# Comprehensive Image Captioning
# via Scene Graph Decomposition

Yiwu Zhong[1]*, Liwei Wang[2], Jianshu Chen[2], Dong Yu[2], and Yin Li[1]

[1] University of Wisconsin-Madison, United States
[2] Tencent AI Lab, Bellevue, United States
{yzhong52, yin.li}@wisc.edu, {liweiwang, jianshuchen, dyu}@tencent.com

**Abstract.** We address the challenging problem of image captioning by revisiting the representation of image scene graph. At the core of our method lies the decomposition of a scene graph into a set of sub-graphs, with each sub-graph capturing a semantic component of the input image. We design a deep model to select important sub-graphs, and to decode each selected sub-graph into a single target sentence. By using sub-graphs, our model is able to attend to different components of the image. Our method thus accounts for accurate, diverse, grounded and controllable captioning at the same time. We present extensive experiments to demonstrate the benefits of our comprehensive captioning model. Our method establishes new state-of-the-art results in caption diversity, grounding, and controllability, and compares favourably to latest methods in caption quality. Our project website can be found at http://pages.cs.wisc.edu/~yiwuzhong/Sub-GC.html.

**Keywords:** Image Captioning, Scene Graph, Graph Neural Networks

## 1 Introduction

It is an old saying that "A picture is worth a thousand words". Complex and sometimes multiple ideas can be conveyed by a single image. Consider the example in Fig. 1. The image can be described by "A boy is flying a kite" when pointing to the boy and the kite, or depicted as "A ship is sailing on the river" when attending to the boat and the river. Instead, when presented with regions of the bike and the street, the description can be "A bike parked on the street". Humans demonstrate remarkable ability to summarize multiple ideas associated with different scene components of the same image. More interestingly, we can easily explain our descriptions by linking sentence tokens back to image regions.

Despite recent progress in image captioning, most of current approaches are optimized for caption quality. These methods tend to produce generic sentences that are minorly reworded from those in the training set, and to "look" at regions that are irrelevant to the output sentence [10,45]. Several recent efforts seek to address these issues, leading to models designed for individual tasks including

---

* Work partially done while Yiwu Zhong was an intern at Tencent AI Lab, Bellevue.

(a) A young **boy** is flying a **kite**.   (b) A **cruise ship** is sailing on the **river**.   (c) A **bike** is parked on the **street**.
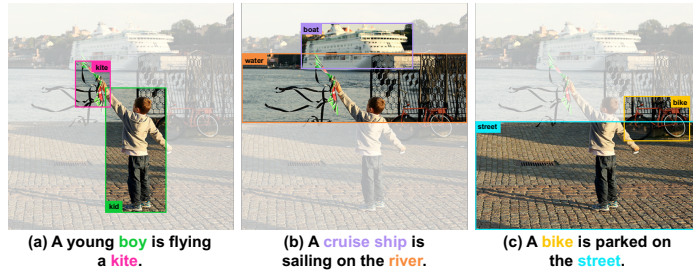
**Fig. 1.** An example image with multiple scene components with each described by a distinct caption. *How can we design a model that can learn to identify and describe different components of an input image?*

diverse [54,11], grounded [46,66] and controllable captioning [34,7]. However, no previous method exists that can address diversity, grounding and controllability at the same time—an ability seemingly effortless for we humans.

We believe the key to bridge the gap is a semantic representation that can better link image regions to sentence descriptions. To this end, we propose to revisit the idea of image captioning using scene graph—a knowledge graph that encodes objects and their relationships. Our core idea is that such a graph can be decomposed into a set of sub-graphs, with each sub-graph as a candidate scene component that might be described by a unique sentence. Our goal is thus to design a model that can identify meaningful sub-graphs and decode their corresponding descriptions. A major advantage of this design is that diversity and controllability are naturally enabled by selecting multiple distinct sub-graphs to decode and by specifying a set of sub-graphs for sentence generation.

Specifically, our method takes a scene graph extracted from an image as input. This graph consists of nodes as objects (nouns) and edges as the relations between pairs of objects (predicates). Each node or edge comes with its text and visual features. Our method first constructs a set of overlapping sub-graphs from the full graph. We develop a graph neural network that learns to select meaningful sub-graphs best described by one of the human annotated sentences. Each of the selected sub-graphs is further decoded into its corresponding sentence. This decoding process incorporates an attention mechanism on the sub-graph nodes when generating each token. Our model thus supports backtracking of generated sentence tokens into scene graph nodes and its image regions. Consequently, our method provides the *first comprehensive model for generating accurate, diverse, and controllable captions that are grounded into image regions.*

Our model is evaluated on MS-COCO Caption [6] and Flickr30K Entities [41] datasets. We benchmark the performance of our model on caption quality, diversity, grounding and controllability. Our results suggest that (1) top-ranked captions from our model achieve a good balance between quality and diversity, outperforming state-of-the-art methods designed for diverse captioning in both quality and diversity metrics and performing on par with latest methods opti-

mized for caption quality in quality metrics; (2) our model is able to link the decoded tokens back into the image regions, thus demonstrating strong results for caption grounding; and (3) our model enables controllable captioning via the selection of sub-graphs, improving state-of-the-art results on controllability. We believe our work provides an important step for image captioning.

## 2   Related Work

There has recently been substantial interest in image captioning. We briefly review relevant work on conventional image captioning, caption grounding, diverse and controllable captioning, and discuss related work on scene graph generation.
**Conventional Image Captioning**. Major progress has been made in image captioning [17]. An encoder-decoder model is often considered, where Convolutional Neural Networks (CNNs) are used to extract global image features, and Recurrent Neural Networks (RNNs) are used to decode the features into sentences [20,52,13,56,63,33,44,31]. Object information has recently been shown important for captioning [62,53]. Object features from an object detector can be combined with encoder-decoder models to generate high quality captions [2].

Several recent works have explored objects and their relationships, encoded in the form of scene graphs, for image captioning [61,60]. The most relevant work is [61]. Their GCN-LSTM model used a graph convolutional network (GCN) [23] to integrate semantic information in a scene graph. And a sentence is further decoded using features aggregated over the full scene graph. Similar to [61], we also use a GCN for an input scene graph. However, our method learns to select sub-graphs within the scene graph, and to decode sentences from ranked sub-graphs instead of the full scene graph. This design allows our model to produce diverse and controllable sentences that are previously infeasible [61,60,2].
**Grounded Captioning**. A major challenge of image captioning is that recent deep models might not focus on the same image regions as a human would when generating each word, leading to undesirable behaviors, e.g., object hallucination [45,10]. Several recent work [56,2,46,66,36,19,59] has been developed to address the problem of grounded captioning—the generation of captions and the alignment between the generated words and image regions. Our method follows the weakly supervised setting for grounded captioning, where we assume that only the image-sentence pairs are known. Our key innovation is to use a sub-graph on an image scene graph for sentence generation, thus constraining the grounding within the sub-graph.

Our work is also relevant to recent work on generating text descriptions of local image regions, also known as dense captioning [19,59,21]. Both our work and dense captioning methods can create localized captions. The key difference is that our method aims to generate sentence descriptions of scene components that spans multiple image regions, while dense captioning methods focused on generating phrase descriptions for local regions [19,59] or pairs of local regions [21].
**Diverse and Controllable Captioning**. The generation of diverse and controllable image descriptions has also received considerable attention. Several

approaches have been proposed for diverse captioning [47,27,8,51,54,11,3]. Wang et al. [54] proposed a variational auto-encoder that can decode multiple diverse sentences from samples drawn from a latent space of image features. This idea was further extended by [3], where every word has its own latent space. More-over, Deshpande et al. [11] proposed to generate various sentences controlled by part-of-speech tags. There is a few recent work on controllable captioning. Lu et al. [34] proposed to fill a generated sentence template with the concepts from an object detector. Cornia et al. [7] selected object regions using grounding anno-tations and then predicted textual chunks to generate diverse and controllable sentences. Similar to [7], we address diversity and controllability within the same model. Different from [7], our model is trained using only image-sentence pairs and can provide additional capacity of caption grounding.

**Scene Graph Generation**. Image scene graph generation has received consid-erable attention, partially driven by large-scale scene graph datasets [25]. Most existing methods [32,65,55,9,29,58,28,64] start from candidate object regions given by an object detector and seek to infer the object categories and their relationships. By using a previous approach [64] to extract image scene graphs, we explore the decomposition of scene graphs into sub-graphs for generating accurate, diverse, and controllable captions. Similar graph partitioning prob-lems have been previously considered in vision for image segmentation [18,15] and visual tracking [49,48], but has not been explored for image captioning.

## 3    Method

Given an input image $I$, we assume an image scene graph $G = (V, E)$ can be extracted from $I$, where $V$ represents the set of nodes corresponding to the detected objects (nouns) in $I$, and $E$ represents the set of edges corresponding to the relationships between pairs of objects (predicates). Our goal is to generate a set of sentences $C = \{C_j\}$ to describe different components of $I$ using the scene graph $G$. To this end, we propose to make use of the sub-graphs $\{G_i^s = (V_i^s, E_i^s)\}$ from $G$, where $V_i^s \subseteq V$ and $E_i^s \subseteq E$. Our method seeks to model the joint probability $P(S_{ij} = (G, G_i^s, C_j)|I)$, where $P(S_{ij}|I) = 1$ indicates that the sub-graph $G_i^s$ can be used to decode the sentence $C_j$. Otherwise, $P(S_{ij}|I) = 0$. We further assume that $P(S_{ij}|I)$ can be decomposed into three parts, given by

$$P(S_{ij}|I) = P(G|I)P(G_i^s|G, I)P(C_j|G_i^s, G, I). \tag{1}$$

Intuitively, $P(G|I)$ extracts scene graph $G$ from an input image $I$. $P(G_i^s|G, I)$ decomposes the full graph $G$ into a diverse set of sub-graphs $\{G_i^s\}$ and selects important sub-graphs for sentence generation. Finally, $P(C_j|G_i^s, G, I)$ decodes a selected sub-graph $G_i^s$ into its corresponding sentence $C_j$, and also associates the tokens in $C_j$ to the nodes $V_i^s$ of the sub-graph $G_i^s$ (the image regions in $I$). Fig. 2 illustrates our method. We now present details of our model.
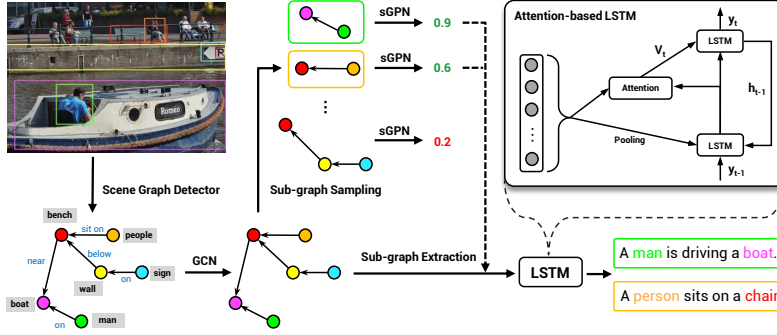
**Fig. 2.** Overview of our method. Our method takes a scene graph extracted from an input image, and decomposes the graph into a set of sub-graphs. We design a sub-graph proposal network (sGPN) that learns to identify meaningful sub-graphs, which are further decoded by an attention-based LSTM for generating sentences and grounding sentence tokens into sub-graph nodes (image regions). By leveraging sub-graphs, our model enables accurate, diverse, grounded and controllable image captioning.

### 3.1   Scene Graph Detection and Decomposition

Our method first extracts scene graph $G$ from image $I$ ($P(G|I)$) using MotifNet [64]. MotifNet builds LSTMs on top of object detector outputs [43] and produces a scene graph $G = (V, E)$ with nodes $V$ for common objects (nouns) and edges $E$ for relationship between pairs of objects (predicates), such as "holding","behind" or "made of". Note that $G$ is a directed graph, i.e., an edge must start from a subject noun or end at an object noun. Therefore, the graph $G$ is defined by a collection of subject-predicate-object triplets, e.g., kid playing ball.

We further samples sub-graphs $\{G_i^s\}$ from the scene graph $G$ by using neighbor sampling [24]. Specifically, we randomly select a set of seed nodes $\{S_i\}$ on the graph. The immediate neighbors of the seed nodes with the edges in-between define a sampled sub-graph. Formally, the sets of sub-graph nodes and edges are $V_i^s = S_i \cup \{N(v)|v \in S_i\}$ and $E_i^s = \{(v, u)|v \in S_i, u \in N(v)\}$ respectively, where $N(v)$ denotes the immediate neighbors of node $v$. Identical sub-graphs are removed to obtain the final set of sub-graphs $\{G_i^s = (V_i^s, E_i^s)\}$, which covers potential scene components in the input image $I$.

### 3.2   Sub-graph Proposal Network

Our next step is to identify meaningful sub-graphs that are likely to capture major scene components in the image ($P(G_i^s|G, I)$). Specifically, our model first combines visual and text features on the scene graph $G$, followed by an integration of contextual information within $G$ using a graph convolutional network, and finally a score function learned to rank sub-graphs $G_i^s$.

**Scene Graph Representation**. Given a directed scene graph $G = (V, E)$, we augment its nodes and edges with visual and text features. For a node $v \in V$, we

use both its visual feature extracted from image regions and the word embedding of its noun label. We denote the visual features as $\boldsymbol{x}_v^v \in \mathbb{R}^{d_v}$ and text features as $\boldsymbol{x}_v^e \in \mathbb{R}^{d_e}$. For an edge $e \in E$, we only use the embedding of its predicate label denoted as $\boldsymbol{x}_e^e \in \mathbb{R}^{d_e}$. Subscripts are used to distinguish node ($v$) and edge ($e$) features and superscripts to denote the feature type, i.e., visual features or text embedding. Visual and text features are further fused by projecting them into a common sub-space. This is done separately for node and edge features by

$$\boldsymbol{x}_v^f = \mathrm{ReLU}(\boldsymbol{W}_f^1 \boldsymbol{x}_v^v + \boldsymbol{W}_f^2 \boldsymbol{x}_v^e), \quad \boldsymbol{x}_e^f = \boldsymbol{W}_f^3 \boldsymbol{x}_e^e, \tag{2}$$

where $\boldsymbol{W}_f^1 \in \mathbb{R}^{d_f \times d_v}$, $\boldsymbol{W}_f^2 \in \mathbb{R}^{d_f \times d_e}$ and $\boldsymbol{W}_f^3 \in \mathbb{R}^{d_f \times d_e}$ are learned projections.
**Graph Convolutional Network (GCN)**. After feature fusion and projection, we further model the context between objects and their relationships using a GCN. The GCN aggregates information from the neighborhood within the graph and updates node and edge features. With an proper ordering of the nodes and edges, we denote the feature matrix for nodes and edges as $\boldsymbol{X}_v^f = [\boldsymbol{x}_v^f] \in \mathbb{R}^{d_f \times |V|}$ and $\boldsymbol{X}_e^f = [\boldsymbol{x}_e^f] \in \mathbb{R}^{d_f \times |E|}$, respectively. The update rule of a single graph convolution is thus given by

$$\begin{aligned}
\hat{\boldsymbol{X}}_v^f &= \boldsymbol{X}_v^f + \mathrm{ReLU}(\boldsymbol{W}_{ps} \boldsymbol{X}_e^f \boldsymbol{A}_{ps}) + \mathrm{ReLU}(\boldsymbol{W}_{po} \boldsymbol{X}_e^f \boldsymbol{A}_{po}), \\
\hat{\boldsymbol{X}}_e^f &= \boldsymbol{X}_e^f + \mathrm{ReLU}(\boldsymbol{W}_{sp} \boldsymbol{X}_v^f \boldsymbol{A}_{sp}) + \mathrm{ReLU}(\boldsymbol{W}_{op} \boldsymbol{X}_v^f \boldsymbol{A}_{op}),
\end{aligned} \tag{3}$$

where $\boldsymbol{W}_{ps}, \boldsymbol{W}_{po}, \boldsymbol{W}_{sp}, \boldsymbol{W}_{op} \in \mathbb{R}^{d_f \times d_f}$ are learnable parameters that link subject or object features (nouns) with predicate features. For example, $\boldsymbol{W}_{ps}$ connects between predicate features and subject features. $\boldsymbol{A}_{ps}, \boldsymbol{A}_{po} \in \mathbb{R}^{|E| \times |V|}$ are the normalized adjacency matrix (defined by $G$) between predicates and subjects, and between predicates and objects, respectively. For instance, a non-zero element in $\boldsymbol{A}_{ps}$ suggests a link between a predicate and a subject on the scene graph $G$. Similarly, $\boldsymbol{A}_{sp}, \boldsymbol{A}_{op} \in \mathbb{R}^{|V| \times |E|}$ are the normalized adjacency matrix between subjects and predicates, and between objects and predicates. $\boldsymbol{A}_{ps} \neq \boldsymbol{A}_{sp}^T$ due to the normalization of adjacency matrix.

Our GCN stacks several graph convolutions, and produces an output scene graph with updated node and edge features. We only keep the final node features ($\boldsymbol{X}_v^u = [\boldsymbol{x}_v^u], v \in V$) for subsequent sub-graph ranking, as the predicate information has been integrated using GCN.
**Sub-graph Score Function**. With the updated scene graph and the set of sampled sub-graphs, our model learns a score function to select meaningful sub-graphs for generating sentence descriptions. For each sub-graph, we index its node features as $\boldsymbol{X}_i^s = [\boldsymbol{x}_v^u], v \in V_i^s$ and construct a score function

$$s_i = \sigma(f(\Phi(\boldsymbol{X}_i^s))), \tag{4}$$

where $\Phi(\cdot)$ is a sub-graph readout function [57] that concatenates the max-pooled and mean-pooled node features on the sub-graph. $f(\cdot)$ is a score function realized by a two-layer multilayer perceptron (MLP). And $\sigma(\cdot)$ is a sigmoid function that normalizes the output score into the range of $[0, 1]$.
**Learning the Score Function**. The key challenge of learning the score function $f$ is the training labels. Our goal is to rank the sampled sub-graphs and select the

best ones to generate captions. Thus, we propose to use ground-truth captions provided by human annotators to guide the learning. A sub-graph with most of the nodes matched to one of the ground-truth sentences should be selected. To this end, we recast the learning of the score function $f$ as training a binary classifier to distinguish between "good" (positive) and "bad" (negative) sub-graphs. Importantly, we design a matching score between a ground-truth sentence and a sampled sub-graph to generate the target binary labels.

Specifically, given a sentence $C_j$ and a scene graph $G$, we extract a reference sub-graph on $G$ by finding the nodes on the graph $G$ that also appears in the sentence $C_j$ and including their immediate neighbor nodes. This is done by extracting nouns from the sentence $C_j$ using a part-of-speech tag parser [5], and matching the nouns to the nodes on $G$ using LCH score [26] derived from WordNet [38]. This matching process is given by $\mathcal{M}(C_j, G)$. We further compute the node Intersection over Union (IoU) score between the reference sub-graph $\mathcal{M}(C_j, G)$ and each of the sampled sub-graph $G_I^s$ by

$$IoU(G_i^s, C_j) = \frac{|G_i^s \cap \mathcal{M}(C_j, G)|}{|G_i^s \cup \mathcal{M}(C_j, G)|}, \tag{5}$$

where $\cap$ and $\cup$ are the intersection and union operation over sets of sub-graph nodes, respectively. The node IoU provides a matching score between the reference sentence $C_j$ and the sub-graph $G_i^s$ and is used to determine our training labels. We only consider a sub-graph as positive for training if its IoU with any of the target sentences is higher than a pre-defined threshold (0.75).

**Training Strategy**. A major issue in training is that we have many negative sub-graphs and only a few positive ones. To address this issue, a mini-batch of sub-graphs is randomly sampled to train our sGPN, where positive to negative ratio is kept as 1:1. If a ground-truth sentence does not match any positive sub-graph, we use the reference sub-graph from $\mathcal{M}(C_j, G)$ as its positive sub-graph.

### 3.3   Decoding Sentences from Sub-graphs

Our final step is to generate a target sentence using features from any selected single sub-graph ($P(C_j|G_i^s, G, I)$). We modify the attention-based LSTM [2] for sub-graph decoding, as shown in Fig. 2 (top right). Specifically, the model couples an attention LSTM and a language LSTM. The attention LSTM assigns each sub-graph node an importance score, further used by the language LSTM to generate the tokens. Specifically, at each time step $t$, the attention LSTM is given by $\boldsymbol{h}_t^A = LSTM_{Att}([\boldsymbol{h}_{t-1}^L, \boldsymbol{e}_t, \boldsymbol{x}_i^s])$, where $\boldsymbol{h}_{t-1}^L$ is the hidden state of the language LSTM at time $t-1$. $\boldsymbol{e}_t$ is the word embedding of the input token at time $t$ and $\boldsymbol{x}_i^s$ is the sub-graph feature. Instead of averaging all region features as [2,61], our model uses the input sub-graph feature, given by $\boldsymbol{x}_i^s = g(\Phi(\boldsymbol{X}_i^s))$, where $g(\cdot)$ is a two-layer MLP, $\Phi(\cdot)$ is the same graph readout unit in Eq. 4.

Based on hidden states $\boldsymbol{h}_t^A$ and the node features $\boldsymbol{X}_i^s = [\boldsymbol{x}_v^u]$ in the sub-graph, an attention weight $a_{v,t}$ at time $t$ for node $v$ is computed by $a_{v,t} = \boldsymbol{w}_a^T \tanh(\boldsymbol{W}_v \boldsymbol{x}_v^u + \boldsymbol{W}_h \boldsymbol{h}_t^A)$ with learnable weights $\boldsymbol{W}_v$, $\boldsymbol{W}_h$ and $\boldsymbol{w}_a$. A softmax function is further used to normalize $\boldsymbol{a}_t$ into $\boldsymbol{\alpha}_t$ defined on all sub-graph nodes

at time $t$. We use $\boldsymbol{\alpha}_t$ to backtrack image regions associated with a decoded token for caption grounding. Finally, the hidden state of the attention LSTM $\boldsymbol{h}_t^A$ and the attention re-weighted sub-graph feature $\boldsymbol{\mathcal{V}}_t = \sum_v \alpha_{v,t} \boldsymbol{x}_v^u$ are used as the input of the language LSTM—a standard LSTM that decodes the next word.

### 3.4   Training and Inference

We summarize the training and inference schemes of our model.

**Loss Functions**. Our sub-graph captioning model has three parts: $P(G|I)$, $P(G_i^s|G,I)$, $P(C_j|G_i^s,G,I)$, where the scene graph generation ($P(G|I)$) is trained independently on Visual Genome [25]. For training, we combine two loss functions for $P(G_i^s|G,I)$ and $P(C_j|G_i^s,G,I)$. Concretely, we use a binary cross-entropy loss for the sub-graph proposal network ($P(G_i^s|G,I)$), and a multi-way cross-entropy loss for the attention-based LSTM model to decode the sentences ($P(C_j|G_i^s,G,I)$). The coefficient between the two losses is set to 1.

**Inference**. During inference, our model extracts the scene graph, samples sub-graphs and evaluates their sGPN scores. *Greedy Non-Maximal Suppression (NMS)* is further used to filter out sub-graphs that largely overlap with others, and to keep sub-graphs with high sGPN scores. The overlapping between two sub-graphs is defined by the IoU of their nodes. We find that using NMS during testing helps to remove redundant captions and to promote diversity.

After NMS, top-ranked sub-graphs are decoded using an attention-based LSTM. As shown in [35], an *optional top-K sampling* [14,42] can be applied during the decoding to further improve caption diversity. We disable top-K sampling for our experiments unless otherwise noticed. The final output is thus a set of sentences with each from a single sub-graph. By choosing which sub-graphs to decode, our model can control caption contents. Finally, we use attention weights in the LSTM to ground decoded tokens to sub-graph nodes (image regions).

## 4   Experiments

We now describe our implementation details and presents our results. We start with an ablation study (4.1) for different model components. Further, we evaluate our model across several captioning tasks, including accurate and diverse captioning (4.2), grounded captioning (4.3) and controllable captioning (4.4).

**Implementation Details**. We used Faster R-CNN [43] with ResNet-101 [16] from [2] as our object detector. Based on detection results, Motif-Net [64] was trained on Visual Genome [25] with 1600/20 object/predicate classes. For each image, we applied the detector and kept 36 objects and 64 triplets in scene graph. We sampled 1000 sub-graphs per image and removed duplicate ones, leading to an average of 255/274 sub-graphs per image for MS-COCO [6]/Flickr30K [41]. We used 2048D visual features for image regions and 300D GloVe [40] embeddings for node and edge labels. These features were projected into 1024D, followed by a GCN with depth of 2 for feature transform and an attention LSTM

**Table 1.** Ablation study on sub-graph/sentence ranking functions, the NMS thresholds and the top-K sampling during decoding. We report results for both accuracy (B4 and C) and diversity (Distinct Caption, 1/2-gram). Our model is trained on the train set of COCO caption and evaluated on the the validation set, following M-RNN split [37].

| Model | Ranking Function | NMS | Top-K Sampling | B4 | C | Distinct Caption | 1-gram (Best 5) | 2-gram (Best 5) |
|---|---|---|---|---|---|---|---|---|
| Sub-GC-consensus | consensus | 0.75 | N/A | 33.0 | 107.6 | 59.3% | 0.25 | 0.32 |
| Sub-GC-sGPN | sGPN | 0.75 | N/A | 33.4 | 108.7 | 59.3% | 0.28 | 0.37 |
| Sub-GC | sGPN+consensus | 0.75 | N/A | 34.3 | 112.9 | 59.3% | 0.28 | 0.37 |
| Sub-GC-consensus | consensus | 0.55 | N/A | 32.5 | 105.6 | 70.5% | 0.27 | 0.36 |
| Sub-GC-sGPN | sGPN | 0.55 | N/A | 33.4 | 108.7 | 70.5% | 0.32 | 0.42 |
| Sub-GC | sGPN+consensus | 0.55 | N/A | 34.1 | 112.3 | 70.5% | 0.32 | 0.42 |
| Sub-GC-S | sGPN+consensus | 0.55 | T=0.6,K=3 | 31.8 | 108.7 | 96.0% | 0.39 | 0.57 |
| Sub-GC-S | sGPN+consensus | 0.55 | T=0.6,K=5 | 30.9 | 106.1 | 97.5% | 0.41 | 0.60 |
| Sub-GC-S | sGPN+consensus | 0.55 | T=1.0,K=3 | 28.4 | 100.7 | 99.2% | 0.43 | 0.64 |

(similar to [2]) for sentence decoding. For training, we used Adam [22] with initial learning rate of 0.0005 and a mini-batch of 64 images and 256 sub-graphs. Beam search was used in decoding with beam size 2, unless otherwise noted.

## 4.1   Ablation Study

We first conduct an ablation study of our model components, including the ranking function, the NMS threshold and the optional top-K sampling. We now describe the experiment setup and report the ablation results.

**Experiment Setup**. We follow the evaluation protocol from [51,54,11,3] and report both accuracy and diversity results using the M-RNN split [37] of MS-COCO Caption dataset [6]. Specifically, this split has 118,287/4,000/1,000 images for train/val/test set, with 5 human labeled captions per image. We train the model on the train set and report the results on the *val* set. For accuracy, we report top-1 accuracy out of the top 20 output captions, using BLEU-4 [39] and CIDEr [50]. For diversity, we evaluate the percentage of distinct captions from 20 sampled output captions, and report 1/2-gram diversity of the best 5 sampled captions using a ranking function. Beam search was disabled for this ablation study. Table 1 presents our results and we now discuss our results.

**Ranking function** is used to rank output captions. Our sGPN provides a socre for each sub-graph and thus each caption. Our sGPN can thus be re-purposed as a ranking function. We compare sGPN with consensus re-ranking [12,37] widely used in the literature [54,11,3]. Moreover, we also experiment with applying consensus on top-scored captions (e.g., top-4) from sGPN (sGPN+consensus). Our sGPN consistently outperforms consensus re-ranking for both accuracy and diversity (+1.1 CIDEr and +12% 1-gram with NMS=0.75). Importantly, consensus re-ranking is computational expensive, while our sGPN incurs little computational cost. Further, combining our sGPN with consensus re-ranking (sGPN+consensus) improves top-1 accuracy (+4.2 CIDEr with NMS=0.75).

sGPN+consensus produces the same diversity scores as sGPN, since only one ranking function (sGPN) is used in diversity evaluation.

**NMS threshold** is used during inference to eliminate similar sub-graphs (see Section 3.4). We evaluate two NMS thresholds (0.55 and 0.75). For all ranking functions, a lower threshold (0.55) increases diversity scores (+8%/+14% 1-gram for consensus/sGPN) and has comparable top-1 accuracy, expect for consensus re-ranking (-2.0 CIDEr). Note that for our sGPN, the top-1 accuracy remains the same as the top-ranked sub-graph stays unchanged.

**Top-K sampling** is optionally applied during caption decoding, where each token is randomly drawn from the top $K$ candidates based on the normalized logits produced by a softmax function with temperature $T$. A small $T$ favors the top candidate and a large $K$ produces more randomness. We evaluate different combinations of $K$ and $T$. Using top-K sampling decreases the top-1 accuracy yet significantly increases all diversity scores (-3.6 CIDEr yet +22% in 1-gram with T=0.6, K=3). The same trend was also observed in [35].

**Our final model** (Sub-GC) combines sGPN and consensus re-ranking for ranking captions. We set NMS threshold to 0.75 for experiments focusing on the accuracy of top-1 caption (Table 3, 4, 5) and 0.55 for experiments on diversity (Table 2). Top-K sampling is only enabled for additional results on diversity.

### 4.2   Accurate and Diverse Image Captioning

**Dataset and Metric**. Moving forward, we evaluate our final model for accuracy and diversity on MS-COCO caption *test* set using M-RNN split [37]. Similar to our ablation study, we report top-1 accuracy and diversity scores by selecting from a pool of top 20/100 output sentences. Top-1 accuracy scores include BLEU [39], CIDEr [50], ROUGE-L [30], METEOR [4] and SPICE [1]. And diversity scores include distinct caption, novel sentences, mutual overlap (mBLEU-4) and n-gram diversity. Beam search was disabled for a fair comparison.

**Baselines**. We consider several latest methods designed for diverse and accurate captioning as our baselines, including Div-BS [51], AG-CVAE [54], POS [11], POS+Joint [11] and Seq-CVAE [3]. We compare our results of Sub-GC to these baselines in Table 2. In addition, we include the results of our model with top-K sampling (Sub-GC-S), as well as human performance for references of diversity.

**Diversity Results**. For the majority of the diversity metrics, our model Sub-GC outperforms previous methods (+8% for novel sentences and +29%/20% for 1/2-gram with 20 samples), except the most recent Seq-CVAE. Upon a close inspection of Seq-CVAE model, we hypothesis that Seq-CVAE benefits from sampling tokens at each time step. It is thus meaningful to compare our model using top-K sampling (Sub-GC-S) with Seq-CVAE. Sub-GC-S outperforms Seq-CVAE in most metrics (+18%/19% for 1/2-gram with 100 samples) and remains comparable for the metric of novel sentences (within 3% difference).

**Accuracy Results**. We notice that the results of our sub-graph captioning models remain the same with increased number of samples. This is because our outputs follow a fixed rank from sGPN scores. Our Sub-GC outperforms all previous methods by a significant margin. Sub-GC achieves +2.6/2.1 in B4 and

**Table 2.** Diversity and top-1 accuracy results on COCO Caption dataset (M-RNN split [37]). Best-5 refers to the top-5 sentences selected by a ranking function. Note that Sub-GC and Sub-GC-S have same top-1 accuracy in terms of sample-20 and sample-100, since we have a sGPN score per sub-graph and global sorting is applied over all sampled sub-graphs. Our models outperform previous methods on both top-1 accuracy and diversity for the majority of the metrics.

| Method | # | Diversity | | | | | Top-1 Accuracy | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Distinct | #novel | mBLEU-4 | 1-gram | 2-gram | B1 | B2 | B3 | B4 | C | R | M | S |
| | | Caption (↑) | (Best 5) (↑) | (Best 5) (↓) | (Best 5) (↑) | (Best 5) (↑) | | | | | | | | |
| Div-BS [51] | | 100% | 3106 | 81.3 | 0.20 | 0.26 | 72.9 | 56.2 | 42.4 | 32.0 | 103.2 | 53.6 | 25.5 | 18.4 |
| AG-CVAE [54] | | 69.8% | 3189 | 66.6 | 0.24 | 0.34 | 71.6 | 54.4 | 40.2 | 29.9 | 96.3 | 51.8 | 23.7 | 17.3 |
| POS [11] | 20 | 96.3% | 3394 | 63.9 | 0.24 | 0.35 | 74.4 | 57.0 | 41.9 | 30.6 | 101.4 | 53.1 | 25.2 | 18.8 |
| POS+Joint [11] | | 77.9% | 3409 | 66.2 | 0.23 | 0.33 | 73.7 | 56.3 | 41.5 | 30.5 | 102.0 | 53.1 | 25.1 | 18.5 |
| Sub-GC | | 71.1% | 3679 | 67.2 | 0.31 | 0.42 | **77.2** | **60.9** | **46.2** | **34.6** | **114.4** | **56.1** | **26.9** | **20.0** |
| Seq-CVAE [3] | 20 | 94.0% | **4266** | 52.0 | 0.25 | 0.54 | 73.1 | 55.4 | 40.2 | 28.9 | 100.0 | 52.1 | 24.5 | 17.5 |
| Sub-GC-S | | **96.2%** | 4153 | **36.4** | **0.39** | **0.57** | 75.2 | 57.6 | 42.7 | 31.4 | 107.3 | 54.1 | 26.1 | 19.3 |
| Div-BS [51] | | 100% | 3421 | 82.4 | 0.20 | 0.25 | 73.4 | 56.9 | 43.0 | 32.5 | 103.4 | 53.8 | 25.5 | 18.7 |
| AG-CVAE [54] | | 47.4% | 3069 | 70.6 | 0.23 | 0.32 | 73.2 | 55.9 | 41.7 | 31.1 | 100.1 | 52.8 | 24.5 | 17.9 |
| POS [11] | 100 | 91.5% | 3446 | 67.3 | 0.23 | 0.33 | 73.7 | 56.7 | 42.1 | 31.1 | 103.6 | 53.0 | 25.3 | 18.8 |
| POS+Joint [11] | | 58.1% | 3427 | 70.3 | 0.22 | 0.31 | 73.9 | 56.9 | 42.5 | 31.6 | 104.5 | 53.2 | 25.5 | 18.8 |
| Sub-GC | | 65.8% | 3647 | 69.0 | 0.31 | 0.41 | **77.2** | **60.9** | **46.2** | **34.6** | **114.4** | **56.1** | **26.9** | **20.0** |
| Seq-CVAE [3] | 100 | 84.2% | **4215** | 64.0 | 0.33 | 0.48 | 74.3 | 56.8 | 41.9 | 30.8 | 104.1 | 53.1 | 24.8 | 17.8 |
| Sub-GC-S | | **94.6%** | 4128 | **37.3** | **0.39** | **0.57** | 75.2 | 57.6 | 42.7 | 31.4 | 107.3 | 54.1 | 26.1 | 19.3 |
| Human | 5 | 99.8% | - | 51.0 | 0.34 | 0.48 | - | - | - | - | - | - | - | - |

**Table 3.** Comparison to accuracy optimized models on COCO caption dataset using Karpathy split [20]. Our Sub-GC compares favorably to the latest methods.

| Method | B1 | B4 | C | R | M | S |
|---|---|---|---|---|---|---|
| Up-Down [2] | 77.2 | 36.2 | 113.5 | 56.4 | 27.0 | 20.3 |
| GCN-LSTM [61] | 77.3 | 36.8 | 116.3 | 57.0 | **27.9** | 20.9 |
| SGAE [60] | **77.6** | **36.9** | **116.7** | **57.2** | 27.7 | **20.9** |
| Full-GC | 76.7 | **36.9** | 114.8 | 56.8 | **27.9** | 20.8 |
| Sub-GC | 76.8 | 36.2 | 115.3 | 56.6 | 27.7 | 20.7 |
| Sub-GC-oracle | 90.7 | 59.3 | 166.7 | 71.5 | 40.1 | 30.1 |

+11.2/9.9 in CIDEr when using 20/100 samples in comparison to previous best results. Moreover, while achieving best diversity scores, our model with top-K sampling (Sub-GC-S) also outperforms previous methods in most accuracy metrics (+0.8/0.9 in B1 and +4.1/2.8 in CIDEr when using 20/100 samples) despite its decreased accuracy from Sub-GC.

**Comparison to Accuracy Optimized Captioning models**. We conduct further experiments to compare the top ranked sentence from ou Sub-GC against the results of latest captioning models optimized for accuracy, including Up-Down [2], GCN-LSTM [61] and SGAE [60]. All these previous models can only generate a single sentence, while our method (Sub-GC) can generate a set of diverse captions. As a reference, we consider a variant of our model (Full-GC) that uses a full scene graph instead of sub-graphs to decode sentences. Moreover, we include an upper bound of our model (Sub-GC-oracle) by assuming that we have an oracle ranking function, i.e., always selecting the maximum scored sentence for each metric. All results are reported on Karpathy split [20] of COCO dataset without using reinforcement learning for score optimization [44].

**Table 4.** Grounded captioning results on Flickr30K Entities [41]. Our method (Sub-GC) outperforms previous weakly supervised methods.

| Method | Grounding Evaluation | | Caption Evaluation | | | | |
|---|---|---|---|---|---|---|---|
| | F1 all | F1 loc | B1 | B4 | C | M | S |
| GVD [66] | 3.88 | 11.70 | 69.2 | 26.9 | 60.1 | 22.1 | 16.1 |
| Up-Down [2] | 4.14 | 12.30 | 69.4 | 27.3 | 56.6 | 21.7 | 16.0 |
| Cyclical [36] | 4.98 | 13.53 | 69.9 | 27.4 | 61.4 | 22.3 | 16.6 |
| Full-GC | 4.90 | 13.08 | 69.8 | **29.1** | **63.5** | **22.7** | **17.0** |
| Sub-GC | **5.98** | **16.53** | **70.7** | 28.5 | 61.9 | 22.3 | 16.4 |
| GVD (Sup.) [66] | 7.55 | 22.20 | 69.9 | 27.3 | 62.3 | 22.5 | 16.5 |

**Table 5.** Controllable captioning results on Flickr30K Entities [41]. With weak supervision, our Sub-GC compares favorably to previous methods. With strong supervision, our Sub-GC (Sup.) achieves the best results.

| Method | B1 | B4 | C | R | M | S | IoU |
|---|---|---|---|---|---|---|---|
| NBT [34] (Sup.) | - | 8.6 | 53.8 | 31.9 | 13.5 | 17.8 | 49.9 |
| SCT [7] (Sup.) | 33.1 | 9.9 | 67.3 | 35.3 | 14.9 | 22.2 | 52.7 |
| Sub-GC | 33.6 | 9.3 | 57.8 | 32.5 | 14.2 | 18.8 | 50.6 |
| Sub-GC (Sup.) | **36.2** | **11.2** | **73.7** | **35.5** | **15.9** | **22.2** | **54.1** |

Our results are shown in Table 3. Our Sub-GC achieves comparable results (within 1-2 points in B4/CIDEr) to latest methods (Up-Down, GCN-LSTM and SGAE). We find that the results of our sub-graph captioning model is slightly worse than those models using the full scene graph, e.g., Full-GC, GCN-LSTM and SGAE. We argue that this minor performance gap does not diminish our contribution, as our model offers new capacity for generating diverse, controllable and grounded captions. Notably, our best case (Sub-GC-oracle) outperforms all other methods for all metrics by a very large margin (+22.4 in B4 and +50.0 in CIDEr). These results suggest that at least one high-quality caption exists among the sentences decoded from the sub-graphs. It is thus possible to generate highly accurate captions if there is a way to select this "good" sub-graph.

### 4.3   Grounded Image Captioning

Moreover, we evaluate our model for grounded captioning. We describe the dataset and metric, introduce our setup and baselines, and discuss our results.
**Dataset and Metric**. We use Flickr30k Entities [41] for grounded captioning. Flickr30k Entities has 31K images, with 5 captions for each image. The dataset also includes 275k annotated bounding boxes associated with the phrases in corresponding captions. We use the data split from [20]. To evaluate the grounding performance, we follow the protocol in GVD [66]. We report both $F1_{all}$ and $F1_{loc}$. $F1_{all}$ considers a region prediction as correct if the object word is correctly predicted and the box is correctly localized. On the other hand $F1_{loc}$ only accounts for localization quality. Moreover, we report the standard BLEU [39], CIDEr [50], METEOR [4] and SPICE [1] scores for caption quality.
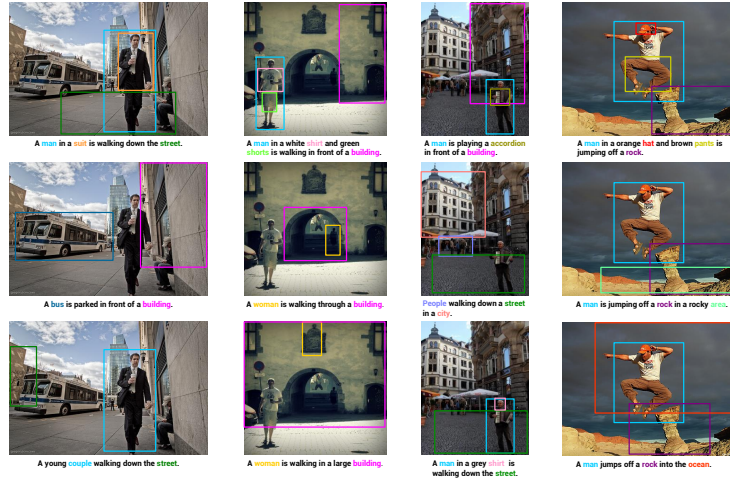
**Fig. 3.** Sample results of our Sub-GC on Flickr30k Entities test set. Each column shows three captions with their region groundings decoded from different sub-graphs for an input image. The first two rows are successful cases and the last row is the failure case. These sentences can describe different parts of the images. Each generated noun and its grounding bounding box are highlighted in the same color.

**Experiment Setup and Baselines**. For this experiment, we only evaluate the top-ranked sentence and its grounding from our model. We select the node on the sub-graph with maximum attention weight when decoding a noun word, and use its bounding box as the grounded region. Our results are compared to a strong set of baselines designed for weakly supervised grounded captioning, including weakly supervised GVD [66], Up-Down [2] and a concurrent work Cyclical [36]. We also include reference results from fully supervised GVD [66] that requires ground-truth matching pairs for training, and our Full-GC that decode a sentence from a full graph.

**Results**. Our results are presented in Table 4. Among all weakly supervised methods, our model achieves the best F1 scores for caption grounding. Specifically, our sub-graph captioning model (Sub-GC) outperforms previous best results by +1.0 for $F1_{all}$ and +3.0 for $F1_{loc}$, leading to a relative improvement of 20% and 22% for $F1_{all}$ and $F1_{loc}$, respectively. Our results also have the highest captioning quality (+1.1 in B4 and +0.5 in CIDEr). We conjecture that constraining the attention to the nodes of a sub-graph helps to improve the grounding. Fig. 3 shows sample results of grounded captions. Not surprisingly, the supervised GVD outperforms our Sub-GC. Supervised GVD can be considered as an upper bound for all other methods, as it uses grounding annotations for training. Comparing to our Full-GC, our Sub-GC is worse on captioning quality (-0.6 in B4 and -1.6 in CIDEr) yet has significant better performance for grounding (+1.1 in $F1_{all}$ and +3.5 in $F1_{loc}$).

### 4.4   Controllable Image Captioning

Finally, we report results on controllable image captioning. Again, we describe our experiments and present the results.

**Dataset and Metric**. Same as grounding, we consider Flickr30k Entities [41] for controllable image captioning and use the data split [20]. We follow evaluation protocol developed in [7]. Specifically, the protocol assumes that an image and a set of regions are given as input, and evaluates a decoded sentence against one or more target ground-truth sentences. These ground-truth sentences are selected from captions by matching the sentences tokens to object regions in the image. Standard captioning metrics are considered (BLEU [39], CIDEr [50], ROUGE-L [30], METEOR [4] and SPICE [1]), yet the ground-truth is different from conventional image captioning. Moreover, the IoU of the nouns between the predicted and the target sentence is also reported as [7].

**Experiment Setup and Baselines**. We consider (1) our Sub-GC trained with only image-sentence pairs; and (2) a supervised Sub-GC trained with ground-truth pairs of region sets and sentences as [7]. Both models follow the same inference scheme, where input controlled set of regions are converted into best matching sub-graphs for sentence decoding. However, supervised Sub-GC uses these matching during training. We compare our results to recent methods developed for controllable captioning, including NBT [34] and SCT [7]. NBT and SCT are trained with matching pairs of region sets and sentences same as our supervised Sub-GC. Results are reported without using reinforcement learning.

**Results**. The results are shown in Table 5. Our models demonstrate strong controllability of the output sentences. Specifically, our supervised Sub-GC outperforms previous supervised methods (NBT and SCT) by a significant margin. Comparing to previous best SCT, our results are +1.3 in B4, +6.4 in CIDEr and +1.4 in IoU. Interestingly, our vanilla model has comparable performance to previous methods, even if it is trained with only image sentence pairs. These results provide further supports to our design of using sub-graphs for image captioning.

## 5   Conclusion

We proposed a novel image captioning model by exploring sub-graphs of image scene graph. Our key idea is to select important sub-graphs and only decode a single target sentence from a selected sub-graph. We demonstrated that our model can generate accurate, diverse, grounded and controllable captions. Our method thus offers the first comprehensive model for image captioning. Moreover, our results established new state-of-the-art in diverse captioning, grounded captioning and controllable captioning, and compared favourably to latest method for caption quality. We hope our work can provide insights into the design of explainable and controllable models for vision and language tasks.

# References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Proceedings of the European Conference on Computer Vision (ECCV). pp. 382–398. Springer (2016)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6077–6086. IEEE (2018)
3. Aneja, J., Agrawal, H., Batra, D., Schwing, A.: Sequential latent spaces for modeling the intention during diverse image captioning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 4261–4270. IEEE (2019)
4. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
5. Bird, S., Loper, E.: NLTK: The natural language toolkit. In: ACL Interactive Poster and Demonstration Sessions. pp. 214–217. Association for Computational Linguistics (2004)
6. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
7. Cornia, M., Baraldi, L., Cucchiara, R.: Show, control and tell: A framework for generating controllable and grounded captions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8307–8316. IEEE (2019)
8. Dai, B., Fidler, S., Urtasun, R., Lin, D.: Towards diverse and natural image descriptions via a conditional gan. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2970–2979. IEEE (2017)
9. Dai, B., Zhang, Y., Lin, D.: Detecting visual relationships with deep relational networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3076–3086. IEEE (2017)
10. Das, A., Agrawal, H., Zitnick, L., Parikh, D., Batra, D.: Human attention in visual question answering: Do humans and deep networks look at the same regions? Computer Vision and Image Understanding **163**, 90–100 (2017)
11. Deshpande, A., Aneja, J., Wang, L., Schwing, A.G., Forsyth, D.: Fast, diverse and accurate image captioning guided by part-of-speech. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10695–10704. IEEE (2019)
12. Devlin, J., Gupta, S., Girshick, R., Mitchell, M., Zitnick, C.L.: Exploring nearest neighbor approaches for image captioning. arXiv preprint arXiv:1505.04467 (2015)
13. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2625–2634. IEEE (2015)
14. Fan, A., Lewis, M., Dauphin, Y.: Hierarchical neural story generation. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 889–898. Association for Computational Linguistics (2018)

15. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision (IJCV) **59**(2), 167–181 (2004)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE (2016)
17. Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CSUR) **51**(6), 1–36 (2019)
18. Jianbo Shi, Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **22**(8), 888–905 (2000)
19. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4565–4574. IEEE (2016)
20. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3128–3137. IEEE (2015)
21. Kim, D.J., Choi, J., Oh, T.H., Kweon, I.S.: Dense relational captioning: Triple-stream networks for relationship-based captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6271–6280. IEEE (2019)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (ICLR) (2015)
23. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (ICLR) (2016)
24. Klusowski, J.M., Wu, Y.: Counting motifs with graph sampling. In: COLT. pp. 1966–2011. Proceedings of Machine Learning Research (2018)
25. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (IJCV) **123**(1), 32–73 (2017)
26. Leacock, C., Miller, G.A., Chodorow, M.: Using corpus statistics and wordnet relations for sense identification. Computational Linguistics **24**(1), 147–165 (1998)
27. Li, D., Huang, Q., He, X., Zhang, L., Sun, M.T.: Generating diverse and accurate visual captions by comparative adversarial learning. arXiv preprint arXiv:1804.00861 (2018)
28. Li, Y., Ouyang, W., Wang, X., Tang, X.: VIP-CNN: Visual phrase guided convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1347–1356. IEEE (2017)
29. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1261–1270. IEEE (2017)
30. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
31. Liu, F., Ren, X., Liu, Y., Wang, H., Sun, X.: simNet: Stepwise image-topic merging network for generating detailed and comprehensive image captions. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 137–149. Association for Computational Linguistics (2018)
32. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Proceedings of the European Conference on Computer Vision (ECCV). pp. 852–869. Springer (2016)

33. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 375–383. IEEE (2017)
34. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7219–7228. IEEE (2018)
35. Luo, R., Shakhnarovich, G.: Analysis of diversity-accuracy tradeoff in image captioning. arXiv preprint arXiv:2002.11848 (2020)
36. Ma, C.Y., Kalantidis, Y., AlRegib, G., Vajda, P., Rohrbach, M., Kira, Z.: Learning to generate grounded image captions without localization supervision. arXiv preprint arXiv:1906.00283 (2019)
37. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (M-RNN). In: International Conference on Learning Representations (ICLR) (2015)
38. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995)
39. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL). pp. 311–318. Association for Computational Linguistics (2002)
40. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics (2014)
41. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2641–2649. IEEE (2015)
42. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI Technical Report (2019)
43. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 91–99. Curran Associates, Inc. (2015)
44. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7008–7024. IEEE (2017)
45. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4035–4045. Association for Computational Linguistics (2018)
46. Selvaraju, R.R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., Parikh, D.: Taking a hint: Leveraging explanations to make vision and language models more grounded. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2591–2600. IEEE (2019)
47. Shetty, R., Rohrbach, M., Anne Hendricks, L., Fritz, M., Schiele, B.: Speaking the same language: Matching machine to human captions by adversarial training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4135–4144. IEEE (2017)
48. Song, J., Andres, B., Black, M.J., Hilliges, O., Tang, S.: End-to-end learning for graph decomposition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 10093–10102. IEEE (2019)

49. Tang, S., Andres, B., Andriluka, M., Schiele, B.: Subgraph decomposition for multi-target tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5033–5041. IEEE (2015)

50. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4566–4575. IEEE (2015)

51. Vijayakumar, A.K., Cogswell, M., Selvaraju, R.R., Sun, Q., Lee, S., Crandall, D., Batra, D.: Diverse beam search for improved description of complex scenes. In: AAAI Conference on Artificial Intelligence (2018)

52. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3156–3164. IEEE (2015)

53. Wang, J., Madhyastha, P.S., Specia, L.: Object counts! bringing explicit detections back into image captioning. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). pp. 2180–2193. Association for Computational Linguistics (2018)

54. Wang, L., Schwing, A., Lazebnik, S.: Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 5756–5766. Curran Associates, Inc. (2017)

55. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5410–5419. IEEE (2017)

56. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning (ICML). pp. 2048–2057 (2015)

57. Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? In: International Conference on Learning Representations (ICLR) (2019)

58. Yang, J., Lu, J., Lee, S., Batra, D., Parikh, D.: Graph R-CNN for scene graph generation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Proceedings of the European Conference on Computer Vision (ECCV). pp. 670–685. Springer (2018)

59. Yang, L., Tang, K., Yang, J., Li, L.J.: Dense captioning with joint inference and visual context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2193–2202. IEEE (2017)

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

61. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Proceedings of the European Conference on Computer Vision (ECCV). pp. 684–699. Springer (2018)

62. Yin, X., Ordonez, V.: Obj2Text: Generating visually descriptive language from object layouts. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 177–187. Association for Computational Linguistics (2017)

63. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4651–4659. IEEE (2016)

64. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5831–5840. IEEE (2018)
65. Zhang, H., Kyaw, Z., Chang, S.F., Chua, T.S.: Visual translation embedding network for visual relation detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5532–5540. IEEE (2017)
66. Zhou, L., Kalantidis, Y., Chen, X., Corso, J.J., Rohrbach, M.: Grounded video description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6578–6587. IEEE (2019)