

Improving Image Captioning with Better Use of Captions

Zhan Shi*, Xu Zhou†, Xipeng Qiu†, Xiaodan Zhu*

*Ingenuity Labs Research Institute, Queen's University

*Department of Electrical and Computer Engineering, Queen's University

†School of Computer Science, Fudan University

{z.shi, xiaodan.zhu}@queensu.ca, {16210240095, xqiu}@fudan.edu.cn

Abstract

Image captioning is a multimodal problem that has drawn extensive attention in both the natural language processing and computer vision community. In this paper, we present a novel image captioning architecture to better explore semantics available in captions and leverage that to enhance both image representation and caption generation. Our models first construct caption-guided visual relationship graphs that introduce beneficial inductive bias using weakly supervised multi-instance learning. The representation is then enhanced with neighbouring and contextual nodes with their textual and visual features. During generation, the model further incorporates visual relationships using multi-task learning for jointly predicting word and object/predicate tag sequences. We perform extensive experiments on the MSCOCO dataset, showing that the proposed framework significantly outperforms the baselines, resulting in the state-of-the-art performance under a wide range of evaluation metrics. The code of our paper has been made publicly available.¹

1 Introduction

Automatically generating a short description for a given image, a problem known as image captioning (Chen et al., 2015), has drawn extensive attention in both the natural language processing and computer vision community. Inspired by the success of encoder-decoder frameworks with the attention mechanism, previous efforts on image captioning adopt variants of pre-trained convolution neural networks (CNN) as the image encoder and recurrent neural networks (RNN) with visual attention as the decoder (Lu et al., 2017; Anderson et al., 2018; Xu et al., 2015; Lu et al., 2018).

Many previous methods translate image representation into natural language sentences without

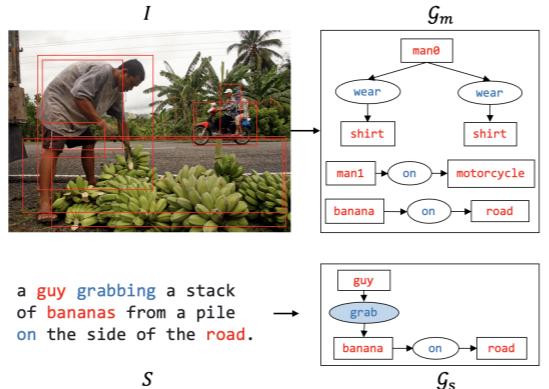


Figure 1: Visual relationship graphs from a pre-trained detection model (Yao et al., 2018) (upper) and from the ground-truth caption (bottom).

explicitly investigating semantic cues from texts and images. To remedy that, some research has also explored to detect high-level semantic concepts presented in images to improve caption generation (Wu et al., 2016; Gan et al., 2017; You et al., 2016; Fang et al., 2015; Yao et al., 2017). It is believed by many that the inductive bias that leverages structured combination of concepts and visual relationships is of importance, which has led to better captioning models (Yao et al., 2018; Guo et al., 2019; Yang et al., 2019). These approaches obtain visual relationship graphs using models pre-trained from visual relationship detection (VRD) datasets, e.g., Visual Genome (Krishna et al., 2017), where the visual relationships capture semantics between pairs of localized *objects* connected by *predicates*, including spatial (e.g., *cake-on-desk*) and non-spatial semantic relationships (e.g., *man-eat-food*) (Lu et al., 2016).

As in many other joint text-image modeling problems, it is crucial to obtain a good semantic representation in image captioning that bridges semantics in language and images. The existing approaches, however, have not yet adequately leveraged the semantics available in captions to con-

Cải thiện chú thích hình ảnh bằng cách sử dụng chú thích tốt hơn

Chiến Thi , Từ Châut , Tây Bằng Khâu, Tiều Đan Chu

Viện nghiên cứu Ingenuity Labs, Đại học Queen

Khoa Kỹ thuật Điện và Máy tính, Đại học Queen

†Khoa Khoa học máy tính, Đại học Fudan

{z.shi, xiaodan.zhu}@queensu.ca,{16210240095, xqiu}@fudan.edu.cn

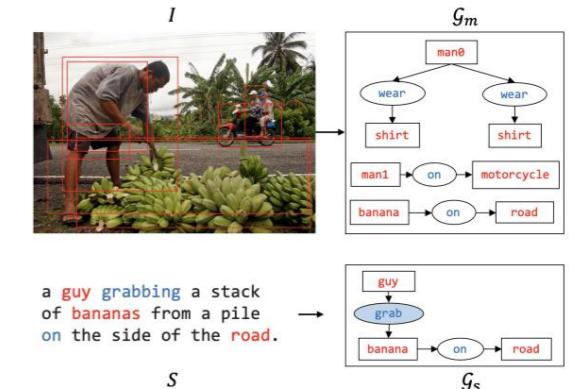
Tóm tắt

Chú thích hình ảnh là một vấn đề đa phương thức đã thu hút sự chú ý rộng rãi ở cả hai cộng đồng xử lý ngôn ngữ tự nhiên và thị giác máy tính. Trong bài báo này, chúng tôi trình bày một kiến trúc chú thích hình ảnh mới lạ để khám phá tốt hơn ngữ nghĩa a có sẵn trong chú thích và tận dụng điều đó để nâng cao cả việc biểu diễn hình ảnh và tạo chú thích. Các mô hình của chúng tôi đầu tiên xây dựng mối quan hệ trực quan hướng dẫn chú thích để thị giới thiệu độ lệch cảm ứng có lợi sử dụng học tập đa trường hợp được giám sát yếu. Sau đó, biểu diễn được tăng cường với các nút lân cận và theo ngữ cảnh với các tính năng văn bản và hình ảnh. Trong quá trình tạo, mô hình tiếp tục kết hợp các mối quan hệ trực quan bằng cách sử dụng học tập đa nhiệm vụ để cùng nhau dự đoán chuỗi từ và thẻ đối tượng/vị ngữ. Chúng tôi thực hiện các thí nghiệm mở rộng trên tập dữ liệu MSCOCO, cho thấy rằng khuôn khổ được đề xuất vượt trội đáng kể các đường cơ sở, dẫn đến tình trạng hiện đại hiệu suất dưới một phạm vi đánh giá rộng
số liệu. Mã của bài báo của chúng tôi đã được thực hiện có sẵn cho công chúng.¹

1 Giới thiệu

Tự động tạo ra một mô tả ngắn cho một hình ảnh nhất định, một vấn đề được gọi là chú thích hình ảnh (Chen và cộng sự, 2015), đã thu hút được nhiều sự chú ý sự chú ý trong cả xử lý ngôn ngữ tự nhiên và cộng đồng thị giác máy tính. Lấy cảm hứng từ thành công của các khuôn khổ mã hóa-giải mã với cơ chế chú ý, những nỗ lực trước đây về hình ảnh chú thích áp dụng các biến thể của mạng nơ-ron tích chập được đào tạo trước (CNN) làm bộ mã hóa hình ảnh và mạng nơ-ron hồi quy (RNN) với hình ảnh sự chú ý như bộ giải mã (Lu et al., 2017; Anderson và cộng sự, 2018; Xu và cộng sự, 2015; Lü và cộng sự, 2018).

Nhiều phương pháp trước đây dịch biểu diễn hình ảnh thành các câu ngôn ngữ tự nhiên mà không cần



Hình 1: Biểu đồ mối quan hệ trực quan từ một hệ thống được đào tạo trước mô hình phát hiện (Yao et al., 2018) (phía trên) và từ chú thích thực tế (phía dưới).

điều tra rõ ràng các tín hiệu ngữ nghĩa a từ các văn bản và hình ảnh. Để khắc phục điều đó, một số nghiên cứu đã cung cấp khám phá để phát hiện các khái niệm ngữ nghĩa a cấp cao được trình bày trong hình ảnh nhằm cải thiện việc tạo chú thích (Wu et al., 2016; Gan et al., 2017; You và cộng sự, 2016; Fang và cộng sự, 2015; Yao và cộng sự, 2017). Nhiều người tin rằng sự thiên vị quy nạp đó tận dụng sự kết hợp có cấu trúc của các khái niệm và mối quan hệ trực quan có tầm quan trọng, điều này đã dẫn đến để có mô hình chú thích tốt hơn (Yao et al., 2018; Guo và cộng sự, 2019; Yang và cộng sự, 2019). Những cách tiếp cận này thu được đô thị mối quan hệ trực quan bằng cách sử dụng các mô hình được đào tạo trước từ phát hiện mối quan hệ trực quan (VRD) các tập dữ liệu, ví dụ, Visual Genome (Krishna et al., 2017), nơi mà các mối quan hệ trực quan nắm bắt ngữ nghĩa a giữa các cặp đối tượng cụ thể được kết nối bởi các vị ngữ, bao gồm không gian (ví dụ, bánh trên bàn) và các mối quan hệ ngữ nghĩa a phi không gian (ví dụ: người - ăn-thức ăn) (Lu và cộng sự, 2016).

Giống như nhiều vấn đề mô hình hóa văn bản-hình ảnh chung khác, điều quan trọng là phải có được sự biểu diễn ngữ nghĩa a tốt trong chú thích hình ảnh, giúp kết nối ngữ nghĩa a trong ngôn ngữ và hình ảnh. Tuy nhiên, các phương pháp tiếp cận hiện tại vẫn chưa tận dụng đầy đủ ngữ nghĩa a có sẵn trong chú thích để

¹ <https://github.com/Gitsamshi/WeakVRD-Captioning>

struct image representation and generate captions. As shown in Figure 1, although VRD detection models present a strong capacity in predicting salient objects and the most common predicates, they often ignore predicates vital for captioning (e.g., “grab” in this example). Exploring better models would still be highly desirable.

A major challenge for establishing a structural connection between captions and images is that the links between predicates and the corresponding object regions are often ambiguous: within the “image-level” label ($obj_1, pred, obj_2$) extracted from captions, there may exist multiple object regions corresponding to obj_1 and obj_2 . In this paper, we propose to use weakly supervised multi-instance learning to detect if a bag of object (region) pairs in an image contain certain predicates, e.g., predicates appearing in ground-truth captions here (or in other applications, they can be any given predicates under concerns). Based on that we can construct caption-guided visual relationship graphs.

Once the visual relationship graphs (VRG) are built, we propose to adapt graph convolution operations (Marcheggiani and Titov, 2017) to obtain representation for object nodes and predicate nodes. These nodes can be viewed as image representation units used for generation.

During generation, we further incorporate visual relationships—we propose multi-task learning for jointly predicting word and tag sequences, where each word in a caption could be assigned with a tag, i.e., *object*, *predicate*, or *none*, which takes as input the graph node features from the above visual relationship graphs. The motivation for predicting a tag in each step is to regularize which types of information should be taken into more consideration for generating words: predicate nodes features, object nodes features, or the current state of language decoder. We study different types of multi-task blocks in our models.

As a result, our models consist of three major components: constructing caption-guided visual relationship graphs (CGVRG) with weakly-supervised multi-instance learning, building context-aware CGVRG, and performing multi-task generation to regularize the network to take into account explicit predicate/predicate constraints. We perform extensive experiments on the MSCOCO (Lin et al., 2014) image captioning dataset with both supervised and Reinforcement

learning strategy (Rennie et al., 2017). The experiment results show that the proposed models significantly outperform the baselines and achieve the state-of-the-art performance under a wide range of evaluation metrics. The main contributions of our work are summarized as follows:

- We propose to construct caption-guided visual relationship graphs that introduce beneficial inductive bias by better bridging captions and images. The representation is further enhanced with neighbouring and contextual nodes with their textual and visual features.
- Unlike existing models, we propose multi-task learning to regularize the network to take into account explicit object/predicate constraints in the process of generation.
- The proposed framework achieves the state-of-the-art performance on the MSCOCO image captioning dataset. We provide detailed analyses on how this is attained.

2 Related Work

Image Captioning A prevalent paradigm of existing image captioning methods is based on the encoder-decoder framework which often utilizes a CNN-plus-RNN architecture for image encoding and text generation (Donahue et al., 2015; Vinyals et al., 2015; Karpathy and Fei-Fei, 2015). Soft or hard visual attention mechanism (Xu et al., 2015; Chen et al., 2017) has been incorporated to focus on the most relevant regions in each generation step. Furthermore, adaptive attention (Lu et al., 2017) has been developed to decide whether to rely on visual features or language model states in each decoding step. Recently, bottom-up attention techniques (Anderson et al., 2018; Lu et al., 2018) have also been proposed to find the most relevant regions based on bounding boxes.

There has been increasing work focusing on filling the gap between image representation and caption generation. Semantic concepts and attributes detected from images have been demonstrated to be effective in boosting image captioning when used in the encoder-decoder frameworks (Wu et al., 2016; You et al., 2016; Gan et al., 2017; Yao et al., 2017). Visual relationship (Lu et al., 2016) and scene graphs (Johnson et al., 2015) have been further employed for image encoder in a unimodal (Yao et al., 2018) or multi-modal (Yang et al., 2019; Guo et al., 2019) manner to improve the over-

Cấu trúc biểu diễn hình ảnh và tạo chú thích. Như thể hiện trong Hình 1, mặc dù phát hiện VRD các mô hình thể hiện khả năng mạnh mẽ trong việc dự đoán các đối tượng nổi bật và các vị ngữ phổ biến nhất, họ thường bỏ qua các vị ngữ quan trọng cho việc chú thích (ví dụ, “lấy” trong ví dụ này). Khám phá tốt hơn các mô hình vẫn được mong muốn cao.

Một thách thức lớn trong việc thiết lập kết nối cấu trúc giữa chú thích và hình ảnh là các liên kết giữa các vị ngữ và các vùng đối tượng tương ứng thường không rõ ràng: trong nhãn “mức hình ảnh” ($obj_1, pred, obj_2$) trích xuất từ chú thích, có thể tồn tại nhiều vùng đối tượng tương ứng với obj_1 và obj_2 . Trong bài báo này, chúng tôi đề xuất sử dụng giám sát yếu học tập đa trường hợp để phát hiện xem một túi vật thể (khu vực) các cặp trong một hình ảnh chứa một số vị ngữ nhất định, ví dụ, các vị ngữ xuất hiện trong sự thật cơ bản chú thích ở đây (hoặc trong các ứng dụng khác, chúng có thể là bất kỳ vị ngữ nào được đưa ra dưới mối quan tâm). Dựa trên rằng chúng ta có thể xây dựng biểu đồ quan hệ trực quan được hướng dẫn bằng chú thích.

Sau khi đồ thị quan hệ trực quan (VRG) được xây dựng, chúng tôi đề xuất điều chỉnh các hoạt động tích chập đồ thị (Marcheggiani và Titov, 2017) để có được biểu diễn cho các nút đối tượng và nút vị ngữ. Các nút này có thể được xem như là biểu diễn hình ảnh đơn vị được sử dụng để tạo ra.

Trong quá trình tạo ra, chúng tôi tiếp tục kết hợp hình ảnh mối quan hệ—chúng tôi đề xuất học tập đa nhiệm vụ cho dự đoán chung các chuỗi từ và thẻ, trong đó mỗi từ trong chú thích có thể được gán với một thẻ, tức là, đối tượng, vị ngữ hoặc không có, được coi là nhập các đặc điểm nút đồ thị từ hình ảnh trên đồ thị mối quan hệ. Động lực để dự đoán một thẻ trong mỗi bước là để chuẩn hóa loại thông tin nào cần được xem xét nhiều hơn để tạo ra các từ: các tính năng của nút vị ngữ, các tính năng của nút đối tượng hoặc trạng thái hiện tại của ngôn ngữ bộ giải mã. Chúng tôi nghiên cứu các loại đa nhiệm khác nhau khác trong mô hình của chúng tôi.

Kết quả là, các mô hình của chúng tôi bao gồm ba thành phần chính: xây dựng đồ thị quan hệ trực quan được hướng dẫn bằng chú thích (CGVRG) với học tập đa trường hợp có giám sát yếu, xây dựng CGVRG nhận thức ngữ cảnh và thực hiện đa nhiệm vụ thẻ để điều chỉnh mạng lưới để đưa vào tài khoản ràng buộc vị ngữ rõ ràng đối tượng/vị ngữ. Chúng tôi thực hiện các thí nghiệm mở rộng trên chú thích hình ảnh MSCOCO (Lin et al., 2014) bộ dữ liệu có cả giám sát và tăng cường

chiến lược học tập (Rennie et al., 2017). Kết quả thí nghiệm cho thấy các mô hình đề xuất vượt trội đáng kể so với các đường cơ sở và đạt được hiệu suất hiện đại trong phạm vi rộng của các số liệu đánh giá. Những đóng góp chính của công việc của chúng tôi được tóm tắt như sau:

- Chúng tôi đề xuất xây dựng các đồ thị quan hệ trực quan được hướng dẫn bằng chú thích, giới thiệu sự thiên vị quy nạp có lợi bằng cách kết nối chú thích và hình ảnh tốt hơn. Biểu diễn được tiếp tục được tăng cường với lân cận và theo ngữ cảnh các nút có các tính năng văn bản và hình ảnh của chúng.
- Không giống như các mô hình hiện có, chúng tôi đề xuất đa nhiệm vụ học cách điều chỉnh mạng để đưa vào ràng buộc đối tượng/vị ngữ rõ ràng của tài khoản trong quá trình tạo ra thẻ hệ.
- Khung đề xuất đạt được hiệu suất tiên tiến nhất trên hình ảnh MSCOCO
- Tập dữ liệu chú thích. Chúng tôi cung cấp các phân tích chi tiết về cách thức đạt được điều này.

2 Công trình liên quan

Chú thích hình ảnh Một mô hình phổ biến của các phương pháp chú thích hình ảnh hiện có dựa trên khung mã hóa-giải mã thường sử dụng Kiến trúc CNN-plus-RNN để mã hóa hình ảnh và tạo văn bản (Donahue et al., 2015; Vinyals et al., 2015; Karpathy và Fei-Fei, 2015). Mềm hoặc cơ chế chú ý thị giác cứng (Xu et al., 2015; Chen et al., 2017) đã được kết hợp để tập trung về các khu vực có liên quan nhất trong mỗi thẻ hệ bước. Hơn nữa, sự chú ý thích ứng (Lu et al., 2017) đã được phát triển để quyết định xem có nên dựa vào các tính năng trực quan hoặc trạng thái mô hình ngôn ngữ trong mỗi bước giải mã. Gần đây, sự chú ý từ dưới lên kỹ thuật (Anderson và cộng sự, 2018; Lu và cộng sự, 2018) cũng đã được đề xuất để tìm ra những điều có liên quan nhất vùng dựa trên hộp giới hạn.

Đã có nhiều công việc tập trung vào lập đầy khoảng cách giữa biểu diễn hình ảnh và tạo chú thích. Các khái niệm ngữ nghĩa và thuộc tính được phát hiện từ hình ảnh đã được chứng minh là có hiệu quả trong việc thúc đẩy chú thích hình ảnh khi được sử dụng trong các khuôn khổ mã hóa-giải mã (Wu và cộng sự, 2016; Bạn và cộng sự, 2016; Gan và cộng sự, 2017; Yao et al., 2017). Mỗi quan hệ trực quan (Lu et al., 2016) và đồ thị cảnh (Johnson và cộng sự, 2015) có thể được sử dụng thêm cho bộ mã hóa hình ảnh ở dạng đơn phương thức (Yao và cộng sự, 2018) hoặc đa phương thức (Yang và cộng sự, 2019; Guo et al., 2019) cách cải thiện quá mức

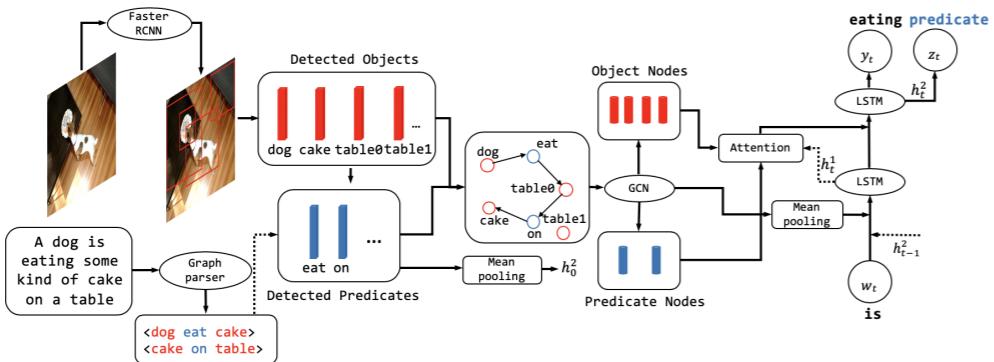


Figure 2: An overview of the proposed image captioning framework.

all performance via the graph convolutional mechanism (Marcheggiani and Titov, 2017). Besides, Kim et al. (2019) proposes a relationship-based captioning task to lead better understanding of images based on relationship. As discussed in introduction, we will further explore the relational semantics available in captions for both constructing image representation and generating caption.

Visual Relationship Detection Visual relations between objects in an image have attracted more studies recently. Conventional visual relation detection have dealt with $\langle \text{subject-predicate-object} \rangle$ triples, including spatial relation and other semantic relation. Lu et al. (2016) detect the triples by performing subject, object, and predicate classification separately. Li et al. (2017) attempt to encode more distinguishable visual features for visual relationships detection. Probabilistic output of object detection (Dai et al., 2017; Zhang et al., 2017) is also considered to reason about the visual relationships.

3 The Models

Given an image I , the goal of image captioning is to generate a visually grounded natural language sentence. We learn our model by minimizing the cross-entropy loss with regard to the ground truth caption $S^* = \{w_1^*, w_2^*, \dots, w_T^*\}$:

$$L_{XE} = -\log p(S^*|I) \quad (1)$$

$$= -\sum_{t=1}^T \log p(w_t^*|w_{<t}^*, I) \quad (2)$$

The model is further tuned with a Reinforcement Learning (RL) objective (Rennie et al., 2017) to maximize the reward of the generated sentence S :

$$J_{RL} = E_{S \sim p(S|I)}(d(S, S^*)) \quad (3)$$

where d is a sentence-level scoring metric.

An overview of our image captioning framework is depicted in Figure 2, with the detail of the components described in the following sections.

3.1 Caption-Guided Visual Relationship Graph (CGVRG) with Weakly Supervised Learning

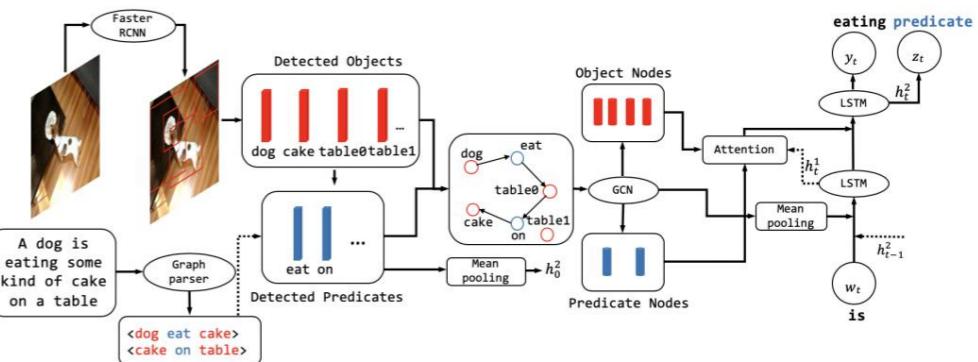
A general challenge of modeling $p(S|I)$ is obtaining a better semantic representation in the multi-modal setting to bridge captions and images. Our framework first focuses on constructing caption-guided visual relationship graphs (CGVRG).

3.1.1 Extracting Visual Relationship Triples and Detecting Objects

The process of constructing CGVRG first extracts relationship triples from captions using textual scene graph parser as described in (Schuster et al., 2015). Our framework employs Faster R-CNN (Ren et al., 2015) to recognize instances of objects and returns a set of image regions for objects: $V = \{v_1, v_2, \dots, v_n\}$.

3.1.2 Constructing CGVRG

The main focus of CGVRG is constructing visual relationship graphs. As discussed in introduction, the existing approaches use pre-trained VRD (visual relationship detection) models, which often ignore key relationships needed for captioning. This gap can be even more prominent if the domain/data used to train image-captioning is farther from where VRD is pretrained. A major challenge to use predicate triples from captions to construct CGVRG is that, the links between predicates and the corresponding object regions are often ambiguous as discussed in introduction. To solve this problem, we use weakly supervised, multi-instance learning.



Hình 2: Tổng quan về khuôn khổ chú thích hình ảnh được đề xuất.

tất cả hiệu suất thông qua cơ chế tích hợp đồ thị

(Marcheggiani và Titov, 2017). Bên cạnh đó, Kim et al. (2019) đề xuất một nhiệm vụ chú thích dựa trên mối quan hệ để dẫn đến sự hiểu biết tốt hơn về hình ảnh dựa trên mối quan hệ. Như đã thảo luận trong phần giới thiệu, chúng ta sẽ khám phá sâu hơn về ngữ nghĩa quan hệ có sẵn trong chú thích cho cả hình ảnh xây dựng biểu diễn và tạo chú thích.

Phát hiện mối quan hệ trực quan Mối quan hệ trực quan giữa các đối tượng trong một hình ảnh đã thu hút nhiều hơn nghiên cứu gần đây. Phát hiện mối quan hệ trực quan thông thường đã giải quyết chủ ngữ-vị ngữ-đối tượng bộ ba, bao gồm mối quan hệ không gian và mối quan hệ ngữ nghĩa khác. Lu et al. (2016) phát hiện bộ ba bằng thực hiện phân loại chủ ngữ, tân ngữ và vị ngữ riêng biệt. Li et al. (2017) cố gắng mã hóa các đặc điểm trực quan để phân biệt hơn để phát hiện mối quan hệ trực quan. Đầu ra xác suất của đối tượng phát hiện (Dai et al., 2017; Zhang et al., 2017) là cũng được coi là lý do về mối quan hệ trực quan-

tàu thuyền.

3 Các mô hình

Với một hình ảnh I , mục tiêu của chú thích hình ảnh là để tạo ra một ngôn ngữ tự nhiên có cơ sở trực quan. Chúng tôi học mô hình của chúng tôi bằng cách giảm thiểu mất mát entropy chéo liên quan đến sự thật cơ bản chú thích $S = \{w_1, w_2, \dots, w_T\}$:

$$L_{XE} = -\log p(S^*|I) \quad (1)$$

$$= -\sum_{t=1}^T \log p(w_t^*|w_{<t}, I) \quad (2)$$

Mô hình được điều chỉnh thêm bằng sự gia cố Mục tiêu học tập (RL) (Rennie et al., 2017) để tối đa hóa phần thưởng của câu được tạo ra S :

$$J_{RL} = E_S p(S|I) (d(S, S^*)) \quad (3)$$

trong đó d là số liệu chấm điểm ở cấp độ câu.

Tổng quan về khuôn khổ chú thích hình ảnh của chúng tôi được mô tả trong Hình 2, với thông tin chi tiết về các thành phần được mô tả trong các phần sau.

3.1 Mối quan hệ trực quan hướng dẫn theo chú thích Đô thị (CGVRG) với Yêu Học có giám sát

Một thách thức chung của mô hình $p(S|I)$ là có được biểu diễn ngữ nghĩa tốt hơn trong bối cảnh đa phương thức để kết nối chú thích và hình ảnh. Khung đầu tiên tập trung vào việc xây dựng biểu đồ quan hệ trực quan có chú thích (CGVRG).

3.1.1 Trích xuất bộ ba quan hệ trực quan và Phát hiện Đối tượng

Quá trình xây dựng CGVRG đầu tiên trích xuất các bộ ba mối quan hệ từ chú thích bằng cách sử dụng trình phân tích đồ thị cảnh văn bản như được mô tả trong (Schuster et al., 2015). Khung của chúng tôi sử dụng Faster R-CNN (Ren et al., 2015) để nhận dạng các trường hợp các đối tượng và trả về một tập hợp các vùng ảnh cho các đối tượng: $V = \{v_1, v_2, \dots, v_n\}$.

3.1.2 Xây dựng CGVRG

Trọng tâm chính của CGVRG là xây dựng hình ảnh đồ thị mối quan hệ. Như đã thảo luận trong phần giới thiệu, các phương pháp tiếp cận hiện tại sử dụng các mô hình VRD (phát hiện mối quan hệ trực quan) được đào tạo trước, thường bỏ qua các mối quan hệ chính cần thiết cho chú thích.

Khoảng cách này thậm chí có thể còn rõ ràng hơn nếu miền dữ liệu được sử dụng để đào tạo chú thích hình ảnh xa hơn từ nơi VRD được đào tạo trước. Một thách thức lớn sử dụng bộ ba vị ngữ từ chú thích để xây dựng CGVRG là các liên kết giữa các vị ngữ và các vùng đối tượng tương ứng thường mơ hồ như đã thảo luận trong phần giới thiệu. Để giải quyết vấn đề này, chúng tôi sử dụng giám sát yếu, đa trường hợp học hỏi.

Obtaining Representation for Object Region Pairs For an image I with a list of salient object regions obtained in object detection $\{v_1, v_2, \dots, v_n\}$, we have a set of region pairs $U = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_N\}$, where $N = n(n - 1)$. As shown in Figure 3(b), the visual features of any two object regions and their union box will be collected to compute $p_{\mathbf{u}_n}^{r_j}$, the probability that a region pair \mathbf{u}_n is associated with the predicate r_j , where $r_j \in R$ and $R = \{r_1, r_2, \dots, r_M\}$ include frequent predicates obtained from the captions in training data. The feed-forward network of Figure 3(b) will be trained in weakly supervised training.

Weakly Supervised Multi-Instance Training As shown in Figure 3(c), during training, one object pair $t = (o_1, o_2)$, e.g., $(woman, hat)$, can correspond to multiple pairs of object regions: the four women-hat combinations between the two women and two hats. To make our description clearer, we refer to $t = (o_1, o_2)$ as an *object pair*, and the four women-hat pairs in the image as *object region pairs*. Accordingly, for a triple we extracted $t = (o_1, r, o_2)$, $r \in R$, e.g., $(woman, in, hat)$, the predicate r (i.e., *in*) can be associated with multiple *object region pairs* (here, $(w0, h0)$, $(w0, h1)$, $(w1, h0)$, and $(w1, h1)$).

To predict predicates over object region pairs, we propose to use Multi-Instance Learning (Fang et al., 2015) as our weakly supervised learning approach. Multi-Instance Learning receives a set of labeled bags, each bag containing a set of instances. A bag would be labeled *negative* if all the instances in it are negative. On the other hand, a bag is labeled *positive* if there is at least one positive instance in the bag.

In our problem, an instance is a region pair. Therefore for a candidate predicate $r \in R$ (e.g., *in*), we use \mathcal{N}_r to denote the object region pairs corresponding to predicate r . If r appears in the caption S , \mathcal{N}_r would be a positive bag. We use $\mathcal{N} \setminus \mathcal{N}_r$ to denote the negative bag for r . When r is not contained in the caption, the entire \mathcal{N} would be the negative bag (the last row of Figure 3(c)). The probability of a bag b having the predicate r_j is measured with “noisy-OR”:

$$p_b^{r_j} = 1 - \prod_{n \in b} (1 - p_{\mathbf{u}_n}^{r_j}) \quad (4)$$

where $p_{\mathbf{u}_n}^{r_j}$ has been introduced above. We adopt the cross-entropy loss on the basis of all predicate

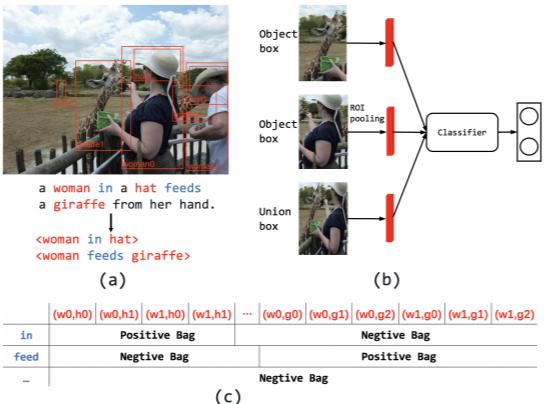


Figure 3: Subcomponents in constructing CGVRG: (a) detecting objects and extracting triples; (b) obtaining representation for object region pairs; (c) examples of positive and negative bags in multi-instance learning for predicate “in” and “feed”, respectively. Here, w , h , and g denote *woman*, *hat*, and *giraffe*, respectively.

probabilities over bags, given an image I and caption S :

$$L(I) = - \sum_{j=1}^M \left[\mathbb{1}_{(r_j \in S)} (\log p_{\mathcal{N}_r}^{r_j} + \log(1 - p_{\mathcal{N} \setminus \mathcal{N}_r}^{r_j})) + \mathbb{1}_{(r_j \notin S)} (\log(1 - p_{\mathcal{N}}^{r_j})) \right] \quad (5)$$

where the indicator function $\mathbb{1}_{condition} = 1$ if the condition is true, otherwise $\mathbb{1}_{condition} = 0$.

Constructing the Graphs Once obtaining the trained module, we can build a CGVRG graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for a given image I , where the node set \mathcal{V} includes two types of nodes: object nodes and predicate nodes. We denote o_i as the i^{th} object node and r_{ij} as a predicate node that connects o_i and o_j (refer to Figure 1 or the middle part of Figure 2). The edges in \mathcal{E} are added based on triples; i.e., (o_i, r_{ij}, o_j) will assign two directed edges from node o_i to r_{ij} and from r_{ij} to o_j , respectively.

Note that due to the use of the proposed weakly supervised models, the acquired graphs can now contain predicates that exist in captions but not in the VRD models used in the previous work that does not explicitly consider predicates in captions. We will show in our experiments that this improves captioning quality.

3.2 Context-Aware CGVRG

We further enhance CGVRG in the context of both modalities, images and text, using graph convolution networks. We first integrate visual and textual features: the textual features for each node are

thu thập biểu diễn cho các cặp vùng đối tượng. Đối với một hình ảnh I có danh sách các vùng đối tượng nổi bật thu được trong quá trình phát hiện đối tượng $\{v_1, v_2, \dots, v_n\}$, chúng ta có một tập hợp các cặp vùng $U = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_N\}$, trong đó $N = n(n - 1)$.

Như thể hiện trong Hình 3(b), các đặc điểm trực quan của bất kỳ hai vùng đối tượng nào và hộp hợp nhất của chúng sẽ được thu thập để tính toán p là xác suất mà một cặp vùng u_n được liên kết với vị trí trong đồ thị R và $R = \{r_1, r_2, \dots, r_M\}$ trong đồ thị, bao gồm các vị trí thường xuyên thu được các chú thích trong dữ liệu đào tạo. Mạng truyền thẳng của Hình 3(b) sẽ được đào tạo trong đào tạo có giám sát yếu.

Huấn luyện đa trường hợp được giám sát yếu. Như thể hiện trong Hình 3(c), trong quá trình huấn luyện, một cặp đối tượng $t = (o_1, o_2)$, ví dụ, (phụ nữ, mũ), có thể tương ứng với nhiều cặp vùng đối tượng: bốn bộ hợp phụ nữ-mũ giữa hai người phụ nữ và hai chiếc mũ. Để làm cho mô tả của chúng tôi rõ ràng hơn, chúng tôi gọi $t = (o_1, o_2)$ là một cặp đối tượng và bốn cặp phụ nữ-mũ trong hình ảnh là các cặp vùng đối tượng. Theo đó, đối với một bộ ba, chúng tôi đã trích xuất $t = (o_1, r, o_2)$, $r \in R$, ví dụ, (phụ nữ, trong, mũ), vị ngữ r (tức là, trong) có thể được liên kết với nhiều cặp vùng đối tượng (ở đây, $(w0, h0)$, $(w0, h1)$, $(w1, h0)$ và $(w1, h1)$).

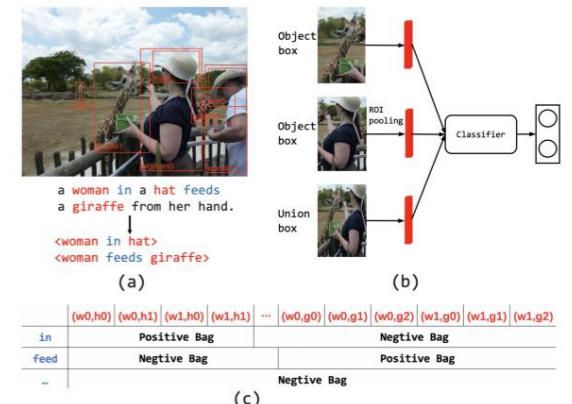
Để dự đoán các vị ngữ trên các cặp vùng đối tượng, chúng tôi đã sử dụng Học đa trường hợp (Fang và cộng sự, 2015) làm phương pháp học có giám sát yếu. Học tập đa trường hợp nhận được một tập hợp các túi được gắn nhãn, mỗi túi chứa một tập hợp các trường hợp. Một túi sẽ được gắn nhãn là tiêu cực nếu tất cả các trường hợp trong đó đều là tiêu cực. Mặt khác, một túi được gắn nhãn là tích cực nếu có ít nhất một trường hợp tích cực trong túi.

Trong bài toán của chúng ta, mỗi thể hiện là một cặp vùng. Do đó, đối với một vị ngữ viên $r \in R$ (ví dụ, *in*), chúng ta sử dụng N_r để biểu thị các cặp vùng đối tượng tương ứng với vị ngữ r . Nếu r xuất hiện trong chủ đề S , N_r sẽ là một túi dương. Chúng ta sử dụng $N \setminus N_r$ để biểu thị túi âm cho r . Khi r không có trong chủ đề, toàn bộ N sẽ là túi âm (hàng cuối cùng của Hình 3(c)).

Xác suất của túi b có vị ngữ r_j được đo bằng “noisy-OR”:

$$p_b^{r_j} = 1 - \prod_{n \in b} (1 - p_{\mathbf{u}_n}^{r_j}) \quad (4)$$

r_j ở đâu p không đã được giới thiệu ở trên. Chúng tôi áp dụng tồn tại entropy chéo trên cơ sở tất cả các vị ngữ



Hình 3: Các thành phần phụ trong việc xây dựng CGVRG: (a) phát hiện các đối tượng và trích xuất các bộ ba; (b) thu thập biểu diễn cho các cặp vùng đối tượng; (c) các ví dụ về các túi dương và âm trong quá trình học đa trường hợp cho vị ngữ “in” và “feed”, tương ứng. Ở đây, w , h và g lần lượt biểu thị phụ nữ, mũ và hươu cao cổ.

xác suất trên các túi, cho một hình ảnh I và chủ đề S :

$$L(I) = \sum_{j=1}^M \left[(r_j \in S) (\log \frac{r_j}{N_r} + \log(1 - \frac{r_j}{N \setminus N_r})) + (r_j \notin S) (\log(1 - \frac{r_j}{N})) \right] \quad (5)$$

trong đó hàm chỉ báo $condition = 1$ nếu điều kiện đúng, nếu không thì $condition = 0$.

Xây dựng đồ thị Sau khi có được mô-đun đã được đào tạo, chúng ta có thể xây dựng đồ thị CGVRG $G = (V, E)$ cho một hình ảnh I cho trước, trong đó tập nút V bao gồm hai loại nút: nút đối tượng và nút vị ngữ. Chúng ta biểu thị oi là nút đối tượng i và rij là nút vị ngữ kết nối oi và oj (tham khảo Hình 1 hoặc phần giữa của Hình 2). Các cạnh trong E được thêm vào dựa trên bộ ba; tức là, (oi, rij, oj) sẽ gắn hai cạnh có hướng từ nút oi đến rij và từ rij đến oj , tương ứng.

Lưu ý rằng do sử dụng các mô hình giám sát yếu được đề xuất nên các đồ thị thu được hiện có thể chứa các vị ngữ tồn tại trong chủ đề nhưng không có trong các mô hình VRD được sử dụng trong công trình trước đó không xem xét r_j ràng buộc các vị ngữ trong chủ đề. Chúng tôi sẽ chứng minh trong các thí nghiệm của mình rằng điều này cải thiện chất lượng phu đề.

3.2 CGVRG nhận thức ngữ cảnh

Chúng tôi tiếp tục nâng cao CGVRG trong bối cảnh của cả hai phương thức, hình ảnh và văn bản, bằng cách sử dụng mạng lưới tích chập đồ thị. Đầu tiên, chúng tôi tích hợp các tính năng trực quan và văn bản: các tính năng văn bản cho mỗi nút là

from a word embedding and the visual features are regional visual representations extracted via ROI pooling from Faster R-CNN. The specific features $\mathbf{g}_{oi}, \mathbf{g}_{rij}$ for object o_i and predicate r_{ij} are shown as follows:

$$\mathbf{g}_{oi} = \phi_o([\mathbf{g}_{oi}^t; \mathbf{g}_{oi}^v]) \quad (6)$$

$$\mathbf{g}_{rij} = \phi_r(\mathbf{g}_{rij}^t) \quad (7)$$

where ϕ_r and ϕ_o are feed-forward networks using ReLU activation; $\mathbf{g}_{oi}^t, \mathbf{g}_{rij}^t$, and \mathbf{g}_{oi}^v denote textual features of o_i, r_{ij} and visual features of o_i , respectively.

We present the process of encoding \mathcal{G} to produce a new set of context-aware representation \mathcal{X} . The representation of predicate r_{ij} and o_i are computed as follows:

$$\mathbf{x}_{rij} = f_r([\mathbf{g}_{oi}; \mathbf{g}_{oj}; \mathbf{g}_{rij}]) \quad (8)$$

$$\mathbf{x}_{oi} = \frac{1}{N_i} \left[\sum_{r \in \mathcal{N}_{out}(o_i)} f_{out}([\mathbf{g}_{oi}; \mathbf{g}_r]) + \sum_{r \in \mathcal{N}_{in}(o_i)} f_{in}([\mathbf{g}_{oi}; \mathbf{g}_r]) \right] \quad (9)$$

where f_r, f_{in}, f_{out} are feed-forward networks using ReLU activation. \mathcal{N}_{in} and \mathcal{N}_{out} denote the adjacent nodes with o_i as head and tail, respectively. N_i is the total number of adjacent nodes.

3.3 Multi-task Caption Generation

Unlike the existing image-captioning models, we further incorporate visual relationships into generation — we propose multi-task learning for jointly predicting word and tag sequences as each word in a caption will be assigned a tag, i.e., *object*, *predicate*, or *none*. The module takes as input the graph node features from the context-aware CGVRG. The output of the generation module is hence the sequence of words $\mathbf{y} = \{y_1, \dots, y_T\}$ as well as the tags $\mathbf{z} = \{z_1, \dots, z_T\}$. Two different approaches are leveraged to train the two tasks jointly.

The bottom LSTM is used to align a textual state to graph node representations:

$$\mathbf{h}_t^1 = \text{LSTM}(\mathbf{h}_{t-1}^1, [\mathbf{h}_{t-1}^2; \bar{\mathbf{x}}; \mathbf{e}_{w_t}]) \quad (10)$$

where LSTM means one step of recurrent unit computation via LSTM; $\bar{\mathbf{x}}$ is the mean-pooled representation of all nodes in the graph; \mathbf{h}_{t-1}^1 and \mathbf{h}_{t-1}^2

denote hidden states of bottom and top LSTM in time step $t-1$, respectively; \mathbf{e} is the word embedding table.

The state \mathbf{h}_t^1 is then used as a query to attend over graph node features $\{\mathbf{x}_o\}$ and $\{\mathbf{x}_r\}$ separately to get attended features $\hat{\mathbf{x}}_t^r$ and $\hat{\mathbf{x}}_t^o$:

$$\hat{\mathbf{x}}_t^r = \text{ATT}(\mathbf{h}_t^1, \{\mathbf{x}_r\}) \quad (11)$$

$$\hat{\mathbf{x}}_t^o = \text{ATT}(\mathbf{h}_t^1, \{\mathbf{x}_o\}) \quad (12)$$

where ATT is a soft-attention operation between a query and graph node features.

The top LSTM works as a language model decoder, in which the hidden state \mathbf{h}_t^2 is initialized with the mean-pooled semantic representation of all detected predicates $\{r\}$. In time step t , the input consists of the output from the bottom LSTM layer \mathbf{h}_t^1 and attended graph features $\hat{\mathbf{x}}_t^r, \hat{\mathbf{x}}_t^o$:

$$\mathbf{h}_t^2 = \text{LSTM}(\mathbf{h}_{t-1}^2, [\mathbf{h}_t^1; \hat{\mathbf{x}}_t^o; \hat{\mathbf{x}}_t^r]) \quad (13)$$

3.3.1 Multi-task Learning

We propose two different blocks to perform the two tasks jointly, as shown in Figure 4. In each step, a multi-task learning block deals with task s_1 as predicting a tag z_t and task s_2 as predicting a word y_t . Specifically MT-I treats the two tasks independent of each other:

$$p(z_t | y_{<t}, \mathbf{I}) = \text{softmax}(f_z(\mathbf{h}_t^2)) \quad (14)$$

$$p(y_t | y_{<t}, \mathbf{I}) = \text{softmax}(f_y(\mathbf{h}_t^2)) \quad (15)$$

where f_z and f_y are feed-forward networks with ReLU activation. Inspired by the adaptive attention mechanism (Lu et al., 2017), MT-II further exploits the probability from $p(z_t | y_{<t}, \mathbf{I})$ to integrate the representation of current hidden state \mathbf{h}_t^2 and attended features from graph $\hat{\mathbf{x}}_t^r, \hat{\mathbf{x}}_t^o$:

$$p(y_t | y_{<t}, \mathbf{I}) = \text{softmax}(f_y(\hat{\mathbf{x}}_t^2)), \quad (16)$$

$$\hat{\mathbf{x}}_t^2 = \mathbf{h}_t^2 p_{na} + \hat{\mathbf{x}}_t^r p_r + \hat{\mathbf{x}}_t^o p_o \quad (17)$$

$$p(z_t | y_{<t}, \mathbf{I}) = \text{softmax}(f_z(\hat{\mathbf{x}}_t^2)) \quad (18)$$

where p_{na}, p_r, p_o denote the probabilities of tag z_t being “none”, “predicate”, and “object”, respectively. The multi-task loss function is as follows:

$$L_{MT}(\mathbf{I}) = - \sum_{t=1}^T \log(p(y_t | y_{<t}, \mathbf{I}) + \gamma \log(p(z_t | y_{<t}, \mathbf{I})) \quad (19)$$

where γ is the hyper-parameter to balance the two tasks.

từ một nhúng từ và các đặc điểm trực quan là các biểu diễn trực quan khu vực được trích xuất thông qua nhóm ROI từ Faster R-CNN. Các đặc điểm cụ thể gọi là g_{rij} cho đối tượng oi và vị ngữ rij được hiển thị như sau:

$$goi = \varphi_o([g_{oi}^t; g_{oi}^v]) \quad (6)$$

$$g_{rij} = \varphi_r(g_{rij}^t) \quad (7)$$

trong đó φ_r và φ_o là mạng truyền thẳng sử dụng ký hiệu văn bản Kích hoạt ReLU; g_{oi}^t, g_{rij}^t , và g_{oi}^v là các đặc điểm của oi, rij và các đặc điểm trực quan của oi , tương ứng.

Chúng tôi trình bày quá trình mã hóa G để tạo ra một tập hợp mới của biểu diễn nhận biết ngữ cảnh X biểu diễn của vị ngữ rij và oi được tính như sau:

$$x_{rij} = fr([goi; goj; g_{rij}]) \quad (8)$$

$$x_{oi} = \frac{1}{N_i} \sum_{r \in Nout(oi)} f_{out}([goi; gr]) + \sum_{r \in Nin(oi)} f_{in}([goi; gr]) \quad (9)$$

trong đó $fr, fin, fout$ là mạng truyền thẳng sử dụng kích hoạt ReLU. Nin và $Nout$ biểu thị các nút liền kề với oi là đầu và đuôi tương ứng.

Ni là tổng số nút liền kề.

3.3 Tạo chú thích đa nhiệm vụ Không giống như

các mô hình chú thích hình ảnh hiện có, chúng tôi tiếp tục kết hợp các mối quan hệ trực quan vào quá trình tạo — chúng tôi đề xuất học tập đa nhiệm vụ để dự đoán chung các chuỗi từ và thẻ vì mỗi từ trong chú thích sẽ được gán một thẻ, tức là đối tượng, vị ngữ hoặc không có. Mô-đun lấy các đặc điểm nút đồ thị từ CGVRG nhận biết ngữ cảnh làm đầu vào.

do đó, đầu ra của mô-đun tạo là chuỗi các từ $y = \{y_1, \dots, y_T\}$ cũng như các thẻ $z = \{z_1, \dots, z_T\}$. Hai cách tiếp cận khác nhau được sử dụng để đào tạo hai nhiệm vụ cùng nhau.

LSTM ở dưới cùng được sử dụng để cân chỉnh trạng thái văn bản với biểu diễn nút đồ thị:

$$\hat{\mathbf{x}}_t^1 = \text{LSTM}(\mathbf{h}_{t-1}^1, [\mathbf{h}_t^2; \bar{\mathbf{x}}; \mathbf{ewt}]) \quad (10)$$

trong đó LSTM có nghĩa là một bước tính toán đơn vị tuần hoàn thông qua LSTM; $\bar{\mathbf{x}}$ là biểu diễn gộp trung bình gửi tất cả các nút trong đồ thị; \mathbf{h}_t^2 và \mathbf{h}_{t-1}^1

biểu thị trạng thái ẩn của LSTM dưới cùng và trên cùng trong bước thời gian $t-1$, tương ứng; e là bảng nhúng từ.

Nhà nước $\hat{\mathbf{x}}_t^1$ sau đó được sử dụng như một truy vấn để tham dự trên các đặc điểm nút đồ thị $\{x_o\}$ và $\{x_r\}$ riêng biệt để có được các đặc điểm tham dự x^r và x^o tại t :

$$x^r_t = \text{ATT}(\mathbf{h}_t^1, \{x_r\}) \quad (11)$$

$$x^o_t = \text{ATT}(\mathbf{h}_t^1, \{x_o\}) \quad (12)$$

trong đó ATT là thao tác chú ý mềm giữa truy vấn và các tính năng nút đồ thị.

LSTM hàng đầu hoạt động như một bộ giải mã hình ngôn ngữ, trong đó trạng thái ẩn h được khởi tạo bằng biểu diễn ngữ nghĩa a được nhóm trung bình $\hat{\mathbf{x}}_t^1$ tất cả các từ được phát hiện (r). Trong bước thời gian t , đầu vào bao gồm đầu ra từ lớp LSTM dưới cùng 1 h

t và các đặc trưng đồ thị tham dự x^r, x^o :

$$x_{giờ}^2 = \text{LSTM}(h_{t-1}^2, [x_{giờ}^r; x_{giờ}^o; \hat{\mathbf{x}}_t^1]) \quad (13)$$

3.3.1 Học tập đa nhiệm vụ

Chúng tôi đề xuất hai khối khác nhau để thực hiện hai nhiệm vụ cùng nhau, như thể hiện trong Hình 4. Trong mỗi bước, một khối học tập đa nhiệm vụ xử lý nhiệm vụ s1 như dự đoán một thẻ zt và nhiệm vụ s2 như dự đoán một từ yt . Cụ thể, MT-I xử lý hai nhiệm vụ độc lập với nhau:

$$p(z_t | y_{<t}, \mathbf{I}) = \text{softmax}(f_z(h_{t-1}^2)) \quad (14)$$

$$p(y_t | y_{<t}, \mathbf{I}) = \text{softmax}(f_y(h_{t-1}^2)) \quad (15)$$

trong đó f_z và f_y là mạng luồng truyền thẳng với kích hoạt ReLU. Lấy cảm hứng từ cơ chế chú ý thích ứng (Lu và cộng sự, 2017), MT-II khai thác thêm xác suất từ $p(z_t | y_{<t}, \mathbf{I})$ đến inté-biểu diễn trạng thái ẩn hiện tại h và các đặc điểm $\hat{\mathbf{x}}_t^1$ được chú ý từ đồ thị x^r, x^o :

$$p(y_t | y_{<t}, \mathbf{I}) = \text{softmax}(f_y(h_{t-1}^2)), \quad (16)$$

$$\hat{\mathbf{x}}_t^2 = h_{t-1}^2 p_{na} + \hat{\mathbf{x}}_t^r p_r + \hat{\mathbf{x}}_t^o p_o \quad (17)$$

$$p(z_t | y_{<t}, \mathbf{I}) = \text{softmax}(f_z(h_{t-1}^2)) \quad (18)$$

trong đó pna, pr, po biểu thị xác suất của thẻ zt lần lượt là “không có”, “thuộc tính” và “đối tượng”. Hàm mất mát đa tác vụ như sau:

$$LMT(\mathbf{I}) = \sum_{t=1}^T \log(p(z_t | y_{<t}, \mathbf{I}) + \gamma \log(p(y_t | y_{<t}, \mathbf{I})) \quad (19)$$

trong đó y là siêu tham số để cân bằng hai nhiệm vụ.

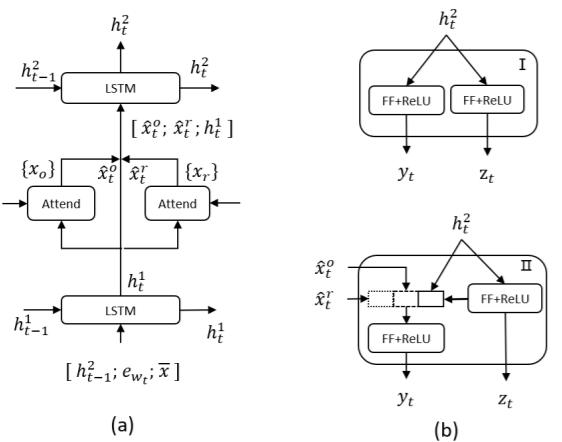


Figure 4: An overview of multi-task caption generation module. Subfigure (a) is a two-layer LSTM; Subfigure (b) depicts two different types of multi-task block.

3.4 Training and Inference

The overall training process can be broken down into two parts: the CGVRG detection module training period and the caption generator training period; the latter includes cross-entropy optimization and the CIDEr-D optimization. For CGVRG detection module training, the detection module is optimized with the multi-instance learning loss in Equation 5. For caption generator training, the model is first optimized with the cross-entropy loss in Equation 19, and then we directly optimize the model with the expected sentence-level reward (CIDEr-D in this work) shown in Equation 3 by self critical sequence learning (Rennie et al., 2017).

In the inference stage, given an image, the CGVRG detection module obtains a graph upon them. The graph convolution network encodes graphs to obtain the context aware multi-modal representations. Then graph object/predicate node features are further provided to the multi-task caption generation module to generate sequences with beam search.

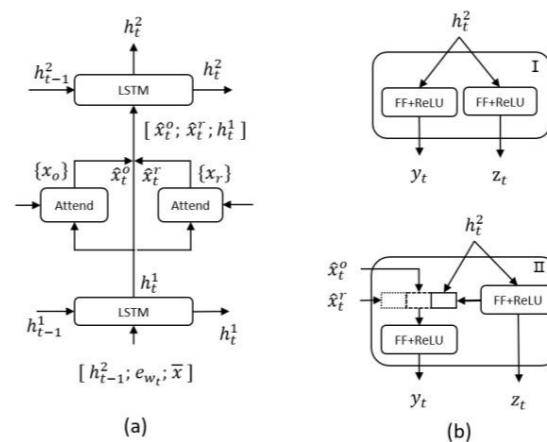
4 Experiments

4.1 Datasets and Experiment Setup

MSCOCO We perform extensive experiments on the MSCOCO benchmark (Lin et al., 2014). The Karpathy split (Karpathy and Fei-Fei, 2015) is adopted for our model selection and offline testing, which contains 113K training images, 5K validation images and 5K testing images. As for the online test server, the result is trained on the entire training and validation set (123K images). To evaluate the generated captions, we employ

standard evaluation metrics: SPICE (Anderson et al., 2016), CIDEr-D (Vedantam et al., 2015), METEOR (Denkowski and Lavie, 2014), ROUGE-L (Lin, 2004), and BLEU (Papineni et al., 2002).

Visual Genome We use the Visual Genome (Krishna et al., 2017) dataset to pre-train our object detection model. The dataset includes 108K images. To pre-train the object detection model with Faster R-CNN, we strictly follow the setting in (Anderson et al., 2018), taking 98K/5K/5K for training, validation, and testing, respectively. The split is carefully selected to avoid contamination of the MSCOCO validation and testing sets, since nearly 51K Visual Genome images are also included in the MSCOCO dataset.



Hình 4: Tổng quan về việc tạo chú thích đa tác vụ mô-đun. Hình phụ (a) là LSTM hai lớp; Hình phụ (b) mô tả hai loại khối đa nhiệm khác nhau.

3.4 Đào tạo và suy luận

Quá trình đào tạo tổng thể có thể được chia nhỏ thành hai phần: giai đoạn đào tạo mô-đun phát hiện CGVRG và giai đoạn đào tạo trình tạo chú thích ; giai đoạn sau bao gồm tối ưu hóa entropy chéo và tối ưu hóa CIDEr-D. Đối với đào tạo mô-đun phát hiện CGVRG, mô-đun phát hiện được tối ưu hóa với mắt mát học tập đa trường hợp trong Phương trình 5. Đối với đào tạo trình tạo chú thích, mô hình là đầu tiên được tối ưu hóa với tồn thắt entropy chéo trong Phương trình 19, và sau đó chúng tôi trực tiếp tối ưu hóa mô hình ngưỡng IoU. Kết quả là, chúng ta có được một tập hợp các phần thưởng ở mức án dụ kiến (CIDEr-D trong tác phẩm này) được thể hiện trong Phương trình 3 bằng cách tự phê bình học trình tự (Rennie và cộng sự, 2017).

Trong giai đoạn suy luận, đưa ra một hình ảnh, Mô-đun phát hiện CGVRG thu được một đồ thị khi chúng. Mạng tích chập đồ thị mã hóa đồ thị để có được ngữ cảnh nhận thức đa phương thức biểu diễn. Sau đó, đối tượng đồ thị/nút vị ngữ các tính năng được cung cấp thêm cho mô-đun tạo chú thích đa nhiệm vụ để tạo ra các chuỗi với tìm kiếm chùm tia.

4 Thí nghiệm

4.1 Bộ dữ liệu và thiết lập thử nghiệm

MSCOCO Chúng tôi thực hiện các thí nghiệm mở rộng trên chuẩn mực MSCOCO (Lin et al., 2014). Sự chia rẽ Karpathy (Karpathy và Fei-Fei, 2015) được áp dụng cho việc lựa chọn mô hình và thử nghiệm ngoại tuyến của chúng tôi , bao gồm 113K hình ảnh đào tạo, 5K hình ảnh xác thực và 5K hình ảnh thử nghiệm. Đối với máy chủ kiểm tra trực tuyến, kết quả được đào tạo trên toàn bộ tập huấn luyện và xác thực (123K hình ảnh). Để đánh giá các chú thích được tạo ra, chúng tôi sử dụng

số liệu đánh giá tiêu chuẩn: SPICE (Anderson và cộng sự, 2016), CIDEr-D (Vedantam và cộng sự, 2015), METEOR (Denkowski và Lavie, 2014), ROUGE-L (Lin, 2004), và BLEU (Papineni và cộng sự, 2002).

Visual Genome Chúng tôi sử dụng bộ dữ liệu Visual Genome (Krishna et al., 2017) để đào tạo trước đối tượng của chúng tôi mô hình phát hiện. Bộ dữ liệu bao gồm 108K hình ảnh . Để đào tạo trước mô hình phát hiện đối tượng với R-CNN nhanh hơn, chúng tôi tuân thủ nghiêm ngặt cài đặt trong (Anderson và cộng sự, 2018), lấy 98K/5K/5K để đào tạo, xác thực và thử nghiệm, tương ứng. Sự phân chia là được lựa chọn cẩn thận để tránh ô nhiễm Bộ kiểm tra và xác thực MSCOCO, vì gần như 51K hình ảnh bộ gen trực quan cũng được bao gồm trong tập dữ liệu MSCOCO.

Chi tiết triển khai Chúng tôi sử dụng Faster R-CNN (Ren et al., 2015) để xác định và định vị các trường hợp của các đối tượng. Giai đoạn phát hiện đối tượng bao gồm hai mô-đun. Mô-đun đầu tiên để xuất vùng đối tượng sử dụng CNN sâu, tức là ResNet-101 (He et al., 2016). Mô-đun thứ hai trích xuất bản đồ đặc điểm sử dụng nhóm vùng quan tâm cho mỗi hộp đề xuất. Thực tế, chúng tôi lấy đầu ra cuối cùng của ResNet-101 và thực hiện việc ngăn chặn không tối đa cho mỗi lớp đối tượng với

Phương trình 19, và sau đó chúng tôi trực tiếp tối ưu hóa mô hình ngưỡng IoU. Kết quả là, chúng ta có được một tập hợp của vùng ảnh, $V = \{v_1, v_2, \dots, v_n\}$, trong đó $n \in [10, 100]$ thay đổi theo hình ảnh đầu vào và ngưỡng tin cậy. Mỗi vùng được biểu diễn dưới dạng vectơ 2.048 chiều thu được từ pool5 lớp sau khi nhóm ROI. Sau đó, chúng tôi áp dụng mạng lưới truyền tiếp với đầu ra 1000 chiều lớp cho phân loại các vị từ. Mạng lưới của cùng kích thước cũng được sử dụng cho phép chiều đặc điểm (ϕ_o, ϕ_i) và GCN (f_r, f_{in}, f_{out}). Trong bộ giải mã LSTM, kích thước nhúng từ được đặt thành là 1.000 và kích thước đơn vị ẩn trong lớp trên cùng và lớp dưới cùng LSTM được đặt là 1.000 và 512, tương ứng. Tham số đánh dồn y trong học tập đa nhiệm là 0,15. Toàn bộ hệ thống là được đào tạo với trình tối ưu hóa Adam. Chúng tôi thiết lập ban đầu tỷ lệ học tập là 0,0005 và kích thước lô nhỏ là 100. Số lượng tối đa của các kỷ nguyên đào tạo là 30 cho tối ưu hóa Cross-entropy và CIDEr-D tương ứng. Đối với việc tạo chuỗi trong giai đoạn suy luận, chúng tôi áp dụng chiến lược tìm kiếm chùm tia và đặt kích thước chùm tia là 3.

Chúng tôi xây dựng các phạm trù đối tượng và vị ngữ cho đào tạo VRD. Tương tự như (Lu et al., 2018), chúng tôi mở rộng thử công 80 danh mục đối tượng ban đầu thành

	Cross entropy				CIDEr-D optimization							
	B1	B4	ME	RG	CD	SP	B1	B4	ME	RG	CD	SP
SCST	-	31.3	26.0	54.3	101.3	-	-	33.3	26.3	55.3	111.4	-
LSTM-A	75.4	35.2	26.9	55.8	108.8	20.0	78.6	35.5	27.3	56.8	118.3	20.8
Up-Down (Baseline)	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
StackCap	76.2	35.2	26.5	-	109.1	-	78.6	36.1	27.4	-	120.4	-
CAVP	-	-	-	-	-	-	-	38.6	28.3	58.5	126.3	21.6
GCN-LSTM	77.3	36.8	27.9	57.0	116.3	20.9	80.5	38.2	28.5	58.3	127.6	22.0
VSUA	-	-	-	-	-	-	-	38.4	28.5	58.4	128.6	22.0
SGAE	77.6	36.9	27.7	57.2	116.7	20.9	80.8	38.4	28.4	58.6	127.8	22.1
This Work (MT-I)	78.1	38.4	28.2	58.0	119.0	21.1	80.8	38.9	28.8	58.7	129.6	22.3
This Work (MT-II)	77.9	38.0	28.1	57.6	117.8	21.3	80.5	38.6	28.7	58.4	128.7	22.4

Table 1: Single-model performances on the MSCOCO dataset (Karpathy split) in both cross-entropy and RL training period. B1, B4, ME, RG, CD, and SP denote BLEU-1, BLEU-4, METEOR, ROUGE, CIDEr-D and SPICE, respectively.

	B4		ME		RG			
	c5	c40	c5	c40	c5	c40		
GCN-LSTM*	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
VSUA	37.4	68.3	28.2	37.1	57.9	72.8	123.1	125.5
SGAE	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
Baseline	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
This Work	38.6	70.1	28.6	37.8	58.8	74.5	125.1	126.7

Table 2: The performance on COCO online test server of various methods that incorporate visual relationships. * denotes that their training batch size and epochs are far beyond average setting in (Anderson et al., 2018; Yang et al., 2019).

413 fine-grained categories by utilizing a list of caption tokens. For example, the object category “*person*” is expanded to a list of fine-grained categories [“*boy*”, “*man*”, · · ·]. Then for all extracted triples that have both objects appearing in the 413 category list, we select the 200 most frequent predicates as our predicate categories.

4.2 Quantitative Analysis

Model Comparison We compare our models with the following state-of-the-art models: (1) SCST (Rennie et al., 2017) employs an improved policy gradient algorithm by utilizing its own inference output to normalize the rewards; (2) LSTM-A (Yao et al., 2017) integrates the detected image attributes into the CNN-plus-RNN image captioning framework; (3) Up-Down (Anderson et al., 2018) uses both a bottom-up and top-down attention mechanism to focus more on salient object regions; (4) GCN-LSTM (Yao et al., 2018) leverages graph convolutional networks over the detected objects and relations; (5) CAVP (Liu et al., 2018) proposes a context-aware policy network by accounting for visual attentions as context for generation; (6) VSUA (Guo et al., 2019) exploits the alignment

between words and different categories of graph nodes; (7) SAGE (Yang et al., 2019) utilizes an additional graph encoder to incorporate language inductive bias into the encoder-decoder framework.

Our baseline is built on Up-Down (Anderson et al., 2018). We propose two variants of final models using different multi-task blocks, namely MT-I and MT-II shown in Fig 4(b). We conduct extensive comparisons on the dataset with the above state-of-the-art techniques. We also perform detailed analysis to demonstrate the impact of different components of our framework.

Table 1 lists the results of various single models on the MSCOCO Karpathy split. Our model outperforms the baseline model significantly, with CIDEr-D scores being improved from 113.5 to 119.0 and 120.1 to 129.6 in the cross-entropy and CIDEr-D optimization period, respectively. In addition, the model with MT-II shows an advantage over that with MT-I on SPICE, which implies that the proposed adaptive visual attention mechanism works in multi-task block II.

Table 2 compares our model with three models that also incorporate VRG, plus the baseline model, on the MSCOCO online test server. Our model improves significantly from the baseline (from 120.5 to 126.7 in CIDEr-D) and has achieved the best results across all evaluation metrics on c40 (40 reference captions).

Figure 5 shows the effect of taking different weights γ in the multi-task loss item (Equation 19). The results indicate that the weight around 0.15 yields the best performance in both multi-task blocks. Meanwhile, Figure 6 shows the ablation analysis by removing the multi-task caption generation and graph convolution operation, respectively, to check the effect of these components. The results

	Entropy chéo								Tối ưu hóa CIDEr-D			
	B1	B4	ME	RG	CD	SP	B1	B4	ME	RG	CD	SP
SCST	-	31.3	26.0	54.3	101.3	-	-	33.3	26.3	55.3	111.4	-
LSTM-A	75.4	35.2	26.9	55.8	108.8	20.0	78.6	35.5	27.3	56.8	118.3	20.8
Up-Xuồng (Cơ sở)	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.1	21.4
StackCap	76.2	35.2	26.5	-	109.1	-	78.6	36.1	27.4	-	120.4	-
CAVP	-	-	-	-	-	-	-	38.6	28.3	58.5	126.3	21.6
GCN-LSTM	77.3	36.8	27.9	57.0	116.3	20.9	80.5	38.2	28.5	58.3	127.6	22.0
VSUA	-	-	-	-	-	-	-	38.4	28.5	58.4	128.6	22.0
SGAE	77.6	36.9	27.7	57.2	116.7	20.9	80.8	38.4	28.4	58.6	127.8	22.1
This Work (MT-I)	78.1	38.4	28.2	58.0	119.0	21.1	80.8	38.9	28.8	58.7	129.6	22.3
This Work (MT-II)	77.9	38.0	28.1	57.6	117.8	21.3	80.5	38.6	28.7	58.4	128.7	22.4

Bảng 1: Hiệu suất của mô hình đơn trên tập dữ liệu MSCOCO (phân tách Karpathy) trong cả entropy chéo và RL thời gian đào tạo. B1, B4, ME, RG, CD và SP biểu thị BLEU-1, BLEU-4, METEOR, ROUGE, CIDEr-D và Tương ứng là SPICE.

	B4		TỐI		RG			
	c5	c40	c5	c40	c5	c40		
GCN-LSTM*	38,7	69,7	28,5	37,6	58,5	73,4	125,3	126,5
VSUA	37,4	68,3	28,2	37,1	57,9	72,8	123,1	125,5
SGAE	38,5	69,7	28,2	37,2	58,6	73,6	123,8	126,5

Bảng 2: Hiệu suất trên máy chủ thử nghiệm trực tuyến COCO của nhiều phương pháp kết hợp các mối quan hệ trực quan.

* biểu thị rằng quy mô lô đào tạo của họ và các thời đại vượt xa bối cảnh trung bình trong (Anderson và cộng sự, 2018; Yang và cộng sự, 2019).

413 danh mục chi tiết bằng cách sử dụng danh sách mã thông báo chú thích. Ví dụ, danh mục đối tượng “người” được mở rộng thành một danh sách các danh mục chi tiết [“cậu bé”, “người đàn ông”, · · ·]. Sau đó, đối với tất cả các mục được trích xuất bộ ba có cả hai đối tượng xuất hiện trong 413 danh sách loại, chúng tôi chọn 200 vị ngữ thường gặp nhất làm loại vị ngữ của mình.

4.2 Phân tích định lượng

So sánh mô hình Chúng tôi so sánh các mô hình của chúng tôi với các mô hình hiện đại sau đây:

(1) SCST (Rennie et al., 2017) sử dụng một cải tiến thuật toán gradient chính xác bằng cách sử dụng đầu ra suy luận của riêng nó để chuẩn hóa phần thưởng; (2) LSTM-A (Yao et al., 2017) tích hợp các thuộc tính hình ảnh được phát hiện vào khuôn khổ chú thích hình ảnh CNN-plus-RNN ; (3) Up-Down (Anderson et al., 2018) sử dụng cả cơ chế chú ý từ dưới lên và từ trên xuống để tập trung nhiều hơn vào các vùng đối tượng nổi bật ; (4) GCN-LSTM (Yao et al., 2018) tận dụng độ thị mạng tách chia trên các đối tượng và mối quan hệ được phát hiện; (5) CAVP (Liu và cộng sự, 2018) đề xuất một mạng chính sách nhận biết ngữ cảnh bằng cách tính đến sự chú ý trực quan như ngữ cảnh để tạo ra;

(6) VSUA (Guo et al., 2019) khai thác sự liên kết

giữa các từ và các loại biểu đồ khác nhau nút; (7) SAGE (Yang và cộng sự, 2019) sử dụng bộ mã hóa đồ thị bổ sung để kết hợp ngôn ngữ độ lệch cảm ứng vào khuôn khổ mã hóa-giải mã.

Đường cơ sở của chúng tôi được xây dựng trên Up-Down (Anderson et al., 2018). Chúng tôi đề xuất hai biến thể của cuối cùng các mô hình sử dụng các khối đa nhiệm vụ khác nhau, cụ thể là MT-I và MT-II được hiển thị trong Hình 4(b). Chúng tôi tiến hành so sánh mở rộng trên tập dữ liệu với các kỹ thuật tiên tiến. Chúng tôi cũng thực hiện phân tích chi tiết để chứng minh tác động của các thành phần khác nhau trong khuôn khổ của chúng tôi.

Bảng 1 liệt kê kết quả của nhiều mô hình đơn lẻ về sự chia tách Karpathy của MSCOCO. Mô hình của chúng tôi vượt trội hơn đáng kể so với mô hình cơ sở, với điểm số CIDEr-D được cải thiện từ 113,5 lên 119,0 và 120,1 đến 129,6 trong entropy chéo và CIDEr-D thời gian tối ưu hóa, tương ứng. Ngoài ra, mô hình với MT-II cho thấy một lợi thế hơn so với MT-I trên SPICE, ngũ ý rằng cơ chế chú ý thị giác thích ứng được đề xuất hoạt động trong khối đa nhiệm II.

Bảng 2 so sánh mô hình của chúng tôi với ba mô hình công kết hợp VRG, cộng với mô hình cơ sở, trên máy chủ thử nghiệm trực tuyến MSCOCO. Mô hình của chúng tôi chứng minh đáng kể từ đường cơ sở (từ 120,5 đến 126,7 trong CIDEr-D) và đã đạt được thành tích tốt nhất kết quả trên tất cả các số liệu đánh giá trên c40 (40 chú thích tham khảo).

Hình 5 cho thấy tác dụng của việc sử dụng các loại khác nhau trọng số y trong mục mắt mèo đa nhiệm vụ (Phương trình 19). Kết quả cho thấy trọng lượng khoảng 0,15 mang lại hiệu suất tốt nhất trong cả hai nhiệm vụ đa nhiệm khôi. Trong khi đó, Hình 6 cho thấy sự cắt bỏ phân tích bằng cách loại bỏ thao tác chú thích đa tác vụ và hoạt động tích chia bộ thị, để kiểm tra tác dụng của

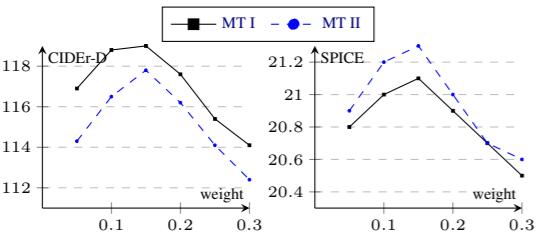


Figure 5: Test results (cross-entropy optimization) on various γ .

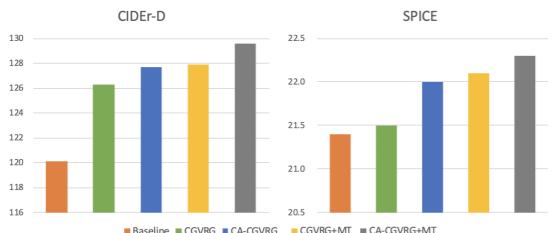


Figure 6: Ablation results (CIDEr-D optimization).

show that both the graph convolution operation and multi-task learning help improve the quality of the generated captions.

Note that the code of our paper has been made publicly available in the webpage provided in the abstract.

Human evaluation We performed human evaluation with three non-author human subjects, using a five-level Likert scale. For each image and each pair of systems in comparison (MT-I vs. Up-Down, MT-I vs. GCN-LSTM, and MT-I vs. SGAE), we show the captions generated by the two systems to the human subjects. We ask each subject if the first caption sentence is: significantly better (2), better (1), equal (0), worse (-1), or significantly worse (-2), compared to the second.

Following (Zhao et al., 2019), we obtain the subjects’ ratings for fidelity (the first caption is superior in terms of making less mistakes?), informativeness (the first caption provides more informative and detailed description?), and fluency (the first caption is more fluent?). For each question asked for an image, we calculate the average of the three subjects’ scores. For each pair of models in comparison, we randomly sampled 50 images from the Karpathy testset.

- MT-I vs. Up-Down: For fidelity, MT-I is better or significantly better on 44% images

(where the average of the three human subjects’ scores is larger than 0.5), equal to Up-Down on 46% images (the average is in range $[-0.5, 0.5]$), and worse or significantly worse on 10% images (average is less than -0.5).

For informativeness, MT-I is better or significantly better on 60% images, equal on 34%, and worse or significantly worse on 6%. For fluency, the numbers are 18%, 72%, and 10%.

- MT-I vs. GCN-LSTM: For fidelity, MT-I is better or significantly better on 40% images, equal to GCN-LSTM on 52%, and worse or significantly worse on 8%. For informativeness, the numbers are 32%, 50%, and 18%, respectively. For fluency, the numbers are 12%, 76%, and 12%.

- MT-I vs. SGAE: For fidelity, MT-I is better or significantly better on 36% images, equal to SGAE on 56%, and worse or significantly worse on 8%. For informativeness, the numbers are 30%, 48%, and 22%, respectively. For fluency, the numbers are 6%, 90%, and 4%.

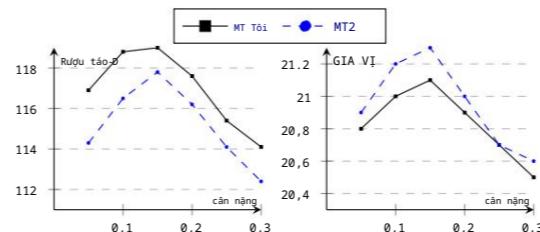
4.3 Qualitative Analysis

Figure 7 shows several specific examples, each including an image, a detected caption guided visual relationship graph, a ground truth sentence, a generated word sequence, and a learned visual relationship composition. We can see that the proposed model generates more accurate captions coherent to the visual relationship detected in the image. Consider the upper middle demo as an example; our model extracts a visual relationship graph covering the critical predicates “filled with” and “in front of” for understanding the image, thus producing a comprehensive description. In addition, we observe that the model generates the triple (*table, filled with, food*), which is a new composition that has not appeared in the training set.

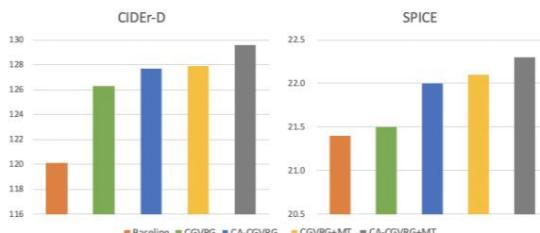
Figure 8 visualizes the effect of our tag sequence generation process. Specifically, we visualize the tag probabilities of the “object”, “predicate”, and “none” category in each generation step. Our model successfully learns to distinguish the correct category for each time step, which is in consistent with the tag of the predicted word. For example, for the generated words “flying over”, the probability for the “predicate” category is the highest, which is also true for words like “bird” and “water”.

5 Conclusions

This paper presents a novel image captioning architecture that constructs caption-guided visual relationship graphs to introduce beneficial inductive



Hình 5: Kết quả thử nghiệm (tối ưu hóa entropy chéo) trên nhiều ý khác nhau.



Hình 6: Kết quả cắt bỏ (tối ưu hóa CIDEr-D).

chứng minh rằng cả hoạt động tích hợp đồ thị và học đa tác vụ đều giúp cải thiện chất lượng chú thích được tạo ra.

Xin lưu ý rằng mã bài báo của chúng tôi đã được công khai trên trang web được cung cấp trong bản tóm tắt.

Đánh giá của con người Chúng tôi đã tiến hành đánh giá của con người với ba đối tượng là con người không phải là tác giả, sử dụng thang đo Likert năm mức. Đối với mỗi hình ảnh và mỗi cặp hệ thống so sánh (MT-I so với Up-Down, MT-I so với GCN-LSTM và MT-I so với SGAE), chúng tôi hiển thị các chủ thích do hai hệ thống tạo ra cho các đối tượng là con người. Chúng tôi hỏi từng đối tượng xem câu chủ thích đầu tiên có: tốt hơn đáng kể (2), tốt hơn (1), bằng (0), tệ hơn (-1) hay tệ hơn đáng kể (-2) so với câu thứ hai không.

Tiếp theo (Zhao và cộng sự, 2019), chúng tôi thu được xếp hạng của các đối tượng về độ trung thực (chú thích đầu tiên vượt trội hơn về mặt ít mắc lỗi hơn?), tính thông tin (chú thích đầu tiên cung cấp nhiều thông tin hơn và mô tả chi tiết hơn?), và tính trôi chảy (chú thích đầu tiên trôi chảy hơn?). Đối với mỗi câu hỏi được yêu cầu cho một hình ảnh, chúng tôi tính điểm trung bình của ba đối tượng. Đối với mỗi cặp mô hình để so sánh, chúng tôi lấy mẫu ngẫu nhiên 50 hình ảnh từ tập kiểm tra Karpathy.

- MT-I so với Up-Down: Về độ trung thực, MT-I tốt hơn hoặc tốt hơn đáng kể trên 44% hình ảnh (trong đó điểm trung bình của ba đối tượng là con người lớn hơn 0.5), bằng với Up- Down trên 46% hình ảnh (điểm trung bình nằm trong khoảng $[0.5, 0.5]$) và kém hơn hoặc kém đáng kể trên 10% hình ảnh (điểm trung bình nhỏ hơn 0.5).

Về tính thông tin, MT-I tốt hơn hoặc tốt hơn đáng kể trên 60% hình ảnh, bằng trên 34% và kém hơn hoặc kém đáng kể trên 6%. Về tính lưu loát, các con số là 18%, 72% và 10%.

- MT-I so với GCN-LSTM: Về độ trung thực, MT-I tốt hơn hoặc tốt hơn đáng kể trên 40% hình ảnh, bằng GCN-LSTM trên 52% và kém hơn hoặc kém đáng kể trên 8%. Về tính thông tin, các con số lần lượt là 32%, 50% và 18%. Về độ trôi chảy, các con số lần lượt là 12%, 76% và 12%.

- MT-I so với SGAE: Về độ trung thực, MT-I tốt hơn hoặc tốt hơn đáng kể trên 36% hình ảnh, bằng SGAE trên 56% và kém hơn hoặc kém đáng kể trên 8%. Về tính thông tin, các con số lần lượt là 30%, 48% và 22%.

Về độ trôi chảy, các con số là 6%, 90% và 4%.

4.3 Phân tích định tính Hình 7

cho thấy một số ví dụ cụ thể, mỗi ví dụ bao gồm một hình ảnh, một biểu đồ quan hệ trực quan được hướng dẫn bằng chú thích đã phát hiện, một câu thực tế, một chuỗi từ được tạo ra và một thành phần quan hệ trực quan đã học. Chúng ta có thể thấy rằng mô hình dễ xuất tạo ra các chủ thích chính xác hơn phù hợp với mối quan hệ trực quan được phát hiện trong hình ảnh. Hãy xem xét bàn demo ở giữa phía trên làm ví dụ; mô hình của chúng tôi trích xuất một biểu đồ quan hệ trực quan bao gồm các vị từ quan trọng “được lắp đầy” và “ở phía trước” để hiểu hình ảnh, do đó tạo ra một mô tả toàn diện. Ngoài ra, chúng tôi quan sát thấy rằng mô hình tạo ra bộ ba (bàn, được lắp đầy, thức ăn), đây là một thành phần mới chưa xuất hiện trong tập huấn luyện.

Hình 8 trực quan hóa hiệu ứng của quá trình tạo chuỗi thẻ của chúng tôi. Cụ thể, chúng tôi trực quan hóa xác suất thẻ của danh mục “đối tượng”, “vị ngữ” và “không có” trong mỗi bước tạo. Mô hình của chúng tôi học thành công cách phân biệt danh mục chính xác cho mỗi bước thời gian, phù hợp với thẻ của từ được dự đoán. Ví dụ, đối với các từ được tạo ra “bay qua”, xác suất cho danh mục “vị ngữ” là cao nhất, điều này cũng đúng với các từ như “chim” và “nước”.

5 Kết luận

Bài báo này trình bày một kiến trúc chú thích hình ảnh mới xây dựng các đồ thị quan hệ trực quan được hướng dẫn bằng chú thích để giới thiệu phương pháp quy nạp có lợi



Figure 7: Several image captioning examples generated by our model.

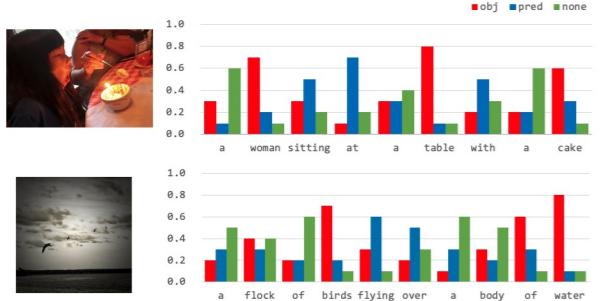


Figure 8: Examples of generated word and tag sequences.

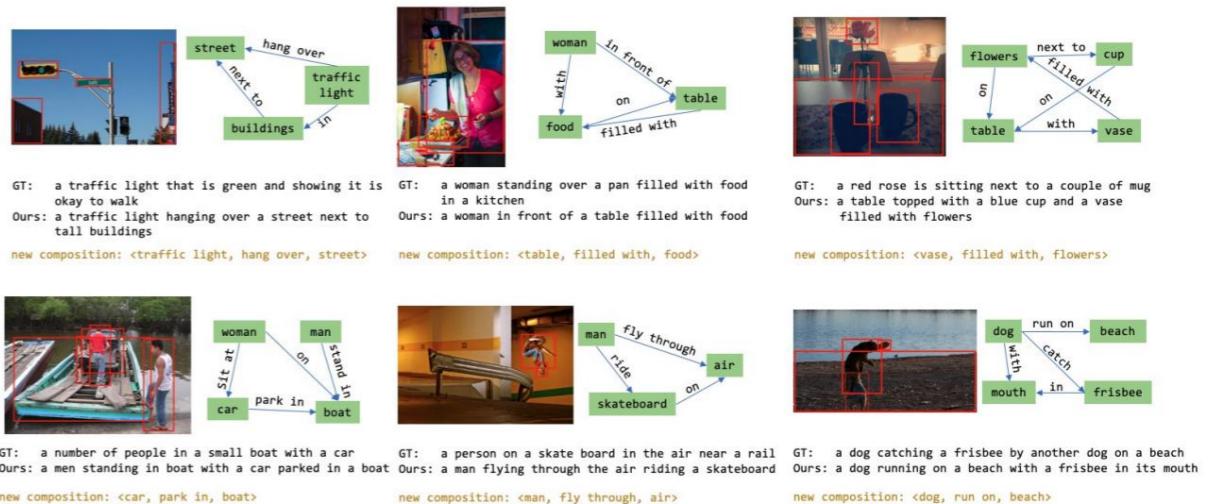
bias to better utilize captions. The representation is further enhanced with text and visual features of neighbouring nodes. During generation, the network is regularized to take into account explicit object/predicate constraints with multi-task learning. Extensive experiments are performed on the MSCOCO dataset, showing that the proposed framework significantly outperforms the baselines, resulting in the state-of-the-art performance under various evaluation metrics. In the near future we plan to extend the proposed approach to several other language-vision modeling tasks.

Acknowledgements

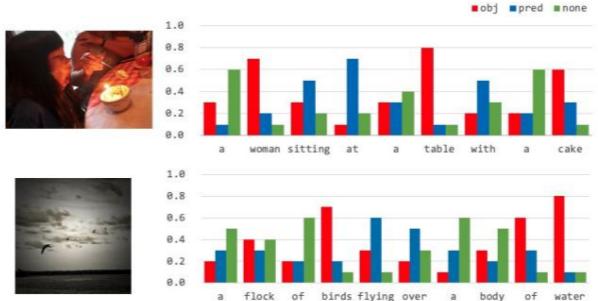
We would like to thank the anonymous reviewers for their valuable comments. This research of the first and last author is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer.
- 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 *CVPR*, pages 6077–6086.
- Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, pages 5659–5667.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv*.
- Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *CVPR*, pages 3076–3086.
- Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.
- Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, pages 2625–2634.
- Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *CVPR*, pages 1473–1482.
- Tài liệu tham khảo
- Peter Anderson, Basura Fernando, Mark Johnson và Stephen Gould. 2016. Spice: Đánh giá chú thích hình ảnh mệnh đề ngữ nghĩa. Trong *ECCV*, trang 382–398. Springer.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould và Lei Zhang. 2018. Sự chú ý từ dưới lên và từ trên xuống để chú thích hình ảnh và trả lời câu hỏi trực quan. Trong *CVPR*, trang 6077–6086.
- Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu và Tat-Seng Chua. 2017. Sca-cnn: Sự chú ý theo không gian và kênh trong mạng lưới tích chập để chú thích hình ảnh. Trong *CVPR*, trang 5659–5667.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár và C Lawrence Zitnick. 2015. Chủ thích coco của Microsoft: Máy chủ thu thập và đánh giá dữ liệu. *arXiv*.
- Bo Dai, Yuqi Zhang và Dahua Lin. 2017. Phát hiện mối quan hệ trực quan với mạng lưới quan hệ sâu sắc. Trong *CVPR*, trang 3076–3086.
- Michael Denkowski và Alon Lavie. 2014. Meteor universal: Đánh giá bản dịch theo ngôn ngữ cụ thể cho bất kỳ ngôn ngữ dịch nào. Trong *Biên bản hội thảo lần thứ chín về dịch máy thông kê*, trang 376–380.
- Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko và Trevor Darrell. 2015. Mạng tích chập hồi quy dài hạn để nhận dạng và mô tả trực quan. Trong *CVPR*, trang 2625–2634.
- Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. Từ chủ thích đến khái niệm trực quan và ngược lại. Trong *CVPR*, trang 1473–1482.



Hình 7: Một số ví dụ về chú thích hình ảnh được tạo ra bởi mô hình của chúng tôi.



Hình 8: Ví dụ về chuỗi từ và thẻ được tạo ra.

thiên vị để sử dụng chú thích tốt hơn. Biểu diễn được cải thiện hơn nữa với các tính năng văn bản và hình ảnh của các nút lân cận. Trong quá trình tạo, mạng được chuẩn hóa để tính đến các ràng buộc đối tượng/vi ngữ rõ ràng với việc học đa tác vụ. Các thí nghiệm mở rộng được thực hiện trên tập dữ liệu MSCOCO, cho thấy khuôn khổ được đề xuất vượt trội đáng kể so với các đường cơ sở, dẫn đến hiệu suất tiên tiến theo nhiều số liệu đánh giá khác nhau. Trong tương lai gần, chúng tôi có kế hoạch mở rộng phương pháp tiếp cận được đề xuất cho một số tác vụ mô hình hóa tầm nhìn ngôn ngữ khác.

Lời cảm ơn

Chúng tôi muốn cảm ơn những người đánh giá ẩn danh vì những bình luận có giá trị của họ. Nghiên cứu này của tác giả đầu tiên và cuối cùng được Hội đồng nghiên cứu khoa học tự nhiên và kỹ thuật Canada (NSERC) hỗ trợ.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko và Trevor Darrell. 2015. Mạng tích chập hồi quy dài hạn để nhận dạng và mô tả trực quan. Trong *CVPR*, trang 2625–2634.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. Từ chủ thích đến khái niệm trực quan và ngược lại. Trong *CVPR*, trang 1473–1482.

- Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *CVPR*, pages 5630–5639.
- Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing Lu. 2019. Aligning linguistic words and visual semantic units for image captioning. *arXiv*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.
- Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *CVPR*, pages 3668–3678.
- Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137.
- Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2019. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6271–6280.
- 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):32–73.
- Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. 2017. Vip-cnn: Visual phrase guided convolutional neural network. In *CVPR*, pages 1347–1356.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer.
- Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yong-dong Zhang, and Feng Wu. 2018. Context-aware visual policy network for sequence-level image captioning. *arXiv*.
- Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In *ECCV*, pages 852–869. Springer.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, pages 375–383.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *CVPR*, pages 7219–7228.
- Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99.
- Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*, pages 7008–7024.
- Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Workshop on Vision and Language (VL15)*, Lisbon, Portugal. ACL.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164.
- Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *CVPR*, pages 203–212.
- 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 *ICML*, pages 2048–2057.
- Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *CVPR*, pages 10685–10694.
- Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *ECCV*, pages 684–699.
- Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *ICCV*, pages 4894–4902.
- Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*, pages 4651–4659.
- Chiết Càn, Trang Càn, Tiều Đông Hà, Văn Thành Bồ, Kenneth Trần, Jianfeng Gao, Lawrence Carin, và Li Deng. 2017. Mạng lưới thành phần ngữ nghĩa để chú thích trực quan. Trong *CVPR*, trang 5630–5639.
- Longteng Guo, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, và Hanqing Lu. 2019. Căn chỉnh ngôn ngữ từ ngữ và đơn vị ngữ nghĩa a trực quan để chú thích hình ảnh. *arXiv*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, và Jian Sun. 2016. Học sâu dựa theo nhận dạng hình ảnh. Trong *CVPR*, trang 770–778.
- Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, và Li Fei-Fei. 2015. Truy xuất hình ảnh bằng đồ thị cảnh. Trong *CVPR*, trang 3668–3678.
- Andrej Karpathy và Li Fei-Fei. 2015. Căn chỉnh ngôn ngữ a thị giác sâu để tạo ra mô tả hình ảnh. Trong *CVPR*, trang 3128–3137.
- Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, và In So Kweon. 2019. Chú thích quan hệ dày đặc: Mạng ba luồng cho chú thích dựa trên mối quan hệ. Trong Biên bản báo cáo của Hội nghị IEEE về Tầm nhìn máy tính và Nhận dạng mẫu, trang 6271–6280.
- 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. Bộ gen trực quan: Kết nối ngôn ngữ và tầm nhìn sử dụng chú thích hình ảnh dày đặc do cộng đồng đóng góp. Tạp chí quốc tế về thị giác máy tính, 123(1):32–73.
- Yikang Li, Wanli Áu Dương, Xiaogang Wang, và Xiao'ou Tang. 2017. Vip-cnn: Hướng dẫn cụm từ trực quan mạng nơ-ron tích chập. Trong *CVPR*, các trang 1347–1356.
- Chin-Yew Lin. 2004. Rouge: Một gói để đánh giá tự động các bản tóm tắt. Tóm tắt văn bản Phân nhánh ra.
- Lin Tsung Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, và C Lawrence Zitnick. 2014. Microsoft dưa: Các đối tượng phổ biến trong ngữ cảnh. Trong *ECCV*, trang 740–755. Mùa xuân.
- Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yong-dong Zhang, và Feng Wu. 2018. Nhận biết bối cảnh mạng chính sách trực quan cho chú thích hình ảnh cấp độ chuỗi. *arXiv*.
- Cewu Lu, Ranjay Krishna, Michael Bernstein, và Li Fei-Fei. 2016. Phát hiện mối quan hệ trực quan với các thông tin trước về ngôn ngữ. Trong *ECCV*, trang 852–869. Mùa xuân.
- Jiasen Lu, Caiming Xiong, Devi Parikh, và Richard Socher. 2017. Biết khi nào cần nhìn: Sự chú ý thích ứng thông qua một línch canh trực quan để chú thích hình ảnh. Trong *CVPR*, trang 375–383.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, và Devi Parikh. 2018. Trò chuyện của trẻ sơ sinh thần kinh. Trong *CVPR*, trang 7219–7228.
- Diego Marcheggiani và Ivan Titov. 2017. Mã hóa câu với mạng lưới tích chập đồ thị để gắn nhãn vai trò ngữ nghĩa a. *arXiv*.
- Kishore Papineni, Salim Roukos, Todd Ward, và Wei-Jing Zhu. 2002. Bleu: một phương pháp đánh giá tự động bản dịch máy. Trong *ACL*, trang 311–318.
- Thiệu Thanh Nhậm, Khải Minh Hà, Ross Girshick, và Jian CN. 2015. R-CNN nhanh hơn: Hướng tới đối tượng thời gian thực phát hiện với mạng lưới đề xuất khu vực. Trong *NIPS*, trang 91–99.
- Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, và Vaibhava Goel. 2017. Tự phê bình đào tạo trình tự cho chú thích hình ảnh. Trong *CVPR*, trang 7008–7024.
- Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, và Christopher D. Manning. 2015. Tạo đồ thị cảnh ngữ nghĩa a chính xác từ mô tả văn bản để cải thiện khả năng truy xuất hình ảnh. Trong Hội thảo về Tầm nhìn và Ngôn ngữ (VL15), Lisbon, Bồ Đào Nha. ACL.
- Ramakrishna Vedantam, C Lawrence Zitnick, và Devi Parikh. 2015. Cider: Đánh giá mô tả hình ảnh dựa trên sự đồng thuận. Trong *CVPR*, trang 4566–4575.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, và Dumitru Erhan. 2015. Hiển thị và kể: Một trình tạo chú thích hình ảnh thần kinh. Trong *CVPR*, trang 3156–3164.
- Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, và Anton Van Den Hengel. 2016. Có giá trị gi các khái niệm cấp cao rõ ràng có trong tầm nhìn về các vấn đề ngôn ngữ? Trong *CVPR*, trang 203–212.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Cho Kyunghyun, Aaron Courville, Ruslan Salakhudinov, Zemel giàu có, và Yoshua Bengio. 2015. Hiển thị, thanh lý và kể lại: Tạo chú thích hình ảnh thần kinh với sự chú ý trực quan. Trong *ICML*, trang 2048–2057.
- Xu Yang, Kaihua Tang, Hanwang Zhang, và Jianfei Cai. 2019. Đồ thị cảnh mã hóa tự động cho hình ảnh chú thích. Trong *CVPR*, trang 10685–10694.
- Ting Yao, Yingwei Pan, Yehao Li, và Tao Mei. 2018. Khám phá mối quan hệ trực quan để chú thích hình ảnh. Trong *ECCV*, trang 684–699.
- Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, và Tao Mei. 2017. Tăng cường chú thích hình ảnh bằng thuộc tính at. Trong *ICCV*, trang 4894–4902.
- Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, và Jiebo Luo. 2016. Chú thích hình ảnh với sự chú ý ngữ nghĩa a. Trong *CVPR*, trang 4651–4659.

Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual translation embedding network for visual relation detection. In *CVPR*, pages 5532–5540.

Sanqiang Zhao, Piyush Sharma, Tomer Levinboim, and Radu Soricut. 2019. Informative image captioning with external sources of information. *arXiv preprint arXiv:1906.08876*.

Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang và Tat-Seng Chua. 2017.
Mạng nhúng dịch trực quan để phát hiện mối quan hệ trực quan.
Trong *CVPR*, trang 5532-5540.

Sanqiang Zhao, Piyush Sharma, Tomer Levinboim và Radu Soricut.
2019. Chủ thích hình ảnh mang tính thông tin với các nguồn
thông tin bên ngoài. Bản in trước *arXiv arXiv:1906.08876*.