PhD Proposal:

Learning Local Representations of Images and Text

PhD candidate: Fuwen Tan
Advisor: Vicente Ordonez

Department of Computer Science
University of Virginia

December 2020

Abstract: Images and text inherently exhibit hierarchical structures, e.g. scenes built from objects, sentences built from words. In many computer vision and natural language processing tasks, learning accurate prediction models requires analyzing the correlation of the local primitives of both the input and output data. In this proposal, we aim to develop techniques for learning local representations of images and text and demonstrate their effectiveness on visual recognition, retrieval, and synthesis. In particular, the proposal includes three primary research projects: (1) Text2Scene, a sequence-to-sequence image synthesis framework which produces a scene depicted in a textual description by sequentially predicting objects, their locations, and their attributes such as sizes, aspect ratios. (2) DrillDown, an interactive image retrieval model which encodes multiple rounds of natural language queries with a region-aware state representation. (3) A newly proposed project which explores the task of instance-level image recognition/retrieval. The key ingredient of this work is a transformer based model that learns the visual similarity of an image-pair by incorporating both the global and local features of the images.

Contents

1 Introduction

1.1 Overview

Learning expressive representations of images and text is a core problem in computer vision and natural language processing. Depending on the desired outputs of the specific tasks, the representation may take on different forms. For category-level recognition, existing algorithms tend to represent the input data with global vector embeddings or discrete codes. However, these compact features are sub-optimal for "rich" prediction problems of which the outputs are beyond symbolic information. For example, prior approaches [70, 100] for the text-to-image synthesis task encode the input text description as a single vector, and leverage conditional Generative Adversarial Networks (GANs) [24] to perform pixel-wise image synthesis. While these works have managed to generate results of increasing quality, there are major challenges when attempting to synthesize images for complex scenes with multiple interacting objects. As another example, traditional image retrieval models [7, 68] make use of global descriptors of the images to perform nearest neighbor search in large-scale databases. Despite their simplicity, global descriptors discard the spatial arrangement of visual elements, which hinders them from learning patch-level matches between images.

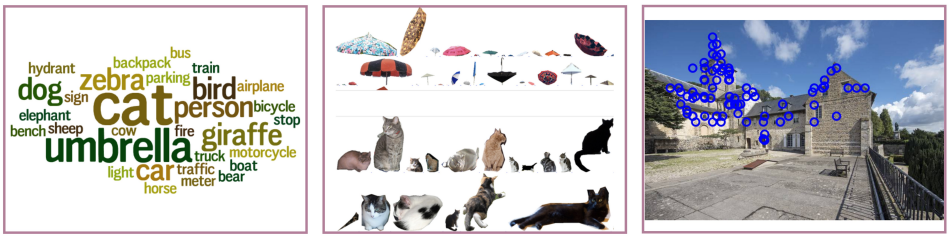

Figure 1: Local primitives of images and text. Complex linguistic expressions can be modeled as the composition of words. Complex visual scenes can be modeled as the composition of key points or segments

One feature shared by the visual and textual data is the capacity of representing complex concepts using a hierarchy of local primitives. As shown in Fig. 1, both images and text can be represented by various forms of constituent elements, e.g. characters/words for text, points/segments for images. In this work, we explore the representations of the local concepts in visual and textual data. We are particularly interested in rich visual prediction tasks that go beyond traditional image classification in the sense that the input and output data comprises fine-grained structures. My completed research demonstrates that modeling the local components of the input and output data and their interactions is crucial for reliable predictions in these tasks.

We discuss three primary research projects in this proposal: (1) Text2Scene, a sequence-to-sequence model that generates various forms of compositional scene representations from natural language descriptions. Unlike recent works for text-to-image synthesis, Text2Scene does not use Generative Adversarial Networks (GANs), but instead learns to sequentially generate objects and their attributes (location, size, appearance, etc) at every time step by attending to different parts of the input text and the current status of the generated scene. This work was published in CVPR 2019 [86]; (2) Drill-down, which explores the task of interactive image retrieval using natural language queries, where a user progressively provides input queries to refine a set of retrieval results. It supports the retrieval of complex image scenes with multiple objects by integrating multiple queries with a region-aware state representation that significantly extends previous methods for single-round image retrieval. This work was published in NeurIPS 2019 [85]; (3) A newly proposed project that studies the problem of instance-level image recog-

nition/retrieval where the goal is searching in a large database for images that match an specific object/scene instance in a query image. To address this task, systems usually rely on a retrieval step that uses global image features, and a subsequent step that performs domain-specific refinements or reranking by leveraging operations such as geometric verification based on local features. We propose Reranking Transformers (RRTs) as a general model to incorporate both global and keypoint based features to learn the matching images in a supervised fashion.

1.2 Local Representations in Vision and Language

Local features has been one of the most common visual representations in computer vision. Much of the progress for visual recognition before the "deep learning revolution" has built on local, or keypoint-based descriptors such as SIFT [53] and HOG [17]. Compared to global signatures [62], these descriptors are believed to be more invariant to image changes such as illumination, translation, occlusion and truncation. They were used in a wide variety of visual prediction tasks such as texture recognition [47], scene recognition [45], image matching [34], 3D reconstruction [1], etc. Part-based features [75, 99] were later introduced to model semantic visual concepts, e.g. classes, attributes, and relations. They were typically used in combination with statistical models for both pure visual recognition tasks and vision-and-language tasks. Famous examples include the Deformable Part Model [23] for object detection and the BabyTalk [43] system for image captioning. With the advent of deep learning, local representations extracted from Convolutional Neural Networks continue to play an important role in building high quality visual prediction models. Both grid-based [33] and region-based [3] features have shown promising performance on various vision-and-language tasks, e.g. image captioning [94], visual question answering [4], referring expression grounding [37], etc.

Compositional (part-based) representations have also been widely studied in linguistics and adjacent fields. Early systems [58, 80] explore incorporating explicit composition operations into vector-based systems. More recent approaches focus on learning distributed representations of natural language from large text corpora. Word2Vec [57] and GloVe [65] proposed to model the co-occurrence of words in both local and global contexts, while Socher et al [81] developed a unified framework to learn the hierarchy of words, phrases, and sentences using recursive neural networks. As a recent breakthrough in natural language processing, the Transformer [56] model learned contextualized text representations using a novel attention based sequence encoder.

1.3 Contributions

This proposal presents my completed research on learning local representations of images and text for visual synthesis/retrieval. It also introduces an on-going project tackling instance-level image recognition/retrieval by learning the correlation of the global and local representations extracted from an image-pair. The potential contributions made in this proposal include

- We develop an end-to-end trainable approach based on a sequence-to-sequence model to generate various forms of scene representations (e.g. abstract scenes, object layouts, composite images) from visually descriptive language. (Chapter 2)
- We present an effective framework for interactive retrieval of specific images of complex scenes. The method explores in depth and addresses several challenges in multiple round retrievals with natural language queries such as the need for region-aware features. (Chapter 3)
- We propose a novel model which learns to predict the visual similarity of an image-pair by analysing the correlation of both the global and keypoint-based representations using a transformer architecture. Preliminary results demonstrate its effectiveness in the context of reranking image search results. (Chapter 4)

2 Learning to Generate Compositional Scene Representations from Text

2.1 Introduction

As the first project in this thesis proposal, we introduce Text2Scene, a model to interpret visually descriptive language in order to generate compositional scene representations. We specifically focus on generating a scene representation consisting of a list of objects, along with their attributes (e.g. location, size, aspect ratio, pose, appearance). We adapt and train models to generate three types of scenes as shown in Figure 2, (1) Cartoon-like scenes as depicted in the Abstract Scenes dataset [102], (2) Object layouts corresponding to image scenes from the COCO dataset [51], and (3) Synthetic scenes corresponding to images in the COCO dataset [51]. Our method, unlike recent approaches, does not rely on Generative Adversarial Networks (GANs) [24]. Instead, we produce an interpretable model that iteratively generates a scene by predicting and adding new objects at each time step.

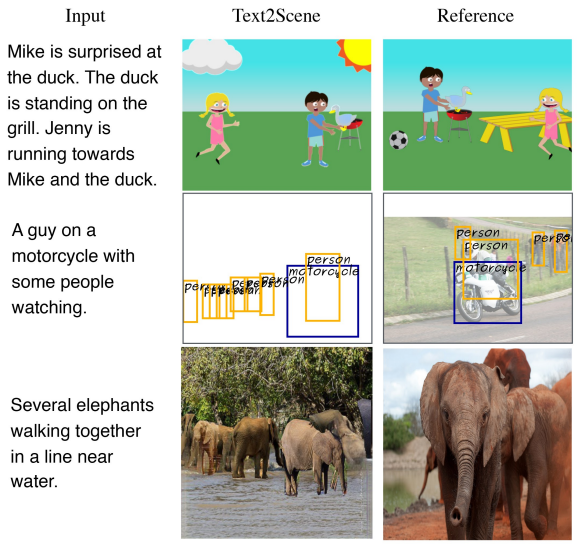

Figure 2: Sample inputs (left) and outputs of our Text2Scene model (middle), along with *ground truth* reference scenes (right) for the generation of abstract scenes (top), object layouts (middle), and synthetic image composites (bottom).

2.2 Related Work

Recent research on text-to-image synthesis [32, 35, 70, 71, 96, 100, 101] mainly leverage conditional Generative Adversarial Networks (GANs) [24]. While these works have managed to generate results of increasing quality, there are major challenges when attempting to synthesize images for complex scenes with multiple interacting objects without explicitly defining such interactions [97]. Most closely related to our approach are [27, 32, 35], and [39], as these works also attempt to predict explicit 2D layout representations. Johnson et al [35] proposed a graph-convolutional model to generate images from structured scene graphs. The presented objects and their relationships were provided as inputs in the scene graphs, while in our work, the presence of objects is inferred from text. Hong et al [32] targeted image synthesis using conditional GANs but unlike prior works, it generated layouts as intermediate representations in a separably trained module. Our work also attempts to predict layouts for photographic image synthesis but unlike [32], we generate pixel-level outputs using a semi-parametric retrieval module without adversarial training and demonstrate superior results. Kim et al [39] performed pictorial generation from chat logs, while our work uses text which is considerably more underspecified. Gupta et al [27] proposed a semi-parametric method to generate cartoon-like pictures. However the presented objects were also provided as inputs to the model, and the predictions of layouts, foregrounds and backgrounds were performed by separably trained modules. Our method is trained end-to-end and goes beyond cartoon-like scenes. To the best of our knowledge, our model is the first work targeting various types of scenes (e.g. abstract scenes, semantic layouts and composite images) under a unified framework.

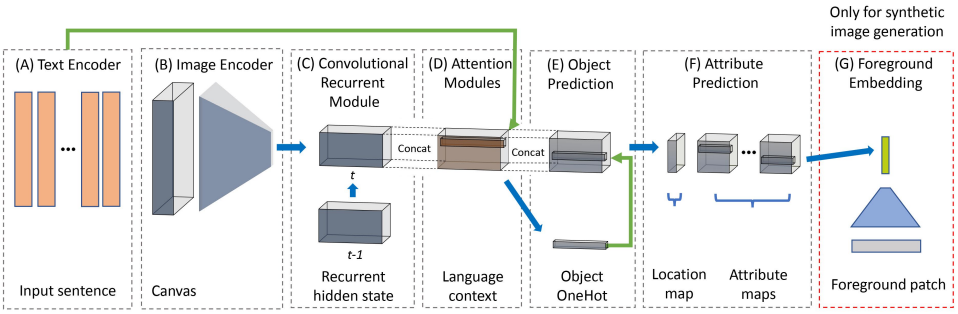

Figure 3: Overview of Text2Scene. Our general framework consists of (A) a Text Encoder that produces a sequential representation of the input, (B) an Image Encoder that encodes the current state of the generated scene, (C) a Convolutional Recurrent Module that tracks, for each spatial location, the history of what have been generated so far, (D-F) two attention-based predictors that sequentially focus on different parts of the input text, first to decide what object to place, then to decide what attributes to be assigned to the object, and (G) an optional foreground embedding step that learns an appearance vector for patch retrieval in the synthetic image generation task.

2.3 Text2Scene

Fig 3 provides an overview of the Text2Scene model. It consists of a text encoder (Fig 3 (A)) that maps the input sentence to a set of latent representations, an image encoder (Fig 3 (B)) which encodes the current generated canvas, a convolutional recurrent module (Fig 3 (C)), which passes the current state to the next step, attention modules (Fig 3 (D)) which focus on different parts of the input text, an object decoder (Fig 3 (E)) that predicts the next object conditioned on the current scene state and attended input text, and an attribute decoder (Fig 3 (F)) that assigns attributes to the predicted object.

The text encoder is a bidirectional recurrent network with Gated Recurrent Units (GRUs), which produces the feature embedding of each word in a given sentence. In the meantime, we use a convolutional network (CNN) to encode the current canvas into a 2D feature map, representing the current scene state. At each step, our model predicts the next object and its attributes from predefined object/attributes vocabularies using the concatenation of the text and scene features as input. For this part, we leverage a convolutional GRU (ConvGRU) to model the history of the scene states. The initial hidden state is created by spatially replicating the last hidden state of the text encoder.

With this 2D scene state, we first predict the next object that is depicted in the input text but missing in the current canvas. In doing this, our model predicts a 2D attention map which is used to reduce the scene state into a single vector. This vector summarizes the spatial context in the scene state where the next object may appear. We develop a text-based attention model to use this vector as the query to attend to the word features from the text encoder. The object is predicted by a simple perceptron model which takes the attended text feature as input.

The attribute set corresponding to the object can be predicted similarly. For each spatial location in the 2D scene state, the model predicts a location likelihood, and a set of attribute likelihoods using another attention module similar as in the object decoder. We predict a particular attribute: an appearance vector, only for the model trained to generate synthetic image composites (i.e. images composed of patches retrieved from other images). As with other attributes, the appearance vector is predicted for every location in the scene state which is used at test time to retrieve similar patches from a precomputed collection of object segments from other images.

2.4 Experiments

We perform experiments on three text-to-scene tasks:

Task (I): Clip-art Generation on Abstract Scenes. The Abstract Scene dataset is introduced by [102]. It comprises pictorial scenes of clip-art objects. The attributes we consider for each clip-art object are the location, size , and the direction the object is facing. For the person objects, we also explicitly model the pose and expression.

Task (II): Semantic Layout Generation on COCO. In this experiment, we make use of the captions and bounding box annotations provided by the COCO [51] dataset to define the semantic layout of a scene. The attributes we consider are location, size, and aspect ratio.

Task (III): Synthetic Image Generation on COCO. We also demonstrate our approach by generating synthetic image composites given input captions in COCO [51]. In addition to the semantic layout as in Task (II), our model also predicts appearance vectors which are used to retrieve object/stuff segments provided by the COCO [51] and COCO-Stuff [31] datasets.

Automatic Metrics. Our tasks pose new challenges on evaluating the models. For the abstract scene generation task, we draw inspiration from the evaluation metrics applied in machine translation [44] and propose to compute the following metrics: precision/recall on single objects (`U-obj`), "bigram" object pairs (`B-obj`); classification accuracies for poses, expressions; Euclidean distances (defined as a Gaussian function with a kernel size of 0.2) for normalized coordinates of `U-obj` and `B-obj`. A "bigram" object pair is defined as a pair of objects with overlapping bounding boxes. The method is evaluated against [102] and variants of our full model, such as Text2Scene (w/o attention), a model without any attention module. In the layout generation experiment, it is harder to define evaluation metrics given the complexity of real world object layouts. Inspired by [32], we employ caption generation as an extrinsic evaluation. We generate captions from the semantic layouts using [98] and compare them back to the original captions used to generate the scenes. We use commonly used metrics for captioning such as BLEU [63], METEOR [8], ROUGE_L [50], CIDEr [91] and SPICE [2]. For synthetic image generation, we adopt the Inception Score (IS) metric [76] which is commonly used in recent text to image generation methods. However, as IS does not evaluate correspondence between images and captions, we also employ an extrinsic evaluation using image captioning using the Show-and-Tell caption generator [94] as in [32]. The baselines for this task cover state-of-the-art text-to-image synthesis methods, such as SG2IM [35], StackGAN [100], HDGAN [101], AttnGAN [96]. We also perform human evaluations using crowdsourcing for the abstract scene and composite image generation tasks.

2.5 Results

Abstract Scenes and Semantic Layouts: Table 1 shows quantitative results on Abstract Scenes. Text2Scene improves over [102] and our variant on all metrics except `U-obj Coord`. Human evaluation results (the last column in Table 1) confirm the quality of our outputs, where the score values presented are the percentage of sentence-scene pairs marked by a human evaluator as a `true` entailment. The results also suggest that our proposed metrics correlate with human judgment on the task.

Table 2 shows an extrinsic evaluation on the layout generation task. We perform this evaluation by generating captions from our predicted layouts. Results show our full method generates the captions that are closest to the captions generated from true layouts.

Synthetic Image Composites: Table 3 shows evaluation of synthetic scenes using automatic metrics. Text2Scene without any post-processing already outperforms all previous methods by large margins except AttnGAN [96]. As our model adopts a composite image generation

Methods	U-obj		B-obj		Pose	Expr	U-obj	B-obj	Human
	Prec	Recall	Prec	Recall			Coord	Coord	Eval.
Zitnick et al. [102]	0.722	0.655	0.280	0.265	0.407	0.370	**0.449**	0.416	0.555
Text2Scene (w/o attention)	0.665	0.605	0.228	0.186	0.305	0.323	0.395	0.338	0.431
Text2Scene (full)	**0.760**	**0.698**	**0.348**	**0.301**	**0.418**	**0.375**	0.409	**0.483**	**0.644**

Table 1: Evaluations on the Abstract Scenes dataset. Our full model performs better in all metrics except U-obj Coord which measures exact object coordinates.

Methods	B1	B2	B3	B4	METEOR	ROUGE	CIDEr	SPICE
Captioning from True Layout [98]	0.678	0.492	0.348	0.248	0.227	0.495	0.838	0.160
Text2Scene (w/o attention)	0.591	0.391	0.254	0.169	0.179	0.430	0.531	0.110
Text2Scene (full)	**0.615**	**0.415**	**0.275**	**0.185**	**0.189**	**0.446**	**0.601**	**0.123**

Table 2: Quantitative evaluation on the layout generation task. Our full model generates more accurate captions from the generated layouts than the baselines. We also include caption generation results from ground truth layouts as an upper bound on this task.

framework without adversarial training, gaps between adjacent patches may result in unnaturally shaded areas. We observe that, after performing a regression-based inpainting [67], the composite outputs achieve consistent improvements on all automatic metrics. We posit that our model can be further improved by incorporating more robust post-processing or in combination with GAN-based methods. On the other hand, human evaluations show that our method significantly outperforms alternative approaches including AttnGAN [96], demonstrating the potential of leveraging realistic image patches for text-to-image generation. It is important to note that SG2IM [35] and Hong et al [32] also use segment and bounding box supervision – as does our method, and AttnGAN [96] uses an Imagenet (ILSVRC) pretrained Inceptionv3 network. In addition, as our model contains a patch retrieval module, it is important that the model does not generate a synthetic image by simply retrieving patches from a single training image. On average, each composite image generated from our model contains 8.15 patches from 7.38 different source images, demonstrating that the model does not simply learn a global image retrieval.

	Ratio
Text2Scene > SG2IM [35]	0.7672
Text2Scene > HDGAN [101]	0.8692
Text2Scene > AttnGAN [96]	0.7588

Table 4: Human evaluation on the synthetic image generation task.

2.6 Summary

In this chapter, we present a novel sequence-to-sequence model for generating compositional scene representations from visually descriptive language. We provide extensive quantitative analysis of our model for different scene generation tasks on datasets from two different domains: Abstract Scenes [102] and COCO [51]. Experimental results demonstrate the capacity of our model to capture finer semantic concepts from visually descriptive text and generate complex scenes.

3 Learning to Retrieve Complex Scenes with Region-aware Representations

3.1 Introduction

In this chapter, we focus on learning region-aware visual and textual representations for text-to-image retrieval. Retrieving images from text-based queries has been an active area of research. Significant improvement has been achieved over the past years with advances in representation

Methods	IS	B1	B2	B3	B4	METEOR	ROUGE	CIDEr	SPICE
Real image	36.00±0.7	0.730	0.563	0.428	0.327	0.262	0.545	1.012	0.188
SG2IM [35]	6.7±0.1	0.504	0.294	0.178	0.116	0.141	0.373	0.289	0.070
StackGAN [100]	10.62±0.19	0.486	0.278	0.166	0.106	0.130	0.360	0.216	0.057
HDGAN [101]	11.86±0.18	0.489	0.284	0.173	0.112	0.132	0.363	0.225	0.060
Hong et al [32]	11.46±0.09	0.541	0.332	0.199	0.122	0.154	–	0.367	–
AttnGAN [96]	**25.89±0.47**	**0.640**	**0.455**	**0.324**	**0.235**	**0.213**	**0.474**	**0.693**	**0.141**
Text2Scene (w/o inpaint.)	22.33±1.58	0.602	0.412	0.288	0.207	0.196	0.448	0.624	0.126
Text2Scene (w inpaint.)	24.77±1.59	0.614	0.426	0.300	0.218	0.201	0.457	0.656	0.130

Table 3: Quantitative evaluation on the synthetic image generation task. Our model is superior on automated metrics than all competing approaches except AttnGan.

Figure 4: An example of the interactive image retrieval with our Drill-down model, where a user generated query (Ut) progressively refines the search results (St) until the target image is among top search results.

learning but finding very specific images with detailed specifications remains challenging. We focus on a scenario where a user is trying to find an exact image, or similarly where the user has a very specific idea of a target image, or is deciding on-the-fly while querying. An example of this type of interaction is shown in Fig. 4. We propose Drill-down, an interactive image search framework for retrieving complex scenes, which learns to capture the fine-grained alignments between images and multiple text queries. Our work is inspired by the observations that: (1) user queries at each turn may not exhaustively describe all the details of the target image, but focus on some local regions, which provide a natural decomposition of the whole scene. Therefore, we explicitly represent images as a set of object/stuff level features; (2) complex scenes contain multiple objects that might share the same feature subspace. Existing state representations of sequential text queries condense all image properties in a single state vector, which makes it difficult to distinguish entities sharing the same feature subspace, such as multiple `person` instances. To address this, we maintain a set of compositional state vectors, encouraging each of the vectors to encode text queries corresponding to a distinct image region.

3.2 Related Work

Text-based image retrieval has been an active research topic for decades [12, 13, 74]. Kovashka et al [40, 41] proposed using user feedback based on individual visual attributes to progressively improve search results. Arandjelovic et al [5] proposed a multiple query retrieval system that was used for querying specific objects within a large set of images. These works show that multiple independent queries generally outperform methods that jointly model the input set with a single query. Since the deep learning revolution, dialog based image search systems using deep-learned features [26, 49] were also introduced. Liao et al [49] proposed to aggregate multi-round user responses from trained agents or human agents in order to iteratively refine a retrieved set of images using a hierarchical recurrent encoder-decoder framework [77]. Guo et al [26] used multiple rounds of natural language queries, and proposed collecting relative image captions as supervision for a product search task. Also relevant to our research are the existing works on learning image-word [28, 36, 46] or region-phrase [60] alignments for vision-language tasks. Karpathy et al [36] proposed to learn a bidirectional image-sentence mapping

by jointly embedding fragments of images (objects) and sentences. Niu et al [60] extended this work by jointly learning hierarchical relations between phrases and image regions in an iterative refinement framework. Lee et al [46] developed a stacked cross attention network for word-region matching. More closely related to our work are Memory Networks [38, 83, 92], which perform `query` and possibly `update` operations on a predefined memory space. In contrast to this line of research, we explore a more challenging scenario where the model needs to `create` and `update` the memory (i.e. the state vectors) on-the-fly so as to maintain the states of the queries.

3.3 Drill-down

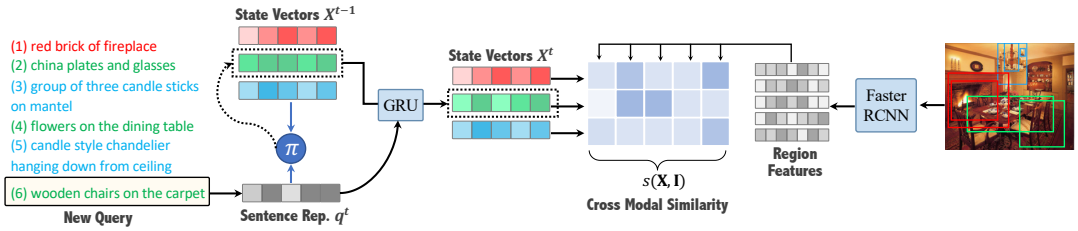

Figure 5: Overview of our model. Drill-down maintains a fixed set of state vectors $\mathbf{X}$, modeling the historical context of the user queries. Given a new query $\mathbf{q}^t$, our model selects and updates one of the state vectors. The updated state vectors $\mathbf{X}^t$ and image region features are then projected to a cross-modal embedding space to measure the fine-grained alignment between each region-state pair.

Our model is inspired by the observation that users naturally underspecify in their queries by referring to local regions of the target image. We aim to capture these region level alignments by learning to map text queries and the target image into two sets of compositional vector representations, and computing the matching score by measuring similarities between them. Figure 5 provides an overview of our model.

Image representation. To identify candidate regions referred in the queries, we follow [3, 46]. For each image, we detect the potential objects and salient regions using the FasterRCNN detector [72]. Corresponding features are extracted from the ROI pooling layer of the detector.

Query representation. Supporting multi-round retrieval requires a state representation for integrating the queries from multiple turns. We propose to maintain a set of latent representations for multiple turn queries. While users might provide a general image description in the first round of querying, subsequent queries typically describe more specific regions. An ideal set of latent vectors should learn to group and encode the input queries into visually discriminative representations referring to distinct image regions.

Cross modal similarity. To measure the similarity of visual and language representations, we leverage a distance metric similar with the cross stacked attention module proposed in [46]. The main difference is, the input text features for [46] are a list of word vectors with a dynamic length. Our method integrates multiple sentences into a set of latent vectors with a fixed length.

Query encoding. Each query sentence is first encoded into an embedding vector via a uni-directional recurrent network with gated recurrent units (GRU). Given the assumption that each text query describes a sub-region of the image, each query only updates a subset of the state vectors. In this work, we focus on a simplified scenario where each query only updates a single state vector. In detail, we develop a policy function that takes that the text query at each time step as input. This function produces a score for each state vector, indicating its probability of being updated by the input query. The state vector with the highest score is

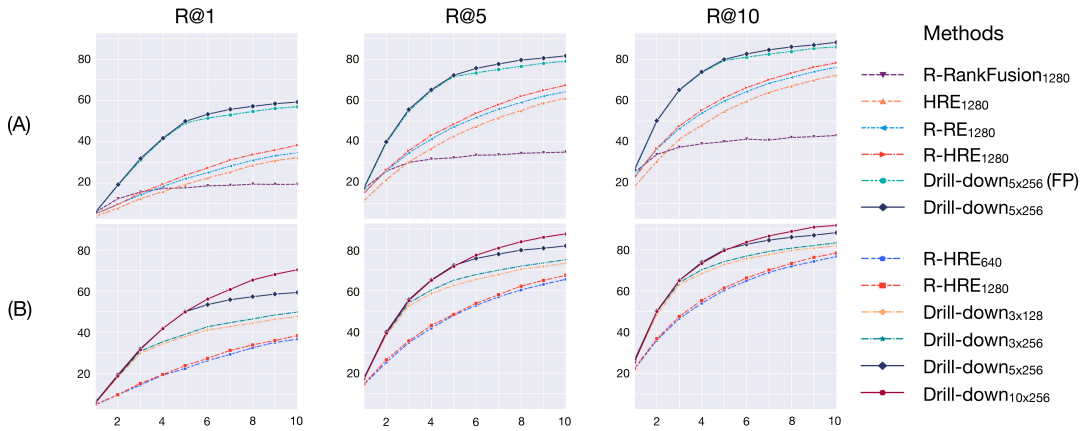

Figure 6: Quantitative evaluation of our models and the baselines. (A) Comparison of models using query representations of the same memory size; (B) Comparison of the models using query representations of different memory sizes. The horizontal axis represents the query turn.

updated by the input query using sing a single uni-directional gated recurrent unit cell (GRU Cell). Note that our formulation is similar to a hard attention module [95].

Training. The full model, including the feature learning/update and the policy function can be trained jointly by the policy gradient algorithm [84] using the negative of the cross modal similarity score as the reward. Following the paradigm proposed in [26, 73], we first pretrain all the modules except the policy function using a supervised embedding loss, then jointly optimize all the modules using policy gradient.

3.4 Experiments

Dataset We evaluate the performance of our method on the Visual Genome dataset [42]. Each image in Visual Genome is annotated with multiple region captions. We preprocess the data by removing duplicate region captions (e.g. multiple captions that are exactly the same), and images with less than 10 region captions. We use region captions as queries to train our model, thus bypassing the challenging issue of data collection for this task.

Baselines We compare our method with four baseline models: (1) **HRE**: a hierarchical recurrent encoder network, which is commonly adopted by recent dialog based approaches [26, 49, 82]. We consider the framework using text queries as context, which consists of a sentence encoder, a context encoder and an image encoder. The sentence encoder has the same word and sentence embedding as the proposed model. The image encoder maps the mean-pooled features of ResNet152 [30] into a one-dimensional feature vector via a linear projection. The ResNet model is pre-trained on ImageNet [18]. The model is trained to optimize the cosine similarity between the text and image features by a triplet loss with hard negatives as in [22]. (2) **R-HRE**: a model similar to baseline (1) but is trained with the region features, as in the proposed method. (3) **R-RE**: a model similar to baseline (2) but instead of using a hierarchical text encoder, this baseline uses a single uni-directional GRU network which encodes the concatenation of the queries. (4) **R-RankFusion**: a model where each query is encoded by a uni-directional GRU network and each image is represented as a set of region features. The ranks of all images are computed separably for each turn. The final ranks of the images are represented as the averages of the per-turn ranks.

Evaluation metrics To measure the retrieval performance, we use the common R@K metric, i.e., recall at K - the ratio of queries for which the target image is among the top-K retrieved images. The R@1, R@5 and R@10 scores at each turn are reported as shown in Fig. 6.

3.5 Results

Simulated user queries. Due to the lack of existing benchmarks for multiple turn image retrieval, we use the annotated region captions in Visual Genome to mimic the user queries. As region captions focus more on invariant information, such as image contents, and convey fewer irrelevant signals, such as different speaking/writing styles, they could be seen as the common "abstracts" of real queries in different forms. While we agree that strong supervisory signals such as real user queries could bridge the domain gap and would like to explore further in this direction, we choose at this stage to use only "weak but free" signals and investigate their potentials of being generalized to real scenarios. First, we compare our method against the baseline models when using query representations of the same memory size. In particular, we use 5 state vectors in our model ($M = 5$), each with a dimension of 256. Accordingly, the baseline models use a 1280-d query vector. Figure 6(A) shows the per-turn performance of the models on the test set. Both the R-RE$_{1280}$ and R-HRE$_{1280}$ baselines perform better than the HRE$_{1280}$ model, demonstrating the benefit of incorporating region features. R-HRE$_{1280}$ is superior to R-RE$_{1280}$, demonstrating the benefit of hierarchical context encoding. R-RankFusion$_{1280}$ performs inferior to all other models. Note that it also requires more memory to store the ranks of all images at each turn. Our model significantly outperforms all baselines by a large margin. On the other hand, we observe that the performance of our model will degrade when different queries have to share the same state vector.

Real user queries. We evaluate our method with the queries from crowdsourced human users via a multi-round interactive system adapted from [11]. Given a target image, a user is asked to search for it by providing descriptions of the image content. The system shows top-5 retrieved images to the user per turn as context so that the user can improve the results by providing additional descriptions. This process is repeated until the image is found or it reaches 5 turns. We sample 80 random images from the test set and evaluate HRED$_{1280}$, R-HRED$_{1280}$ and Drill-down$_{3\times256}$ on these images respectively. Each image is viewed by 3 different users. For each model, the best result on each image is selected across users to ensure high quality responses. As shown in Figure 7, most users ($> 80\%$) successfully find the target image within 5 turns, demonstrating the effectiveness of the multi-round search paradigm and the quality of using region captions for training. In particular, Drill-down$_{3\times256}$ consistently outperforms HRE$_{1280}$ and R-HRE$_{1280}$ on all evaluation metrics. On the other hand, as real user queries have more flexible forms, e.g. longer sentences, repeated descriptions of the same region, etc, we also observe smaller performance gaps between our method and the baselines. We believe further efforts such as real query data collection are needed to systematically fill this domain gap.

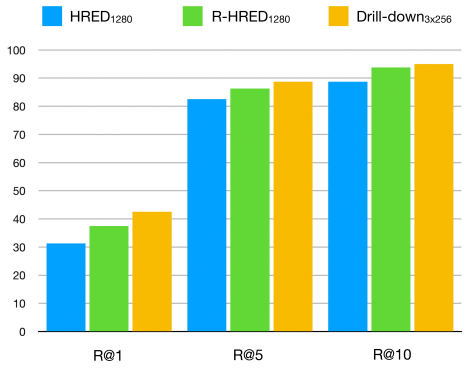

Figure 7: Human subject evaluation of the HRE$_{1280}$, R-HRE$_{1280}$ baselines and our Drill-down$_{3\times256}$ model.

3.6 Summary

In this chapter, we present an efficient and effective framework for interactive retrieval of complex scene images. Particularly, we propose a novel region-aware representation of multiple round text queries and demonstrate its effectiveness using both simulated and real user inputs.

4 Learning Visual Similarity of Images using Global and Local Descriptors

4.1 Introduction

In the Text2Scene and Drill-down projects, we leverage part-based representations to model the correlation between visual and language data. To complete the dissertation, I plan to perform research on learning visual correlation of an image-pair using local-based representations. Particularly, we study the problem of instance-level recognition/retrieval.

Distinct from category-level recognition where the goal is to identify an object class, instance recognition aims to identify a particular object instance. As the number of possible labels can be very large, instance recognition is typically cast as image retrieval instead of classification , and usually involves both metric learning and local feature matching strategies for reranking. Specifically, recent approaches incorporate both global and local descriptors extracted from deep learning models [7, 61], where the global descriptor is used to reduce the search space and the local descriptors are used to *re-rank* the nearest images. The dominant solution for the reranking task is geometric verification [66].

We propose a *Reranking Transformer* (RRT), which learns to predict the similarity of an image-pair using global and local features. Similar with geometric verification, Reranking Transformers aim to learn the visual relation of an image-pair but with a more straightforward pipeline: it directly predicts a similarity score of the matching images, instead of estimating a homography. The proposed method leverages the transformer architecture [90] which has led to significant improvements in several natural language processing [19, 52] and vision-and-language tasks [14, 48, 54]. Most recently, it has also been used for pure visual tasks, notably for image synthesis [64], recognition [20] and object detection [10]. To the best of our knowledge, our proposed work is the first to adopt transformers for a purely visual task involving the analysis of image pairs in the context of reranking image search results. The proposed method has several potential advantages:

- It is lightweight. Compared with the CNN feature extractors which have over 20 million parameters (e.g. 25 million in ResNet 50), the proposed model has 2.2 million parameters.
- It can be easily parallelized such that re-ranking the top-100 neighbors requires only a single forward-pass, allowing for more efficient model inference.
- It can potentially be jointly optimized with the CNN feature extractor, which may lead to feature representations tailored to downstream tasks and further accuracy improvements.

4.2 Related Work

Global/Local features for instance recognition/retrieval. Local descriptors, either handcrafted [53] or extracted from from convolution neural networks (CNN) [21, 61, 78], are widely used for instance image recognition/retrieval. Pioneering systems include [59, 79]. More recent approaches propose to detects local features from a CNN based feature extractor by performing a non-local maximum suppression [21, 87], or leveraging an extra attention module [9, 61]. On the other hand, a global descriptor has the advantage of providing a compact representation of an image, facilitating large-scale retrieval. Most of the existing global descriptors are extracted from CNN based models [7, 25, 68, 89] by spatially pooling two-dimensional feature responses [6, 68, 89]. These pooling operations may hinder global descriptors from modeling complex spatial relationship among image regions.

Reranking for instance recognition/retrieval. Compared with global/local feature learning, image search by reranking is less explored. The classic geometric verification approach is widely used in both traditional [66] and the most recent work [9, 61, 78]. Geometric

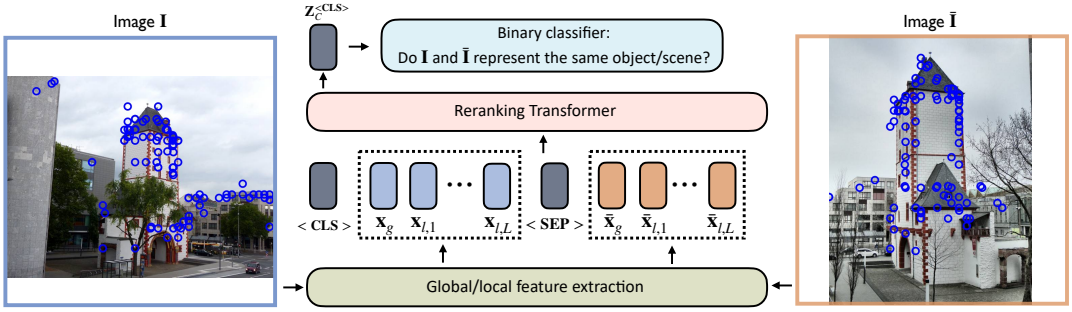

Figure 8: Illustration of the proposed *Reranking Transformer* (RRT) model. The input of RRT is a sequence of global and local descriptors (circled in blue) extracted from an image-pair. This sequence is fed into a multi-layer transformer model which produces a similarity score.

verification assumes rigid objects and seeks to estimate a linear transformation between images by iteratively aligning local descriptors from each image. Inspired by text retrieval, query expansion techniques have also been introduced for image retrieval [15, 16, 88]. These methods differ from geometric verification and our work as they rely on analysing the nearest neighbor graph built on the query and gallery images during testing.

Transformers for visual tasks. Transformers have become the dominant model architecture in natural language processing [19, 52]. Recently, it has also been introduced to vision-and-language [48, 54] and pure vision tasks [10, 64]. Parmar et.al. [64] develop a transformer based autoregressive model for image synthesis. Carion et.al. [10] casts object detection as a direct set prediction problem using transformers. All these prior works explore the application of transformers for making single image predictions while we leverage transformers to learn the visual relation of an image-pair in the context of reranking image search results.

4.3 The Proposed Method

Fig. 8 provides an illustration of the proposed model. We follow the transformer architecture introduced in [90], which takes a sequence of feature vectors as input. In the proposed model, the feature vectors are the global and local descriptors extracted from the image-pair.

Image representations. An image $\mathbf{I}$ is represented by a global descriptor: $\mathbf{x}_g \in \mathbb{R}^{d_g}$ and a set of L local descriptors: $\mathbf{x}_l = \{\mathbf{x}_{l,i} \in \mathbb{R}^{d_l}\}_{i=1}^{L}$. Both $\mathbf{x}_g$ and $\mathbf{x}_l$ are extracted from a CNN backbone. Optionally, each $\mathbf{x}_{l,i}$ is associated with a coordinate tuple $\mathbf{p}_{l,i} = (u, v) \in \mathbb{R}^2$ and a scale factor $s_{l,i} \in \mathbb{R}$, indicating the pixel location and image scale where $\mathbf{x}_{l,i}$ is extracted from. In our case, $s_{l,i}$ is an integer, representing the index of a set of pre-defined image scales.

Input. Following the transformer encoder proposed in [19], we define the input sequence of features, extracted from an image-pair $(\mathbf{I}, \bar{\mathbf{I}})$, as:

$$\begin{aligned}\mathbf{X}(\mathbf{I}, \bar{\mathbf{I}}) := [\ &\langle\text{CLS}\rangle; f_g(\mathbf{x}_g); f_l(\mathbf{x}_{l,1}); \cdots; f_l(\mathbf{x}_{l,L}); \\ &\langle\text{SEP}\rangle; \bar{f}_g(\bar{\mathbf{x}}_g); \bar{f}_l(\bar{\mathbf{x}}_{l,1}); \cdots; \bar{f}_l(\bar{\mathbf{x}}_{l,L})\ ],\end{aligned} \tag{1}$$

where:

$$\begin{aligned}f_g(\mathbf{x}_g) &:= \mathbf{x}_g + \alpha; \\ f_l(\mathbf{x}_{l,i}) &:= \mathbf{x}_{l,i} + \varphi(\mathbf{p}_{l,i}) + \psi(s_{l,i}) + \beta \\ \bar{f}_g(\bar{\mathbf{x}}_g) &:= \bar{\mathbf{x}}_g + \bar{\alpha}; \\ \bar{f}_l(\bar{\mathbf{x}}_{l,i}) &:= \bar{\mathbf{x}}_{l,i} + \varphi(\bar{\mathbf{p}}_{l,i}) + \psi(\bar{s}_{l,i}) + \bar{\beta}.\end{aligned} \tag{2}$$

Here, $\langle\text{CLS}\rangle$ a special token used for summarizing the signals from both images. $\langle\text{SEP}\rangle$ is an extra separator token. $\alpha, \bar{\alpha}, \beta, \bar{\beta}$ are one dimensional segment embeddings, being used to

distinguish the global and local descriptors of $\mathbf{I}$ and $\bar{\mathbf{I}}$. φ is a linear position embedding, as used in [10]. ψ is a linear embedding taking the scale index $s_{l,i}$ as input.

Model architecture. We leverage the standard architecture defined in [55], which comprises C transformer layers. The output of the model is a set of new vectors $\mathbf{Z}_{C,k}$, which is of the same length as the input sequence.

Training objective. The proposed model is trained to optimize a binary cross entropy loss:

$$\mathrm{E}(\mathbf{I}, \bar{\mathbf{I}}) = \mathrm{BCE}(\textsc{sigmoid}(\mathbf{Z}_{C,\langle\mathrm{CLS}\rangle} W_z^T), \mathbb{1}(\mathbf{I}, \bar{\mathbf{I}})), \tag{3}$$

where $\mathbf{Z}_{C,\langle\mathrm{CLS}\rangle}$ is a one-dimensional feature vector, corresponding to the input token $\langle\mathrm{CLS}\rangle$. It is extracted from the last transformer layer. $W_z^T \in \mathbb{R}^{(hd)\times 1}$ is a linear function mapping $\mathbf{Z}_C^{\langle\mathrm{CLS}\rangle}$ into a logit scalar. $\mathbb{1}(\mathbf{I}, \bar{\mathbf{I}})$ is an indicator function which equals to one when $\mathbf{I}$ and $\bar{\mathbf{I}}$ represent the same object/scene, or zero otherwise.

4.4 Preliminary Experiments

Dataset. We perform preliminary experiments on three instance recognition benchmarks: Google Landmarks v2 [93], Revisited Oxford/Paris [69]. Google Landmarks v2 (GLDv2) [93] is a large-scale dataset which includes over five millions images from 200k natural landmarks. We sample a small subset of the images from the "v2-clean" split of GLDv2 for training. Revisited Oxford ($\mathcal{R}$Oxf) and Paris ($\mathcal{R}$Par) [69] are typically used as the evaluation sets for instance recognition/retrieval, which have 4,993 and 6,322 gallery images respectively. They both have 70 query images. An extra distractor set ($\mathcal{R}$1M) with 1,001,001 images is also included for large-scale experiments. We report mean Average Precision (mAP) on the Medium (+$\mathcal{R}$1M) and Hard (+$\mathcal{R}$1M) setups.

Implementation. At this stage, we mainly focuses on similarity learning rather than feature learning, we leverage image descriptors obtained from state-of-the-art feature extractors. In particular, we use the DELG models provided by [9] with ResNet50 [29] as the CNN backbone. DELG provides a unified framework for global/local feature extraction. The local descriptors are extracted from 7 image scales ranging from 0.25 to 2.0. The global descriptor is extracted from 3 image scales: $\{\frac{1}{\sqrt{2}}, 1, \sqrt{2}\}$. In the original DELG model, each local descriptor comes with an attention score. The top 1000 local descriptors with the highest attention scores are selected for image reranking. In our experiments we choose the top 500 local descriptors. During training, the positive image for the query is randomly sampled from the images sharing the same label as the query. the negative image is randomly sampled from the top-100 neighbors returned by the global retrieval, which have a different label from the query.

Figure 9: Qualitative examples from Revisited Oxford [69]. For each query, the top-3 neighbors predicted by geometry verification and the proposed Reranking Transformer are presented. Correct/incorrect neighbors are marked with green/red borders.

Comparison with Geometric Verification. As the main baseline, geometric verification (GV) has a very similar goal as our approach. We perform experiments on comparing GV and RRT using the same set of descriptors. We follow the protocol in DELG [9]: given a query, we use its global descriptor to retrieve a set of top-ranked images. The top-100 neighbors are reranked by GV and RRT. On $\mathcal{R}$Oxf and $\mathcal{R}$Par, both GV and RRT significantly outperform

Method	Medium				Hard			
	$\mathcal{R}$Oxf	+$\mathcal{R}$1M	$\mathcal{R}$Par	+$\mathcal{R}$1M	$\mathcal{R}$Oxf	+$\mathcal{R}$1M	$\mathcal{R}$Par	+$\mathcal{R}$1M
DELG global	69.7	55.0	81.6	59.7	45.1	27.8	63.4	34.1
GV	75.4	61.1	82.3	60.5	54.2	36.8	64.9	34.8
RRT (ours)	**75.53**	**61.23**	**82.68**	**60.70**	**56.35**	**37.02**	**68.56**	**37.54**

Table 5: Comparison to geometric verification on Revisited Oxford/Paris [69]. The mAP scores on the Medium (+$\mathcal{R}$1M) and Hard (+$\mathcal{R}$1M) setups are reported.

global-only retrieval, as shown in Table 5. RRT shows further advantages over GV. On $\mathcal{R}$Oxf (+$\mathcal{R}$1M), RRT performs on par with GV on the Medium setup and consistently better on the Hard setup. On $\mathcal{R}$Par (+$\mathcal{R}$1M), RRT consistently outperforms GV. The largest performance gap appears on the Hard setup. RRT obtains 2.15 (3.66) absolute improvements over GV on $\mathcal{R}$Oxf ($\mathcal{R}$Par). We posit that, while GV is very effective for sufficiently similar images, it has difficulty handling challenging cases, e.g. large variations in viewpoint. Fig. 9 provides two qualitative examples comparing the reranking results of geometry verification and the proposed method. The queries in example (A) and (B) represent the same landmark but exhibit a large viewpoint change. While geometry verification predicts two different sets of top neighbors, our model predicts the same set of top ranked images for the two queries.

5 Research Plan

To complete the dissertation, I plan to include more experimental results on the newly proposed project introduced in Chapter 4. Potential experiments include:

- More comparisons with geometry verification with different settings on different evaluation sets. For example, using another set of feature descriptors, reranking different numbers of top neighbors for each query, ablation study on the number of local descriptors used for each image, evaluations on the Google landmarks v2 retrieval task, etc.
- Comparisons with other image reranking approaches, such as query expansion based methods [88] over the same settings.
- Studying the test time behaviors of the proposed method and the baselines, e.g. inference time on CPU/GPU based machines.
- Exploring the potential benefit of jointly optimizing the feature extractor and the proposed model, which may lead to better feature representations and further accuracy improvements.

With more solid evaluations, we hope the proposed method could be a peer-reviewed publication.

References

[1] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski. Building rome in a day. In *2009 IEEE 12th International Conference on Computer Vision*, pages 72–79, 2009.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Eur. Conf. Comput. Vis.*, pages 382–398. Springer, 2016.

[3] 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 visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Int. Conf. Comput. Vis.*, 2015.

[5] Relja Arandjelovic and Andrew Zisserman. Multiple queries for large scale specific object retrieval. In *Brit. Mach. Vis. Conf.*, pages 1–11, 2012.

[6] Artem Babenko and Victor S. Lempitsky. Aggregating deep convolutional features for image retrieval. In *Int. Conf. Comput. Vis.*, 2015.

[7] Artem Babenko, Anton Slesarev, Alexander Chigorin, and Victor S. Lempitsky. Neural codes for image retrieval. In *Eur. Conf. Comput. Vis.*, 2014.

[8] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[9] Bingyi Cao, Andre Araujo, and Jack Sim. Unifying deep local and global features for image search. In *Eur. Conf. Comput. Vis.*, 2020.

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Eur. Conf. Comput. Vis.*, 2020.

[11] Paola Cascante-Bonilla, Xuwang Yin, Vicente Ordonez, and Song Feng. Chat-crowd: A dialog-based platform for visual layout composition. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019.

[12] Ning-San Chang and King-Sun Fu. Query-by-pictorial-example. *IEEE Trans. Softw. Eng.*, 6(6):519–524, November 1980.

[13] Ning-San Chang and King-Sun Fu. A relational database system for images. In *Pictorial Information Systems*, 1980.

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

[15] O. Chum, A. Mikulík, M. Perdoch, and J. Matas. Total recall ii: Query expansion revisited. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 889–896, 2011.

[16] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *Int. Conf. Comput. Vis.*, 2007.

[17] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[21] Mihai Dusmanu, Ignacio Rocco, Tomás Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable CNN for joint detection and description of local features. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[22] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Brit. Mach. Vis. Conf.*, 2018.

[23] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, September 2010.

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, pages 2672–2680, 2014.

[25] Albert Gordo, Jon Almazán, Jérôme Revaud, and Diane Larlus. End-to-end learning of deep visual representations for image retrieval. *Int. J. Comput. Vis.*, 2017.

[26] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. In *Adv. Neural Inform. Process. Syst.*, pages 676–686, 2018.

[27] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *Eur. Conf. Comput. Vis.*, 2018.

[28] Tanmay Gupta, Kevin J. Shih, Saurabh Singh, and Derek Hoiem. Aligned image-word representations improve inductive transfer across vision-language tasks. In *Int. Conf. Comput. Vis.*, 2017.

[29] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[31] Jasper Uijlings Holger Caesar and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[32] Seunghoon Hong, Dingdong Yang, Jongwook Choi, and Honglak Lee. Inferring semantic

layout for hierarchical text-to-image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[33] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[34] Yuhe Jin, Dmytro Mishkin, A. Mishchuk, Jiri Matas, P. Fua, K. Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. In *Int. J. Comput. Vis.*, 2020.

[35] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[36] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Adv. Neural Inform. Process. Syst.*, pages 1889–1897, 2014.

[37] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, 2014.

[38] Chloé Kiddon, Luke S. Zettlemoyer, and Yejin Choi. Globally coherent text generation with neural checklist models. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[39] Jin-Hwa Kim, Devi Parikh, Dhruv Batra, Byoung-Tak Zhang, and Yuandong Tian. Codraw: Visual dialog for collaborative drawing. *arXiv preprint arXiv:1712.05558*, 2017.

[40] Adriana Kovashka and Kristen Grauman. Attribute pivots for guiding relevance feedback in image search. In *Int. Conf. Comput. Vis.*, December 2013.

[41] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Interactive image search with relative attribute feedback. *Int. J. Comput. Vis.*, 115(2):185–210, 2015.

[42] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, May 2017.

[43] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1601–1608, 2011.

[44] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[45] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 2, pages 2169–2178, 2006.

[46] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Eur. Conf. Comput. Vis.*, 2018.

[47] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *Int. J. Comput. Vis.*, 43(1):29–44, June 2001.

[48] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[49] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-Seng Chua. Knowledge-aware multimodal dialogue systems. In *ACM Int. Conf. Multimedia*, pages 801–809, 2018.

[50] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[51] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. *Eur. Conf. Comput. Vis.*, 2014.

[52] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *1907.11692*, 2019.

[53] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 2004.

[54] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Adv. Neural Inform. Process. Syst.*, pages 13–23. Curran Associates, Inc., 2019.

[55] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, 2015.

[56] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, 2015.

[57] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119, 2013.

[58] Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[59] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *IEEE Conf. Comput. Vis. Pattern Recog.*, volume 2, pages 2161–2168, 2006.

[60] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Hierarchical multimodal lstm for dense visual-semantic embedding. In *Int. Conf. Comput. Vis.*, 2017.

[61] Hyeonwoo Noh, Andre Araujo, Jack Sim, and Bohyung Han. Image retrieval with deep local features and attention-based keypoints. In *Int. Conf. Comput. Vis.*, 2017.

[62] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.*, 42:145–175, 2004.

[63] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for*

Computational Linguistics (ACL), pages 311–318. Association for Computational Linguistics, 2002.

[64] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, and Alexander Ku. Image transformer. In *Int. Conf. Mach. Learn.*, 2018.

[65] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[66] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1–8, 2007.

[67] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[68] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[69] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[70] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Int. Conf. Mach. Learn.*, volume 48 of *Proceedings of Machine Learning Research*, pages 1060–1069. PMLR, 20–22 Jun 2016.

[71] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. Learning what and where to draw. In *Adv. Neural Inform. Process. Syst.*, pages 217–225, 2016.

[72] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst.*, 2015.

[73] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[74] Yong Rui, Thomas S. Huang, and Shih-Fu Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 10(1):39 – 62, 1999.

[75] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1745–1752, 2011.

[76] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Adv. Neural Inform. Process. Syst.*, 2016.

[77] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3783, 2016.

[78] Oriane Siméoni, Yannis Avrithis, and Ondrej Chum. Local features and visual words emerge in activations. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[79] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Int. Conf. Comput. Vis.*, pages 1470–1477 vol.2, 2003.

[80] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[81] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[82] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *ACM International on Conference on Information and Knowledge Management (CIKM)*, pages 553–562, 2015.

[83] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Adv. Neural Inform. Process. Syst.*, 2015.

[84] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, pages 1057–1063, 2000.

[85] Fuwen Tan, Paola Cascante-Bonilla, Xiaoxiao Guo, Hui Wu, Song Feng, and Vicente Ordonez. Drill-down: Interactive retrieval of complex scenes using natural language queries. In *Adv. Neural Inform. Process. Syst.*, December 2019.

[86] Fuwen Tan, Song Feng, and Vicente Ordonez. Text2scene: Generating compositional scenes from textual descriptions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2019.

[87] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *Eur. Conf. Comput. Vis.*, 2020.

[88] Giorgos Tolias and Hervé Jégou. Visual query expansion with or without geometry: Refining local descriptors by feature aggregation. *Pattern Recognition*, 47(10):3466 – 3476, 2014.

[89] Giorgos Tolias, R. Sicre, and H. Jégou. Particular object retrieval with integral max-pooling of cnn activations. In *Int. Conf. Learn. Represent.*, 2016.

[90] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Adv. Neural Inform. Process. Syst.*, pages 5998–6008. Curran Associates, Inc., 2017.

[91] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4566–4575, 2015.

[92] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. In *Int. Conf. Learn. Represent.*, 2015.

[93] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[94] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Int. Conf. Mach. Learn.*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057. PMLR, 07–09 Jul 2015.

[95] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Int. Conf. Mach. Learn.*, volume 37, pages 2048–2057, 2015.

[96] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[97] Mark Yatskar, Vicente Ordonez, and Ali Farhadi. Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 193–198, 2016.

[98] Xuwang Yin and Vicente Ordonez. Obj2text: Generating visually descriptive language from object layouts. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[99] Alan L. Yuille. Deformable templates for face recognition. *Journal of Cognitive Neuroscience*, 3(1):59–70, 1991.

[100] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Int. Conf. Comput. Vis.*, pages 5907–5915, 2017.

[101] Zizhao Zhang, Yuanpu Xie, and Lin Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[102] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *Int. Conf. Comput. Vis.*, 2013.