Wang, Tzu Jui Julius; Tavakoli, Hamed R.; Sjöberg, Mats; Laaksonen, Jorma

Geometry-aware relational exemplar attention for dense captioning

Published in:
MULEA 2019 - 1st International Workshop on Multimodal Understanding and Learning for Embodied Applications, co-located with MM 2019

DOI:
10.1145/3347450.3357656

Published: 15/10/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Wang, T. J. J., Tavakoli, H. R., Sjöberg, M., & Laaksonen, J. (2019). Geometry-aware relational exemplar attention for dense captioning. In *MULEA 2019 - 1st International Workshop on Multimodal Understanding and Learning for Embodied Applications, co-located with MM 2019* (pp. 3-11). ACM.
https://doi.org/10.1145/3347450.3357656

Geometry-aware Relational Exemplar Attention for Dense Captioning

Tzu-Jui Julius Wang
Department of Computer Science, Aalto University
Espoo, Finland
tzu-jui.wang@aalto.fi

Hamed R. Tavakoli
Nokia Technologies
Espoo, Finland
hamed.rezazadegan_tavakoli@nokia.com

Mats Sjöberg
CSC – IT Center for Science
Espoo, Finland
mats.sjoberg@csc.fi

Jorma Laaksonen
Department of Computer Science, Aalto University
Espoo, Finland
jorma.laaksonen@aalto.fi

ABSTRACT

Dense captioning (DC), which provides a comprehensive context understanding of images by describing all salient visual groundings in an image, facilitates multimodal understanding and learning. As an extension of image captioning, DC is developed to discover richer sets of visual contents and to generate captions of wider diversity and increased details. The state-of-the-art models of DC consist of three stages: (1) region proposals, (2) region classification, and (3) caption generation for each proposal. They are typically built upon the following ideas: (a) guiding the caption generation with image-level features as the context cues along with regional features and (b) refining locations of region proposals with caption information. In this work, we propose (a) a joint visual-textual criterion exploited by the region classifier that further improves both region detection and caption accuracy, and (b) a Geometry-aware Relational Exemplar attention (GREatt) mechanism to relate region proposals. The former helps the model learn a region classifier by effectively exploiting both visual groundings and caption descriptions. Rather than treating each region proposal in isolation, the latter relates regions in complementary relations, i.e. *contextually dependent*, *visually supported* and *geometry* relations, to enrich context information in regional representations. We conduct an extensive set of experiments and demonstrate that our proposed model improves the state-of-the-art by at least +5.3% in terms of the mean average precision on the Visual Genome dataset.

CCS CONCEPTS

• **Computing methodologies → Scene understanding**.

KEYWORDS

dense captioning, attention, relationship modeling

MULEA '19, October 25, 2019, Nice, France

ACM ISBN 978-1-4503-6918-3/19/10…$15.00
https://doi.org/10.1145/3347450.3357656

ACM Reference Format:
Tzu-Jui Julius Wang, Hamed R. Tavakoli, Mats Sjöberg, and Jorma Laaksonen. 2019. Geometry-aware Relational Exemplar Attention for Dense Captioning. In *1st International Workshop on Multimodal Understanding and Learning for Embodied Applications (MULEA '19), October 25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3347450.3357656

1 INTRODUCTION

Advancements in computer vision applications, such as object detection and segmentation, have laid a strong foundation of comprehensive context understanding in images. Besides learning on visual domain, tasks such as image captioning (IC) [3, 7, 24] and visual question answering (VQA) [1] are the iconic examples that connect vision and language modalities to not only provide better visual reasoning, but also enable multimodal context understanding. The IC task is to generate a human understandable sentence from a given image. Such a sentence should be grammatically correct, adequately expressive, and capture holistic view of the image content. The VQA task is to generate a sentence to answer a given question targeting at an image. While such a multimodal model (e.g. an IC model) is able to describe an image, it continues to express varying image contents with a sentence that can hardly capture multiple perspectives of the image content.

To extend the capability of a captioning model, Johnson et al. introduced the *Dense Captioning* (DC) task where the aim is to describe as many as possible regions of interest (RoIs) in an image [9]. More specifically, DC comprises two joint tasks: (a) localizing the RoIs (e.g. by bounding boxes) and (b) generating a sentence describing each grounded region. These tasks introduce two more challenges to image captioning: (1) detecting and proposing meaningful ROIs for captions and (2) understanding the relations between the region proposals. For example, in Figure 1, two visual groundings surrounding the man are closely related in visual contents and captions. Besides, the larger RoI surrounding the whole body of the man provides the most informative context for the smaller RoI captioned with "blue jeans of *man*". This indicates that the captioning process can benefit from a DC model that is capable of capturing relationships between regions.

We address the aforementioned challenges by (1) introducing a joint visual-textual criterion for detecting RoIs and (2) proposing a Geometry-aware Relational Exemplar attention (GREatt) module

Figure 1: An instance in the Visual Genome dataset [11] that reveals how the visual and caption information from other regions can be particularly informative to some regions. Here, the informative context for region captioned with *blue jeans of man* is the region captioned with *man playing guitar on street*, which is cued by the street area captioned with *leaves scattered on sidewalk and street*.

for capturing relations between the RoIs. Utilizing the caption embedding along with the visual representations enforces the model to learn a better alignment between the visual content and the corresponding captions by projecting them into a shared subspace. Simultaneously, the optimality of the proposed RoIs is improved. GREatt accounts for three intrinsically distinct types of region-level relationships, including (a) *spatially correlated*, (b) *contextually dependent*, and (c) *visually similar and supported* relations. The spatially correlated relation considers regions that are correlated by their locations and sizes. The contextually dependent relation considers if a region provides contextual information for another region. The visually similar and supported relation focuses on visually similar contexts to enhance the evidence of the existence of a specific context.

To summarize, our contributions are (1) a new geometry-aware relational exemplar attention module and (2) a joint visual-textual region classification criteria. which together lead to a new state-of-the-art in the dense captioning task.

2 RELATED WORK

2.1 Image Captioning

Understanding image captioning is essential because it is the fundamental building block of any captioning pipeline. We, thus, briefly overview some of the most relevant works and refer the readers to [7] for further reading.

The classical image captioning methods such as [3] relied on linking a sentence to an image via feature mapping and were limited to retrieving a pre-existing sentence from a corpus of sentences. The techniques which utilize a language model, however, show more flexibility in generating a sentence from a feature vector representing the image. The most successful of such methods are neural-based techniques [10, 16, 21]. Many of the recent image captioning pipelines follow a similar path.

The most relevant works to us are, in particular, the attention-based image captioning methods. For example [24] defined a soft-attention mechanism (also known as top-down attention) that learns to align the visual features with textual features dynamically over time while generating a sentence. Pedersoli et al. [13] extended the same idea by employing geometrical transformations to the regions used for captioning. The top-down attention mechanism often loses its effectiveness after the visual features are fine-tuned for the captioning task [13]. In contrast to the top-down mechanism, R. Tavakoli et al. [18] investigated the bottom-up attention mechanism. While they demonstrated that bottom-up attention cannot help much improving the caption qualities, they showed such a mechanism enhances the robustness of captioning models. Recently, He et al. [6] proposed an effective approach for combining both bottom-up and top-down attention.

Our proposed approach follows a similar path to attention-based image captioning, specifically using top-down attention. Nevertheless, we focus on dense captioning and try to encode the relations between regions for building powerful context features.

2.2 Dense Captioning

Dense captioning was introduced along with the Visual Genome dataset [11], which aims to promote vision and language research in conjunctions with a range of perceptual reasoning and question answering tasks. The dataset provides 5.4 million region annotations with bounding boxes and captions for 108,077 images, averaging ~50 annotations per image.

The first dense captioning model was introduced by the pioneering work of Johnson et al. [9]. Their framework consists of three components: (1) an image feature extractor (e.g. implemented by a VGG net [17]), (2) a region detector, and (3) a caption generator. Given an image, it first projects the image into the feature space. Then, it detects a series of RoIs using the region proposal mechanism. Finally, each RoI is described with a sentence using the caption generator language model based on recurrent neural networks (RNN) [12] and image features corresponding to that RoI. They tested their model on Visual Genome version 1 [11] and established the first baseline for this task.

Yang et al. [26] extended the idea by replacing the localization layer with Faster-RCNN [14], using captions for improving the localization of region proposals generated by Faster-RCNN, and exploiting both regional and image-level features for the language model. They demonstrated that each of these modifications and their combinations significantly improve the dense captioning.

Nevertheless, image-level features as context can mislead the caption generator towards describing the global context rather than the region of interest [26]. In contrast, our proposed GREatt mechanism learns the context features from the proposed regions by considering distinct types of pairwise relationships between the RoIs. Hence, our pipeline uses features which are more contextually dependent yet region-specific and improve caption quality. In addition, to further capitalize on the idea of engaging captions in the proposal process, we propose a region classifier (which determines the likelihood of a proposal being a genuine RoI) learned on a subspace shared by textual features and their visual counterparts. Developing these two novel designs on top of the pipeline proposed in [26] further enhances the performance in both region classification and caption generation.

2.3 Attention and Relation Reasoning

Reasoning about the relation of two feature vectors which represent objects, entities, and elements with neural networks has gained a recent interest and has been a core module in wide range of applications, such as image captioning [27], object detection [8], and visual question answering (VQA) [15], and scene graph generation (SGG) [23, 25].

Many existing works have proposed different means to associate two feature vectors (e.g. $\mathbf{v}_i$ and $\mathbf{v}_j$) and capture their mutual importance as $\alpha_{i,j}$. Introducing the notion of importance, one can link relation reasoning to attention and interpret $\alpha_{i,j}$ as a quantity of how much one should also pay attention to $\mathbf{v}_j$ during inference about $\mathbf{v}_i$ given a task. The most notable work for our purpose in this annals is transformer networks [19] (originally for natural language processing (NLP) tasks) in which the attention weights are defined by the function of scaled dot-product (SDP) between $\mathbf{v}_i$ and $\mathbf{v}_j$, emphasizing similarity of representations.

In the context of object detection, Hu et al. [8] proposed a revised SDP attention, which additionally considers the geometry relationship between object proposals, allowing them to be refined and classified jointly rather than in isolation. Yao et al. [27] constructed a directed graph over the object proposals, in which each node of the graph is represented by the visual features of the proposals, in order to do image captioning. The refined object-level representation which embeds with the graph structure is then calculated through graph convolutional networks (GCN). Yang et al. [25] capitalized on a similar idea to relate the region proposals for scene graph generation.

Two other relevant ideas are graph attention networks [20] and Neural Turing Machine [4]. The first one was originally proposed for the graph classification task, and in it two features interact through concatenation followed by a multi-layer perceptron (MLP). The second one extends the same line of research with external memory modules and employs the cosine similarity function to capture the interaction between entities.

Even though many works have proposed different attention mechanisms for the downstream tasks, most of them learn the attention embodied by *single* relation (e.g. by SDP attention [8, 19]). What remains less studied is can *multiple* attentions formulated in different computational forms benefit each other for a given computer vision task. This work addresses 1) *do different attention mechanisms work better in isolation?* and 2) *are they complementary to each other?* By examining and exploiting the complementary relations captured by visual and geometry features, we propose a novel attention mechanism built upon distinct types of relations which improve the dense captioning task.

3 METHOD

In this section, we describe the problem formulation, our proposed architecture and each component in the pipeline. The code is publicly available at https://github.com/aalto-cbir/greatt_densecap.

3.1 Problem Formulation

We devise the dense captioning problem to consist of four sub-tasks: 1) region proposal (RP), 2) region classification (RC), 3) proposal refinement (PR), and 4) region caption generation (CG). Region proposal firstly generates a set of region proposals which are then classified by a region classifier. The locations of region proposals are refined gradually as the caption generation process proceeds. The objectives of each task are formulated as follows:

Region proposal (RP). Region proposal is to learn to generate a set of proposals $\hat{\mathcal{B}} = \{\hat{B}_i\}_{i=1}^{N_r}$ that well match to the ground-truth proposals $\mathcal{B} = \{B_i\}_{i=1}^{N}$, where N_r is the number of the generated proposals and N is the number of proposals in an image. Each proposal is characterized by a rigid box, defined by its center coordinate, width and height. Note that, here we use N and N_r for notational simplicity, though they may be different for each image.

Region classification (RC). Region classification decides whether a region proposal is good enough to be captioned or should be ignored. We classify the regions by additionally conditioning them on the captions $\hat{\mathcal{S}} = \{\hat{S}_i\}_{i=1}^{N_r}$ (which are generated by the model learned on the ground-truth captions $\mathcal{S} = \{S_i\}_{i=1}^{N}$) and the relationships between proposals. For an image $\mathcal{I}$ we build a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ over the representations of N_r proposed regions, denoted by $\mathcal{V} = \{\mathbf{v}_i, \mathbf{b}_i\}_{i=1}^{N_r}$, where $\mathbf{v}_i$ refers to the visual representation and $\mathbf{b}_i$ to the geometry representation, which are defined later in Sec. 3.3. The edges $\mathcal{E}$ correspond to the relationships. We, thus, minimize

$$E_{cls} = -\sum_i \log P(c_i|\hat{B}_i, \hat{S}_i, \mathcal{G}), \tag{1}$$

where E_{cls} is the energy function for region classification and c_i indicates the class label, i.e. captioned ($c_i = 1$) or non-captioned ($c_i = 0$) region.

Proposal refinement (PR). We further refine the proposed regions by leveraging the caption information, akin to [26]. That is, we minimize the following energy function:

$$E_{box} = \sum_{i \in \mathbf{pos}} E_i^{box}(\Delta\hat{B}_i|\hat{B}_i, \hat{S}_i), \tag{2}$$

where $\Delta\hat{B}_i$ is the offsets to the proposal $\hat{B}_i$ estimated in the region proposal task and **pos** denotes the set of positive proposals.

Region caption generation (CG). To generate a caption for each region, we consider the relation graph $\mathcal{G}$ to minimize

$$E_{cap} = \sum_{i \in \mathbf{pos}} E_i^{cap}(S_i|\mathcal{G}). \tag{3}$$

3.2 Overview of the Framework

Figure 2 depicts a high-level sketch of the proposed framework for dense captioning. The input image is first processed by a region proposal network (RPN) [14] to attain proposals from which the regional visual representations $\{\mathbf{v}_i\}_{i=1}^{N_r}$ are extracted. A graph $\mathcal{G}$, whose edge weights are calculated by GREatt, is constructed over $\mathbf{v}_i$ and employed to obtain a relational representation $\mathbf{g}_i$. Both $\mathbf{v}_i$ and $\mathbf{g}_i$ are then fed into the captioning module to generate a caption embedding. Finally, the caption embedding along with $\mathbf{v}_i$ and $\mathbf{g}_i$ are used to classify the region as captioned or non-captioned class. In the following subsections, we introduce the formation of $\mathbf{g}_i$ and describe the proposal refinement and caption nets in detail.

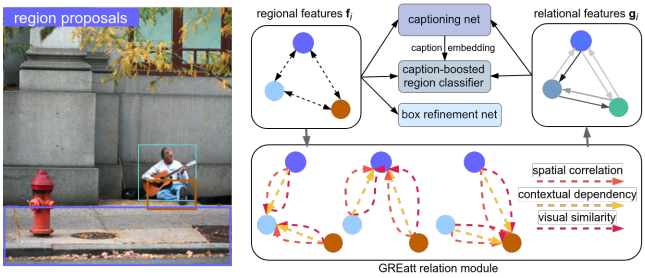

Figure 2: The proposed framework divides the dense captioning problem into four sub-tasks tackled by four sub-modules, i.e. a) region proposal network, b) region classifier, c) proposal refinement net, and d) captioning net. It features Geometry-aware Relational Exemplar attention mechanism (GREatt), a relation module which is constructed on different types of relationships among region proposals and learns region-specific features which account for the most relevant context in the image. In addition, the proposed region classifier is learned on relational features delivered by GREatt and additionally on the caption information.

3.3 Geometry-aware Relational Exemplar Attention

In this section, we discuss the construction of the graph structure between the proposed regions and demonstrate how one can learn a powerful representation by considering the latent relationships between the proposed regions. To this end, we propose the Geometry-aware Relational Exemplar Attention (GREatt) module.

Having the region proposals $\hat{B}$ and their representations $\mathcal{V}$ generated by the RPN, we aim to learn contextual representations which are constructed on different types of relationships, namely, visual relationships and geometry relationships. The visual relationships account for the contextual dependency and visual similarity. The geometry relationships explain the spatial correlation and arrangement between any two region proposals (i.e. bounding boxes).

Given the individual regional features $\mathbf{v}_i \in \mathbb{R}^{D_v}, i = 1, ..., N_r$, GREatt calculates the relational features $\mathbf{g}_i$ by

$$\mathbf{g}_i = \mathbf{v}_i + \sum_{j=1}^{N_r} \alpha_{i,j} \mathbf{v}_j, \ \forall i, \tag{4}$$

$$\alpha_{i,j} = f_\alpha(\alpha^g_{i,j}, \alpha^v_{i,j}, \alpha^\omega_{i,j}), \tag{5}$$

where $\alpha_{i,j}$ reflects how much $\mathbf{v}_j$ should be associated with $\mathbf{v}_i$ in region classification and caption generation. $f_\alpha(\cdot)$ is GREatt *contextual function* (details provided in Sec. 3.3.4) that learns to embed three different relationships into $\alpha_{i,j}$. These relationships are 1) contextually dependent relation $\alpha^g_{i,j}$, 2) visually similar relation $\alpha^v_{i,j}$, and 3) geometry relation $\alpha^\omega_{i,j}$. The first two relations are based on the visual representation and the third relation is based on the geometry representation. In the following paragraphs, we describe how $\alpha^g_{i,j}$, $\alpha^v_{i,j}$, and $\alpha^\omega_{i,j}$ can be addressed computationally and discuss the possible options to implement f_α.

3.3.1 Contextually Dependent Relations $\alpha^g_{i,j}$. Used in [8, 22] for the object detection task, and in [20] for aggregating representations in graphical structures for graph classification, concatenating one representation (e.g. $\mathbf{v}_j$) to another (e.g. $\mathbf{v}_i$) augments the information that might be missing in $\mathbf{v}_i$, but can be provided by $\mathbf{v}_j$. Specifically, we define $\alpha^g_{i,j}$ as

$$\alpha'^g_{i,j} = W^g_\alpha(\mathbf{v}'_i \,||\, \mathbf{v}'_j), \ \mathbf{v}'_i = \tanh(W^g_v \mathbf{v}_i), \tag{6}$$

$$\alpha^g_{i,j} = \frac{\exp(\alpha'^g_{i,j})}{\sum_{j=1}^{N_r} \exp(\alpha'^g_{i,j})}, \ i = 1, ..., N_r, \tag{7}$$

where $||$ denotes concatenation, $\tanh(\cdot)$ is the hyperbolic tangent activation function, $W^g_v \in \mathbb{R}^{D_w \times D_v}$, and $W^g_\alpha \in \mathbb{R}^{1 \times 2D_w}$. Concatenation is used to associate any two feature vectors, i.e. $\mathbf{v}'_i$ and $\mathbf{v}'_j$ to learn how much importance $\mathbf{v}_j$ has to $\mathbf{v}_i$ through W^g_α and W^g_v. It is worth noting that applying concatenation imposes a *directedness* assumption on the link between any two regional features $\mathbf{v}_i$ and $\mathbf{v}_j$ since, in general, $\alpha_{i,j} \neq \alpha_{j,i}$, when $i \neq j$.

3.3.2 Visually Similar Relations $\alpha^v_{i,j}$. We introduce two visual relations based on dot-product and cosine distance. We categorize the relation modules based on these two operations together because they naturally capture the similarity between two representations and can help enhance the visual signals by identifying other similar ones.

Scaled Dot-Product: Firstly introduced in [19], scaled dot-product (SDP) attention mechanism calculates $\alpha^s_{i,j}$ as

$$\alpha^s_{i,j} = \frac{(W^s_{v_1} \mathbf{v}_i) \cdot (W^s_{v_2} \mathbf{v}_j)}{\sqrt{D_w}}, \tag{8}$$

where $W^s_{v_1}, W^s_{v_2} \in \mathbb{R}^{D_w \times D_v}$. What is worth noting is that $\alpha^s_{i,j}$ in our framework is used to weight $\mathbf{v}_i$ directly, whereas it is used to weight another embedding projected from $\mathbf{f}_i$ in [19].

Cosine Similarity: Eq. (8) learns the attention weights according to the correlation of $W^s_{v_1} \mathbf{v}_i$ and $W^s_{v_2} \mathbf{v}_i$ measured by the dot-product. Used for learning the attention weighting in Neural Turing Machine [4], cosine similarity measures the angle between vectors:

$$\alpha^c_{i,j} = \frac{(W^s_{v_1} \mathbf{v}_i) \cdot (W^s_{v_2} \mathbf{v}_j)}{||W^s_{v_1} \mathbf{v}_i|| \cdot ||W^s_{v_2} \mathbf{v}_j||}. \tag{9}$$

We model the relational weight $\alpha^v_{i,j}$, which is determined by visual similarity between two vectors in Eq. (5), with either $\alpha^s_{i,j}$ or $\alpha^c_{i,j}$, i.e.

$$\alpha^v_{i,j} = \gamma^s \alpha^s_{i,j} + \gamma^c \alpha^c_{i,j}, \tag{10}$$

where $\gamma^s, \gamma^c \in \{0, 1\}$ are hyperparameters deciding either $\alpha^s_{i,j}$ or $\alpha^c_{i,j}$ to be adopted. This marks the difference between $\alpha^v_{i,j}$ and $\alpha^g_{i,j}$ where the latter learns to identify dependent context with respect to the representation $\mathbf{v}_i$.

3.3.3 Geometry Relations $\alpha^\omega_{i,j}$. Relative geometry relation that encodes the spatial relationship between two proposals has shown to be important when modeling contextual information [8, 27]. We model it with $\alpha^\omega_{i,j}$ [8], where

$$\alpha^\omega_{i,j} = f^\omega(W^\omega_2 \sigma^\omega(W^\omega_1 \mathbf{b}_{i,j})), \tag{11}$$

$$\mathbf{b}_{i,j} = [\log(\frac{|x_i - x_j|}{w_i}), \log(\frac{|y_i - y_j|}{h_i}), \log(\frac{w_i}{w_j}), \log(\frac{h_i}{h_j})]^T. \tag{12}$$

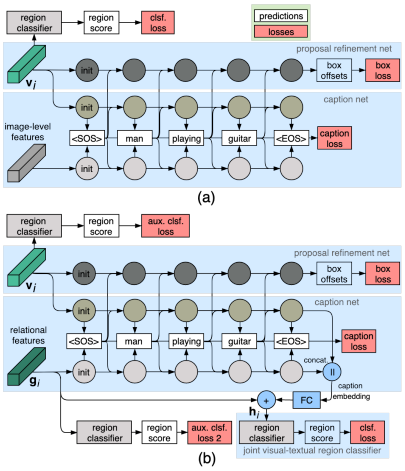

Figure 3: Architectures of (a) Yang et al. [26] and (b) our proposed model. Both architectures consist of three RNN branches which comprise the proposal refinement and caption nets. The proposed model is empowered by the features learned with GREatt and a joint visual-textual region classifier.

$\mathbf{b}_{i,j}$ is the geometry features encoded by center coordinates (x_*, y_*), width and height of the bounding box w_*, and h_*. Since $x_i - x_j$ or $y_i - y_j$ can be zero, we set a lower bound (i.e. 10^{-3}) on them. $f^\omega : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ can be: 1) $\max(x, 10^{-3})$ similar to ReLU or 2) a softmax operation. We empirically find that $\omega_{i,j}$ tends to be rather uniformly distributed when learning it with ReLU for any fixed i and $j = 1, ..., N_r$. Hence, we adopt softmax in f^ω throughout the experiments.

3.3.4 Contextual Function f_α. Before introducing how one can define the contextual function f_α in Eq. (5), we would like to emphasize the differences in three visual relationships, defined in Eqs. (7)-(9). We hypothesize that the first scheme (i.e. concatenation-based) learns how to identify the essential contextual cues with respect to each proposal, while the latter two (similarity-based) learn how to enhance the evidence on the existence of the similar content to be recognized. This further leads to the assumption that these two types of interactions in visual domain can potentially provide distinct contextual information. With this hypothesis, we write the contextual function f_α as

$$f_\alpha(\alpha^g_{i,j}, \alpha^v_{i,j}, \alpha^\omega_{i,j}) = \frac{\alpha^\omega_{i,j} \exp(\gamma^g \alpha^g_{i,j} + \alpha^v_{i,j})}{\sum_{j=1}^{N_r} \alpha^\omega_{i,j} \exp(\gamma^g \alpha^g_{i,j} + \alpha^v_{i,j})}, \tag{13}$$

where γ^g is predefined hyperparameters. $\alpha^\omega_{i,j}$, $\alpha^g_{i,j}$ $\alpha^v_{i,j}$ are defined in Eqs. (11), (7), and (10), respectively. We empirically validate the hypothesis by studying the quantities of the attentions (provided in Figure 5) estimated from different schemes.

3.4 Proposal Refinement and Caption Nets

Yang et al. [26] proposed a triple-stream RNN architecture (shown in Fig. 3(a)) for refining the proposals generated by the region proposal network (RPN) [14] and generating the captions. We mainly follow a similar architecture, i.e. the proposal refinement net RNN^r_t, and the caption nets composed by RNN^v_t and RNN^g_t, $t = 0, ..., T+1$, where t indexes the RNN steps with $T+1$ being the maximal length of a caption including the start (<SOS>) and end (<END>) symbols. At step t, each RNN^*_t receives a word predicted in step $(t-1)$ and updates its hidden states $h^*_t \in \mathbb{R}^{D_r}$ and cell states $c^*_t \in \mathbb{R}^{D_r}$.

The main difference between the proposed architecture and that in [26] can be seen in Figure 3. While RNN^r_0 and RNN^v_0 take $\mathbf{v}_i$ as input, RNN^g_0 is fed with the context features $\mathbf{g}_i$ learned with GREatt instead of image-level features. The hidden state h^r_τ is used to predict the offsets to the x and y coordinates, width and height of the region proposals with a MLP, where τ is the step that predicts (<END>). As for caption branches, h^v_t and h^g_t are concatenated to make a prediction on the distribution of the next word through another MLP. The proposed context features $\mathbf{g}_i$, adapted with respect to each region, are endowed with contextual relationships captured in the scene. By contrast, the image-level features devised by [26] in Figure 3(a) can only provide a fixed and generic guidance to all the regions to be captioned.

3.5 Joint Visual-Textual Region Classifier

Conventionally, the region classifier estimates $P(c_i|\mathcal{V})$ which indicates that the prediction is purely conditioned on corresponding regional features. In this work, we aim to improve the classifier by replacing the target of estimation with $P(c_i|\hat{B}_i, S_i, \mathcal{G}, \mathcal{V})$, as shown in Eq. (1), which additionally considers the learned relationships among the proposals and the caption information. Specifically, we estimate $P(c_i|\cdot)$ with

$$P(c_i|\mathcal{I}, \hat{\mathcal{B}}_i, \mathcal{S}_i, \mathcal{G}, \mathcal{V}) = \text{MLP}_{rc}(\mathbf{h}_i), \tag{14}$$

$$\mathbf{h}_i = \mathbf{g}_i + W^r(\mathbf{h}^v_\tau || \mathbf{h}^g_\tau). \tag{15}$$

In the above equation, the relational representation $\mathbf{g}_i$ is defined in Eq. (4), c_i is the class label defined in Eq. (1), $\text{MLP}_{rc}(\cdot)$ represents a MLP with a sigmoid activation function placed at the output, and $W^r \in \mathbb{R}^{D^v \times 2D^r}$ is learned to project the caption embedding to the same domain in which the visual features reside. The rationale behind this approach is two-fold:

1) **Better vision-caption consistency:** Projecting (or "translating") caption embedding back to the visual domain in which the classification is performed can potentially improve the model's consistency between the generated caption embedding and the embedding of the visual counterpart.

2) **Mimicking human annotator's behavior:** We hypothesize that two actions in the annotation process, i.e. 1) sizing up the bounding boxes around the interesting contents and 2) captioning, are bonded in both directions. A human annotator's attention may be drawn to a relatively salient object, caption it, and then refine the bounding area and the caption. This indicates that the caption information can as well provide evidence to infer the region saliency.

3.6 The Losses

The proposed model is trained by minimizing the total loss L addressing all sub-tasks, i.e. the region proposal (**RP**), region classification (**RC**), proposal refinement (**PR**), and caption generation

(**CG**) sub-tasks as presented in Sec. 3.1. Specifically,

$$L = L^{RP} + L^{RC} + L^{PR} + L^{CG}, \tag{16}$$

$$L^{RP} = \alpha_1 L^{RP}_{det} + \alpha_2 L^{RP}_{box}, \tag{17}$$

$$L^{RC} = \beta(L^{RC}_{v} + L^{RC}_{g} + L^{RC}_{h}), \tag{18}$$

$$L^{PR} = \gamma L^{PR}_{box}, \tag{19}$$

$$L^{CG} = L^{cap}, \tag{20}$$

where

$$L^{RP}_{det} = \alpha_r \sum_{i=1}^{N_r} L^{RP}_{det,i},\ L^{RP}_{box} = \alpha_r \sum_{i=1}^{N_r} L^{RP}_{box,i}, \tag{21}$$

$$L^{RC}_{v} = \alpha_r \sum_{i=1}^{N_r} L^{RC}_{v,i},\ L^{RC}_{g} = \alpha_r \sum_{i=1}^{N_r} L^{RC}_{g,i},\ L^{RC}_{h} = \alpha_r \sum_{i=1}^{N_r} L^{RC}_{h,i}, \tag{22}$$

$$L^{PR}_{box} = \frac{1}{|\mathbf{pos}|} \sum_{i \in \mathbf{pos}} L^{PR}_{box,i},\ L^{cap} = \frac{1}{|\mathbf{pos}|} \sum_{i \in \mathbf{pos}} L^{cap}_{i}, \tag{23}$$

$\alpha_r = \frac{1}{N_r}$ is a normalization factor, **pos** represents the set of positive regions in the batch of N_r regions, and $|\mathbf{pos}|$ denotes the size of the set. α_1, α_2, β, and γ are hyperparameters.

RP Losses. Per-sample losses for training RPN are the detection loss $L^{RP}_{det,i}$ and regression loss $L^{RP}_{box,i}$. The former is defined as the cross-entropy function over the predicted and the ground-truth classes, in which the classes refer to either $c_i = 0$, negative non-captioned regions, or $c_i = 1$, positive captioned regions. The latter loss is defined by the smooth L1 function used in [14].

RC Losses. Region classification involves three losses with respect to $\mathbf{v}_i$, $\mathbf{g}_i$, and $\mathbf{h}_i$, respectively. These three losses are defined as the cross-entropy function over the predicted and the ground-truth classes. $L^{RC}_{v,i}$, $L^{RC}_{g,i}$, and $L^{RC}_{h,i}$ are evaluated based on the ground-truth classes and the predicted classes given by $\mathrm{MLP}_{rc}(\mathbf{v}_i)$, $\mathrm{MLP}_{rc}(\mathbf{g}_i)$, and $\mathrm{MLP}_{rc}(\mathbf{h}_i)$, respectively. As we take the predictions from $\mathrm{MLP}_{rc}(\mathbf{h}_i)$ during evaluation, $\mathrm{MLP}_{rc}(\mathbf{v}_i)$ and $\mathrm{MLP}_{rc}(\mathbf{g}_i)$ are treated as auxiliary predictions which are meant for enhancing the discriminative power of individual $\mathbf{v}_i$ and $\mathbf{g}_i$. Note that these three predictions share the same set of parameters from $\mathrm{MLP}_{rc}(\cdot)$. Minimizing L^{RC} corresponds to minimizing E_{cls} in Eq. (1).

PR Loss. Proposal refinement loss $L^{PR}_{box,i}$ in Eq. (23), same as $L^{RP}_{box,i}$, is defined by the smooth L1 function over coordinates of the predicted box and the ground-truth box. Note that minimizing L^{PR}_{box} corresponds to minimizing E_{box} in Eq. (2).

CG Loss. Caption generation loss L^{cap}_{i}, defined over word distributions in i^{th} ground-truth caption and predicted word distribution, is measured by the cross-entropy function. Minimizing L^{cap} corresponds to minimizing E_{cap} in Eq. (3).

4 EXPERIMENTS

4.1 Dataset

All the experiments are conducted on the Visual Genome dataset [11], created for various vision-language tasks such as dense captioning, VQA, and SGG. For the DC task, the annotations with region bounding boxes and corresponding captions are provided. Even though three versions, V1.0, V1.2, and V1.4 are available, we compare different DC models on V1.2 since the changes in V1.4 do not affect the data used in the DC task, and the state-of-the-art models are extensively evaluated mainly on V1.2 [26].

4.2 Experimental Setting

Following the split protocol provided in [9, 26], the images are divided into training, validation, and test sets, comprising 77398, 5000, and 5000 images, respectively. The provided bounding box annotations are often highly overlapping, hence all the annotations with IoU > 0.7 of their bounding boxes are merged into one [26]. Accordingly, each merged region across all sets can contain multiple reference captions, in which a caption for a merged region is randomly drawn during training. The parameter settings in the RPN strictly follow those in [26].

4.3 Hyperparameter Setting and Model Training

The hyperparameters defined in Eqs. (21)–(23) are given by $\alpha_1 = 0.1$, $\alpha_2 = 0.05$, $\beta = 0.1$, and $\gamma = 0.01$. The input image is resized so that the longer side is of 720 pixels. The most frequent 10,000 words are used and those excluded are replaced with an <UNK> (unknown word) symbol. Hence, this amounts to 10,003 words (10,000 most frequent words plus <SOS>, <END>, and <UNK>) available for the caption model. Regions with captions longer than 10 words are discarded, and each caption of the remaining ones is padded with <SOS> in the beginning and <EOS> at the tail. The proposal refinement and caption nets adopt three seperate LSTMs with 512 hidden units. The experiments with three visual features: VGG16 [17], which has two fully-connected layers both consisting of 4096 units at the output, extracts 4096-dimensional features for each region proposal. ResNet50 and ResNet101 [5] extract 1024-dimensional features. The training batch size is set to be 1 (i.e. a single image) with $N_r = 256$ (referred in Eqs. (21)–(23)) region proposals evenly sampled from positive and negative proposals in the RPN.

All the models throughout the experiments are trained with stochastic gradient descent with momentum set at 0.98. The initial learning rate is 0.001, reduced by half every 100,000 steps ($\approx$ 1.3 epochs). Models with VGG16 are trained only with `Conv4_*` and `Conv5_*` being fine-tuned in the periods of 1.5–4 epochs and 5.5–10 epochs. Models with ResNet50 and ResNet101 are trained with 4^{th} residual block being fine-tuned in 0–1.5 epochs and 4–5.5 epochs, and 3^{rd} residual block as well being fine-tuned in the periods of 1.5–4 epochs and 5.5–10 epochs. We follow the stage-wise training scheme suggested in [26] to train the proposed models. Firstly, we train the RPN, the proposal refinement net, and the caption net end-to-end. Here at this stage, only one caption LSTM (i.e. RNN^{v}_{t}, but not RNN^{g}_{t}) which receives the regional features $\mathbf{v}_i$ is trained. Secondly, we add the second LSTM stream RNN^{g}_{t} with the context features $\mathbf{g}_i$ into the models and fine-tune the other parts. Finally, we fine-tune the models and feed the region classifier MLP_{rc} with $\mathbf{h}_i$ of Eq. (15), the features containing both visual and caption embedding. This training scheme helps the models in which the performance of each component is based upon each other, e.g., the proposal refinement net can only start to refine the proposals generated by the RPN once the RPN has learned to produce reasonable proposals.

4.4 Evaluation Metric

The main metric adopted to evaluate the DC models is the mean average precision (mAP) that jointly considers the goodness of the region proposals and the generated captions in terms of IoU and METEOR [2] scores with the ground-truth annotations [9]. mAP is calculated by averaging the average precision scores evaluated at different IoU thresholds, {0.3, 0.4, 0.5, 0.6, 0.7}, and METEOR thresholds, {0, 0.05, 0.1, 0.15, 0.2, 0.25}. Besides, we also adopt mAP@{IoU=0.3,0.4,0.5,0.6,0.7} and mAP@{*small*,*medium*,*large*} (evaluated at proposals smaller than 48^2, between $48^2 - 108^2$, and larger than 108^2 pixels) to facilitate a deeper comparison between models.

Table 1: The representation of different attention modules defined by γ^g and γ^v in Eqs. (13) and (10). The geometry relationship captured by $\alpha^{\omega}_{i,j}$ is considered by all different modules listed.

models	ctx	sim(sdp)	sim(cos)	ctx+sim(sdp)	ctx+sim(cos)
(γ^g, γ^s, γ^c)	(1,0,0)	(0,1,0)	(0,0,1)	(1,1,0)	(1,0,1)

4.5 Quantitative Comparison

We compare the proposed framework with the state-of-the-art DC models [26]. The pioneer DC framework from Johnson et al. [9] reported the performance of their models on Visual Genome V1.0, and thus a direct comparison with their results is not possible. It is difficult to compare results also from many other different DC models since, to the best of our knowledge, the only notable and reliable results one can compare against are from [26]. In the following subsections, we compare different models of our own with configurations listed in Table 1 and those described in [26].

4.5.1 Comparing with State of the Art. We have tried our best to replicate the best performing architecture reported in [26], and the highest mAP we can obtain is 9.72, which is reasonably close to 9.96 reported in their work. First, we study whether the models with added geometry relation and a single visual attention mechanism can improve over those without. The results in the second to the fourth rows (against those in the first row) in Table 2 highlight the effect of a model that considers a single visual relationship (implemented by either $\alpha^g_{i,j}$, $\alpha^s_{i,j}$, or $\alpha^c_{i,j}$, referred in Sec. 3.3.1 and 3.3.2) and the geometry relationship captured by $\alpha^{\omega}_{i,j}$ (referred in Sec. 3.3.3). We observe the consistent improvement made by the proposed models in the mAP across VGG16, ResNet50, and ResNet101 visual features.

Moving to the fifth row onwards in Table 2, one can observe the best mAP is obtained from the proposed architecture when GREatt (with geometry, concatenation-based, cosine distance based attention modules simultaneously employed) and caption-boosted classifier (described in Sec. 3.5) are used. The best result with VGG16 achieves 10.23, which, to date, surpasses the state-of-the-art number that has been reported. A greater margin of improvement in mAP can be observed (+5.3%, +5.4%, +6.23% with VGG16, ResNet50, and ResNet101, respectively) when comparing the best performing models of ours and those in [26].

We also report the mAP at different proposal sizes in Table 3. One can easily observe a similar trend where our architectures bring steady improvement for all proposal size groups. This shows that our models do not favor proposals of certain sizes, but provide all-around improvement over arbitrary sizes of proposals. Moreover, the largest improvement often comes from the mAP@small, indicating that our context modeling scheme has the largest positive impact on making inference on the small region proposals.

4.5.2 Comparing Models with Different Attention Modules. Here, we study the effect on varying computational attention modules proposed. The aim of the study is to answer whether (1) models with GREatt employing one geometry and two visual attention mechanisms (out of three presented in Sec. 3.3.1 and 3.3.2), improves the results over those with one geometry and a single visual attention mechanisms, and (2) models equipped with the region classifier exposed with caption information improves the results over those without.

Fusing attentions. From Table 2, one can also compare two types of models: (1) those with combined visual attentions (presented in the fifth to sixth rows) and (2) those with single visual attention (presented in the second to fourth rows). We compare them by picking the best result (e.g. mAP) that a model in each type can achieve. One can observe the improvement in mAP made by the models with combined visual attentions on VGG16 and ResNet50, but not on ResNet101.

Classifying regions with captions. From Table 2, one can observe a significant improvement made by the models with the caption-boosted region classifier based on all visual feature extractors. From Table 3, we see that the largest improvements are made on mAP@small, demonstrating that the caption information is crucial to make smaller RoIs detectable.

4.6 Qualitative Results

We compare qualitative results from our model (i.e. the best performing one, "ctx+sim(cos)" model listed in Table 1) and the one from [26] with ResNet101 features in Figure 4. Clearly shown, Yang's model tends to ignore the relationship (Figure 4(a): missing "on a cutting board"), or fail to encode the context (e.g. Figure 4(e): missing "laptop" in the caption). By contrast, our proposed model not only captures the correct relationships, but also correctly recognizes and names the objects in the context.

Next, we study attention weights (i.e. $\alpha^{\omega}_{i,j}$, $\alpha^g_{i,j}$, and $\alpha^c_{i,j}$) learned to capture different relationships in Figure 5. One can observe that three types of weights attend to quite distinct and sometimes complementary sets of areas with respect to each proposal. While the $\alpha^{\omega}_{i,j}$ and $\alpha^g_{i,j}$ tend to capture the necessary context (i.e. the tennis field in this example), cosine distance based visual attention $\alpha^c_{i,j}$ tends to capture visually similar context. For example, while the subject in the proposal is the tennis player in the distance, it tries to retrieve similar person-like objects. The combined attention is able to capture the most relevant context, e.g. in Figure 5(c), it identifies who is holding the racket, and in Figure 5(d), it captures almost the whole tennis court to be able to recognize that the clock is in the court.

5 CONCLUSIONS

In this paper, we visited the dense captioning task, which serves as a powerful means to facilitate multimodal context understanding

Table 2: Quantitative results of models with VGG16, ResNet50, and ResNet101, respectively, on Visual Genome V1.2. `models` column shows models with varying visual attention modules named in Table 1. `cap` indicates if the caption embedding is added when classifying the region proposals. The best model with respect to each metric is highlighted in bold, and the second best is underlined. (*) indicates the figure reported in [26] while the other figures are obtained from our implementation. @n indicates the mAP score evaluated at IoU=n, $n = \{0.3, 0.4, 0.5, 0.6, 0.7\}$.

		VGG16						ResNet50						ResNet101					
models	cap	mAP	@0.3	@0.4	@0.5	@0.6	@0.7	mAP	@0.3	@0.4	@0.5	@0.6	@0.7	mAP	@0.3	@0.4	@0.5	@0.6	@0.7
Yang et al. [26]	-	9.72 (9.96*)	15.13	13.16	10.25	6.77	3.28	10.89	16.85	14.62	11.55	7.73	3.70	11.92	18.16	15.83	12.58	8.68	4.37
ctx	-	9.85	15.22	13.25	10.41	6.96	3.39	11.00	16.48	14.52	11.70	8.12	4.14	12.51	17.73	15.76	12.84	9.14	4.82
sim(sdp)	-	9.88	15.29	13.32	10.44	6.96	3.39	11.00	16.51	14.58	11.68	8.11	4.12	11.79	17.95	15.67	12.43	8.59	4.30
sim(cos)	-	9.73	15.10	13.15	10.29	6.78	3.33	11.07	17.09	14.85	11.77	7.90	3.75	11.73	17.91	15.64	12.40	8.49	4.20
ctx+sim(sdp)	-	9.97	15.33	13.40	10.55	**7.06**	**3.48**	11.03	16.54	14.63	11.70	8.14	4.14	12.14	18.37	16.09	12.87	8.88	4.96
ctx+sim(cos)	-	9.93	15.90	13.59	10.36	6.68	3.11	11.10	16.62	14.73	11.77	8.20	4.19	12.15	18.37	16.09	12.88	8.90	4.48
ctx+sim(sdp)	✓	10.22	16.30	14.00	10.71	6.91	3.14	11.39	17.03	14.98	12.09	8.43	4.40	12.52	**18.72**	**16.37**	13.23	9.34	4.93
ctx+sim(cos)	✓	**10.23**	**16.39**	**14.04**	**10.76**	6.85	3.13	**11.48**	**17.14**	**15.08**	**12.15**	**8.56**	**4.45**	**12.67**	18.39	16.32	**13.44**	**9.79**	**5.40**

Table 3: Results on comparing models on mAP@{small, medium, large}, denoted by @S, @M, @L.

		VGG16			ResNet50			ResNet101		
models	cap	@S	@M	@L	@S	@M	@L	@S	@M	@L
Yang et al. [26]		3.99	8.15	14.22	4.09	9.08	16.03	4.78	9.94	17.35
ctx	-	4.03	8.39	14.46	4.19	8.82	16.26	4.61	9.98	17.82
sim(sdp)	-	4.00	8.25	14.36	4.33	8.97	16.24	4.28	9.57	17.44
sim(cos)	-	4.14	8.23	14.19	4.46	9.21	16.16	4.36	9.68	17.42
ctx+sim(sdp)	-	3.83	8.34	**14.63**	4.35	9.13	16.22	4.46	9.86	17.79
ctx+sim(cos)	-	3.95	8.53	14.24	4.48	9.11	16.27	4.52	10.15	17.71
ctx+sim(sdp)	✓	4.19	**8.63**	14.46	4.25	**9.52**	16.44	4.83	**10.55**	18.14
ctx+sim(cos)	✓	**4.39**	8.61	14.42	**4.68**	9.34	**16.80**	**4.94**	10.48	**18.39**

Figure 4: Qualitative comparison between the proposed method and that proposed by Yang et al. [26]. More relationships and context information are revealed in the captions generated by our method. Captions (ours / [26]): (a) two pieces of cheese on a cutting board / a slice of yellow cheese, (b) a blue bus on the road / a blue and white bus, (c) green trees on the side of the tracks / green leaves on the tree, (d) a person skiing on the snow / person wearing blue pants, (e) screen of laptop computer / a computer monitor.

and learning. We proposed an improved architecture which features (1) a Geometry-aware Relational Exemplar attention (GREatt) mechanism and (2) a joint visual-textual relational region classifier, for the dense captioning problem. Our proposed methods bring significant improvements over the state-of-the-art results. In addition,

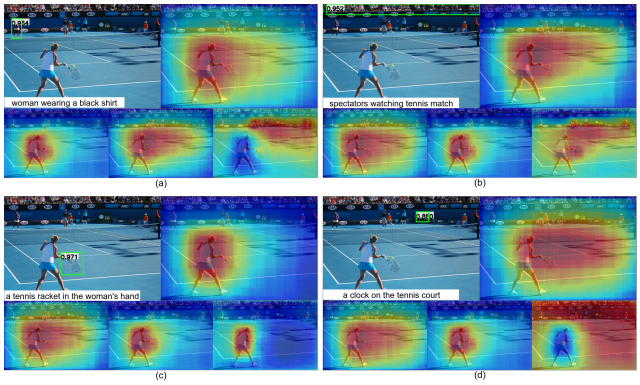

Figure 5: Different attention mechanisms jointly learned with model "ctx+sim(cos)" (referred in Table 1). Each set of image, from top to bottom, left to right, shows 1) detection and caption results, 2) combined attention, $\alpha_{i,j}$, 3) geometry attention, $\alpha^{\omega}_{i,j}$, 4) contextual dependent visual attention, $\alpha^{g}_{i,j}$, and 5) visually similar and supported attention, $\alpha^{c}_{i,j}$.

we demonstrated that GREatt captures varying and meaningful contexts for different regions to construct contextually dependent and region-specific features. The proposed region classifier which learns on the subspace shared with visual and textual embeddings has also demonstrated its effectiveness and led to improvements in almost all metrics. Qualitatively, our proposed models, comparing to the prior arts, are more capable of generating captions that capture relationships between objects and are able to accurately recognize and name the objects in the context. However, how to optimally combine the heterogeneous types of attention still remains an open question, and we leave it as a future avenue of research.

ACKNOWLEDGMENTS

This work has been funded by the Academy of Finland project number 313988 (DeepGraph), and the European Union's Horizon 2020 research and innovation programme under grant agreement No. 780069 (MeMAD). We also acknowledge the Aalto University's Aalto Science IT project and CSC âĂŞ IT Center for Science Ltd. for providing computer resources and NVIDIA Corporation for donation of GPU for this research.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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*. 6077–6086.

[2] Satanjeev Banerjee and Alon Lavie. 2005. 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*. 65–72.

[3] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*. Springer, 15–29.

[4] Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401* (2014).

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[6] Sen He, Hamed R. Tavakoli, Ali Borji, and Nicolas Pugeault. 2019. A Synchronized Multi-Modal Attention-Caption Dataset and Analysis. *CoRR* abs/1903.02499 (2019). arXiv:1903.02499 http://arxiv.org/abs/1903.02499

[7] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv.* 51, 6, Article 118 (Feb. 2019), 36 pages. https://doi.org/10.1145/3295748

[8] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. 2018. Relation networks for object detection. In *Computer Vision and Pattern Recognition (CVPR)*, Vol. 2.

[9] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4565–4574.

[10] A. Karpathy and L. Fei-Fei. 2017. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (April 2017), 664–676. https://doi.org/10.1109/TPAMI.2016.2598339

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

[12] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association*.

[13] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob J. Verbeek. 2017. Areas of Attention for Image Captioning. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 1251–1259.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[15] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in neural information processing systems*. 4967–4976.

[16] Rakshith Shetty, Hamed Rezazadegan Tavakoli, and Jorma Laaksonen. 2018. Image and Video Captioning with Augmented Neural Architectures. *IEEE MultiMedia* 25, 2 (2018), 34–46. https://doi.org/10.1109/MMUL.2018.112135923

[17] K. Simonyan and A. Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[18] Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying Attention to Descriptions Generated by Image Captioning Models. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 2506–2515.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[20] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations* (2018). https://openreview.net/forum?id=rJXMpikCZ

[21] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3156–3164. https://doi.org/10.1109/CVPR.2015.7298935

[22] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7794–7803.

[23] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. 2018. LinkNet: Relational Embedding for Scene Graph. In *Advances in Neural Information Processing Systems*. 558–568.

[24] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[25] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 670–685.

[26] Linjie Yang, Kevin D Tang, Jianchao Yang, and Li-Jia Li. 2017. Dense Captioning with Joint Inference and Visual Context.. In *CVPR*. 1978–1987.

[27] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *European Conference on Computer Vision (ECCV)*. 684–699.