | 95.2 | 50.4 | 78.7 | 86.1 |
| VSRN (Li et al., 2019) | 62.8 | 89.7 | 95.1 | 54.7 | 81.8 | 88.2 |
| IMRAM (Chen et al., 2020) | 61.7 | 89.1 | 95.0 | 53.9 | 79.4 | 87.2 |
| SGRAF (Haiwen Diao et al., 2021) | 63.2 | 90.7 | 96.1 | 58.5 | 83.0 | 88.8 |
| Ours (MMG-GAN: 200 Vocabularies) | 68.4 | 78.9 | 85.0 | 69.1 | 72.0 | 87.7 |

IV. EXPERIMENT AND RESULTS

A. DATASET DETAILS

To train our proposed model, we used the MSCOCO dataset [13], a vast dataset that is one of the main datasets used for a multitude of tasks such as image classification, object detection and masking, and image captioning tasks. Our main focus for this dataset was to utilize the numerous captions paired with various objects and images. Because these same images are also used for object detection/masking tasks, we were able to gain the information about the objects existing in each image. This allowed us to train our search engine with cross-evaluation.

However, augmentation of the data was necessary to prepare it for use. For the given caption data, we utilized the BERT embedding and attention map against each object within the sentence, along with lexical simplification. Lexical simplification was performed to reduce the number of individual vocabularies to process, whereas BERT embedding and attention mapping were used to detect the word that indicated the main object within the sentence, and the way all other words interacted with it was used to describe the details of the main object. Simultaneously, we did not use the raw image to represent itself; we instead used a pretrained inception-v4 CNN [14] to extract visual representations.

For this experiment, we trained a total of 200 generators, each assigned to a common noun or verb. While these vocabularies are mostly nouns and verbs associated to the 91 object categories from the MSCOCO dataset. Due to our own computational and physical limitations while training our model, the 200 vocabularies do not cover all of the vocabulary present within the dataset. All other types of vocabulary were excluded as they typically serve to either link words within a sentence or are too rare and specific to assign a generator to it (e.g., proper nouns). In the following section, we discuss the change in the model’s performance with progress in its training and compare it with other similar search engines currently in the market.

B. EXPERIMENTAL DESIGN

For this experiment, we compared our proposed model with other image retrieval techniques, especially state-of-the-art models that use content-based image retrieval techniques. This allowed us to verify the effectiveness of our model.

Using the same split as that used by Diao et al. [7], the MSCOCO [13] dataset was split into 113,287 images for training, 5000 images for validation, and 5000 images for testing. We used 5000 images by averaging 1000 test images over five folds first and then tested it on the full 5000 images. In addition, we used 1000 images for validation and 1000 images for testing from The Flikr30K dataset, which contains 31,783 images.

We also prepared an additional experiment to demonstrate the performance of our model based on the number of details that it interacts with. For this experiment, we used different layers of the attention map, where the object of each input sentence was associated with a different number of details to show how increasing the specification for image retrieval affects the overall model in general.

C. EVALUATION METRICS

For evaluation, we measured each model’s performance based on recall at K, which measures accuracy as the proportion of relevant items found in the top-K recommendations of the tested model. In short, we adopted the same evaluation metrics for both datasets used by previous studies such as Diao et al. [7], Chen et al. [21], and Li et al. [9].

In addition, because of the novel way that our model works in, we next evaluated the effect of the minimum number of considered details on our model. Note that the number of details is determined using different layers from the attention map to determine the reach of attention across other vocabularies listed on the x axis. While measuring the numbers in Fig. 3, we ensured that the attention layers that we used had the number of attended objects written on the x axis. Next, we discuss the results achieved.
**D. MAIN RESULTS**

As seen in Table 1, our proposed model outperforms other models in retrieving images on R@1 (68.4% on the MSCOCO data and 69.1% on the Flikr30K data). Even when considering the proposed model’s lower R@5 and R@10 performances, the proposed model performs within an acceptable margin as it tends to be more strict depending on the number of details described within the input sentence. It should be noted that while many of the test sets designate each sentence with a group of associated images, our model tended to emphasize less on objects that differ significantly in detail. Our model was trained only on 200 words, mostly nouns and verbs that represent objects and specific activities and statuses. Naturally we suspect that our model may...
have missed a good portion of objects and statuses simply because it did not recognize them. It can be observed that this has especially affected the R@5 and R@10 performances. As the model relies on searching images by recognizing objects and associated details through vocabulary, the lack of recognizable vocabulary has affected the discriminator’s precision in ranking images. And while this does not affect the results when retrieving a single image that perfectly aligns with the given text description, it has affected the model’s performance when retrieving images that partially align with the given text input. Which lead to the result of showing underwhelming performances on both the R@5 and R@10 performances compared to the R@1 performance. Nonetheless, even with a disadvantage in recognizing vocabulary, our model’s performances are still comparable to state-of-the-art models. This proves that using our proposed method is certainly a viable way to retrieve specific images by giving specific details.

Furthermore, notice that the accuracy of R@1 increases significantly as more details are considered; on the contrary, the performances of R@5 and R@10 degrade when more details are considered. This is because of two reasons. First, most sentences do not have those many words that serve as meaningful details for each object. Thus, considering excessive amounts of unneeded details only causes disruption in generating an appropriate visual representation. Second, even when there are three or four words that can act as meaningful details, the generated representation for each object becomes too specific to retrieve images. Thus, while the R@1 performances are fairly remarkable when incorporating multiple details, the model loses robustness, making it unsuitable for real-life applications.

We found that utilizing two vocabularies as additional details was the most practical approach, as it seems to fare best on average when dealing with both short texts and long texts simultaneously. Of course, if the input was to have objects associated with multiple meaningful details, incorporating more details during processing will indeed increase our models performance to retrieve that specific image. But on average, incorporating only two vocabularies as additional details gave the best performance in most situations. We also estimated that with additional vocabularies included in the model, the proposed model will be able to achieve even better performances.

While the performance when using no additional details is inferior, it is acceptable, and it proves that even with just the base object representations, it is possible to retrieve suitable images. This implies that the object representation itself is indeed independent of the details added later.

V. CONCLUSION AND FUTURE WORK

Herein, we presented a MMG-GAN designed to learn individual vocabularies to perform image retrieval with a library utilizing grounded vocabulary. Through the use of multiple generators designated to specific vocabularies and utilizing a discriminator that can identify and locate objects within images, we demonstrated the idea of image retrieval using grounded vocabulary and made two main contributions. Our first contribution is the introduction of a viable method that can be used to process both text and images in applications, showing that even with only 200 trained vocabularies, the proposed model can compete with state-of-the-art models. Our second more significant contribution is the verification that each visual representation generated from the vocabulary works independently from its details, indicating that we have successfully found a way to introduce structure into images by utilizing multiple generators. Since multiple generators would act independently when processing individual vocabularies, it would avoid the problem of two closely associated instances collapsing into each other and warping the visual output. Further, by controlling the number of details considered, the results confirm that our approach did retain some elements of grounded vocabulary representation.

For our future work, We first plan to further train our proposed model to expand the number of vocabulary it can process. For this paper we were only able to train 200 meaningful vocabularies, and we expect to see better performances as we expand its dictionary. We will also attempt to reverse the process to test whether it is possible to use these representations to extract information from images (information that includes not only objects but also details such as status, location, and interaction between objects). Once this is possible, seamless multiprocessing between visual information and natural language will become much easier as one would only need to add vocabularies to the set of libraries in our method. We will also try to fine-tune the proposed method by training it with more vocabularies and speeding up the training process.

REFERENCES

[1] 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. 262–282, 2007.
[2] A. D’Amour et al., “Underspecification presents challenges for credibility in modern machine learning,” 2020, arXiv:2011.03395. [Online]. Available: https://arxiv.org/abs/2011.03395
[3] M.-L. Bourguet, “Designing and prototyping multimodal commands,” Tech. Rep., 2003.
[4] Q. Hoang, T. D. Nguyen, T. Le, and Q. D. Phung, “Multi-generator generative adversarial nets,” 2017, arXiv:1708.02556. [Online]. Available: https://arxiv.org/abs/1708.02556
[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, arXiv:1810.04805. [Online]. Available: https://arxiv.org/abs/1810.04805
[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), vol. 25, Dec. 2012, pp. 1097–1105.
[7] H. Diao, Y. Zhang, L. Ma, and H. Lu, “Similarity reasoning and filtration for image-text matching,” 2021, arXiv:2101.01368. [Online]. Available: https://arxiv.org/abs/2101.01368
[8] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, and A. S. Marchand-Maillet, “Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders,” 2020, arXiv:2008.05231. [Online]. Available: https://arxiv.org/abs/2008.05231
[9] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, “Visual semantic reasoning for image-text matching,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4654–4662.
[10] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, “Stacked cross attention for image-text matching,” 2018, arXiv:1803.08024. [Online]. Available: https://arxiv.org/abs/1803.08024

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” Tech. Rep., 2014.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” 2016, arXiv:1611.07004. [Online]. Available: https://arxiv.org/abs/1611.07004

[13] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common objects in context,” 2014, arXiv:1405.0312. [Online]. Available: https://arxiv.org/abs/1405.0312

[14] C. Szegedy, S. Ioffe, and V. Vanhoucke, “Inception-v4, inception-ResNet and the impact of residual connections on learning,” 2016, arXiv:1602.07261. [Online]. Available: https://arxiv.org/abs/1602.07261

[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” 2014, arXiv:1411.4555. [Online]. Available: https://arxiv.org/abs/1411.4555

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, N. A. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017, arXiv:1706.03762. [Online]. Available: https://arxiv.org/abs/1706.03762

[17] J. Vig, “A multiscale visualization of attention in the transformer model,” 2019, arXiv:1906.05714. [Online]. Available: https://arxiv.org/abs/1906.05714

[18] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” 2015, arXiv:1505.04597. [Online]. Available: https://arxiv.org/abs/1505.04597

[19] C. Wu, Y. Zou, and Z. Yang, “U-GAN: Generative adversarial networks with U-Net for retinal vessel segmentation,” in Proc. 14th Int. Conf. Comput. Sci. Educ. (ICCSE), Aug. 2019, pp. 642–646.

[20] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” 2016, arXiv:1604.07379. [Online]. Available: https://arxiv.org/abs/1604.07379

[21] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, “IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval,” 2020, arXiv:2003.03772. [Online]. Available: https://arxiv.org/abs/2003.03772

KUEKYEONG KIM received the B.S. degree in engineering in computer science and cognitive science from Korea University, Seoul, South Korea, in 2016, where he is currently pursuing the Ph.D. degree in computer science and engineering. Since 2017, he has been a Research Assistant with the Natural Language Processing and Artificial Intelligence Laboratory, Korea University.

CHANJUN PARK received the B.S. degree in computer science and cognitive science from the Busan University of Foreign Studies, Busan, South Korea, in 2019. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Korea University, Seoul, South Korea. From 2018 to 2019, he worked as a Research Engineer at SYSTRAN.

JAEHYUNG SEO received the B.S. degree in english language and literature from Korea University, Seoul, South Korea in 2020, where he is currently pursuing an integrated master’s and Ph.D. degrees in computer science and engineering with the Natural Language Processing and Artificial Intelligence Laboratory.

HEUISEOK LIM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science and engineering from Korea University, Seoul, South Korea, in 1992, 1994, and 1997, respectively. He is currently a Professor with the Department of Computer Science and Engineering, Korea University.

VOLUME 9, 2021 144623