Scene Graph Embeddings Using Relative Similarity Supervision

Paridhi Maheshwari[1*], **Ritwick Chaudhry**[2*†], **Vishwa Vinay**[1]
[1] Adobe Research
[2] Carnegie Mellon University
parimahe@adobe.com, rchaudhr@andrew.cmu.edu, vinay@adobe.com

Abstract

Scene graphs are a powerful structured representation of the underlying content of images, and embeddings derived from them have been shown to be useful in multiple downstream tasks. In this work, we employ a graph convolutional network to exploit structure in scene graphs and produce image embeddings useful for semantic image retrieval. Different from classification-centric supervision traditionally available for learning image representations, we address the task of learning from relative similarity labels in a ranking context. Rooted within the contrastive learning paradigm, we propose a novel loss function that operates on pairs of similar and dissimilar images and imposes relative ordering between them in embedding space. We demonstrate that this Ranking loss, coupled with an intuitive triple sampling strategy, leads to robust representations that outperform well-known contrastive losses on the retrieval task. In addition, we provide qualitative evidence of how retrieved results that utilize structured scene information capture the global context of the scene, different from visual similarity search.

Introduction

In recent times, the advancement of deep convolutional neural networks (CNNs) has led to significant improvements in image classification (He et al. 2016; Krizhevsky, Sutskever, and Hinton 2012). Their intermediate image representations have also proven to be powerful visual descriptors in a variety of other tasks. One such application is content-based image retrieval, where a given image is used to query a database and retrieve other similar images. Models trained for image classification capture visually discriminative features and the retrieved images are, therefore, visually similar to the query image. On the other hand, semantic image retrieval enables the search of *visual situations* where multiple objects interact in different ways such as 'man talking on the phone while sipping coffee'. This requires an understanding of the semantics of the scene as well as bridging the *semantic gap* (Smeulders et al. 2000; Hare et al. 2006) between visual features and high-level concepts. To address this, numerous visual-semantic embedding models have been proposed that incorporate semantic information from the text modality into image representations. While natural language enables rich and detailed descriptions of visual content, it is unstructured and requires explicit grounding into images (Krishnamurthy and Kollar 2013; Lin et al. 2014a). This has led to the recent development of image representations using graph-based formulations that capture detailed scene semantics in a structured format.

Scene graphs (Johnson et al. 2015) have emerged as one such popular and powerful representation of the underlying content of images. This construct encapsulates the constituent objects and their relationships, and also encodes object attributes and spatial information. Their success can be ascribed to the numerous downstream applications that they facilitate, including visual question answering (Teney, Liu, and van Den Hengel 2017; Norcliffe-Brown, Vafeias, and Parisot 2018), scene classification (Schroeder, Tripathi, and Tang 2019), image manipulation (Dhamo et al. 2020) and visual relationship detection (Lu et al. 2016).

In the context of image retrieval, there has been work on grounding scene graphs into images to obtain the likelihood of the scene graph-image pair (Johnson et al. 2015; Schuster et al. 2015). 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.

We leverage Graph Convolutional Networks to map an image's scene graph into an embedding. In the existing literature, the training of such models is set up in one of the following paradigms: (1) Self-supervised, where scene graphs are jointly embedded with corresponding supplementary information (like visual features or text) about the image (Wang, Li, and Lazebnik 2016; Belilovsky et al. 2017), and (2) Task-dependent, where learning of the scene graph representation is driven by supervision from specific downstream applications (Teney, Liu, and van Den Hengel 2017; Schroeder, Tripathi, and Tang 2019). In contrast, we con-

[*]These authors contributed equally
[†]Work done while at Adobe Research

Figure 1: Comparison of image retrieval results for exemplar queries using scene graph embeddings from our proposed model, and classification features of ResNet-152 (He et al. 2016). While the latter retrieves visually similar images, our embeddings capture the global structure of the scene, i.e, 'man throwing frisbee' and 'man playing with dog' respectively. Notice that classification features do not distinguish various outdoor sports in (i), and fail to depict the human-animal interaction in (ii).

sider a pairwise similarity matrix as our supervision signal where every value represents a noisy notion of similarity between the corresponding image pair. We do not make any assumptions on this similarity (for example, it may not obey properties of a metric) and hence argue that our supervision is less strict. In this work, we define these similarities using the text modality, specifically image captions, but it can also be derived from other sources.

We show how robust representations can be learnt from scene graphs by leveraging caption similarities in a ranking context. This is enabled by a novel loss function that extracts signal from pairs of similar and dissimilar images, as in the contrastive learning approach. Contrary to the commonly-used class labels per image, we introduce soft target labels based on relative similarities to appropriately weigh the extent of similarity or dissimilarity. Further, we show improved retrieval performance with the learnt representations.

We summarize our key contributions as follows:

1. A graph convolutional network to process scene graphs into a visual-semantic embedding space for images, which combines benefits of both the structured information present in graphs and distributed representations.
2. A novel ranking loss that incorporates relative ranking constraints and outperforms other contrastive learning losses. Furthermore, a comparative analysis of different triplet sampling strategies is presented.
3. The embeddings are demonstrated to be robust to noise via a retrieval experiment with incomplete scene graph queries. We also provide qualitative insights into the use of these representations for content-based image retrieval.

Related Work

The relevant literature is introduced from the following two aspects: (1) Scene Graphs and their Applications (2) Visual-Semantic Embedding Models.

Scene Graphs and their Applications

The availability of large image datasets with detailed graph representations (Krishna et al. 2017; Lu et al. 2016) has led to a surge in research on scene graphs. A large fraction of this work focuses on generating scene graphs from images (Xu et al. 2020). They have also proven to be effective in a range of visual tasks such as image retrieval (Johnson et al. 2015; Wang et al. 2020), image generation (Johnson, Gupta, and Fei-Fei 2018) and captioning (Yang et al. 2019).

There has been some work on learning scene graph representations for downstream applications. Raboh et al. propose intermediate representations, called differentiable scene graphs, that can be trained end-to-end with supervision from a reasoning task. Schroeder, Tripathi, and Tang introduce triplet supervision and data augmentation to learn scene graph embeddings for layout generation. Distinct from their work, we aim to learn semantic embeddings for images using scene graphs, leveraging the text modality for weak supervision rather than specific downstream tasks.

The closest related work (Belilovsky et al. 2017) considers joint representations of scene graphs and images by aligning scene graphs with pre-trained image features using neural networks. However, we do not wish to match scene graphs to visual features as they are object-centric and do not attend to the semantics of the scene. Our focus is to learn scene graph embeddings that capture the global context.

Another line of relevant work involves image-text matching using graph structures. Li et al. propose a reasoning model to identify the key objects and relationships of a scene. Shi et al. incorporate common-sense knowledge from aggregate scene graphs to improve the matching process. Wang et al. take a more direct approach by developing graphs for both modalities and computing their similarity via custom distance functions. In this work, however, we leverage the text modality for weak supervision rather than embedding multiple modalities in the same space. In our retrieval setup, a scene graph query is used to run a nearest neighbor search over embeddings of the indexed images.

Visual-Semantic Embedding Models

Image representations obtained by deep convolutional networks have had tremendous success in a range of vision tasks. Early works (He et al. 2016; Krizhevsky, Sutskever, and Hinton 2012) focused on image classification using image-level labels. Follow up works include triplet formulations (Wang et al. 2014) 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 (Frome et al. 2013; Li et al. 2017), exploiting class structure for classification (Yan et al. 2015) and leveraging WordNet ontology for class hierarchies (Deng, Berg, and Fei-Fei 2011; Barz and Denzler 2019). 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 (Kiros, Salakhutdinov, and Zemel 2014; Wang, Li, and Lazebnik 2016) in a ranking formulation for representation learning. Additionally, similarity networks (Garcia and Vogiatzis 2019) 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 embeddings from scene graphs (rather than pixel information or unstructured text) by utilizing a Graph Convolutional Network (2) we leverage a weak pairwise similarity supervision from the text modality and design appropriate objective functions to drive the training of our model.

Model

We consider the task of learning image embeddings from a structured representation of its content. Each image $\mathcal{I}$ has a corresponding scene graph $\mathcal{G}_\mathcal{I} = (V_\mathcal{I}, E_\mathcal{I})$, where the vertices $V_\mathcal{I}$ represent objects and directed edges $E_\mathcal{I}$ denote the relationships between them. Therefore, $\mathcal{G}_\mathcal{I}$ comprises of <*subject, predicate, object*> triples such as <*cat, on, bed*> or <*man, driving, car*>. We wish to learn a mapping $\Phi : \mathcal{G}_\mathcal{I} \rightarrow \mathbf{f}_\mathcal{I}$ where $\mathbf{f}_\mathcal{I} \in \mathbb{R}^D$ is the embedding of image $\mathcal{I}$. Inspired by recent work (Johnson, Gupta, and Fei-Fei 2018) on learning intermediate scene graph representations, we model Φ as a Graph Convolutional Network (GCN). It performs a series of *convolution* operations on the graph, followed by an aggregation layer to pool context from different entities in the image.

Each vertex u and edge e_{uv} is encoded as a vector, $\mathbf{\Lambda}_u \in \mathbb{R}^d$ and $\mathbf{\Lambda}_{uv} \in \mathbb{R}^d$ respectively, using separate learnable embedding layers. These vectors are updated by convolution operations from their respective immediate neighborhoods. For nodes, this update step is a function of all its one-hop neighbor nodes. And edge representations are updated based on the source and target node. Hence, the context is propagated throughout the graph via its edges. Mathematically, each convolutional layer in the GCN relays information across entities by applying the following operations in order:

1. **Message Passing**: Each edge in the graph generates a "message" for its source and target nodes. For edge $e_{uv} \in E_\mathcal{I}$, a message $\mathbf{m}^s_{uv} \in \mathbb{R}^h$ is sent to the source node u and another message $\mathbf{m}^t_{uv} \in \mathbb{R}^h$ is sent to the target node v. These messages gather information from the edge state $\mathbf{\Lambda}_{uv}$ and the node states $\mathbf{\Lambda}_u$ and $\mathbf{\Lambda}_v$ and are denoted by
$$\mathbf{m}^s_{uv} \leftarrow \psi_s(\mathbf{\Lambda}_u, \mathbf{\Lambda}_v, \mathbf{\Lambda}_{uv})$$
$$\mathbf{m}^t_{uv} \leftarrow \psi_t(\mathbf{\Lambda}_u, \mathbf{\Lambda}_v, \mathbf{\Lambda}_{uv})$$
2. **State Update for Edges**: The state vector for an edge $\mathbf{\Lambda}_{uv}$ is updated to $\hat{\mathbf{\Lambda}}_{uv} \in \mathbb{R}^D$ by combining the most recent node states with the edge's prior state as
$$\hat{\mathbf{\Lambda}}_{uv} \leftarrow \psi_e(\mathbf{\Lambda}_u, \mathbf{\Lambda}_{uv}, \mathbf{\Lambda}_v)$$
3. **State Update for Nodes**: The state for every node $\mathbf{\Lambda}_u$ is updated to an intermediate representation which is obtained by pooling all the messages it receives via its edges
$$\mathbf{\Gamma}_u \leftarrow \frac{\sum_{w|(u,w)\in E_\mathcal{I}} \mathbf{m}^s_{uw} + \sum_{w|(w,u)\in E_\mathcal{I}} \mathbf{m}^t_{wu}}{\sum_{w|(u,w)\in E_\mathcal{I}} 1 + \sum_{w|(w,u)\in E_\mathcal{I}} 1}$$
This intermediate pooled representation is passed through another non-linear transformation and normalized to produce the updated node state $\hat{\mathbf{\Lambda}}_u \in \mathbb{R}^D$ as
$$\hat{\mathbf{\Lambda}}_u \leftarrow \frac{\psi_n(\mathbf{\Gamma}_u)}{\|\psi_n(\mathbf{\Gamma}_u)\|_2}$$
ℓ_2-normalization results in unit length vectors and has been shown to provide superior results (Chen et al. 2020).

The state vectors $\mathbf{\Lambda}_u$ and $\mathbf{\Lambda}_{uv}$ are iteratively updated via a series of graph convolutional layers such that the resulting state vectors of nodes capture information from the entire graph. Finally, we define the embedding of the scene graph (and image) as the average over all learnt node state vectors
$$\mathbf{f}_\mathcal{I} \leftarrow \frac{\sum_{u\in V_\mathcal{I}} \hat{\mathbf{\Lambda}}_u}{\sum_{u\in V_\mathcal{I}} 1}$$

All non-linear transformations – $\psi_s, \psi_t, \psi_e, \psi_n$ – are implemented as multilayer perceptrons. Specifically, the functions ψ_s, ψ_t and ψ_e are modeled using a single network that concatenates the inputs $\mathbf{\Lambda}_u, \mathbf{\Lambda}_{uv}, \mathbf{\Lambda}_v$ and computes 3 outputs using separate fully connected heads. Weight sharing across all neighborhoods allows the layer to operate on graphs of arbitrary shapes. To ensure that a scene graph is connected, we augment a trivial node $\_image\_$ and trivial edges $\_in\_image\_$ from every other node to this node.

Weak Supervision from the Text Modality

The previous section described our GCN architecture that maps the scene graph for each image $\mathcal{A}$ in a collection of N images into its corresponding embedding $\mathbf{f}_{\mathcal{A}}$. Our supervision signal for training the network is an $N \times N$ similarity matrix where entries $\mathrm{s}_{\mathcal{XY}}$ represent the measure of similarity between images $\mathcal{X}$ and $\mathcal{Y}$. In the current work, these similarities are computed using textual captions of corresponding images as natural language is key in conveying semantics. Further, single-sentence, user-generated captions tend to focus on the entirety of the scene.

A strict criterion would be to set $\mathrm{sim}(\mathbf{f}_{\mathcal{X}}, \mathbf{f}_{\mathcal{Y}}) \approx \mathrm{s}_{\mathcal{XY}}$. Our work follows similar lines, but instead of treating the similarities $\mathrm{s}_{\mathcal{XY}}$ as direct regression targets, we employ a contrastive approach to impose only ordering or ranking constraints. Our approach is motivated by the nature of the data, shown in Figure 2a, where each curve corresponds to the sorted similarity values $\mathrm{s}_{\mathcal{AX}}$ of all images $\mathcal{X}$ with respect to an anchor image $\mathcal{A}$. We observe that the image captions are mostly equally distant from each other - represented by the narrow range from 0.6 to 0.8 in the middle of the plots. This is also corroborated in Figure 2b where the distribution of absolute differences in similarity peaks at 0 and steadily declines with the 99^{th} percentile occurring at 0.16. Thus, learning embeddings $\mathbf{f}_*$ with regression objectives using Siamese or Triplet architectures (Chopra, Hadsell, and LeCun 2005) are likely to lead to degenerate solutions.

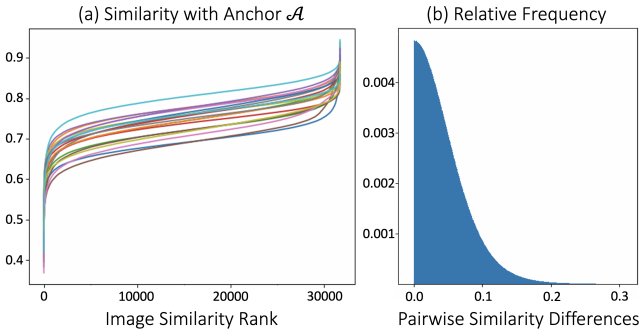

Figure 2: (a) Each of the 20 plots show the sorted similarities $\mathrm{s}_{\mathcal{AX}}\ \forall \mathcal{X}$ for randomly chosen anchor images $\mathcal{A}$ (b) Relative frequencies of the absolute values of all pairwise similarity differences $|\mathrm{s}_{\mathcal{AX}} - \mathrm{s}_{\mathcal{AY}}|\ \forall \mathcal{X}, \mathcal{Y}$ of the 20 selected anchors.

We rely on the text modality to only provide weak supervision, i.e., we expect that the image scene graphs contain complementary information, with the text captions only providing a guiding signal for the model's training. To this end, we impose a lenient requirement that $\mathrm{sim}(\mathbf{f}_{\mathcal{A}}, \mathbf{f}_{\mathcal{P}}) > \mathrm{sim}(\mathbf{f}_{\mathcal{A}}, \mathbf{f}_{\mathcal{N}})$ if $\mathrm{s}_{\mathcal{AP}} > \mathrm{s}_{\mathcal{AN}}$. This formulation invokes a set of three images $\langle \mathcal{A}, \mathcal{P}, \mathcal{N} \rangle$ similar to well-known losses in contrastive learning. However, we need to account for the fact that the similarity of a *positive* image $\mathcal{P}$ with respect to the *anchor* $\mathcal{A}$ might be very close to that of *negative* image $\mathcal{N}$. That is, $\mathrm{s}_{\mathcal{AP}}$ and $\mathrm{s}_{\mathcal{AN}}$ might occupy very similar regions in the density plot of $\mathrm{s}_{\mathcal{A}*}$. We therefore design a loss function that is tolerant to the selection of such samples during training.

Ranking Loss for Similarity Supervision

In this section, we describe a novel loss function to learn image embeddings that is specifically designed to optimize our model given a continuous space of similarities (or distances) between images, rather than discrete labels which is common in classification literature. Inspired by RankNet (Burges et al. 2005), we model the posterior probability $\hat{\mathrm{P}}$ of having similarities in the correct order as

$$\hat{\mathrm{P}}(\ \mathbf{f}_{\mathcal{A}}^T \mathbf{f}_{\mathcal{P}} > \mathbf{f}_{\mathcal{A}}^T \mathbf{f}_{\mathcal{N}}\ ) = \sigma\left(\frac{\mathbf{f}_{\mathcal{A}}^T \mathbf{f}_{\mathcal{P}} - \mathbf{f}_{\mathcal{A}}^T \mathbf{f}_{\mathcal{N}}}{\nu}\right)$$

where σ is the sigmoid function, ν is a temperature hyperparameter and we have utilized the inner product $\mathbf{f}_{\mathcal{A}}^T \mathbf{f}_{\mathcal{X}}$ for the similarity function $\mathrm{sim}(\mathbf{f}_{\mathcal{A}}, \mathbf{f}_{\mathcal{X}})$. For a given anchor $\mathcal{A}$, $\mathcal{P}$ (positive) and $\mathcal{N}$ (negative) are such that the pair $(\mathcal{A}, \mathcal{P})$ are expected to be more similar than $(\mathcal{A}, \mathcal{N})$. Since the corresponding embeddings $\mathbf{f}_*$ are ℓ_2-normalized, the inner products above correspond to using cosine similarity.

To reflect the constraints from relative similarities, we define the desired target value P as

$$\mathrm{P}(\ \mathrm{s}_{\mathcal{AP}} > \mathrm{s}_{\mathcal{AN}}\ ) = \frac{\mathrm{s}_{\mathcal{AP}}}{\mathrm{s}_{\mathcal{AP}} + \mathrm{s}_{\mathcal{AN}}}$$

where $\mathrm{s}_{\mathcal{AP}}$ and $\mathrm{s}_{\mathcal{AN}}$ denote the caption similarity of the anchor with the positive and negative respectively. If both negative and positive are sampled with high confidence such that $\mathrm{s}_{\mathcal{AP}} \gg \mathrm{s}_{\mathcal{AN}}$, then $\mathrm{P} \approx 1$. However, as explained in Figure 2, such samples are uncommon in our dataset. The proposed setup is efficient as it allows the use of samples where $\mathrm{s}_{\mathcal{AP}}$ is only marginally more than $\mathrm{s}_{\mathcal{AN}}$ with an appropriately weighted contribution to the objective. Hence, the use of non-binary targets offers an alternative to the explicit mining of positive and negative samples.

The final loss function, termed as Ranking loss hereafter, takes the form of a cross entropy and is given by

$$\mathcal{L} = -\mathrm{P} \log \hat{\mathrm{P}} - (1 - \mathrm{P}) \log\ (1 - \hat{\mathrm{P}})$$

Optimizing this loss function allows us to learn an embedding space in which the similarity between scene graph embeddings respects the ordering or ranking indicated in the similarity matrix. The advantages of this setup are: (1) the similarity values $\mathrm{s}_{\mathcal{A}*}$ are not assumed to be transitive or obey triangle inequalities, and (2) the actual magnitude of the similarities are not part of the supervision, only the relative values. Therefore, the Ranking loss imposes minimal requirements from the supervision signal.

Name	Loss Function
Triplet	$\max\left(\ \mathbf{f}_{\mathcal{A}}^T \mathbf{f}_{\mathcal{N}} - \mathbf{f}_{\mathcal{A}}^T \mathbf{f}_{\mathcal{P}} + m, 0\ \right)$
InfoNCE	$-\log \dfrac{\exp\left(\mathbf{f}_{\mathcal{A}}^T \mathbf{f}_{\mathcal{P}}/\lambda\right)}{\exp\left(\mathbf{f}_{\mathcal{A}}^T \mathbf{f}_{\mathcal{P}}/\lambda\right) + \exp\left(\mathbf{f}_{\mathcal{A}}^T \mathbf{f}_{\mathcal{N}}/\lambda\right)}$

Table 1: Loss functions used in contrastive learning

We compare the proposed loss function with other commonly used losses in the contrastive learning paradigm,

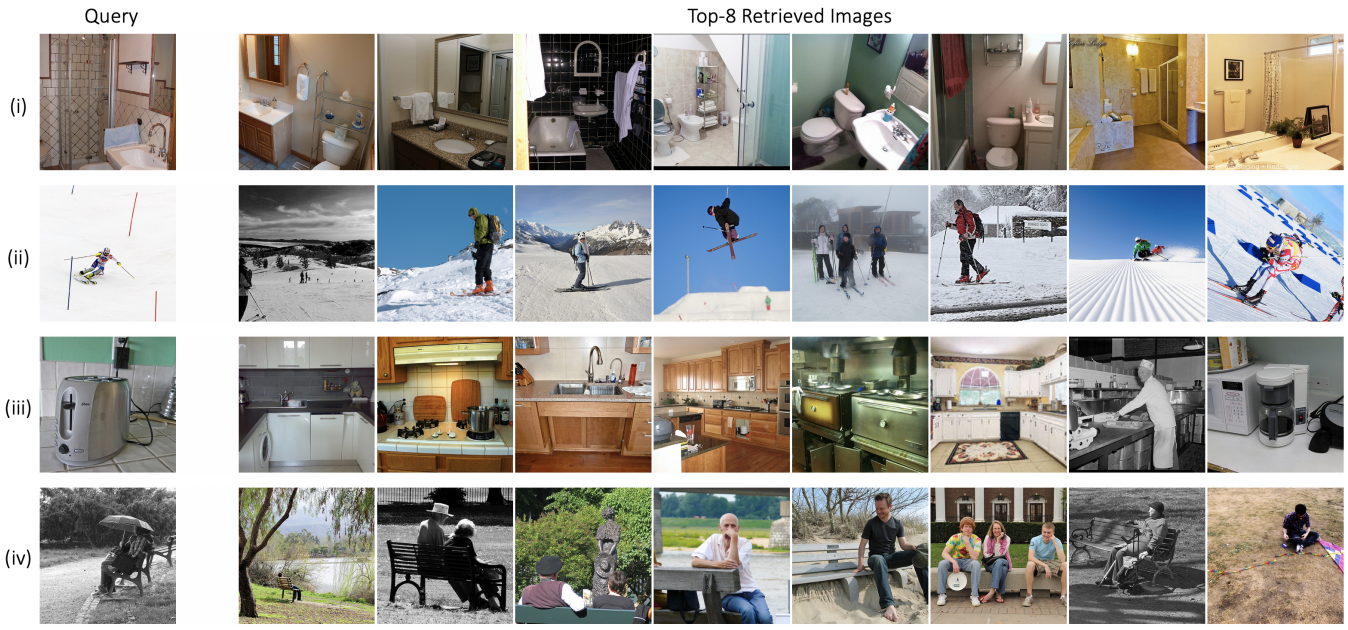

Figure 3: Qualitative examples showing Top-8 retrieved images of a given query using the proposed scene graph embeddings. Notice how the model is able to retrieve images of a kitchen in (iii) while the query image only contains a toaster. This can be attributed to the GCN framework which captures object co-occurrences in the scene graphs. In example (iv), the embeddings capture the global context of the query image - people sitting on a bench in an outdoor setting - while providing visual diversity.

namely Triplet or Margin loss (Weinberger, Blitzer, and Saul 2006; Schroff, Kalenichenko, and Philbin 2015) and InfoNCE (Oord, Li, and Vinyals 2018), shown in Table 1. Note that contrastive learning as defined by Chen et al. is equivalent to our formulation if: (i) Class labels are used to set $s_{\mathcal{AX}} = 1$ if the two images $\mathcal{A}$ & $\mathcal{X}$ belong to the same class and 0 otherwise (ii) The $\langle \mathcal{A}, \mathcal{P}, \mathcal{N} \rangle$ triples are chosen such that $s_{\mathcal{AP}} = 1$ and $s_{\mathcal{AN}} = 0$. We believe that our setup is naturally more robust to alternate ways of sampling the triples, and we experiment with different options described next.

Sampling Techniques

We provide different strategies to sample a positive $\mathcal{P}$ and negative $\mathcal{N}$ for a given anchor image $\mathcal{A}$. We do this by leveraging the caption similarities $\mathrm{s}_{\mathcal{AX}}$ of the anchor $\mathcal{A}$ with every other image $\mathcal{X}$. The sampling alternatives are:

1. **Random**: Given an anchor $\mathcal{A}$, we sample uniformly at random a positive-negative pair $\langle \mathcal{P}, \mathcal{N} \rangle$ from the set of all correctly-ordered pairs given by
$$\{ \langle \mathcal{P}', \mathcal{N}' \rangle \mid \mathrm{s}_{\mathcal{AP}'} > \mathrm{s}_{\mathcal{AN}'} \}$$
While this ensures that the positive is closer to the anchor than the negative, it does not consider the relative distances between them.
2. **Extreme**: For every anchor image $\mathcal{A}$, we pick the most similar image as the positive $\mathcal{P}$ and the most dissimilar image as the negative $\mathcal{N}$. Mathematically
$$\mathcal{P} = \arg\max_{\mathcal{P}'} \mathrm{s}_{\mathcal{AP}'} \qquad \mathcal{N} = \arg\min_{\mathcal{N}'} \mathrm{s}_{\mathcal{AN}'}$$
Note that this is a deterministic method, i.e., same positive and negative examples for a given anchor across epochs.
3. **Probability**: We sample the positive and negative based on their caption similarities with the anchor as
$$\mathrm{P}(\mathcal{P}) = \frac{\mathrm{s}_{\mathcal{AP}}}{\sum_{\mathcal{P}'} \mathrm{s}_{\mathcal{AP}'}} \qquad \mathrm{P}(\mathcal{N}) = \frac{1 - \mathrm{s}_{\mathcal{AN}}}{\sum_{\mathcal{N}'} (1 - \mathrm{s}_{\mathcal{AN}'})}$$
The upper limit of caption similarities $\mathrm{s}_{\mathcal{AX}}$ is 1 and therefore, $1 - \mathrm{s}_{\mathcal{AX}}$ is a measure of distance between images $\mathcal{A}$ and $\mathcal{X}$. This sampling captures the intuition that images closer (farther) to the anchor should have a higher probability of being sampled as the positive (negative).
4. **Reject**: Though infrequent, sampling based on similarities might lead to cases where $\mathcal{N}$ is closer to $\mathcal{A}$ than $\mathcal{P}$. In this method, we follow Probability sampling with an added constraint on the direction of pairwise similarities, i.e., rejecting samples where $\mathrm{s}_{\mathcal{AP}} < \mathrm{s}_{\mathcal{AN}}$.

The loss functions in Table 1 typically utilize strategies where *hard negatives* are coupled with a positive (Wang and Gupta 2015), or conversely *easy positives* alongside negatives (Levi et al. 2020) to aid the learning. The Ranking loss and sampling techniques described are designed to leverage as many of the N^2 positive-negative pairs for a given anchor as possible. Given the observed benefits of multiple negatives (Chen et al. 2020) and multiple positives (Khosla et al. 2020) per anchor, future work will look into adapting the methods above to handle the exhaustive range of triples.

Experiments

We describe the dataset, experimental setup and both qualitative and quantitative evaluation of our approach.

Model		Per Image (Row-wise) Evaluation			All Pairs Evaluation		
Objective	**Sampling**	**Kendall τ**	**Spearman ρ**	**Pearson r**	**Kendall τ**	**Spearman ρ**	**Pearson r**
Triplet	Random	0.258	0.375	0.402†	0.251	0.369	0.389†
	Extreme	0.133	0.197	0.264	0.148	0.220	0.274
	Probability	0.235	0.345	0.358	0.224	0.333	0.335
	Reject	0.269†	0.392†	0.388	0.254†	0.375†	0.364
InfoNCE	Random	0.263†	0.382†	0.407†	0.253†	0.372†	0.388†
	Extreme	0.034	0.048	0.049	0.062	0.081	0.081
	Probability	0.187	0.275	0.336	0.197	0.290	0.333
	Reject	0.225	0.329	0.376	0.228	0.337	0.367
Ranking	Random	0.381	0.539	0.548	0.366	0.524	0.525
	Extreme	0.207	0.302	0.361	0.209	0.309	0.359
	Probability	**0.391**†	**0.554**†	**0.549**†	**0.377**†	**0.540**†	**0.529**†
	Reject	0.380	0.537	0.538	0.362	0.520	0.511
Normal Features		0.003	0.004	0.024	0.001	0.002	0.017
Classification Features		0.260	0.378	0.417	0.247	0.363	0.382

Table 2: Evaluation of the models trained using different objective functions and sampling techniques. *Normal Features* denotes embeddings drawn at random from a normal distribution and *Classification Features* are pre-trained visual embeddings (He et al. 2016). The best results are highlighted in boldface. † indicates the best results for a given objective function.

Dataset: We work on the Visual Genome (Krishna et al. 2017) dataset which is a collection of $108,077$ images and their scene graphs. We use a subset of $51,498$ images which have a set of 5 user-provided textual captions in MS-COCO (Lin et al. 2014b). We also filter object and relationship types that occur at least 25 times, resulting in 2416 object and 478 relationship categories. We use images with minimum 3 and maximum 40 objects, and at least one relationship. This results in $45,358$ images with an average of 21 objects and 15 relationships per image. We divide the data into train, validation and test sets with a $70:20:10$ split. Further, we consider the 5 captions available for each image and embed them by taking the average of the constituent word embeddings (Pennington, Socher, and Manning 2014). The image similarities $\mathrm{s}_{\mathcal{XY}}$ are defined as the average of the $5\times5=25$ pairwise inner products over caption embeddings.

Implementation Details: Both objects and relationships are first embedded into a $d=300$ dimensional space using separate learnable layers. These are initialized with the average of constituent GloVe embeddings (Pennington, Socher, and Manning 2014). The intermediate messages for nodes are $h=512$ size vectors, while the final node and edge states of each layer are $D=300$ size vectors. For all multilayer perceptrons, we use ReLU activation and batch normalization (Ioffe and Szegedy 2015). The model consists of 5 GCN layers and is trained using Adam optimizer (Kingma and Ba 2014) for 100 epochs with learning rate 10^{-4} and batch size 16. The temperature parameter in InfoNCE and Ranking loss has been set as $\lambda=1$ and $\nu=1$ and the margin in Triplet loss as $m=0.5$. Training is performed on a Ubuntu 16.01 machine, using a single Tesla V100 GPU and PyTorch framework.

Evaluation: We compute the GCN output $\mathbf{f}_{\mathcal{X}}$ for every image X in the test set and derive the pairwise similarities as $\mathrm{sim}(\mathbf{f}_{\mathcal{X}},\mathbf{f}_{\mathcal{Y}})=\mathbf{f}_{\mathcal{X}}^T\mathbf{f}_{\mathcal{Y}}$. These scene graph similarities are compared against the corresponding caption equivalents $\mathrm{s}_{\mathcal{XY}}$ using Kendall rank correlation coefficient τ, Spearman's rank correlation coefficient ρ and Pearson correlation coefficient r. The 2 rank correlations are of primary importance as our model is trained on relative similarities, and not absolute values. We compute the metrics at two levels - per image (or row-wise) and across all pairs. The micro-averaged option of computing row-wise correlation between model-derived scene graph similarities $\mathrm{sim}(\mathbf{f}_{\mathcal{X}},\mathbf{f}_{*})$ and caption similarities $\mathrm{s}_{\mathcal{X}*}$ reflects the retrieval focus in this work.

The results are tabulated in Table 2 and we make the following observations: (a) Comparing across the 3 objective functions, it is evident that the proposed Ranking loss consistently outperforms the Triplet and InfoNCE alternatives for any sampling. The magnitude of improvements is particularly noticeable in the *All Pairs Evaluation*. (b) Amongst the 4 sampling methods, Random is a robust alternative across loss functions, while the deterministic sampling strategy, Extreme, performs the worst. Comparing between Probability and Reject sampling - the Triplet and InfoNCE losses that utilize binary labels perform better when coupled with Reject strategy. In Ranking loss, however, Probability sampling outperforms as it is based on soft target labels and can handle invalid triples (when the positive is further away from the anchor than the negative). (c) The best performing model, trained using Ranking loss and Probability sampling combination, has a Kendall τ of 0.391. Note that a perfect value of 1 would be undesirable as it indicates that the scene graph modality contains redundant information with respect to the textual captions. (d) The classification features (He et al. 2016) perform very competitively. This is particularly impressive given that they are pre-trained and not customized to the current task. The qualitative comparison in Figure 1 provides some intuition on how the two embeddings differ. Further, Figure 3 shows nearest neighbors of query images and indicates the viability of the scene graph embeddings.

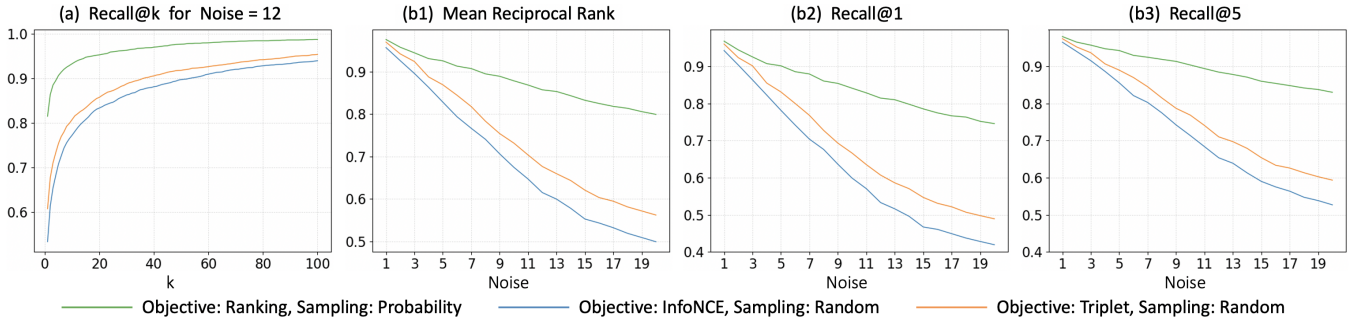

Figure 4: Retrieval performance for models trained on different objectives and corresponding best sampling: (a) Recall@k versus k for noise = 12 (b) Variation of the ranking metrics with increasing noise levels, from 1 to 20, in the query scene graph.

Image Retrieval

We design a retrieval setup to highlight the robustness of our scene graph embeddings to missing objects and relationships. We believe that our model trained over observed scene graphs produces embeddings that implicitly contain contextual information about objects that tend to co-occur and their relationships. To illustrate this, we consider every image in the test set and introduce increasing levels of noise in the scene graphs. We eliminate a set of M edges chosen at random from the scene graph and subsequently drop all isolated objects which are disconnected from the rest of the graph. This *noisy scene graph* is passed through the GCN model to get a query embedding which is issued against the test set. We examine the ranked list of items and evaluate the model's ability to retrieve the known image. The objective and sampling combination where the relevant image continues to be returned in top ranks despite noisy input is deemed to be most immune to incomplete information in the query.

The results for noise level $M = 12$ (median number of edges across scene graphs) are shown in Table 3. We compute the retrieval performance using standard metrics - Mean Reciprocal Rank (MRR), Recall@1 and Recall@5. It can be seen that Ranking loss generates embeddings that are significantly more effective than Triplet and InfoNCE losses. In fact, even the worst performing model of Ranking loss outperforms all variants of the other two losses. The best combination is the Ranking loss alongside the Probability sampling based on similarities. The target image is returned in top-5 ranks (out of 4537 in test set) in over 90% of the cases. The increased levels of recall are observed even as we go further down the ranked list, as shown in Figure 4(a).

For generalization, we sweep the noise parameter by progressively removing a chosen number of edges (and isolated objects) from the query scene graph up to a maximum of 20 edges (3^{rd} quartile for the number of edges across scene graphs). We compute the same metrics as before but restrict our attention to the best sampling strategy for each objective (marked with † in Table 3). The results are provided in Figure 4(b). It can be observed that the Ranking loss alongside Probability sampling performs the best across all three metrics. Despite removing 75% of edges in the query scene graph, a nearest neighbor search in the proposed embedding space places the target image at rank 1 in over 70% of the cases. This indicates the robustness of our scene graph representations and validates the model setup - a graph convolutional network to compute image embeddings, and the novel ranking loss that effectively utilizes pairwise similarity constraints as a weak supervision signal.

Model		Retrieval Metrics		
Objective	**Sampling**	**MRR**	**R@1**	**R@5**
Triplet	Random	0.676†	0.607†	0.753†
	Extreme	0.431	0.370	0.496
	Probability	0.208	0.138	0.265
	Reject	0.337	0.242	0.431
InfoNCE	Random	0.615†	0.533†	0.707†
	Extreme	0.005	0.002	0.005
	Probability	0.145	0.094	0.185
	Reject	0.576	0.502	0.655
Ranking	Random	0.710	0.643	0.789
	Extreme	0.700	0.648	0.758
	Probability	**0.857**†	**0.815**†	**0.906**†
	Reject	0.712	0.639	0.793

Table 3: Retrieval performance for different models. Queries are noisy scene graphs (noise = 12) and images are ranked on their similarities in the embedding space. The best results are in boldface. † indicates the best results for a given objective.

Conclusion and Discussion

We have considered the setting of retrieval of images based on their scene content. To do this, we obtained embeddings from ground-truth scene graphs of images using a graph convolutional network. This model was trained using a weak supervision signal of pairwise similarity preferences obtained from the text modality. We have proposed a loss function based on relative similarity labels, and have shown superior performance of the derived embeddings in a retrieval task. There are at least 2 promising future directions: (1) leveraging progress in scene graph generation literature allows the representation learning method to be applicable in a wide set of scenarios (2) the training objective based on weaker supervision requirements is general and relevant in other situations where classification-based labels are not available.

References

Barz, B.; and Denzler, J. 2019. Hierarchy-based image embeddings for semantic image retrieval. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 638–647. IEEE.

Belilovsky, E.; Blaschko, M.; Kiros, J.; Urtasun, R.; and Zemel, R. 2017. Joint embeddings of scene graphs and images. In *International Conference On Learning Representations-Workshop*.

Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; and Hullender, G. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, 89–96.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, 539–546. IEEE.

Deng, J.; Berg, A. C.; and Fei-Fei, L. 2011. Hierarchical semantic indexing for large scale image retrieval. In *CVPR 2011*, 785–792. IEEE.

Dhamo, H.; Farshad, A.; Laina, I.; Navab, N.; Hager, G. D.; Tombari, F.; and Rupprecht, C. 2020. Semantic Image Manipulation Using Scene Graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2121–2129.

Garcia, N.; and Vogiatzis, G. 2019. Learning non-metric visual similarity for image retrieval. *Image and Vision Computing* 82: 18–25.

Hare, J. S.; Lewis, P. H.; Enser, P. G.; and Sandom, C. J. 2006. Mind the Gap: Another look at the problem of the semantic gap in image retrieval. In *Multimedia Content Analysis, Management, and Retrieval 2006*, volume 6073, 607309. International Society for Optics and Photonics.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456. PMLR.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Johnson, J.; Krishna, R.; Stark, M.; Li, L.-J.; Shamma, D.; Bernstein, M.; and Fei-Fei, L. 2015. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3668–3678.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. *Advances in Neural Information Processing Systems* 33.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* .

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123(1): 32–73.

Krishnamurthy, J.; and Kollar, T. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics* 1: 193–206.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Levi, E.; Xiao, T.; Wang, X.; and Darrell, T. 2020. Reducing Class Collapse in Metric Learning with Easy Positive Sampling. *arXiv preprint arXiv:2006.05162* .

Li, D.; Lee, H.-Y.; Huang, J.-B.; Wang, S.; and Yang, M.-H. 2017. Learning structured semantic embeddings for visual recognition. *arXiv preprint arXiv:1706.01237* .

Li, K.; Zhang, Y.; Li, K.; Li, Y.; and Fu, Y. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4654–4662.

Lin, D.; Fidler, S.; Kong, C.; and Urtasun, R. 2014a. Visual semantic search: Retrieving videos via complex textual queries. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2657–2664.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014b. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *European conference on computer vision*, 852–869. Springer.

Norcliffe-Brown, W.; Vafeias, S.; and Parisot, S. 2018. Learning conditioned graph structures for interpretable visual question answering. In *Advances in neural information processing systems*, 8334–8343.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* .

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Raboh, M.; Herzig, R.; Berant, J.; Chechik, G.; and Globerson, A. 2020. Differentiable scene graphs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1488–1497.

Schroeder, B.; Tripathi, S.; and Tang, H. 2019. Triplet-aware scene graph embeddings. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Schuster, S.; Krishna, R.; Chang, A.; Fei-Fei, L.; and Manning, C. D. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, 70–80.

Shi, B.; Ji, L.; Lu, P.; Niu, Z.; and Duan, N. 2019. Knowledge Aware Semantic Concept Expansion for Image-Text Matching. In *IJCAI*, volume 1, 2.

Smeulders, A. W.; Worring, M.; Santini, S.; Gupta, A.; and Jain, R. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on pattern analysis and machine intelligence* 22(12): 1349–1380.

Teney, D.; Liu, L.; and van Den Hengel, A. 2017. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; and Wu, Y. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1386–1393.

Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5005–5013.

Wang, S.; Wang, R.; Yao, Z.; Shan, S.; and Chen, X. 2020. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *The IEEE Winter Conference on Applications of Computer Vision*, 1508–1517.

Wang, X.; and Gupta, A. 2015. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2794–2802.

Weinberger, K. Q.; Blitzer, J.; and Saul, L. K. 2006. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, 1473–1480.

Xu, P.; Chang, X.; Guo, L.; Huang, P.-Y.; Chen, X.; and Hauptmann, A. G. 2020. A Survey of Scene Graph: Generation and Application. *EasyChair Preprint* (3385).

Yan, Z.; Zhang, H.; Piramuthu, R.; Jagadeesh, V.; DeCoste, D.; Di, W.; and Yu, Y. 2015. HD-CNN: hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE international conference on computer vision*, 2740–2748.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10685–10694.