VvsS-Net : Image-to-Image Retrieval with Controlling Visual versus Semantic Similarity

Seongeun Lee
Seoul National University
ryuha96@snu.ac.kr

Eunseok Yang
Seoul National University
mayth24@snu.ac.kr

Jonghyeon Seon
Seoul National University
sunutf@snu.ac.kr

Abstract

Image retrieval has been studied in two approaches, with visual similarity and semantic similarity. As image searches are carried out in various fields and situations, searching under a single criterion has a limited ability to respond flexibly to user intentions. We proposed a model that allows users to freely adjust their weight according to their purpose by considering both perspectives and meanings and to search flexibly according to the user's intention. We also show that a similarity based on human-annotated region-level captions is highly correlated with the human ranking and constitutes a good computable surrogate. Following this observation, we learn a visual embedding of the images where the similarity in the visual space is correlated with their semantic similarity surrogate. Finally, our model can consider both visual and semantic factors at the same time, so that the ranking can be determined according to the user's purpose.

1. Introduction

The task of image retrieval aims at, given a query image, retrieving all images relevant to that query within a potentially very large database of images. This topic has been heavily studied over the years. Initially tackled with bag of-features representations, large vocabularies, and inverted files [31], and then with feature encodings such as the Fisher vector or the VLAD descriptors [14], the retrieval task has recently benefited from the success of deep learning representations such as convolutional neural networks that were shown to be both effective and computationally efficient for this task [33, 28, 11]. Among previous retrieval methods, many have focused on retrieving the exact same instance as in the query image, such as a particular landmark [27].

Another group of methods have concentrated on retrieving semantically-related images. Specifically, there has been work on grounding scene graphs into images to obtain the likelihood of the scene graph-image pair [15, 30, 22].

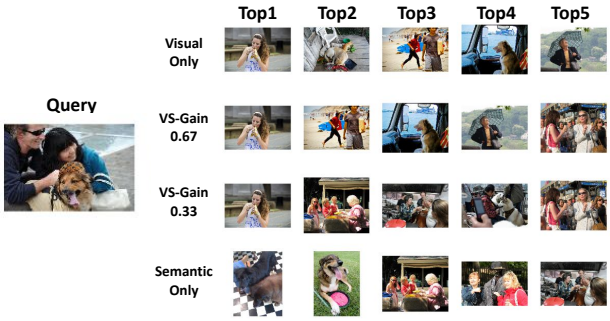

Figure 1: Image retrieval examples from Visual dependent model, Semantic dependant model, and Ours VvsS-Net. Both Visual dependent model and Semantic dependent model look at the user's intentions only from one point of view and set a standard, while our model is an adjustable model that can properly reflect both perspectives. e.g., VS gain(Visual semantic gain) controls visual and semantic weights, if the user wants to find an image with more similar visual elements, the gain can be increased, and if the user wants to find an image with more similar semantic elements, the gain can be lowered.

Alternately, we propose to utilize distributed representations derived from scene graphs of images alongside standard measures of similarity such as cosine similarity or inner product. Embeddings derived from the scene graphs capture the information present in the scene and this allows us to combine the advantages of structured representations like graphs and continuous intermediate representations. Further, we demonstrate in Figure 1 that similarity search over these embeddings captures the overall context of the scene, offering an alternative to visual similarity provided by traditional image embeddings.

As image searches are carried out in various fields and situations, searching under a single criterion has a limited ability to respond flexibly to user intention. Therefore, we propose a VvsS-Net that can adjust the weight of visual similarity and semantic similarity on scene graph level.

As described above, by extending the scene graph generation technique that showed reliable performance in semantic similarity among the existing approaches, the semantic expression, and the visual expression are solved by proceeding with scene graph generation. We train and optimize the scene graph embedding created by the scene graph generation model for visual and semantic aspects. Unlike a simple approach in which an extracted vector by passing an image through a network for direct search, each scene graph is embedded based on a each feature vector so that semantic expressions and visual expressions are compared at the same level as a scene graph to give equal representation. According to VS gain(Visual-Semantic ratio gain), we adjusted the final candidates by internalizing the results of two scene graph models for visual and semantic aspects.

The contribution of this paper is three-fold as follows:

- We propose a novel framework, VvsS-Net, for the image to image retrieval framework, which can control visual and semantic aspects by using the additional controllable unit module to existed model.
- We suggest a new approach to compare visual and semantic similarity by embedding it on each scene graph.
- Our algorithm outperforms the state-of-the-art methods on MS-COCO[20] and Visual-Genome[17].

2. Related Works

In this section, we provide an overview of two approach for image retrieval.

2.1. Image Retrieval

Image retrieval has been mostly tackled as the problem of instance-level image retrieval [31, 14, 33, 28, 11], which focuses on the retrieval of the exact same instance as defined in standard benchmark datasets [27]. Moving away from instances, some works have tackled visual search as the retrieval of images that share the same category label [3, 4] or a set of tags [10]. These works still have a crude understanding of the semantics of a scene. On their synthetic dataset of abstract scenes, Zitnick and Parikh have shown that image retrieval can be greatly improved when detailed semantics is available [38]. Explicit modeling of a scene can be done with attributes [25], object co-occurrences [23], or pairwise relationship between objects [5, 7, 21]. As the interaction between objects in a scene can be highly complex, going beyond simple pairwise relations, one extreme interface proposed by Johnson et al. [11] is to compare explicit scene graph representations instead of visual representations. One shortcoming of their method is that it requires the user to query with a full scene graph, which is a tedious process. We believe that querying with an image is a more intuitive interface. A number of approaches have cast the task of image captioning as a retrieval problem, first retrieving similar images, and then transferring caption annotations from the retrieved images to the query image [13, 32, 24]. Yet these methods use features that are not trained for the task, either simple global features [13], features pre-trained on ImageNet [32] or complex features relying on object detectors, scene classifiers, etc. [24]. We believe that the representation should be free of assumptions about the list of objects, attributes, and interactions one might encounter in the scene, and therefore, we learn these representations directly from the training data.

2.2. Visual Semantic Embedding Models

Image representations obtained by deep convolutional networks have had tremendous success in a range of vision tasks. Early works [12, 18] focused on image classification using image-level labels. Follow-up works include triplet formulations [34] that produce more generally useful visual representations with reduced data requirements. Some common directions to learn visual-semantic representations for images include the use of word embeddings of class labels [8, 19], exploiting class structure for classification[36] and leveraging WordNet ontology for class hierarchies [6, 2]. These methods work for simple images but cannot be trivially extended to complex scenes with multiple objects and relationships. More recent work considers the multimodal setting where pairwise ordering constraints are placed on both image and text modalities [16, 35] in a ranking formulation for representation learning. Additionally, similarity networks [9] have been proposed that take as input a pair of images and train a network using regression objectives over pairwise similarity values. We build on the above in two directions: (1) we derive both visual and semantic feature embeddings from scene graphs by utilizing a Graph Convolutional Network (2) we derivie the model to adjust the search results according to the user's intention, taking into account both visual similarity and semantic similarity.

3. Method

In this section, we describe our framework, VvsS-Net. VvsS-Net first receives a single query image and generates a scene graph. The visual similarity and semantic similarity between scene graphs are computed using graph visual embedding and graph semantic embedding, respectively. Each embedding is trained by a graph neural network with corresponding surrogate relevance. Finally, VvsS-Net retrieves images with total similarity reflecting visual and semantic similarities. The critical point is that the ratio of visual and semantic information to use is adjustable according to the user's purpose.

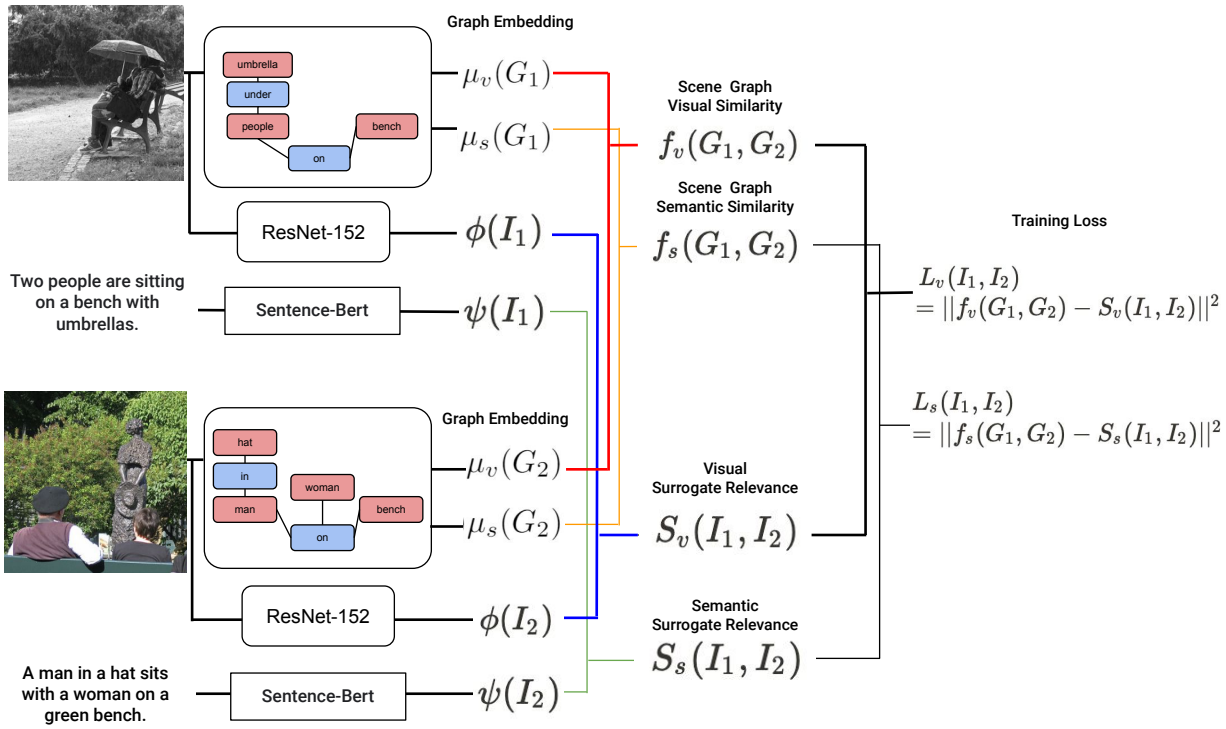

Figure 2: An overview of the proposed algorithm VvsS-Net. Images I_1, I_2 are converted into two vector representations, *i)* visual vector ϕ and *ii)* semantic vector ψ , through scene graph generation and graph embedding. Two parametrized graph embedding function is learned to minimize loss to each aspects of surrogate relevance. Note that, the training objective of our model is to embed the generated graph for each of the semantic and visual features.

3.1. Image Retrieval with Scene Graph

A scene graph is structured representations of images, where the nodes in scene graphs correspond to object categories and attributes of objects, and the undirected edges correspond to the paired relationships between objects. Each of them is associated with a word label, for example, 'apple(object)', 'red(attribute)', and 'behind(relation)'. Each word label is converted into pretrained 300-dimensional GloVe vector[26] and treated as node feature. In our model, the attributes are excluded from the scene graph because the attributes mainly represent the visual information rather than semantic information. This process makes a scene graph to be neutral between the visual and semantic aspects.

When a query image I_q is given, an image retrieval framework ranks candidate images by the similarity to the query image. Our model uses scene graphs to measure the similarity between images $sim(I_i, I_j) = f(G_i, G_j)$ where G_i and G_j are the scene graphs of images I_i and I_j. the similarity function f is obtained by training the graph neural network with surrogate relevance. The details are described in the following section.

It is possible to include generating a scene graph from the image to our framework for end-to-end training, however, the predefined scene graph is used to our experiment for training in order to avoid the high computational cost. During the inference phase, the pre-trained algorithm is used for generation.

3.2. Learning to Predict Surrogate Relevance

We define two surrogate relevance measures between two images. One measures the visual similarity $S_v(I_i, I_j)$, and the other measures the semantic similarity $S_s(I_i, I_j)$ of images I_i and I_j.

For the visual surrogate relevance, the visual features $\phi(I_i)$ and $\phi(I_j)$ of images I_i and I_j are used. The visual feature vector is created by averaging the pretrained ResNet-152[12] node features of the image. For the object node, the bounding box of it is used. For the relation node, the union of the corresponding object and subject bounding

boxes is used. The visual surrogate relevance of two images is then defined by the dot product of two representation unit vectors.

$$S_v(I_i, I_j) = \phi(I_i) \cdot \phi(I_j)$$

For semantic surrogate relevance, human-annotated captions c_i and c_j of image I_i and I_j are used. The captions describing the semantic information of images are given in the form of one or several sentences. We used Sentence-BERT(SBERT) [29] to convert each caption c_i into the unit vector $\psi(c_i)$. The semantic surrogate relevance of two images is then defined by the inner product of two embedded unit vectors.

$$S_s(I_i, I_j) = \psi(c_i) \cdot \psi(c_j)$$

When there are more than one caption in one image, the semantic surrogate relevance of every caption pair is averaged.

We train the visual and semantic scene graph similarity models by minimizing mean squared error from the visual and semantic surrogate relevance measures. When the scene graphs G_i and G_j of Image I_i and I_j are given, the loss functions for images are defined as

$$L_v(I_i, I_j) = ||f_v(G_i, G_j) - S_v(I_i, I_j)||^2$$
$$L_s(I_i, I_j) = ||f_s(G_i, G_j) - S_s(I_i, I_j)||^2$$

where f_v and f_s are the visual and semantic scene graph similarities.

Finally we define the VS scene graph similarity as $f_{vs}(I_i, I_j) = \lambda f_v(I_i, I_j) + (1 - \lambda) f_s(I_i, I_j)$ where $0 \leq \lambda \leq 1$. λ is VS gain(Visual Semantic ratio gain), the factor to control the importance of visual or semantic aspect when the similarity of two images is measured. The visual information is highlighted when $\lambda = 1$, and the semantic information is emphasized when $\lambda = 0$.

3.3. Human agreement score

We measure the human agreement score to compare the decisions of choice in user-to-user and user-to-algorithm sense. The score is the proportion of annotators who made the same choice as the other annotators as proposed in [37]. In detail, let s_1, s_2, s_3, and s_4 be the number of human annotators who chose the first, second, both, and none of the given target images. The the human agreement score is defined as $w_i/(s_1 + s_2 + s_3 + s_4)$ where $w_i = (s_1 + 0.5s_3)\mathbb{1}_{i=1} + (s_2 + 0.5s_3)\mathbb{1}_{i=2} + (0.5s_1 + 0.5s_2 + s_3)\mathbb{1}_{i=3}$.

Given the query images, the retrieval algorithm choose the similar one between the target images on the relative or absolute criteria. The relative criteria refers to the selection of a more similar image between the two target images. In this case, one and only one image is chosen regardless of what the similarity between the query image and target image is. On the other hand, the absolute criteria has a threshold to choose the image. Therefore, none or both of target images might be chosen by the threshold setting. We used the absolute criteria with several thresholds to make the algorithm more human-like decisions unlike the previous studies using the relative criteria. Finally, the agreement scores are compared to the human decision for the evaluation.

4. Experiments

In this section, we evaluate our VvsS-Net and other baselines on the retrieval task with two metrics. First, we compute normalized discounted cumulative gain(nDCG) with the surrogate relevance as gain using pre-trained ResNet visual similarity and SBERT semantic similarity respectively. A larger nDCG indicates more relevant images are ranked higher in the retrieval result. Second, the agreement score between a retrieval algorithm and human decision is measured for evaluation. Human agreement score without adoption of threshold setting are used to compare our model with baselines. Then, within our model or other interpolation based baseline models, human agreement score with threshold is calculated.

4.1. Experimental Setup

Data In experiments, we decide to use datasets in a similar manner to [37]. We have image dataset with captions and scene graphs and human annotation similarity dataset about image triplets. The image dataset is VG-COCO, which is the intersection of Visual Genome[17] and MS-COCO[20]. Visual Genome is a dataset with a scene graph containing object, attribute, relationship, and bounding box information for each image. MS-COCO is a dataset with five human annotated captions for each image. We obtain fully annotated 35,017 training images and 13,203 test images. For query images, we randomly select a fixed 1000 images among test images as the query set and retrieve images among the other 13,202 test images for each query image.

We decide to use the VG-COCO to utilize both semantic annotations, caption and scene graph. However, since it is impractical to assume human annotated scene graphs for images are given in the evaluation phase, we also use automatically generated scene graphs for each image. The human-annotated similarity dataset is a dataset of which humans label which image is semantically closer to the query image, or whether they are similar or neither given an image triplet. The implementation detail of generating scene graphs and collection detail of human annotation similarity is illustrated in [37].

Scene graph generation follows the works [1], which the bounding boxes of objects in images are detected by Faster R-CNN method, and the name is predicted based on the

Method	Data	nDCG						Human Agreement
		5	10	20	30	40	50	
Inter Human	-	-	-	-	-	-	-	0.728 ± 0.05
ResNet	I	1	1	1	1	1	1	0.494
Caption SBERT	Cap(GT)	0.972	0.976	0.980	0.983	0.986	0.988	0.646
Gen. Cap. SBERT	Cap(Gen)	0.971	0.974	0.979	0.982	0.985	0.987	0.473
Object Count	I+SG	0.971	0.974	0.979	0.982	0.985	0.987	0.506
VvsS-Visual	I+SG	**0.979**	**0.981**	**0.984**	**0.986**	**0.988**	**0.990**	0.527
VvsS-Semantic	I+SG	0.971	0.975	0.979	0.982	0.985	0.987	0.509
VvsS-Half	I+SG	0.973	0.976	0.980	0.983	0.986	0.988	0.510
VvsS-Best	I+SG	0.978	0.981	0.984	0.986	0.988	0.990	**0.528**

Table 1: Image retrieval on VG-COCO with human-annotated scene graphs. VvsS-Visual: our visual graph embedding. VvsS-Semantic: our semantic graph embedding. VvsS-Half: interpolation of our visual model and semantic model, weighted with 0.5 and 0.5. VvsS-Best: interpolation of our visual model and semantic model, weighted with the highest human agreement score. Data columns indicates which data are used to inference. Cap(GT): human annotated caption. Cap(Gen): machine-generated captions. I: image. SG: scene graphs. The nDCG scores are calculated using the surrogate relavance of visual ResNet feature. Human agreement score is calculated without threshold setting so that model only distinguishes which one is more relevant to a query image.

ResNet-101 features from the detected bounding boxes. We keep up to 100 objects in a image with a confidence threshold of 0.3. To extract relation between objects, we used the frequency prior information constructed from the GQA dataset that covers 309 kinds of relations. For the detected pairs of objects, relations are predicted with a confidence threshold of 0.2. Human-annotated similarity dataset contains 10,712 human annotations from 29 human labelers for 1,752 image triplets. The triplet consists of a query image from the query set and two candidate images with selection criterion. Two candidate images are selected with two criteria, visual similarity with the query and exceeding margin distances between candidates.

Two-step retrieval using pre-trained ResNet feature We use two-step approach in the experiment, roughly retrieving relevant images and reranking over them using relevance. We employ this approach for two reasons, to reduce the computational burden and to make a set of good candidate images. For a query image, we first retrieve 100 images that are closest to the query with pre-trained ResNet-152 feature. After making candidates, we rerank over them using the retrieval algorithms. We should note that although there is large flexibility for designing this step, we employ the same approach with [37]. Unless mentioned, the results of our paper is based on the reranking setting.

Training details We use Adam optimizer with the initial learning rate of 0.0001 for both visual and semantic model, multiplying 0.9 to the learning rate each epoch. We set batch size as 16, and models are trained for 20 epochs. We pair images using an oversampling approach making total three images for each anchor image. Among three images, the first one is drawn from 100 most relevant samples from visual surrogate relevance score, the second one is drawn from semantic surrogate relevance score, and final one is drawn from the other. By employing oversampling scheme, we design training more suitable for the image retrieval by reinforcing the learning of more similar images through the oversampling approach.

Our VvsS-Net applies GCN to scene graphs of visual and semantic and the final node embeddings are aggregated through average pooling and scaled to the unit norm, making a graph embedding representation vector. We use three graph convolution layers with 1024 hidden neurons in each layer.

4.2. Baselines

ResNet-152 Features Image retrieval is performed based on the cosine similarity of pre-trained ResNet-152 feature, which implicits visual charateristic of the images.
Object Count(OC) Using only objects in scene graph, we compute the cosine similarity between object count vectors.
Caption SBERT To implicit semantic meaning of the images, we obtain SBERT representations of both ground truth captions and generated captions.
Interpolation of ResNet and SBERT To tackle the framework which puts different importances on visual and semantic similarity, we set the explicit baseline with cosine similarity of features using interpolation of ResNet feature of the image and SBERT feature of the captions. The difference from our model is that they are trained for different purpose and one is for image and the other is for sentence,

Method	Data	nDCG						Human Agreement
		5	10	20	30	40	50	
Inter Human	-	-	-	-	-	-	-	0.728±0.05
ResNet	I	0.821	0.838	0.859	0.874	0.887	0.898	0.494
Caption SBERT	Cap(GT)	1	1	1	1	1	1	0.646
Gen. Cap. SBERT	Cap(Gen)	0.823	0.836	0.857	0.872	0.886	0.898	0.473
Object Count	I+SG	0.806	0.827	0.850	0.865	0.879	0.895	0.506
VvsS-Visual	I+SG	0.805	0.825	0.848	0.864	0.878	0.891	0.527
VvsS-Semantic	I+SG	**0.822**	**0.837**	**0.856**	**0.870**	**0.882**	**0.894**	0.509
VvsS-Half	I+SG	0.822	0.837	0.856	0.870	0.883	0.895	0.510
VvsS-Best	I+SG	0.809	0.829	0.851	0.867	0.880	0.893	**0.528**

Table 2: Image retrieval on VG-COCO with human-annotated scene graphs. The nDCG scores are calculated using the surrogate relavance using semantic SBERT feature of groundtruth caption.

on the other hand, our embedding is trained on the same graph using different features and surrogate relevance.

4.3. Quantitative Results

Table 1 and Table 2 are the quantitative results of retrieval using ground truth scene graphs. The nDCG scores of Table 1 are calculated using surrogate relevance through similarity between ResNet features of the images as the ground truth label of relevance. Table 2 is computed using SBERT feature of ground truth captions as the label. VvsS-Visual, VvsS-Semantic, VvsS-Half, and VvsS-Best indicate our visual embedding model, semantic model, interpolation of both using weights 0.5, and best combination in respect to human agreement score, respectively. Due to the difficulty of considering all methods using threshold approach to output human agreement score, human agreement scores of Table 1 and Table 2 are calculated without thresholds so that algorithms selecting which one is more similar to a query image, not none or both of them.

From Table 1, our VvsS-Visual model shows larger nDCG score than other models. Here, we should note that the experiment is conducted in reranking setting of candidates using ResNet feature. Although overall nDCG score of all algorithms are high, our VvsS-Visual embedding could capture overall image visual similarity over other models. The nDCG score of VvsS-Semantic is lower than other models and interpolations with visual embedding yield higher nDCG score. From Table 2, our VvsS-Semantic model shows one of the largest nDCG score than other models and VvsS-Visual model shows lower nDCG score than other models. Although interpolations of both model produce slightly better result, this may be due to the reranking setting. From Table 1 and Table 2, our model successfully divide visual and semantic aspect of the scene graph. Table 3 is the result of image retrieval using machine-generated scene graph and calculate nDCG with SBERT features as label. Comparing Table 2 and Table 3,

Method	nDCG				Human Agreement
	5	10	20	40	
Inter Human	-	-	-	-	0.728
ResNet	0.821	0.838	0.859	0.887	0.494
Caption SBERT	1	1	1	1	0.646
Gen. Cap. SBERT	0.823	0.836	0.857	0.886	0.473
Object Count	0.795	0.801	0.822	0.865	0.511
VvsS-Visual	0.799	0.815	0.839	0.872	0.501
VvsS-Semantic	**0.800**	**0.820**	**0.842**	**0.874**	0.523
VvsS-Half	0.798	0.812	0.842	0.874	0.537
VvsS-Best	0.798	0.812	0.842	0.874	**0.537**

Table 3: Image retrieval on VG-COCO with machine-generated scene graphs using SBERT feature as groundtruth label. The results of baseline methods not using scene graphs are same as Table 2. The nDCG scores are calculated using the surrogate relavance using semantic SBERT feature of groundtruth caption.

although only the groundtruth scene graphs are used without generated during training, it can be seen that the performance does not deteriorate significantly even for the generated graph, which would have slightly different structure from groundtruth scene graph. If the training proceeds with the generated scene graph and the inference is also conducted with the generated graph, the performance may be better.

Figure 3 illustrates human agreement score across interpolation of visual features and semantic features. Semantic feature with caption is linearly interpolated with visual feature with ResNet feature of whole image. Semantic feature with scene graph is linearly interpolated with visual feature of scene graph. Interpolation using SBERT feature of ground truth caption and ResNet feature of whole image yields the best performance among them. However, ground

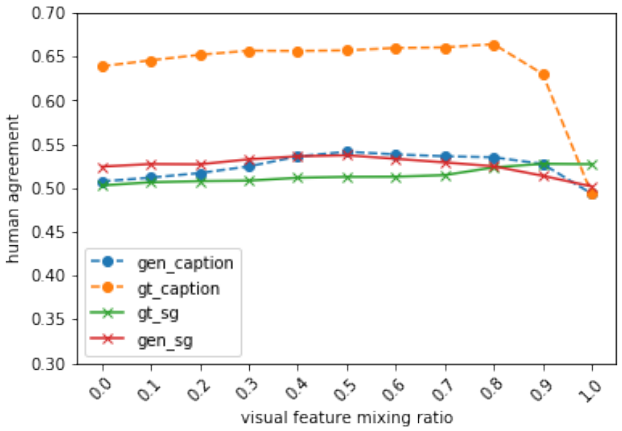

Figure 3: Human agreement score across interpolation of visual features and semantic features.

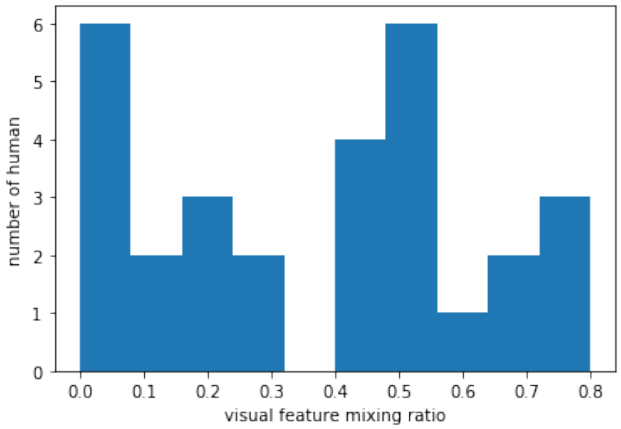

Figure 4: Number of humans across different weights of visual feature.

truth caption and human annotated scene graph are almost available in practice. Comparing generated scene graph and generated caption, interpolation with generated scene graph has slightly better result. Note that model is not trained using generated scene graph, then using state-of-the-art scene graph generation and training using them can improve the result.

We employ threshold setting for calculating human agreement score. There are two thresholds, which the first one is the minimum similarity between anchor and chosen candidate and the second one is the minimum margin between chosen candidate and the other. We explore various values for thresholds, and the first theshold of 0.4 and the second one of 0.05 are chosen. There are no significant differences among using different thresholds.

We also perform an additional experiment that shows biases of individuals toward visual and semantic aspect of images, calculating using visual feature and semantic feature of generated scene graphs. We test weights from 0.0 to 1.0 by 0.1 scale and get min and max human agreement score along weights. By applying different weights to each individual, we can get better results in respect to maximum human agreement score by 0.543, which is 0.006 higher than applying same weight to all and minimum score by 0.457, which is 0.08 lower. Mean inter individual standard deviation is 0.03. Figure 4 shows number of humans with highest human agreement score for each weight, indicating human evaluate retrieval results by different patterns between individuals.

4.4. Quantitative Results

Figure 5 show the example retrieved result from the query images we test. In this section, our goal is not showing our model outperforms other baselines, but showing the possibility that VvsS-Net can control weights of visual and semantic aspects. In the each block, the first row is retrieved with only visual features and increasing semantic ratio as the row to the final row could be with only visual features. For example, on the left instance, query images is people hugging dog but the top image retrieved with visual feature is the woman with poodle-like hair. ResNet feature is sensitive to the pattern of the image and dog and woman has the large importance in scene graphs respectively. By comparing the visual scene graph embedding, her poodle-like hair get importance and retrieved. Besides the top 1 image, images with dog or with backgroud of ocean are retrieved for visual similarity. On the contrary, semantic feature puts the importance on the dogs and people, them retrieves image with dogs or humans. There would be improvement toward solving the pattern-sensitive issues in using visual similarity. The right instance shows ability of our VvsS-Net well. The query image is the cat with laptop and retrieved results with visual similarities often ignore cats and concentrate on visual similarity of objects. However, the results with semantic similarities retrieved the images that a cat lay on laptop, although patterns or directions the cat lay is different from the query image.

5. Conclusion

We tackle the limitation of the existing image retrieval approach that can not adjust the weight of importance between visual and semantic. Moreover, we check the problem: Humans utilize both semantic and visual information with images, and individuals have biases toward their importance from the baseline method, as interpolates between the ResNet feature and caption SBERT. Therefore, we propose the framework VvsS-Net, which aims to retrieve images based on different importance to visual and semantic similarity. We implemented data pipelines and base-

Figure 5: Image retrieval with query image in the first column for each block with VvsS-Net. The first row is retrieved with visual similarity only. The second row is more with visual similarity(0.67, 0.33 for visual and semantic), and the third row is more with semantic similarity(0.33, 0.67 for visual and semantic). The fourth row is retrieved with semantic similarity only.

lines using pre-trained models. Then, We train on the scene graph without attributes using surrogate relevance with caption and image. The ablation studies that replace each embedding with ResNet and caption SBERT show that comparing the same level to train graph embedding on visual and semantic features is fairer than the naive approach. Our algorithm outperforms on MS-COCO[20] and Visual-Genome[17].

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, 2017. 4

[2] Björn Barz and Joachim Denzler. Hierarchy-based image embeddings for semantic image retrieval. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 638–647. IEEE, 2019. 2

[3] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *European Conference on Computer Vision*, pages 663–676. Springer, 2010. 2

[4] Alessandro Bergamo, Lorenzo Torresani, and Andrew W Fitzgibbon. Picodes: Learning a compact code for novel-category recognition. In *NIPS*. Citeseer. 2

[5] Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 129–136. IEEE, 2010. 2

[6] Jia Deng, Alexander C Berg, and Li Fei-Fei. Hierarchical semantic indexing for large scale image retrieval. In *CVPR 2011*, pages 785–792. IEEE, 2011. 2

[7] Chaitanya Desai, Deva Ramanan, and Charless C Fowlkes. Discriminative models for multi-class object layout. *International journal of computer vision*, 95(1):1–12, 2011. 2

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013. 2

[9] Noa Garcia and George Vogiatzis. Learning non-metric visual similarity for image retrieval. *Image and Vision Computing*, 82:18–25, 2019. 2

[10] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210–233, 2014. 2

[11] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016. 1, 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 2, 3

[13] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 2

[14] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *European conference on computer vision*, pages 304–317. Springer, 2008. 1, 2

[15] 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. 1

[16] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2, 4, 8

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 2

[19] Dong Li, Hsin-Ying Lee, Jia-Bin Huang, Shengjin Wang, and Ming-Hsuan Yang. Learning structured semantic embeddings for visual recognition. *arXiv preprint arXiv:1706.01237*, 2017. 2

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 4, 8

[21] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016. 2

[22] Paridhi Maheshwari, Ritwick Chaudhry, and Vishwa Vinay. Scene graph embeddings using relative similarity supervision. *arXiv preprint arXiv:2104.02381*, 2021. 1

[23] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2441–2448, 2014. 2

[24] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011. 2

[25] Devi Parikh and Kristen Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510. IEEE, 2011. 2

[26] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3

[27] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008. 1, 2

[28] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pages 3–20. Springer, 2016. 1, 2

[29] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 4

[30] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. 1

[31] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, IEEE International Conference on*, volume 3, pages 1470–1470. IEEE Computer Society, 2003. 1, 2

[32] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014. 2

[33] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015. 1, 2

[34] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1386–1393, 2014. 2

[35] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016. 2

[36] Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2740–2748, 2015. 2

[37] Sangwoong Yoon, Woo Young Kang, Sungwook Jeon, SeongEun Lee, Changjin Han, Jonghun Park, and Eun-Sol Kim. Image-to-image retrieval by learning similarity between scene graphs. *arXiv preprint arXiv:2012.14700*, 2020. 4, 5

[38] C Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016, 2013. 2