

CIDEr: Consensus-based Image Description Evaluation

Ramakrishna Vedantam
Virginia Tech
vrama91@vt.edu

C. Lawrence Zitnick
Microsoft Research
larryz@microsoft.com

Devi Parikh
Virgina Tech
parikh@vt.edu

Abstract

Automatically describing an image with a sentence is a long-standing challenge in computer vision and natural language processing. Due to recent progress in object detection, attribute classification, action recognition, etc., there is renewed interest in this area. However, evaluating the quality of descriptions has proven to be challenging. We propose a novel paradigm for evaluating image descriptions that uses human consensus. This paradigm consists of three main parts: a new triplet-based method of collecting human annotations to measure consensus, a new automated metric (CIDEr) that captures consensus, and two new datasets: PASCAL-50S and ABSTRACT-50S that contain 50 sentences describing each image. Our simple metric captures human judgment of consensus better than existing metrics across sentences generated by various sources. We also evaluate five state-of-the-art image description approaches using this new protocol and provide a benchmark for future comparisons. A version of CIDEr named CIDEr-D is available as a part of MS COCO evaluation server to enable systematic evaluation and benchmarking.

1. Introduction

Recent advances in object recognition [15], attribute classification [23], action classification [26, 9] and crowdsourcing [40] have increased the interest in solving higher level scene understanding problems. One such problem is generating human-like descriptions of an image. In spite of the growing interest in this area, the evaluation of novel sentences generated by automatic approaches remains challenging. Evaluation is critical for measuring progress and spurring improvements in the state of the art. This has already been shown in various problems in computer vision, such as detection [13, 7], segmentation [13, 28], and stereo [39].

Existing evaluation metrics for image description attempt to measure several desirable properties. These include grammaticality, saliency (covering main aspects), correctness/truthfulness, etc. Using human studies, these prop-

erties may be measured, e.g. on separate *one to five* [29, 37, 43, 11] or *pairwise* scales [44]. Unfortunately, combining these various results into one measure of sentence quality is difficult. Alternatively, other works [22, 18] ask subjects to judge the overall quality of a sentence.

An important yet non-obvious property exists when image descriptions are judged by humans: What humans like often does not correspond to what is human-like.¹ We introduce a novel consensus-based evaluation protocol, which measures the similarity of a sentence to the majority, or *consensus* of how most people describe the image (Fig. 1). One realization of this evaluation protocol uses human subjects to judge sentence similarity between a candidate sentence and human-provided ground truth sentences. The question “Which of two sentences is more similar to this other sentence?” is posed to the subjects. The resulting quality score is based on how often a sentence is labeled as being *more* similar to a human-generated sentence. The relative nature of the question helps make the task objective. We encourage the reader to review how a similar protocol has been used in [41] to capture human perception of image similarity. These annotation protocols for similarity may be understood as instantiations of 2AFC (two alternative forced choice) [3], a popular modality in psychophysics.

Since human studies are expensive, hard to reproduce, and slow to evaluate, automatic evaluation measures are commonly desired. To be useful in practice, automated metrics should agree well with human judgment. Some popular metrics used for image description evaluation are BLEU [33] (precision-based) from the machine translation community and ROUGE [45] (recall-based) from the summarization community. Unfortunately, these metrics have been shown to correlate weakly with human judgment [22, 11, 4, 18]. For the task of judging the overall quality of a description, the METEOR [11] metric has shown better correlation with human subjects. Other metrics rely on the ranking of captions [18] and cannot evaluate novel

¹This is a subtle but important distinction. We show qualitative examples of this in the appendix. That is, the sentence that is most similar to a typical human generated description is often not judged to be the “best” description. In this paper, we propose to directly measure the “human-likeness” of automatically generated sentences.

CIDEr: Đánh giá mô tả hình ảnh dựa trên sự đồng thuận

Ramakrishna Vedantam
Viện Công nghệ Virginia
vrama91@vt.edu

C. Lawrence Zitnick
Nghiên cứu của Microsoft
larryz@microsoft.com

Devi Parikh
Viện Công nghệ Virginia
parikh@vt.edu

Tóm tắt

Tự động mô tả một hình ảnh bằng một câu là một thách thức lâu dài trong tầm nhìn máy tính và xử lý ngôn ngữ tự nhiên. Do tiến bộ gần đây trong đối tượng phát hiện, phân loại thuộc tính, nhận dạng hành động, v.v., có sự quan tâm mới trong lĩnh vực này. Tuy nhiên, việc đánh giá chất lượng mô tả đã được chứng minh là một thách thức. Chúng tôi đề xuất một mô hình mới để đánh giá mô tả hình ảnh sử dụng sự đồng thuận của con người. Mô hình này bao gồm ba phần chính: một phương pháp mới dựa trên bộ ba để thu thập chủ thích của con người nhằm đo lường sự đồng thuận, một số liệu tự động mới (CIDEr) ghi lại sự đồng thuận và hai bộ dữ liệu mới: PASCAL-50S và ABSTRACT-50S chứa 50 câu mô tả từng hình ảnh. Chỉ số đơn giản của chúng tôi nắm bắt được phán đoán của con người về sự đồng thuận tốt hơn so với các số liệu hiện có trên các câu được tạo ra bởi nhiều nguồn khác nhau. Chúng tôi cũng đánh giá năm phương pháp mô tả hình ảnh tiên tiến nhất bằng cách sử dụng giao thức mới này và cung cấp một chuẩn mực để so sánh trong tương lai. Một phiên bản của CIDEr có tên là CIDEr-D có sẵn như một phần của máy chủ đánh giá MS COCO để cho phép đánh giá và so sánh một cách có hệ thống.

1. Giới thiệu

Những tiến bộ gần đây trong nhận dạng đối tượng [15], thuộc tính phân loại [23], phân loại hành động [26, 9] và huy động cộng đồng [40] đã làm tăng sự quan tâm trong việc giải quyết các vấn đề cao hơn vấn đề hiểu cảnh cấp độ. Một trong những vấn đề như vậy là tạo ra các mô tả giống như con người về một hình ảnh. Mặc dù của sự quan tâm ngày càng tăng trong lĩnh vực này, việc đánh giá tiêu chuẩn các câu được tạo ra bởi các phương pháp tiếp cận tự động vẫn còn là thách thức. Đánh giá là rất quan trọng để đo lường tiến trình và thúc đẩy cải tiến trong tình trạng nghệ thuật. Điều này đã được thể hiện trong nhiều vấn đề khác nhau về tầm nhìn máy tính, chẳng hạn như phát hiện [13, 7], phân đoạn [13, 28] và âm thanh nói [39].

Các số liệu đánh giá hiện có cho mô tả hình ảnh có gắng do lưỡng một số thuộc tính mong muốn. Chúng bao gồm ngữ pháp, độ nổi bật (bao gồm các khía cạnh chính), tính chính xác/trung thực, v.v. Sử dụng các nghiên cứu trên con người, các prop-

có thể đo được các giá trị khác nhau, ví dụ trên riêng biệt một đến năm [29, 37, 43, 11] hoặc thang đo từng cặp [44]. Thật không may, việc kết hợp những kết quả khác nhau này thành một thước đo chất lượng câu là khó khăn. Ngoài ra, các tác phẩm khác [22, 18] yêu cầu các đối tượng đánh giá chất lượng chung của một câu.

Một đặc tính quan trọng nhưng không rõ ràng tồn tại khi mô tả hình ảnh được con người đánh giá: Con người thích gì thuong không tương ứng với những gì giống con người. Chúng tôi giới thiệu một giao thức đánh giá dựa trên sự đồng thuận mới, đo lường mức độ giống nhau của một câu với đa số, hoặc sự đồng thuận về cách hầu hết mọi người mô tả hình ảnh (Hình 1). Một thực hiện giao thức đánh giá này sử dụng đối tượng là con người để đánh giá sự giống nhau của câu giữa một câu ứng viên và các câu sự thật cơ bản do con người cung cấp. Câu hỏi “Trong hai câu, câu nào giống với câu kia hơn?” “Tence?” được đặt ra cho các đối tượng. Điểm chất lượng kết quả dựa trên tần suất một câu được đánh giá là nhiều hơn tư duy tự như một câu do con người tạo ra. Bản chất tư duy đối của câu hỏi giúp làm cho nhiệm vụ khách quan. Chúng tôi khuyến khích người đọc xem xét cách một giao thức tư duy tự đã được sử dụng trong [41] để nắm bắt nhận thức của con người về sự giống nhau của hình ảnh. Các giao thức chủ yếu này cho sự giống nhau có thể được hiểu là các thể hiện của 2AFC (hai thay thế bắt buộc lựa chọn) [3], một phương thức phổ biến trong tâm lý học.

Vì nghiên cứu trên người xác tồn kém, khó tái tạo, và chậm đánh giá, các biện pháp đánh giá tự động là thường được mong muốn. Để hữu ích trong thực tế, tự động số liệu thông kê phải phù hợp với phán đoán của con người. Một số các số liệu phổ biến được sử dụng để đánh giá mô tả hình ảnh là BLEU [33] (dựa trên độ chính xác) từ cộng đồng dịch máy và ROUGE [45] (dựa trên khả năng thu hồi) từ cộng đồng tóm tắt. Thực không may, các số liệu này đã được chứng minh là có mối tương quan yếu với phán đoán của con người [22, 11, 4, 18]. Đối với nhiệm vụ đánh giá chất lượng tổng thể của một mô tả, số liệu METEOR [11] đã cho thấy tư duy quan trọng hơn với các đối tượng là con người. Các số liệu khác dựa vào thứ hạng của chủ yếu [18] và không thể đánh giá tiêu chuẩn.

Đây là một sự phân biệt tinh tế như quan trọng. Chúng tôi trình bày các ví dụ định tính về điều này trong phần phụ lục. Nghĩa là, câu giống nhất với một mô tả điển hình do con người tạo ra thường không được đánh giá là “tốt nhất” mô tả. Trong bài báo này, chúng tôi đề xuất do trực tiếp “mức độ giống con người” của các câu được tạo tự động.



Figure 1: Images from our PASCAL-50S (left) and ABSTRACT-50S (right) datasets with a subset of corresponding (human) sentences. Sentences shown in **bold** are representative of the consensus descriptions for these images. We propose to capture such descriptions with our evaluation protocol.

image descriptions.

We propose a new automatic *consensus* metric of image description quality – CIDEr (Consensus-based Image Description Evaluation). Our metric measures the similarity of a generated sentence against a set of ground truth sentences written by humans. Our metric shows high agreement with consensus as assessed by humans. Using sentence similarity, the notions of grammaticality, saliency, importance and accuracy (precision and recall) are inherently captured by our metric.

Existing datasets popularly used to evaluate image description approaches have a maximum of only five descriptions per image [35, 18, 32]. However, we find that five sentences are not sufficient for measuring how a “majority” of humans would describe an image. Thus, to accurately measure consensus, we collect two new evaluation datasets containing 50 descriptions per image – PASCAL-50S and ABSTRACT-50S. The PASCAL-50S dataset is based on the popular UIUC Pascal Sentence Dataset, which has 5 descriptions per image. This dataset has been used for both training and testing in numerous works [29, 22, 14, 37]. The ABSTRACT-50S dataset is based on the dataset of Zitnick and Parikh [46]. While previous methods have only evaluated using 5 sentences, we explore the use of 1 to ~50 reference sentences. Interestingly, we find that most metrics improve in performance with more sentences.² Inspired by this finding, the MS COCO testing dataset now contains 5K images with 40 reference sentences to boost the accuracy of automatic measures [5].

Contributions: In this work, we propose a consensus-based evaluation protocol for image descriptions. We introduce a new annotation modality for human judgment, a new automated metric, and two new datasets. We compare the performance of five state-of-the-art machine generation approaches [29, 22, 14, 37]. Our code and datasets are available on the author’s webpages. Finally, to facilitate the adoption of this protocol, we have made CIDEr available as a metric on the newly released MS COCO caption evaluation server [5].

²Except BLEU computed on unigrams

2. Related Work

Vision and Language: Numerous papers have studied the relationship between language constructs and image content. Berg *et al.* [2] characterize the relative importance of objects (nouns). Zitnick and Parikh [46] study relationships between visual and textual features by creating a synthetic Abstract Scenes Dataset. Other works have modeled prepositional relationships [16], attributes (adjectives) [23, 34], and visual phrases (*i.e.* visual elements that co-occur) [38]. Recent works have utilized techniques in deep learning to learn joint embeddings of text and image fragments [20].

Image Description Generation: Various methods have been explored for generating full descriptions for images. Broadly, the techniques are either retrieval- [14, 32, 18] or generation-based [29, 22, 44, 37]. While some retrieval-based approaches use global retrieval [14], others retrieve text phrases and stitch them together in an approach inspired by extractive summarization [32]. Recently, generative approaches based on combination of Convolutional and Recurrent Neural Networks [19, 6, 10, 42, 27, 21] have created a lot of excitement. Other generative approaches have explored creating sentences by inference over image detections and text-based priors [22] or exploiting word co-occurrences using syntactic trees [29]. Rohrbach *et al.* [37] propose a machine translation approach that goes from an intermediate semantic representation to sentences. Some other approaches include [17, 24, 43, 44]. Most of the approaches use the UIUC Pascal Sentence [14, 22, 29, 37, 17] and the MS COCO datasets [19, 6, 10, 42, 27, 21] for evaluation. In this work we focus on the problem of evaluating image captioning approaches.

Automated Evaluation: Automated evaluation metrics have been used in many domains within Artificial Intelligence (AI), such as statistical machine translation and text summarization. Some of the popular metrics in machine translation include those based on precision, such as BLEU [33] and those based on precision as well as recall, such as METEOR [1]. While BLEU (BiLingual Evaluation Understudy) has been the most popular metric, its effective-



Hình 1: Hình ảnh từ bộ dữ liệu PASCAL-50S (trái) và ABSTRACT-50S (phải) của chúng tôi với một tập hợp con tư ứng (con ngựa). Các câu được in đậm là đại diện cho các mô tả đồng thuận cho những hình ảnh này. Chúng tôi đề xuất để nắm bắt những mô tả như vậy với giao thức đánh giá của chúng tôi.

mô tả hình ảnh.

Chúng tôi đề xuất một thư ứng do đồng thuận tự động mới về hình ảnh chất lượng mô tả – CIDEr (Đánh giá mô tả hình ảnh dựa trên sự đồng thuận). Số liệu của chúng tôi do lưỡng sự giống nhau của một câu được tạo ra dựa trên một tập hợp các câu chân lý cơ bản được viết bởi con người. Số liệu của chúng tôi cho thấy sự đồng thuận cao với sự đồng thuận được đánh giá bởi con người. Sử dụng sự tương đồng của câu, các khái niệm về ngữ pháp, tính nổi bật, tầm quan trọng và độ chính xác (độ chính xác và độ thu hồi) vốn được nắm bắt bởi số liệu của chúng tôi.

Các tập dữ liệu hiện có thường được sử dụng để đánh giá các phương pháp mô tả hình ảnh có tối đa chỉ năm mô tả cho mỗi hình ảnh [35, 18, 32]. Tuy nhiên, chúng tôi thấy rằng năm những câu không đủ để đo lường mức độ “đa số” của con người sẽ mô tả một hình ảnh. Vì vậy, để chính xác do lưỡng sự đồng thuận, chúng tôi thu thập hai tập dữ liệu đánh giá mới chứa 50 mô tả cho mỗi hình ảnh – PASCAL-50S và TÓM TẮT-50S. Bộ dữ liệu PASCAL-50S dựa trên Bộ dữ liệu Pascal UIUC phổ biến, có 5 mô tả cho mỗi hình ảnh. Bộ dữ liệu này đã được sử dụng cho cả đào tạo và thử nghiệm trong nhiều tác phẩm [29, 22, 14, 37]. Bộ dữ liệu ABSTRACT-50S dựa trên bộ dữ liệu của Zitnick và Parikh [46]. Trong khi các phương pháp trước đây chỉ đánh giá bằng 5 câu, chúng tôi khám phá việc sử dụng 1 đến 50 câu tham chiếu. Điều thú vị là chúng tôi thấy rằng hầu hết các số liệu cải thiện hiệu suất với nhiều câu hơn. Lấy cảm hứng từ phát hiện này, tập dữ liệu thử nghiệm MS COCO hiện chứa 5K hình ảnh với 40 câu tham khảo để tăng độ chính xác của biện pháp tự động [5].

Đóng góp: Trong công trình này, chúng tôi đề xuất một giao thức đánh giá dựa trên sự đồng thuận cho các mô tả hình ảnh. Chúng tôi giới thiệu một phương thức chú thích mới cho sự phán đoán của con người, một số liệu tự động mới và hai tập dữ liệu mới. Chúng tôi so sánh hiệu suất của năm phương pháp tạo máy tiên tiến [29, 22, 14, 37]. Mô và tập dữ liệu của chúng tôi có sẵn trên trang web của tác giả. Cuối cùng, để tạo điều kiện thuận lợi việc áp dụng giao thức này, chúng tôi đã làm cho CIDEr có sẵn như một số liệu trên tiêu đề MS COCO mới phát hành máy chủ đánh giá [5].

²Ngoài trừ BLEU được tính trên unigram



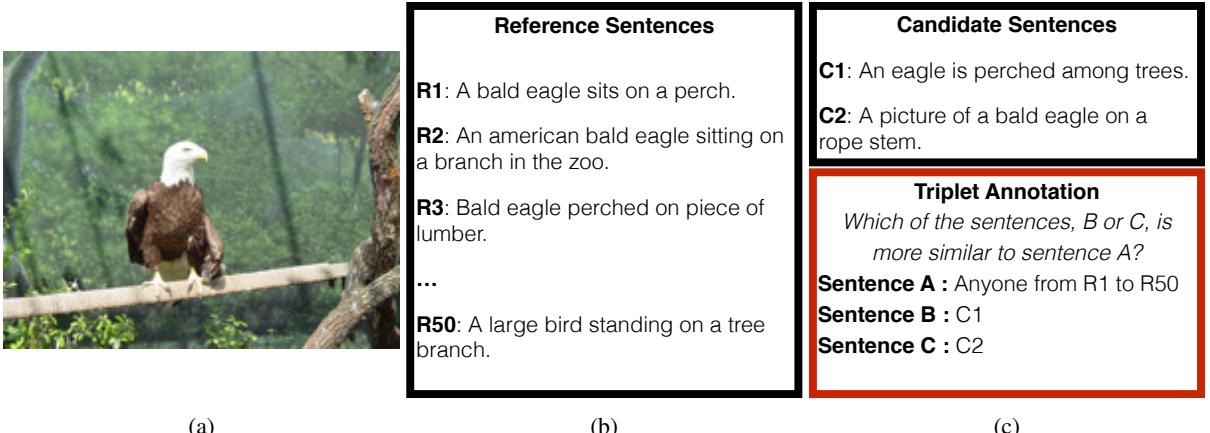
Jenny đang cầm một quả bóng zô và Mike đang cầm một quả bóng chày. Jenny đang chơi bóng zô và Mike đang chơi bóng chày. Jenny mang theo một quả bóng lớn hơn Mike. Mike buồn vì Jenny sẽ rời đi sau năm ngày nữa.

2. Công trình liên quan

Tầm nhìn và Ngôn ngữ: Nhiều bài báo đã nghiên cứu mối quan hệ giữa các cấu trúc ngôn ngữ và nội dung hình ảnh. Berg *et al.* [2] mô tả tầm quan trọng tương đối của các đối tượng (danh từ). Zitnick và Parikh [46] nghiên cứu mối quan hệ giữa các đặc điểm trực quan và văn bản bằng cách tạo ra một Bộ dữ liệu Cảnh trùu tương hỗ. Các tác phẩm khác đã các mối quan hệ giới từ được mô hình hóa [16], các thuộc tính (tính từ) [23, 34] và cụm từ trực quan (tức là các yếu tố trực quan cùng xảy ra) [38]. Các công trình gần đây đã sử dụng các kỹ thuật trong học sâu để học những chung của văn bản và hình ảnh mankind vở [20].

Mô tả hình ảnh Thẻ hệ: Nhiều phương pháp khác nhau có đã được khám phá để tạo ra mô tả đầy đủ cho hình ảnh. Nói chung, các kỹ thuật là truy xuất- [14, 32, 18] hoặc dựa trên thẻ hệ [29, 22, 44, 37]. Trong khi một số phương pháp tiếp cận dựa trên truy xuất sử dụng truy xuất toàn cầu [14], những phương pháp khác truy xuất cụm từ văn bản và ghép chúng lại với nhau theo cách tiếp cận lấy cảm hứng từ tóm tắt trích xuất [32]. Gần đây, các cách tiếp cận tạo sinh dựa trên sự kết hợp của Tích hợp và Mạng nơ-ron hồi quy [19, 6, 10, 42, 27, 21] có tạo ra rất nhiều sự phản kháng. Các cách tiếp cận tạo ra khác đã khám phá việc tạo ra câu bằng cách suy luận qua hình ảnh phát hiện và các tiên nghiệm dựa trên văn bản [22] hoặc khai thác sự đồng xuất hiện của từ bằng cách sử dụng cây cú pháp [29]. Rohrbach *et al.* [37] đề xuất một phương pháp dịch máy đi từ biểu diễn ngữ nghĩa trung gian cho các câu. Một số các cách tiếp cận khác bao gồm [17, 24, 43, 44]. Hầu hết các cách tiếp cận sử dụng Câu Pascal của UIUC [14, 22, 29, 37, 17] và các tập dữ liệu MS COCO [19, 6, 10, 42, 27, 21] để đánh giá. Trong công trình này, chúng tôi tập trung vào vấn đề đánh giá phương pháp chú thích hình ảnh.

Đánh giá tự động: Số liệu đánh giá tự động đã được sử dụng trong nhiều lĩnh vực trong Trí tuệ nhân tạo (AI), chẳng hạn như dịch máy thông kê và tóm tắt văn bản. Một số số liệu phổ biến trong dịch máy bao gồm số liệu dựa trên độ chính xác, chẳng hạn như BLEU [33] và những cái dựa trên độ chính xác cũng như khả năng thu hồi, chẳng hạn như METEOR [1]. Trong khi BLEU (Đánh giá song ngữ Nghiên cứu sinh) là thư ứng do phổ biến nhất, hiệu quả của nó-



(a)

(b)

(c)

Figure 2: Illustration of our triplet annotation modality. Given an image (a), with reference sentences (b) and a pair of candidate sentences (c, top), we match them with a reference sentence one by one to form triplets (c, bottom). Subjects are shown these 50 triplets on Amazon Mechanical Turk and asked to pick which sentence (B or C) is more similar to sentence A.

ness has been repeatedly questioned [22, 11, 4, 18]. A popular metric in the summarization community is ROUGE [45] (Recall Oriented Understudy of Gisting Evaluation). This metric is primarily recall-based and thus has a tendency to reward long sentences with high recall. These metrics have been shown to have weak to moderate correlation with human judgment [11]. Recently, METEOR has been used for image description evaluation with more promising results [12]. Another metric proposed by Hodosh *et al.* [18] can only evaluate ranking-based approaches, it cannot evaluate novel sentences. We propose a consensus-based metric that rewards a sentence for being similar to the majority of human written descriptions. Interestingly, similar ideas have been used previously to evaluate text summarization [31].

Datasets: Numerous datasets have been proposed for studying the problem of generating image descriptions. The most popular dataset is the UIUC Pascal Sentence Dataset [35]. This dataset contains 5 human written descriptions for 1,000 images. This dataset has been used by a number of approaches for training and testing. The SBU captioned photo dataset [32] contains one description per image for a million images, mined from the web.

These are commonly used for training image description approaches. Approaches are then tested on a query set of 500 images with one sentence each. The Abstract Scenes dataset [46] contains cartoon-like images with two descriptions. The recently released MS COCO dataset [25] contains five sentences for a collection of over 100K images. This dataset is gaining traction with recent image description approaches [19, 6, 10, 42, 27, 21]. Other datasets of images and associated descriptions include ImageClef [30] and Flickr8K [18]. In this work, we introduce two new datasets. First is the PASCAL-50S dataset where we collected 50 sentences per image for the 1,000

images from UIUC Pascal Sentence dataset. The second is the ABSTRACT-50S dataset where we collected 50 sentences for a subset of 500 images from the Abstract Scenes dataset. We demonstrate that more sentences per image are essential for reliable automatic evaluation.

The rest of this paper is organized as follows. We first give details of our triplet human annotation modality (Sec. 3). Then we provide the details of our consensus-based automated metric, CIDEr (Sec. 4). In Sec. 5 we provide the details of our two new image-sentence datasets, PASCAL-50S and ABSTRACT-50S. Our contributions of triplet annotation, metric and dataset make consensus-based image description evaluation feasible. Our results (Sec. 7) demonstrate that our automated metric and our proposed datasets capture consensus better than existing choices.

All our human studies are performed on the Amazon Mechanical Turk (AMT). Subjects are restricted to the United States, and other qualification criteria are imposed based on worker history.³

3. Consensus Interface

Given an image and a collection of human generated *reference* sentences describing it, the goal of our consensus-based protocol is to measure the similarity of a *candidate* sentence to a majority of how most people describe the image (*i.e.* the *reference* sentences). In this section, we describe our human study protocol for generating ground truth consensus scores. In Sec. 7, these ground truth scores are used to evaluate several automatic metrics including our proposed CIDEr metric.

An illustration of our human study interface is shown in Fig. 2. Subjects are shown three sentences: A, B and C. They are asked to pick which of two sentences (B or C)

³Approval rate greater than 95%, minimum 500 HITs approved



Hình 2: Minh họa về phương thức chú thích bộ ba của chúng tôi. Cho một hình ảnh (a), với các câu tham chiếu (b) và một cặp câu ứng viên (c, trên cùng), chúng tôi ghép chúng với một câu tham chiếu từng cái một để tạo thành bộ ba (c, dưới cùng). Chủ ngữ là đã cho thấy 50 bộ ba này trên Amazon Mechanical Turk và được yêu cầu chọn câu nào (B hoặc C) giống với câu hơn A.

nhà đã được đặt câu hỏi nhiều lần [22, 11, 4, 18]. Một số liệu phổ biến trong cộng đồng tóm tắt là ROUGE [45]

(Nhớ lại hướng nghiên cứu của Đánh giá Gisting). Điều này phép do chủ yếu dựa trên sự nhớ lại và do đó có xu hướng thưa ở cho những câu dài với khả năng nhớ lại cao. Những số liệu này có đã được chứng minh là có mối tương quan yếu đến trung bình với phản đoán của con người [11]. Gần đây, METEOR đã được sử dụng để đánh giá mô tả hình ảnh với kết quả hứa hẹn hơn [12]. Một số liệu khác đã được đề xuất bởi Hodosh *et al.* [18]

chỉ có thể đánh giá các phương pháp tiếp cận dựa trên xếp hạng, nó không

thể đánh giá các câu mới. Chúng tôi đã đề xuất một phép đo dựa trên sự đồng thuận để thử nghiệm cho một câu vì nó giống với phần lớn các mô tả đã được viết bởi con người. Thật thú vị, tương tự

các ý tưởng đã được sử dụng trước đây để đánh giá tóm tắt văn bản [31].

Bộ dữ liệu: Nhiều bộ dữ liệu đã được đề xuất cho nghiên cứu vẫn để tạo ra mô tả hình ảnh.

Bộ dữ liệu phổ biến nhất là UIUC Pascal Sentence

Bộ dữ liệu [35]: Bộ dữ liệu này chứa 5 mô tả được viết bởi con người cho

1.000 hình ảnh. Bộ dữ liệu này đã được sử dụng

bằng một số phương pháp đào tạo và thử nghiệm.

Bộ dữ liệu ảnh có chủ đề của SBU [32] chứa một mô tả cho mỗi hình ảnh

trong một triệu hình ảnh, được khai thác từ web.

Chúng thử nghiệm đã được sử dụng để đào tạo mô tả hình ảnh

các cách tiếp cận. Các cách tiếp cận sau đó được thử nghiệm trên một tập truy vấn

500 hình ảnh với mỗi hình ảnh có một câu. Các cách thử nghiệm

tập dữ liệu [46] chứa hình ảnh giống phim hoạt hình với hai mô tả.

Tập dữ liệu MS COCO mới phát hành [25] chứa năm câu cho bộ sưu tập

hơn 100K ảnh. Tập dữ liệu này đang thu hút sự chú ý với các phương

pháp mô tả ảnh gần đây [19, 6, 10, 42, 27, 21]. Khác

các tập dữ liệu hình ảnh và mô tả liên quan bao gồm ImageClef [30] và Flickr8K [18]. Trong công trình này, chúng tôi giới thiệu hai tập dữ liệu mới. Đầu tiên là tập dữ liệu PASCAL-50S

nơi chúng tôi thu thập 50 câu cho mỗi hình ảnh trong 1.000

hình ảnh từ bộ dữ liệu Câu Pascal của UIUC. Thứ hai là tập dữ liệu ABSTRACT-50S trong đó chúng tôi đã thu thập 50 câu cho một tập hợp con gồm 500 hình ảnh từ Abstract Scenes tập dữ liệu. Chúng tôi chứng minh rằng nhiều câu hơn trên mỗi hình ảnh cần thiết cho việc đánh giá tự động đáng tin cậy.

Phản hồi lại của bài báo này đã được tổ chức như sau. Chúng tôi đầu tiên cung cấp thông tin chi tiết về phương thức chú thích con người bộ ba của chúng tôi (Phần 3). Sau đó, chúng tôi cung cấp thông tin chi tiết về số liệu tự động dựa trên sự đồng thuận của chúng tôi, CIDEr (Phần 4). Trong Phần 5, chúng tôi cung cấp thông tin chi tiết về hai tập dữ liệu câu-hình ảnh mới của chúng tôi, PASCAL-50S và ABSTRACT-50S. Những đóng góp của chúng tôi về chủ đề bộ ba, số liệu và tập dữ liệu tạo ra sự đồng thuận đánh giá mô tả hình ảnh khá thi. Kết quả của chúng tôi (Phần 7) chứng minh rằng số liệu tự động của chúng tôi và đề xuất của chúng tôi tập dữ liệu nắm bắt sự đồng thuận tốt hơn so với các lựa chọn hiện có.

Tất cả các nghiên cứu trên người của chúng tôi đều được thực hiện trên Amazon Mechanical Turk (AMT). Các đối tượng bị giới hạn ở Hoa Kỳ. Các tiêu bang và các tiêu chuẩn điều kiện khác được áp dụng dựa trên lịch sử công nhận.3

3. Giao diện đồng thuận

Với một hình ảnh và một tập hợp các câu tham chiếu do con người tạo ra để mô tả hình ảnh đó, mục tiêu của giao thức dựa trên sự đồng thuận của chúng tôi là do lường mức độ giống nhau của một ứng viên câu nói này đại diện cho cách mà hầu hết mọi người mô tả hình ảnh (tức là các câu tham chiếu). Trong phần này, chúng tôi mô tả giao thức nghiên cứu của con người chúng tôi để tạo ra một danh sách số đồng thuận về sự thật. Trong Mục 7, những điểm số về sự thật cơ bản này được sử dụng để đánh giá một số số liệu tự động bao gồm đề xuất chỉ số CIDEr.

Một hình ảnh minh họa về giao diện nghiên cứu của con người được thể hiện trong Hình 2. Các chủ ngữ được hiển thị trong ba câu: A, B và C. Họ được yêu cầu chọn một trong hai câu (B hoặc C)

3Tỷ lệ phê duyệt lớn hơn 95%, tối thiểu 500 HIT được phê duyệt

is most similar to sentence A. Sentences B and C are two candidate sentences, while sentence A is a reference sentence. For each choice of B and C, we form triplets using all the reference sentences for an image. We provide no explicit concept of “similarity”. Interestingly, even though we do not say that the sentences are image descriptions, some workers commented that they were imagining the scene to make the choice. The relative nature of the task – “Which of the two sentences, B or C, is more similar to A?” – helps make the assessment more objective. That is, it is easier to judge if one sentence is more similar than another to a sentence, than to provide an absolute rating from 1 to 5 of the similarity between two sentences [3].

We collect three human judgments for each triplet. For every triplet, we take the majority vote of the three judgments. For each pair of candidate sentences (B, C), we assign B the winner if it is chosen as more similar by a majority of triplets, and similarly for C. These pairwise relative rankings are used to evaluate the performance of the automated metrics. That is, when automatic metrics give both sentences B and C a score, we check whether B received a higher score or C. Accuracy is computed as the proportion of candidate pairs on which humans and the automatic metric agree on which of the two sentences is the winner.

4. CIDEr Metric

Our goal is to automatically evaluate for image I_i how well a candidate sentence c_i matches the consensus of a set of image descriptions $S_i = \{s_{i1}, \dots, s_{im}\}$. All words in the sentences (both candidate and references) are first mapped to their stem or root forms. That is, “fishes”, “fishing” and “fished” all get reduced to “fish.” We represent each sentence using the set of n -grams present in it. An n -gram ω_k is a set of one or more ordered words. In this paper we use n -grams containing one to four words.

Intuitively, a measure of consensus would encode how often n -grams in the candidate sentence are present in the reference sentences. Similarly, n -grams not present in the reference sentences should not be in the candidate sentence. Finally, n -grams that commonly occur across all images in the dataset should be given lower weight, since they are likely to be less informative. To encode this intuition, we perform a Term Frequency Inverse Document Frequency (TF-IDF) weighting for each n -gram [36]. The number of times an n -gram ω_k occurs in a reference sentence s_{ij} is denoted by $h_k(s_{ij})$ or $h_k(c_i)$ for the candidate sentence c_i . We compute the TF-IDF weighting $g_k(s_{ij})$ for each n -gram ω_k using:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log \left(\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))} \right), \quad (1)$$

where Ω is the vocabulary of all n -grams and I is the set of all images in the dataset. The first term measures the TF of each n -gram ω_k , and the second term measures the rarity of ω_k using its IDF. Intuitively, TF places higher weight on n -grams that frequently occur in the reference sentence describing an image, while IDF reduces the weight of n -grams that commonly occur across all images in the dataset. That is, the IDF provides a measure of word saliency by discounting popular words that are likely to be less visually informative. The IDF is computed using the logarithm of the number of images in the dataset $|I|$ divided by the number of images for which ω_k occurs in any of its reference sentences.

Our CIDEr $_n$ score for n -grams of length n is computed using the average cosine similarity between the candidate sentence and the reference sentences, which accounts for both precision and recall:

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m} \sum_j \frac{\mathbf{g}^n(c_i) \cdot \mathbf{g}^n(s_{ij})}{\|\mathbf{g}^n(c_i)\| \|\mathbf{g}^n(s_{ij})\|}, \quad (2)$$

where $\mathbf{g}^n(c_i)$ is a vector formed by $g_k(c_i)$ corresponding to all n -grams of length n and $\|\mathbf{g}^n(c_i)\|$ is the magnitude of the vector $\mathbf{g}^n(c_i)$. Similarly for $\mathbf{g}^n(s_{ij})$.

We use higher order (longer) n -grams to capture grammatical properties as well as richer semantics. We combine the scores from n -grams of varying lengths as follows:

$$\text{CIDEr}(c_i, S_i) = \sum_{n=1}^N w_n \text{CIDEr}_n(c_i, S_i), \quad (3)$$

Empirically, we found that uniform weights $w_n = 1/N$ work the best. We use $N = 4$.

5. New Datasets

We propose two new datasets – PASCAL-50S and ABSTRACT-50S – for evaluating image caption generation methods. Both the datasets have 50 reference sentences per image for 1,000 and 500 images respectively. These are intended as “testing” datasets, crafted to enable consensus-based evaluation. For a list of training datasets, we encourage the reader to explore [25, 32]. The PASCAL-50S dataset uses all 1,000 images from the UIUC Pascal Sentence Dataset [35] whereas the ABSTRACT-50S dataset uses 500 random images from the Abstract Scenes Dataset [46]. The Abstract Scenes Dataset contains scenes made from clipart objects. Our two new datasets are different from each other both visually and in the type of image descriptions produced.

Our goal was to collect image descriptions that are objective and representative of the image content. Subjects were shown an image and a text box, and were asked to “Describe what is going on in the image”. We asked subjects to

giống nhất với câu A. Câu B và C là hai câu ứng viên, trong khi câu A là câu tham chiếu. Đối với mỗi lựa chọn B và C, chúng tôi tạo thành bộ bằng cách sử dụng tất cả các câu tham chiếu cho một hình ảnh. Chúng tôi không cung cấp khái niệm rõ ràng về “sự tương đồng”. Điều thú vị là, mặc dù chúng tôi không nói rằng các câu là mô tả hình ảnh, một số công nhân đã bình luận rằng họ đang tự ứng cảnh để đưa ra lựa chọn. Bản chất tương đối của nhiệm vụ - “Cái nào

trong hai câu, B hay C, giống A hơn?” - giúp đánh giá khách quan hơn. Nghĩa là, dễ dàng đánh giá xem một câu có giống câu khác hơn so với một câu hay không, hơn là đưa ra xếp hạng tuyệt đối từ 1 đến 5 về mức độ giống nhau giữa hai câu [3].

Chúng tôi thu thập ba phần đoán của con người cho mỗi bộ ba. Đối với mỗi bộ ba, chúng tôi lấy phiếu bầu đa số của ba phần đoán. Đối với mỗi cặp câu ứng viên (B, C), chúng tôi gán B là câu chiến thắng nếu nó được chọn là giống nhau hơn bởi đa số các bộ ba, và tương tự đối với C. Các xếp hạng tương đối theo cặp này được sử dụng để đánh giá hiệu suất của các số liệu tự động. Nghĩa là, khi các số liệu tự động cho cả hai câu B và C một điểm, chúng tôi kiểm tra xem B có nhận được điểm cao hơn hay C. Độ chính xác được tính là tỷ lệ các cặp ứng viên mà con người và số liệu tự động đồng ý về câu nào trong hai câu là câu chiến thắng.

4. Hệ thống đo lường sự ưu tú

Mục tiêu của chúng tôi là tự động đánh giá cho hình ảnh I_i mức độ câu ứng viên có khớp với sự đồng thuận của một tập hợp các mô tả hình ảnh $S_i = \{s_{i1}, \dots, s_{im}\}$. Tất cả các từ trong câu (câu ứng viên và tham chiếu) trước tiên được ánh xạ thành dạng gốc hoặc dạng gốc của chúng. Nghĩa là, “fishes”, “fish-ing” và “fished” đều được rút gọn thành “fish”. Chúng tôi biểu diễn từng câu bằng tập hợp n -gram có trong đó. Một n -gram ω_k là tập hợp một hoặc nhiều từ được sắp xếp. Trong bài báo này, chúng tôi sử dụng n -gram chứa một đến bốn từ.

Theo trực giác, một biện pháp đồng thuận sẽ mã hóa tần suất n -gram trong câu ứng viên có mặt trong các câu tham chiếu. Tương tự như vậy, n -gram không có trong các câu tham chiếu thì không nên có trong câu ứng viên. Cuối cùng, các n -gram thường xuất hiện trên tất cả các hình ảnh trong tập dữ liệu nên được đưa ra trọng số thấp hơn, vì chúng có khả năng ít thông tin hơn. Để mã hóa trực giác này, chúng tôi thực hiện trọng số tần suất thuật ngữ nghịch đảo tần suất tài liệu (TF-IDF) cho mỗi n -gram [36]. Số lần một n -gram ω_k xuất hiện trong câu tham chiếu s_i được ký hiệu là $h_k(s_{ij})$ hoặc $h_k(c_i)$ cho câu ứng viên c_i . Chúng tôi tính toán trọng số TF-IDF $g_k(s_{ij})$ cho mỗi n -gram ω_k bằng cách sử dụng:

$$g_k(s_{ij}) = \frac{h_k(s_{ij})}{\sum_{\omega_l \in \Omega} h_l(s_{ij})} \log \frac{1}{\frac{|I|}{\sum_{I_p \in I} \min(1, \sum_q h_k(s_{pq}))}}, \quad (1)$$

trong đó Ω là từ vựng của tất cả n -gram và I là tập hợp tất cả các hình ảnh trong tập dữ liệu. Thuật ngữ đầu tiên do TF của mỗi n -gram ω_k , và thuật ngữ thứ hai do độ hiếm của ω_k bằng IDF của nó. Theo trực giác, TF đặt trọng số cao hơn vào n -gram thường xuất hiện trong câu tham chiếu mô tả một hình ảnh, trong khi IDF giảm trọng số của n -gram thường xuất hiện trên tất cả các hình ảnh trong tập dữ liệu.

Nghĩa là, IDF cung cấp thước đo mức độ nổi bật của từ bằng cách loại trừ các từ phổ biến có khả năng ít thông tin trực quan hơn. IDF được tính bằng cách sử dụng logarit của số hình ảnh trong tập dữ liệu $|I|$ chia cho số hình ảnh mà ω_k xuất hiện trong tất cả các câu tham chiếu nào của nó.

Điểm CIDEr $_n$ của chúng tôi cho n -gram có độ dài n được tính bằng cách sử dụng độ tương đồng cosin trung bình giữa câu ứng viên và câu tham chiếu, tính đến cả độ chính xác và khả năng nhớ lại:

$$\text{CIDEr}(c_i, S_i) = \frac{1}{tôi} \sum_j \frac{g_n(c_i) \cdot g_n(s_{ij})}{\|g_n(c_i)\| \|g_n(s_{ij})\|}, \quad (2)$$

trong đó $g_n(c_i)$ là một vectơ được tạo thành bởi $g_k(c_i)$ tương ứng với tất cả n -gram có độ dài n và $g_n(c_i)$ là độ lớn của vectơ $g_n(c_i)$. Tương tự đối với $g_n(s_{ij})$.

Chúng tôi sử dụng n -gram bậc cao hơn (đài hơn) để nắm bắt các thuộc tính ngữ pháp cũng như nghĩa phong phú hơn. Chúng tôi kết hợp các điểm số từ n -gram có độ dài khác nhau như sau:

$$\text{CIDEr}(c_i, S_i) = \frac{1}{N} \sum_{n=1}^N w_n \text{CIDEr}_n(c_i, S_i), \quad (3)$$

Theo kinh nghiệm, chúng tôi thấy rằng trọng số đồng đều $w_n = 1/N$ hoạt động tốt nhất. Chúng tôi sử dụng $N = 4$.

5. Bộ dữ liệu mới

Chúng tôi đã xuất hai tập dữ liệu mới - PASCAL-50S và ABSTRACT-50S - để đánh giá các phương pháp tạo chú thích hình ảnh. Cả hai tập dữ liệu đều có 50 câu tham chiếu cho mỗi hình ảnh tương ứng với 1.000 và 500 hình ảnh.

Chúng tôi coi là các tập dữ liệu “thử nghiệm”, được tạo ra để cho phép đánh giá dựa trên sự đồng thuận. Để biết danh sách các tập dữ liệu đào tạo, chúng tôi khuyến khích người đọc khám phá [25, 32]. Tập dữ liệu PASCAL-50S sử dụng tất cả 1.000 hình ảnh từ Tập dữ liệu câu Pas-cal của UIUC [35] trong khi Tập dữ liệu ABSTRACT-50S sử dụng 500 hình ảnh ngẫu nhiên từ Tập dữ liệu cảnh trữ tình [46]. Tập dữ liệu cảnh trữ tình chứa các cảnh được tạo từ các đối tượng clipart. Hai tập dữ liệu mới của chúng tôi khác nhau về mặt trực quan và loại mô tả hình ảnh được tạo ra.

Mục tiêu của chúng tôi là thu thập các mô tả hình ảnh khách quan và đại diện cho nội dung hình ảnh. Các đối tượng được xem một hình ảnh và một hộp văn bản, và được yêu cầu “Mô tả những gì đang diễn ra trong hình ảnh”. Chúng tôi yêu cầu các đối tượng

capture the main aspects of the scene and provide descriptions that others are also likely to provide. This includes writing descriptions rather than “dialogs” or overly descriptive sentences. Workers were told that a good description should help others recognize the image from a collection of similar images. Instructions also mentioned that work with poor grammar would be rejected. Snapshots of our interface can be found in the appendix. Overall, we had 465 subjects for ABSTRACT-50S and 683 subjects for PASCAL-50S datasets. We ensure that each sentence for an image is written by a different subject. The average sentence length for the ABSTRACT-50S dataset is 10.59 words compared to 8.8 words for PASCAL-50S.

6. Experimental Setup

The goals of our experiments are two-fold:

- Evaluating how well our proposed metric CIDEr captures human judgement of consensus, as compared to existing metrics.
- Comparing existing state-of-the-art automatic image description approaches in terms of how well the descriptions they produce match human consensus of image descriptions.

We first describe how we select candidate sentences for evaluation and the metrics we use for comparison to CIDEr. Finally, we list the various automatic image description approaches and our experimental set up.

Candidate Sentences: On ABSTRACT-50S, we use 48 of our 50 sentences as reference sentences (sentence A in our triplet annotation). The remaining 2 sentences per image can be used as candidate sentences. We form 400 pairs of candidate sentences (B and C in our triplet annotation). These include two kinds of pairs. The first are 200 human–human correct pairs (HC), where we pick two human sentences describing the same image. The second kind are 200 human–human incorrect pairs (HI), where one of the sentences is a human description for the image and the other is also a human sentence but describing some other image from the dataset picked at random.

For PASCAL-50S, our candidate sentences come from a diverse set of sources: human sentences from the UIUC Pascal Sentence Dataset as well as machine-generated sentences from five automatic image description methods. These span both retrieval-based and generation-based methods: Midge [29], Babytalk [22], Story [14], and two versions of Translating Video Content to Natural Language Descriptions [37] (Video and Video+).⁴ We form 4,000 pairs of candidate sentences (again, B and C for our triplet annotation). These include four types of pairs (1,000 each).

⁴We thank the authors of these approaches for making their outputs available to us.

The first two are human–human correct (HC) and human–human incorrect (HI) similar to ABSTRACT-50S. The third are human–machine (HM) pairs formed by pairing a human sentence describing an image with a machine generated sentence describing the same image. Finally, the fourth are machine–machine (MM) pairs, where we compare two machine generated sentences describing the same image. We pick the machine generated sentences randomly, so that each method participates in roughly equal number of pairs, on a diverse set of images. Ours is the first work to perform a comprehensive evaluation across these different kinds of sentences.

For consistency, we drop two reference sentences for the PASCAL-50S evaluations so that we evaluate on both datasets (ABSTRACT-50S and PASCAL-50S) with a maximum of 48 reference sentences.

Metrics: The existing metrics used in the community for evaluation of image description approaches are BLEU [33], ROUGE [45] and METEOR [1]. BLEU is precision-based and ROUGE is recall-based. More specifically, image description methods have used versions of BLEU called BLEU₁ and BLEU₄, and a version of ROUGE called ROUGE₁. A recent survey paper [12] has used a different version of ROUGE called ROUGE_S, as well as the machine translation metric called METEOR [1]. We now briefly describe these metrics. More details can be found in the appendix. **BLEU** (BiLingual Evaluation Understudy) [33] is a popular metric for Machine Translation (MT) evaluation. It computes an n -gram based precision for the candidate sentence with respect to the references. The key idea of BLEU is to compute precision by *clipping*. Clipping computes precision for a word, based on the maximum number of times it occurs in any reference sentence. Thus, a candidate sentence saying “The The The”, would get credit for saying only one “The”, if the word occurs at most once across individual references. BLEU computes the geometric mean of the n -gram precisions and adds a brevity-penalty to discourage overly short sentences. The most common formulation of BLEU is BLEU4, which uses 1-grams up to 4-grams, though lower-order variations such as BLEU1 (unigram BLEU) and BLEU2 (unigram and bigram BLEU) are also used. Similar to [12, 18] for evaluating image descriptions, we compute BLEU at the sentence level. For machine translation BLEU is most often computed at the corpus level where correlation with human judgment is high; the correlation is poor at the level of individual sentences. In this paper we are specifically interested in the evaluation of accuracies on individual sentences. **ROUGE** stands for Recall Oriented Understudy of Gisting Evaluation [45]. It computes n -gram based recall for the candidate sentence with respect to the references. It is a popular metric for summarization evaluation. Similar to BLEU, versions of ROUGE can be computed by varying the n -gram count. Two other versions

năm bắt các khía cạnh chính của cảnh và cung cấp các mô tả mà người khác cũng có thể cung cấp. Điều này bao gồm viết mô tả thay vì “đối thoại” hoặc câu mô tả quá mức. Người lao động được cho biết rằng một mô tả tốt nên giúp người khác nhận ra hình ảnh từ một bộ sưu tập hình ảnh tự ứng tự. Hư ứng dẫn cũng đề cập rằng làm việc với ngữ pháp kém sẽ bị từ chối. Ảnh chụp nhanh giao diện của chúng tôi có thể được tìm thấy trong phần phụ lục. Nhìn chung, chúng tôi có 465 chủ đề cho ABSTRACT-50S và 683 chủ đề cho các tập dữ liệu PASCAL-50S. Chúng tôi đảm bảo rằng mỗi câu cho một hình ảnh là được viết bởi một chủ đề khác. Độ dài câu trung bình đối với tập dữ liệu ABSTRACT-50S là 10,59 từ so với đến 8,8 từ cho PASCAL-50S.

6. Thiết lập thử nghiệm

Mục tiêu của thí nghiệm của chúng tôi có hai mặt:

- Đánh giá mức độ mà số liệu CIDEr được đề xuất của chúng tôi nắm bắt được đánh giá của con người về sự đồng thuận, so với số liệu hiện có.
- So sánh hình ảnh tự động hiện đại cách tiếp cận mô tả dựa trên mức độ phù hợp của các mô tả mà chúng đưa ra với sự đồng thuận của con người về mô tả hình ảnh.

Đầu tiên chúng tôi mô tả cách chúng tôi chọn câu ứng viên cho đánh giá và các số liệu chúng tôi sử dụng để so sánh với CIDEr. Cuối cùng, chúng tôi liệt kê các phương pháp mô tả hình ảnh tự động khác nhau và thiết lập thử nghiệm của chúng tôi.

Câu ứng cử viên: Trên ABSTRACT-50S, chúng tôi sử dụng 48 trong số 50 câu của chúng tôi làm câu tham chiếu (câu A trong chú thích bộ ba của chúng tôi). 2 câu còn lại cho mỗi hình ảnh có thể được sử dụng làm câu ứng viên. Chúng tôi tạo ra 400 cặp của các câu ứng cử viên (B và C trong chú thích bộ ba của chúng tôi). Bao gồm hai loại cặp. Loại đầu tiên là 200 con người-cặp đúng của con người (HC), trong đó chúng ta chọn hai câu của con người mô tả cùng một hình ảnh. Loại thứ hai là 200 cặp câu sai giữa người và người (HI), trong đó một câu là mô tả của con người về hình ảnh và câu còn lại cũng là một câu của con người như mô tả một số hình ảnh khác từ tập dữ liệu được chọn ngẫu nhiên.

Đối với PASCAL-50S, các câu ứng viên của chúng tôi đến từ một tập hợp đa dạng các nguồn: câu của con người từ UIUC Bộ dữ liệu câu Pascal cũng như các câu do máy tạo ra từ năm phương pháp mô tả hình ảnh tự động. Chúng bao gồm cả phương pháp dựa trên truy xuất và dựa trên thẻ hệ: Midge [29], Babytalk [22], Story [14] và hai phiên bản của Dịch nội dung video thành mô tả ngôn ngữ tự nhiên [37] (Video và Video+).⁴ Chúng tôi tạo thành 4.000 cặp của các câu ứng viên (một lần nữa, B và C cho chú thích bộ ba của chúng tôi). Chúng bao gồm bốn loại cặp (mỗi loại 1.000).

⁴Chúng tôi cảm ơn các tác giả của những cách tiếp cận này đã đưa ra kết quả của họ có sẵn cho chúng tôi.

Hai cái đầu tiên là con người-con người đúng (HC) và con người-con người không chính xác (HI) tương tự như ABSTRACT-50S. Thứ ba là cặp người-máy (HM) được hình thành bằng cách ghép một câu người-người mô tả một hình ảnh với một câu do máy tạo ra mô tả cùng một hình ảnh. Cuối cùng, câu thứ tư là cặp máy-máy (MM), trong đó chúng ta so sánh hai câu do máy tạo ra mô tả cùng một hình ảnh. Chúng tôi chọn các câu do máy tạo ra một cách ngẫu nhiên, để mỗi phương pháp tham gia vào số lượng cặp gần nhau, bằng nhau, trên một tập hợp hình ảnh đa dạng. Tác phẩm của chúng tôi là tác phẩm đầu tiên thực hiện một đánh giá toàn diện trên các loại khác nhau này câu.

Để thống nhất, chúng tôi bỏ hai câu tham chiếu cho đánh giá PASCAL-50S để chúng tôi đánh giá trên cả hai bộ dữ liệu (ABSTRACT-50S và PASCAL-50S) với tối đa 48 câu tham chiếu.

Số liệu: Các số liệu hiện có được sử dụng trong cộng đồng để đánh giá các phương pháp mô tả hình ảnh là BLEU [33], ROUGE [45] và METEOR [1]. BLEU dựa trên độ chính xác và ROUGE dựa trên việc nhớ lại. Cụ thể hơn, hình ảnh phương pháp mô tả đã sử dụng các phiên bản của BLEU được gọi là BLEU1 và BLEU4, và một phiên bản của ROUGE được gọi là ROUGE1. Một bài báo khảo sát gần đây [12] đã sử dụng một phiên bản của ROUGE được gọi là ROUGES, cũng như máy

phiên bản của BLEU được gọi là METEOR [1]. Bởi giờ chúng tôi mô tả tóm tắt các phép đo này. Có thể tìm thấy thêm thông tin chi tiết trong phần phụ lục. BLEU (Sinh viên đánh giá song ngữ) [33] là thư ớc do phổ biến để đánh giá Dịch máy (MT). Nó

tính toán độ chính xác dựa trên n -gram cho câu ứng viên liên quan đến các tham chiếu. Ý tưởng chính của BLEU là tính độ chính xác bằng cách cắt xén. Cắt xén tính độ chính xác cho một từ, dựa trên số lần tối đa nó xảy ra trong bất kỳ câu tham chiếu nào. Do đó, một câu ứng cử viên nói rằng “The The The”, sẽ được ghi nhận vì đã nói chỉ một “The”, nếu từ này xuất hiện nhiều nhất một lần trong các tham chiếu riêng lẻ. BLEU tính toán giá trị trung bình hình học của độ chính xác n -gram và thêm một hình phạt ngắn gọn để ngăn chặn các câu quá ngắn. Công thức phổ biến nhất của BLEU là BLEU4, sử dụng 1-gram lên đến 4-

gram, mặc dù các biến thể bậc thấp hơn như BLEU1 (BLEU unigram)

và BLEU2 (BLEU unigram và bigram) là

cũng được sử dụng. Tương tự như [12, 18] để đánh giá mô tả hình ảnh, chúng tôi tính toán BLEU ở cấp độ câu.

Đối với máy

BLEU thường được tính toán ở cấp độ ngữ liệu

nói mà mỗi từ ống quan với phân đoán của con người là cao; mỗi

từ ống quan là kém ở cấp độ của từng câu. Trong bài báo này

chúng tôi đặc biệt quan tâm đến việc đánh giá độ chính xác

trên các câu riêng lẻ. ROUGE là viết tắt của Recall Oriented

Understudy of Gisting Evaluation [45]. Nó tính toán

n -gram dựa trên việc thu hồi câu ứng viên liên quan đến

đến các tài liệu tham khảo. Đây là một thước đo phổ biến để tóm tắt

đánh giá. Tương tự như BLEU, các phiên bản của ROUGE có thể được

được tính bằng cách thay đổi số lượng n -gram. Hai phiên bản khác

of ROUGE are ROUGE_S and ROUGE_L. These compute an F-measure with a recall bias using *skip-bigrams* and *longest common subsequence* respectively, between the candidate and each reference sentence. Skip-bigrams are all pairs of ordered words in a sentence, sampled non-consecutively. Given these scores, they return the maximum score across the set of references as the judgment of quality. METEOR stands for Metric for Evaluation of Translation with Explicit ORdering [1]. Similar to ROUGE_L and ROUGE_S, it also computes the F-measure based on matches, and returns the maximum score over a set of references as its judgment of quality. However, it resolves word-level correspondences in a more sophisticated manner, using exact matches, stemming and semantic similarity. It optimizes over matches minimizing *chunkiness*. Minimizing chunkiness implies that matches should be consecutive, wherever possible. It also sets parameters favoring recall over precision in its F-measure computation. We implement all the metrics, except for METEOR, for which we use [8] (version 1.5). Similar to BLEU, we also aggregate METEOR scores at the sentence level.

Machine Approaches: We comprehensively evaluate which machine generation methods are best at matching consensus sentences. For this experiment, we select a subset of 100 images from the UIUC Pascal Sentence Dataset for which we have outputs for all the five machine description methods used in our evaluation: Midge [29], Babytalk [22], Story [14], and two versions of Translating Video Content to Natural Language Descriptions [37] (Video and Video+). For each image, we form all 5C_2 pairs of machine–machine sentences. This ensures that each machine approach gets compared to all other machine approaches on each image. This gives us 1,000 pairs. We form triplets by “tripling” each pair with 20 random reference sentences. We collect human judgement of consensus using our triplet annotation modality as well as evaluate our proposed automatic consensus metric CIDEr using the same reference sentences. In both cases, we count the fraction of times a machine description method beats another method in terms of being more similar to the reference sentences. To the best of our knowledge, we are the first work to perform an exhaustive evaluation of automated image captioning, across retrieval- and generation-based methods.

7. Results

In this section we evaluate the effectiveness of our consensus-based metric CIDEr on the PASCAL-50S and ABSTRACT-50S datasets. We begin by exploring how many sentences are sufficient for reliably evaluating our consensus metric. Next, we compare our metric against several other commonly used metrics on the task of matching human consensus. Then, using CIDEr we evaluate several existing automatic image description approaches. Finally,

we compare performance of humans and CIDEr at predicting consensus.

7.1. How many sentences are enough?

We begin by analyzing how the number of reference sentences affects the accuracy of automated metrics. To quantify this, we collect 120 sentences for a subset of 50 randomly sampled images from the UIUC Pascal Sentence Dataset. We then pool human–human correct, human–machine, machine–machine and human–human incorrect sentence pairs (179 in total) and get triplet annotations. This gives us the ground truth consensus score for all pairs. We evaluate BLEU₁, ROUGE₁ and CIDEr₁ with up to 100 reference sentences used to score the candidate sentences. We find that the accuracy improves for the first 10 sentences (Fig. 7) for all metrics. From 1 to 5 sentences, the agreement for ROUGE₁ improves from 0.63 to 0.77. Both ROUGE₁ and CIDEr₁ continue to improve until reaching 50 sentences, after which the results begin to saturate somewhat. Curiously, BLEU₁ shows a decrease in performance with more sentences. BLEU does a max operation over sentence level matches, and thus as more sentences are used, the likelihood of matching a lower quality reference sentence increases. Based on this pilot, we collect 50 sentences per image for our ABSTRACT-50S and PASCAL-50S datasets. For the remaining experiments we report results using 1 to 50 sentences.

7.2. Accuracy of Automated Metrics

We evaluate the performance of CIDEr, BLEU, ROUGE and METEOR at matching the human consensus scores in Fig. 11. That is, for each metric we compute the scores for two candidate sentences. The metric is correct if the sentence with higher score is the same as the sentence chosen by our human studies as being more similar to the reference sentences. The candidate sentences are both human and machine generated. For BLEU and ROUGE we show both their popular versions and the version we found to give best performance. We sample METEOR at fewer points due to high run-time. For a more comprehensive evaluation across different versions of each metric, please see the appendix.

At 48 sentences, we find that CIDEr is the best performing metric, on both ABSTRACT-50S as well as PASCAL-50S. It is followed by METEOR on each dataset. Even using only 5 sentences, both CIDEr and METEOR perform well in comparison to BLEU and ROUGE. CIDEr beats METEOR at 5 sentences on ABSTRACT-50S, whereas METEOR does better at five sentences on PASCAL-50S. This is because METEOR incorporates soft-similarity, which helps when using fewer sentences. However, METEOR, despite its sophistication does a max across reference scores, which limits its ability to utilize larger numbers of reference sentences. Popular metrics like

của ROUGE là ROUGES và ROUGEL. Những thứ này tính toán một độ lưỡng F với độ lệch thu hồi bằng cách sử dụng bỏ qua các bigram và dài nhất dãy con chung tự ơng ứng, giữa ứng viên và mỗi câu tham chiếu. Skip-bigrams là tất cả các cặp các từ được sắp xếp theo thứ tự trong một câu, được lấy mẫu không liên tiếp. Với những điểm số này, họ trả về điểm số tối đa trên bộ tài liệu tham khảo như là sự phán đoán về chất lư ơng. METEOR viết tắt của Metric for Evaluation of Translation with Explicit ORdering [1]. Tự ơng tự như ROUGEL và ROUGES, nó cũng tính toán F-measure dựa trên các kết quả khớp và trả về điểm số tối đa trên một tập hợp các tham chiếu như của nó phán đoán về chất lư ơng. Tuy nhiên, nó giải quyết các phản hồi ở cấp độ từ theo cách tinh vi hơn, sử dụng chính xác khớp, bắt nguồn và sự tự ơng đồng về mặt ngữ nghĩa. Nó tối ưu hóa trên các trận đấu giảm thiểu độ thô. Giảm thiểu độ thô ngụ ý rằng các trận đấu phải liên tiếp, bắt cứ nơi i nào có thể. Nó cũng thiết lập các tham số ưu tiên thu hồi hơn độ chính xác trong phép tính F-measure của nó. Chúng tôi triển khai tất cả số liệu, ngoại trừ METEOR, mà chúng tôi sử dụng [8] (phiên bản 1.5). Tự ơng tự như BLEU, chúng tôi cũng tổng hợp điểm METEOR ở cấp độ câu.

Phương pháp tiếp cận của máy móc: Chúng tôi đánh giá toàn diện phương pháp tạo máy nào là tốt nhất để phù hợp câu đồng thuận. Đối với thí nghiệm này, chúng tôi chọn một tập hợp con của 100 hình ảnh từ Bộ dữ liệu câu Pascal của UIUC cho mà chúng tôi có đầu ra cho tất cả năm mô tả máy phương pháp được sử dụng trong đánh giá của chúng tôi: Midge [29], Babytalk [22], Câu chuyện [14] và hai phiên bản của Nội dung Video dịch đến Mô tả Ngôn ngữ Tự nhiên [37] (Video và Video+). Đối với mỗi hình ảnh, chúng tôi tạo thành tất cả 5C_2 cặp máy-máy câu. Điều này đảm bảo rằng mỗi cách tiếp cận máy móc nhận đủ góc so với tất cả các phương pháp tiếp cận máy khác trên mỗi hình ảnh. Điều này cho chúng ta 1.000 cặp. Chúng ta tạo thành bộ ba bằng cách “gắn ba” mỗi cặp với 20 câu tham khảo ngẫu nhiên. Chúng tôi thu thập phán đoán của con người về sự đồng thuận bằng cách sử dụng chú thích bộ ba của chúng tôi phương thức cũng như đánh giá thư ớc do sự đồng thuận tự động CIDEr mà chúng tôi đề xuất bằng cách sử dụng cùng các câu tham chiếu. Trong cả hai trường hợp, chúng tôi đếm phần thời gian mà phương pháp mô tả máy đánh bại phương pháp khác về mặt tự ơng tự hơn với các câu tham khảo. Theo ý kiến tốt nhất của chúng tôi kiến thức, chúng tôi là công trình đầu tiên thực hiện một cách toàn diện đánh giá việc thêm chủ thích cho hình ảnh tự động, thông qua các phương pháp dựa trên truy xuất và tạo.

7. Kết quả

Trong phần này chúng tôi đánh giá hiệu quả của chúng tôi số liệu dựa trên sự đồng thuận CIDEr trên PASCAL-50S và TÓM TẮT-50S bộ dữ liệu. Chúng tôi bắt đầu bằng cách khám phá cách nhiều câu là đủ để đánh giá đáng tin cậy của chúng tôi thư ớc do đồng thuận. Tiếp theo, chúng tôi so sánh thư ớc do của mình với một số thư ớc do thư ờng dùng khác trong nhiệm vụ khớp sự đồng thuận của con người. Sau đó, sử dụng CIDEr, chúng tôi đánh giá một số phương pháp mô tả hình ảnh tự động hiện có. Cuối cùng,

chúng tôi so sánh hiệu suất của con người và CIDEr trong việc dự đoán sự đồng thuận.

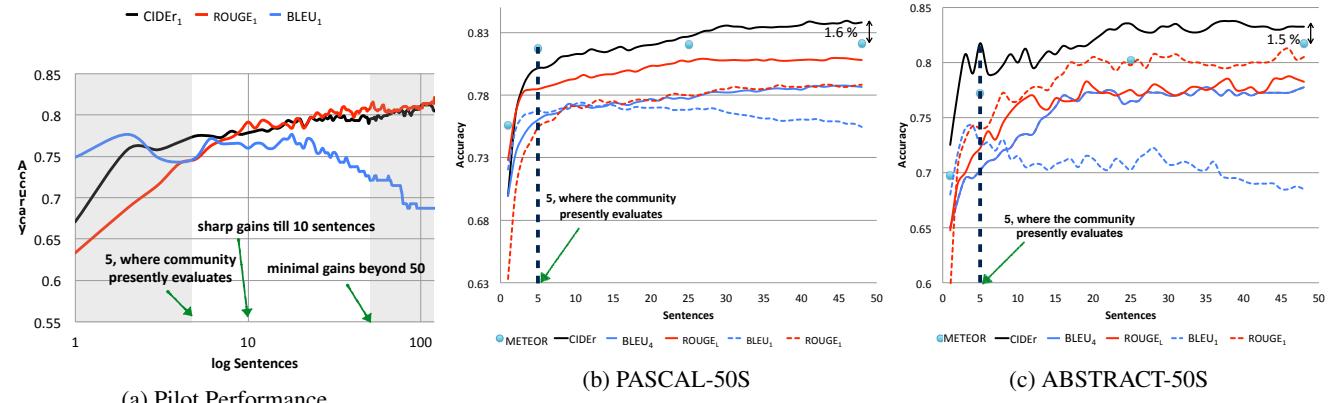
7.1. Bao nhiêu câu là đủ?

Chúng tôi bắt đầu bằng cách phân tích số lượng tham chiếu các câu ảnh hưởng đến độ chính xác của số liệu tự động. Để định lượng điều này, chúng tôi thu thập 120 câu cho một tập hợp con gồm 50 hình ảnh được lấy mẫu ngẫu nhiên từ Câu Pascal của UIUC Bộ dữ liệu. Sau đó chúng tôi tập hợp con người-con người chính xác, con người-máy, máy-máy và con người-con người không đúng cặp câu (tổng cộng 179) và nhận xét thích bộ ba. Điều này cung cấp cho chúng ta điểm số đồng thuận thực tế cho tất cả các cặp. Chúng tôi đánh giá BLEU1, ROUGE1 và CIDEr1 với tối đa 100 câu tham khảo được sử dụng để chấm điểm các câu ứng viên. Chúng tôi thấy xứng đáng độ chính xác được cải thiện cho 10 câu đầu tiên (Hình 7) cho tất cả các số liệu. Từ 1 đến 5 câu, sự đồng thuận cho ROUGE1 cải thiện từ 0,63 lên 0,77. Cả hai ROUGE1 và CIDEr1 tiếp tục cải thiện cho đến khi đạt được 50 câu, sau đó kết quả bắt đầu bão hòa phần nào. Thật kỳ lạ, BLEU1 cho thấy hiệu suất giảm với nhiều câu hơn. BLEU thực hiện thao tác tối đa trên các câu khớp cấp độ, và do đó khi sử dụng nhiều câu hơn, khả năng khớp với câu tham chiếu chất lư ơng thấp hơn tăng lên. Dựa trên thí điểm này, chúng tôi thu thập 50 câu cho mỗi hình ảnh cho các tập dữ liệu ABSTRACT-50S và PASCAL-50S của chúng tôi. Đối với các thí nghiệm còn lại, chúng tôi báo cáo kết quả bằng cách sử dụng 1 đến 50 câu.

7.2. Độ chính xác của số liệu tự động

Chúng tôi đánh giá hiệu suất của CIDEr, BLEU, ROUGE và METEOR khi khớp với điểm số đồng thuận của con người trong Hình 11. Nghĩa là, đối với mỗi số liệu, chúng tôi tính toán điểm số cho hai câu ứng cử viên. Thư ớc do là đúng nếu câu có điểm cao hơn giống với câu đã chọn bởi các nghiên cứu của chúng tôi trên con người giống với tài liệu tham khảo hơn câu. Các câu ứng viên đều do con người và máy tạo ra. Đối với BLEU và ROUGE, chúng tôi hiển thị cả hai phiên bản phổ biến của họ và phiên bản chúng tôi thấy mang lại hiệu quả tốt nhất hiệu suất. Chúng tôi lấy mẫu METEOR tại ít điểm hơn do thời gian chạy cao. Để đánh giá toàn diện hơn trên Các phiên bản khác nhau của từng số liệu, vui lòng xem phần phụ lục.

Ở 48 câu, chúng tôi thấy rằng CIDEr là số liệu có hiệu suất tốt nhất, trên cả ABSTRACT-50S cũng như PASCAL-50S. Tiếp theo là METEOR trên mỗi tập dữ liệu. Ngay cả khi chỉ sử dụng 5 câu, cả CIDEr và METEOR đều có hiệu suất tốt khi so sánh với BLEU và ROUGE. CIDEr đánh bại METEOR ở 5 câu trên ABSTRACT-50S, trong khi METEOR làm tốt hơn ở năm câu trên PASCAL-50S. Điều này là do METEOR kết hợp tính tự ơng tự mèm, điều này giúp ích khi sử dụng ít câu hơn. Tuy nhiên, METEOR, mặc dù tinh vi nhưng vẫn đạt được điểm tối đa trên các điểm tham chiếu, điều này hạn chế khả năng sử dụng các câu lớn hơn số lượng câu tham chiếu. Các số liệu phổ biến như



(a) Pilot Performance

Figure 3: (a): We show accuracy (y-axis) versus *log* number of sentences (x-axis) for our pilot study. We note that the gains saturate after 50 sentences. (b) and (c): Accuracy of automated metrics (y-axis) plotted against number of reference sentences (x-axis) for PASCAL-50S (b) and ABSTRACT-50S (c). Metrics currently used for evaluating image descriptions are shown in *dashed* lines. Other existing metrics and our proposed metric are in *solid* lines. CIDEr is the best performing metric on both datasets followed by METEOR. METEOR is sampled at fewer points, due to high run-time. Note that more reference sentences that we collect clearly help.

ROUGE₁ and BLEU₁ are not as good at capturing consensus. CIDEr provides consistent performance across both the datasets, giving 84% and 84% accuracy on PASCAL-50S and ABSTRACT-50S respectively.

Considering previous papers only used 5 reference sentences per image for evaluation, the relative boost in performance is substantial. Using BLEU₁ or ROUGE₁ at 5 sentences, we obtained 76% and 74% accuracy on PASCAL-50S. With CIDEr at 48 sentences, we achieve 84% accuracy. This brings automated evaluation much closer to human performance (90%, details in Sec. 7.4). On the Flickr8K dataset [18] with human judgments on 1-5 ratings, METEOR has a correlation (Spearman's ρ) of 0.56 [12], whereas CIDEr achieves a correlation of 0.58 with human judgments.⁵

We next show the best performing versions of the metrics CIDEr, BLEU, ROUGE and METEOR on PASCAL-50S and ABSTRACT-50S, respectively, for different kinds of candidate pairs (Table 1). As discussed in Sec. 5 we have four kinds of pairs: (human–human correct) HC, (human–human incorrect) HI, (human–machine) HM, and (machine–machine) MM. We find that out of six cases, our proposed automated metric is best in five. We show significant gains on the challenging MM and HC tasks that involve differentiating between fine-grained differences between sentences (two machine generated sentences and two human generated sentences). This result is encouraging because it indicates that the CIDEr metric will continue to perform well as image description methods continue to improve. On the easier tasks of judging consensus on HI and HM pairs, all methods perform well.

⁵We thank Desmond Elliot for the result.

Metric	PASCAL-50S			ABSTRACT-50S		
	HC	HI	HM	MM	HC	HI
BLEU ₄	64.8	97.7	93.8	63.6	65.5	93.0
ROUGE	66.3	98.5	95.8	64.4	71.5	91.0
METEOR	65.2	99.3	96.4	67.7	69.5	94.0
CIDEr	71.8	99.7	92.1	72.2	71.5	96.0

Table 1: Results on four kinds of pairs for PASCAL-50S and two kinds of pairs for ABSTRACT-50S. The best performing method is shown in **bold**. Note: we use ROUGE_L for PASCAL-50S and ROUGE₁ for ABSTRACT-50S

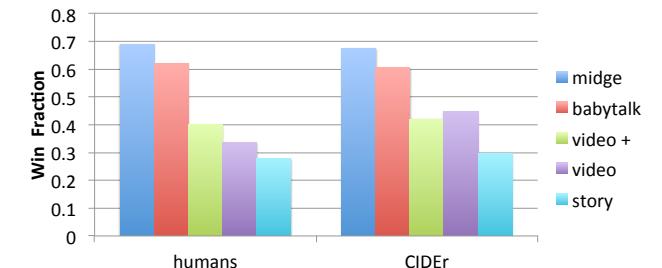
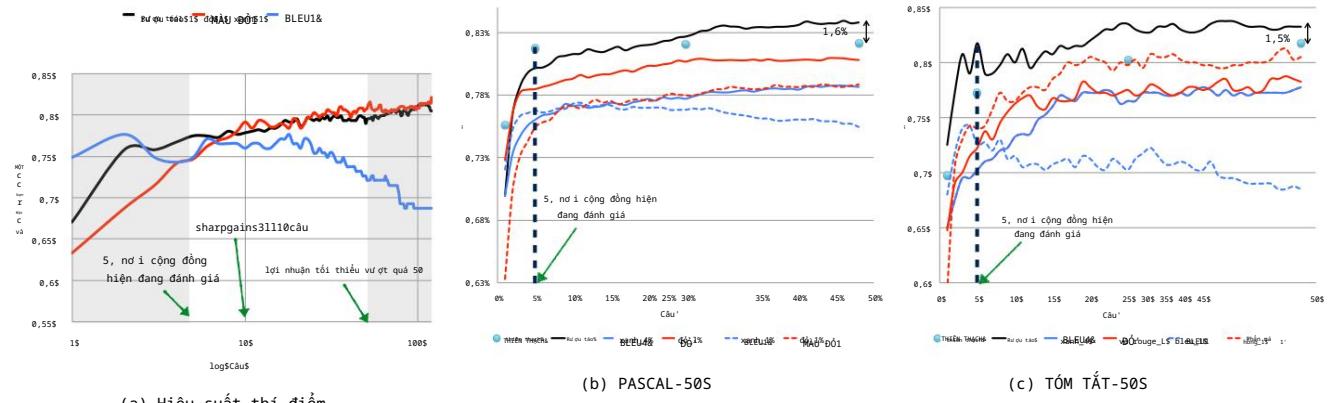


Figure 4: Fraction of times a machine generation approach wins against the other four (y-axis), plotted for human annotations and our automated metric, CIDEr.

7.3. Which automatic image description approaches produce consensus descriptions?

We have shown that CIDEr and our new datasets containing 50 sentences per image provide a more accurate metric over previous approaches. We now use it to evaluate some existing automatic image description approaches. Our methodology for conducting this experiment is described in Sec. 6. Our results are shown in Fig. 12. We show the fraction of times an approach is rated better than other ap-



(a) Hiệu suất thí điểm

Hình 3: (a): Chúng tôi hiển thị độ chính xác (trục y) so với số lư ơng câu logarit (trục x) cho nghiên cứu thí điểm của chúng tôi. Chúng tôi lưu ý rằng đạt đư ợc bao hòa sau 50 câu. (b) và (c): Độ chính xác của số liệu tự động (trục y) đư ợc biểu diễn theo số tham chiếu câu (trục x) cho PASCAL-50S (b) và ABSTRACT-50S (c). Các số liệu hiện đang đư ợc sử dụng để đánh giá mô tả hình ảnh đư ợc hiển thị bằng đường nét đứt. Các số liệu hiện có khác và số liệu đê xuất của chúng tôi đư ợc hiển thị bằng đường nét liền. CIDEr là hiệu suất tốt nhất số liệu trên cả hai tập dữ liệu theo sau là METEOR. METEOR đư ợc lấy mẫu tại ít điểm hơn, do thời gian chạy cao. Lưu ý rằng nhiều hơn những câu tham khảo mà chúng tôi thu thập đư ợc sđ giúp ích rõ ràng.

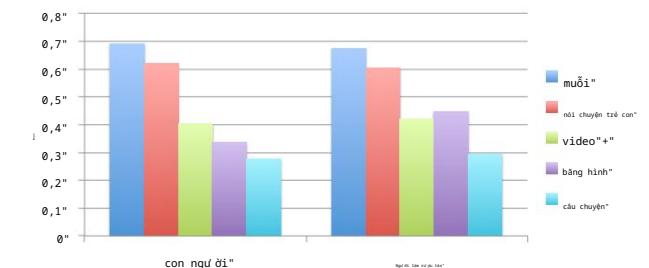
ROUGE1 và BLEU1 không giỏi trong việc nắm bắt sự đồng thuận. CIDEr cung cấp hiệu suất nhât quán trên cả hai bộ dữ liệu, cho độ chính xác 84% và 84% trên PASCAL-50S và ABSTRACT-50S tư ơng ứng.

Xem xét các bài báo trước đây chỉ sử dụng 5 câu tham chiếu cho mỗi hình ảnh để đánh giá, sự gia tăng tư ơng đối về hiệu suất là đáng kể. Sử dụng BLEU1 hoặc ROUGE1 ở 5 câu, chúng tôi đạt đư ợc độ chính xác 76% và 74% trên PASCAL-50S. Với CIDEr ở 48 câu, chúng tôi đạt đư ợc độ chính xác 84%. Điều này đư a đánh giá tự động đến gần hơ n nhiều với hiệu suất của con người (90%, chi tiết trong Mục 7.4). Trên Bộ dữ liệu Flickr8K [18] với các đánh giá của con người về xếp hạng 1-5, METEOR có hệ số tư ơng quan (Spearman's ρ) là 0,56 [12], trong khi CIDEr đạt đư ợc mối tư ơng quan là 0,58 với con người phán đoán.

Tiếp theo, chúng tôi sđ trình bày các phiên bản có hiệu suất tốt nhất của các hệ mét CIDEr, BLEU, ROUGE và METEOR trên PASCAL-50S và ABSTRACT-50S, tư ơng ứng, cho các loại khac nhau của các cặp ứng viên (Bảng 1). Như đã thảo luận trong Mục 5, chúng tôi có bốn loại cặp: (con người–con người đúng) HC, (con người–con người không đúng) HI, (con người–máy móc) HM, và (máy–máy) MM. Chúng tôi thấy rằng trong sáu trứ ơng hợp, đê xuất số liệu tự động là tốt nhât trong năm. Chúng tôi cho thấy những tiến bộ đáng kể trong các nhiệm vụ MM và HC đầy thách thức bao gồm việc phân biệt giữa những khac biệt chi tiết giữa các câu (hai câu do máy tạo ra và hai câu (hai câu do con người tạo ra). Kết quả này rất đáng khich lè vì nó chỉ ra rằng số liệu CIDEr sđ tiếp tục thực hiện tốt khi các phương pháp mô tả hình ảnh tiếp tục đư ợc cải thiện. Về các nhiệm vụ đê dàng hơ n là đánh giá sự đồng thuận vè HI và Cập HM, tất cả các phương pháp đều thực hiện tốt.

Hệ mét	PASCAL-50S			TÔM TẮT-50S		
	HC	HI	HM	MM	HC	HI
BLEU4	64,8	97,7	93,8	63,6	65,5	93,0
Đỗ	66,3	98,5	95,8	64,4	71,5	91,0
SAO TINH	65,2	99,3	96,4	67,7	69,5	94,0
Rú ợu táo	71,8	99,7	92,1	72,2	71,5	96,0

Bảng 1: Kết quả trên bốn loại cặp cho PASCAL-50S và hai loại cặp cho ABSTRACT-50S. Phu ơng pháp thực hiện tốt nhất đư ợc hiển thị bằng chữ in đậm. Lưu ý: chúng tôi sử dụng ROUGEL cho PASCAL-50S và ROUGE1 cho ABSTRACT-50S



Hình 4: Phân số thời gian tiếp cận thê hệ máy chiến thắng bốn chiến thắng còn lại (trục y), đư ợc vẽ biểu đồ để chú thích của con người và số liệu tự động của chúng tôi, CIDEr.

7.3. Phu ơng pháp mô tả hình ảnh tự động nào tạo ra mô tả đồng thuận?

Chúng tôi đã chứng minh rằng CIDEr và các tập dữ liệu mới của chúng tôi chứa 50 câu trên mỗi hình ảnh cung cấp số liệu chính xác hơ n so với các phu ơng pháp tiếp cận trứ ơng. Bởi giờ chúng tôi sử dụng nó để đánh giá một số phu ơng pháp mô tả hình ảnh tự động hiện có. Của chúng tôi phu ơng pháp thực hiện thí nghiệm này đư ợc mô tả trong Mục 6. Kết quả của chúng tôi đư ợc thể hiện trong Hình 12. Chúng tôi trình bày một phần nhỏ thời gian một cách tiếp cận đư ợc đánh giá hơ n các ứng dụng khac

5Chúng tôi cảm ơn Desmond Elliot vì kết quả này.

proaches on the y-axis. We note that Midge [29] is rated as having the best consensus by both humans and CIDEr, followed by Babytalk [22]. Story [14] is the lowest ranked, by both humans and CIDEr. Humans and CIDEr differ on the ranking of the two video approaches (Video and Video+) [37]. We calculate the Pearson’s correlation between the fraction of wins for a method on human annotations and using CIDEr. We find that humans and CIDEr agree with a high correlation (0.98).

7.4. Human Performance

In our final set of experiments we measure human performance at predicting which of two candidate sentences better matches the consensus. Human performance puts into context how clearly consensus is defined, and provides a loose bound on how well we can expect automated metrics to perform. We evaluate both human and machine performance at predicting consensus on all 4,000 pairs from PASCAL-50S dataset and 400 pairs from the ABSTRACT-50S dataset described in Sec. 6. To create the same experimental set up for both humans and machines, we obtain ground truth consensus for each of the pairs using our triplet annotation on 24 references out of 48. For predicting consensus, humans (via triplet annotations) and machines both use the remaining 24 sentences as reference sentences. We find that the best machine performance is 82% on PASCAL-50S using CIDEr, in contrast to human performance which is at 90%. On the ABSTRACT-50S dataset, CIDEr is at 82% accuracy, whereas human performance is at 83%.

8. Gameability and Evaluation Server

Gameability When optimizing an algorithm for a specific metric undesirable results may be achieved. The “gaming” of a metric may result in sentences with high scores, yet produce poor results when judged by a human. To help defend against the future gaming of the CIDEr metric, we propose several modifications to the basic CIDEr metric called CIDEr-D.

First, we propose the removal of stemming. When performing stemming the singular and plural forms of nouns and different tenses of verbs are mapped to the same token. The removal of stemming ensures the correct forms of words are used. Second, in some cases the basic CIDEr metric produces higher scores when words of higher confidence are repeated over long sentences. To reduce this effect, we introduce a Gaussian penalty based on the difference between candidate and reference sentence lengths. Finally, the sentence length penalty may be gamed by repeating confident words or phrases until the desired sentence length is achieved. We combat this by adding clipping to the n -gram counts in the CIDEr_n numerator. That is, for a specific n -gram we clip the number of candidate occurrences to the number of reference occurrences. This penalizes the

repetition of specific n -grams beyond the number of times they occur in the reference sentence. These changes result in the following equation (analogous to Equation 2):

$$\text{CIDEr-D}_n(c_i, S_i) = \frac{10}{m} \sum_j e^{\frac{-(l(c_i) - l(s_{ij}))^2}{2\sigma^2}} * \frac{\min(g^n(c_i), g^n(s_{ij})) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|}, \quad (4)$$

Where $l(c_i)$ and $l(s_{ij})$ denote the lengths of candidate and reference sentences respectively. We use $\sigma = 6$. A factor of 10 is added to make the CIDEr-D scores numerically similar to other metrics.

The final CIDEr-D metric is computed in a similar manner to CIDEr (analogous to Equation 3):

$$\text{CIDEr-D}(c_i, S_i) = \sum_{n=1}^N w_n \text{CIDEr-D}_n(c_i, S_i), \quad (5)$$

Similar to CIDEr, uniform weights are used. We found that this version of the metric has a rank correlation (Spearman’s ρ) of 0.94 with the original CIDEr metric while being more robust to gaming. Qualitative examples of ranking can be found in the appendix.

Evaluation Server To enable systematic evaluation and benchmarking of image description approaches based on consensus, we have made CIDEr-D available as a metric in the MS COCO caption evaluation server [5].

9. Conclusion

In this work we proposed a consensus-based evaluation protocol for image description evaluation. Our protocol enables an objective comparison of machine generation approaches based on their “human-likeness”, without having to make arbitrary calls on weighing content, grammar, saliency, etc. with respect to each other. We introduce an annotation modality for measuring consensus, a metric CIDEr for automatically computing consensus, and two datasets, PASCAL-50S and ABSTRACT-50S with 50 sentences per image. We demonstrate CIDEr has improved accuracy over existing metrics for measuring consensus.

Acknowledgements: We thank Chris Quirk, Margaret Mitchell and Michel Galley for helpful discussions in formulating CIDEr-D. This work was supported in part by The Paul G. Allen Family Foundation Allen Distinguished Investigator award to D.P.

tíền gần trên trục y. Chúng tôi lưu ý rằng Midge [29] được đánh giá là có sự đồng thuận tốt nhất của cả con người và CIDEr, tiếp theo là Babytalk [22]. Story [14] được xếp hạng thấp nhất, bởi cả con người và CIDEr. Con người và CIDEr khác nhau về thứ hạng của hai phương pháp tiếp cận video (Video và Video+) [37]. Chúng tôi tính toán mối tương quan của Pearson giữa tỷ lệ chiến thắng cho một phương pháp về chủ thích của con người và sử dụng CIDEr. Chúng tôi thấy rằng con người và CIDEr đồng ý với mối tương quan cao (0.98).

7.4. Hiệu suất của con người

Trong tập hợp các thí nghiệm cuối cùng của chúng tôi, chúng tôi do lường hiệu suất của con người trong việc dự đoán câu nào trong hai câu ứng cử viên tốt hơn phù hợp với sự đồng thuận. Hiệu suất của con người đặt vào bối cảnh sự đồng thuận được định nghĩa rõ ràng như thế nào và cung cấp một bí quyết buộc vào mức độ chúng ta có thể mong đợi các số liệu tự động thực hiện tốt như thế nào. Chúng tôi đánh giá cả hiệu suất của con người và máy móc tại dự đoán sự đồng thuận trên tất cả 4.000 cặp từ PASCAL-50S tập dữ liệu và 400 cặp từ tập dữ liệu ABSTRACT-50S được mô tả trong Phần 6. Để tạo ra một thiết lập thử nghiệm đối với cả con người và máy móc, chúng tôi thu được sự đồng thuận thực tế cho từng cặp bằng cách sử dụng chủ thích bộ ba của chúng tôi trên 24 tài liệu tham khảo trong số 48. Để dự đoán sự đồng thuận, con người (thông qua chủ thích bộ ba) và máy móc đều sử dụng 24 câu còn lại làm câu tham chiếu. Chúng tôi thấy rằng hiệu suất máy tốt nhất là 82% trên PASCAL-50S khi sử dụng CIDEr, trái ngược với hiệu suất của con người chỉ đạt 90%. Trên tập dữ liệu ABSTRACT-50S, CIDEr có độ chính xác là 82%, trong khi hiệu suất của con người chỉ đạt 83%.

8. Máy chủ đánh giá và khả năng chơi game

Khả năng chơi game Khi tôi ưu hóa một thuật toán cho một mục đích cụ thể kết quả không mong muốn về mặt số liệu có thể đạt được. “Trò chơi” của một số liệu có thể dẫn đến các câu có điểm cao, tuy nhiên tạo ra kết quả kém khi được đánh giá bởi con người. Để giúp bảo vệ chống lại việc chơi trò chơi trong tương lai của số liệu CIDEr, chúng tôi đã xuất một số sửa đổi đối với số liệu CIDEr cơ bản được gọi là Rugged-D.

Đầu tiên, chúng tôi đã xuất loại bỏ gốc. Khi thực hiện gốc các dạng số ít và số nhiều của danh từ và các thi khác nhau của động từ được ánh xạ tới cùng một token. Việc loại bỏ gốc từ đảm bảo các hình thức chính xác của các từ được sử dụng. Thứ hai, trong một số trường hợp, CIDEr cơ bản phép đo tạo ra điểm số cao hơn khi các từ có độ tin cậy cao hơn được lặp lại trong các câu dài. Để giảm thiểu điều này hiệu ứng, chúng tôi giới thiệu một hình phạt Gaussian dựa trên sự khác biệt giữa độ dài câu ứng viên và câu tham chiếu. Cuối cùng, hình phạt độ dài câu có thể được chơi bằng cách lặp lại các từ hoặc cụm từ chắc chắn cho đến khi câu mong muốn dài đạt được. Chúng tôi chống lại điều này bằng cách thêm cắt vào n-gram đếm trong tử số CIDEr. Nghĩa là, đối với một n-gram cụ thể, chúng tôi cắt số lần xuất hiện của ứng viên với số lần xuất hiện tham chiếu. Điều này phạt

sự lặp lại của n-gram cụ thể vượt quá số lần chúng xảy ra trong câu tham chiếu. Những thay đổi này dẫn đến trong chương trình sau (tương tự như Chương trình 2):

$$\text{CIDEr-D}_n(c_i, S_i) = \frac{10}{m} \sum_j e^{\frac{-(l(c_i) - l(s_{ij}))^2}{2\sigma^2}} * \frac{\min(g^n(c_i), g^n(s_{ij})) \cdot g^n(s_{ij})}{\|g^n(c_i)\| \|g^n(s_{ij})\|}, \quad (4)$$

Trong đó $l(c_i)$ và $l(s_{ij})$ biểu thị độ dài của ứng viên và các câu tham khảo tương ứng. Chúng tôi sử dụng $\sigma = 6$. Một hệ số 10 được thêm vào để làm cho điểm số CIDEr-D về mặt số tương tự như các số liệu khác.

Chỉ số CIDEr-D cuối cùng được tính theo cách tương tự như CIDEr (tương tự như Chương trình 3):

$$\text{CIDEr-D}(c_i, S_i) = \sum_{n=1}^N w_n \text{CIDEr-D}_n(c_i, S_i), \quad (5)$$

Tương tự như CIDEr, trọng lượng đồng đều được sử dụng. Chúng tôi thấy rằng phiên bản này của số liệu có tương quan thứ hạng (Spearman’s ρ) là 0.94 với số liệu CIDEr ban đầu trong khi nhiều hơn mạnh mẽ để chơi game. Các ví dụ định tính về xếp hạng có thể là có trong phần phụ lục.

Máy chủ đánh giá Để cho phép đánh giá có hệ thống và chuẩn mực của các phương pháp mô tả hình ảnh dựa trên sự đồng thuận, chúng tôi đã đưa CIDEr-D vào sử dụng như một thư viện trong máy chủ đánh giá chủ đề MS COCO [5].

9. Kết luận

Trong công trình này chúng tôi đã xuất một đánh giá dựa trên sự đồng thuận giao thức để đánh giá mô tả hình ảnh. Giao thức của chúng tôi cho phép so sánh khách quan các phương pháp tiếp cận tạo máy dựa trên “tính giống con người” của chúng, mà không cần phải đưa ra các quyết định tùy ý về nội dung cảm nhận, ngữ pháp, sự nổi bật, v.v. liên quan đến nhau. Chúng tôi giới thiệu một phương thức chủ đề để lường sự đồng thuận, một CIDEr metric-ric để tự động tính toán sự đồng thuận và hai bộ dữ liệu, PASCAL-50S và ABSTRACT-50S với 50 câu cho mỗi hình ảnh. Chúng tôi chứng minh CIDEr đã cải thiện độ chính xác so với các số liệu hiện có để do lường sự đồng thuận.

Lời cảm ơn: Chúng tôi xin cảm ơn Chris Quirk, Margaret Mitchell và Michel Galley đã thảo luận hữu ích trong việc xây dựng công thức CIDEr-D. Công trình này được hỗ trợ một phần bởi The Quỹ Gia đình Paul G. Allen Giải thưởng Nhà điều tra xuất sắc Allen cho DP

References

- [1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. pages 65–72, 2005. 2, 5, 6
- [2] A. C. Berg, T. L. Berg, H. D. III, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, and K. Yamaguchi. Understanding and predicting importance in images. In *CVPR*. IEEE, 2012. 2
- [3] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev*, 113(4):700–765, Oct. 2006. 1, 4
- [4] C. Callison-burch and M. Osborne. Re-evaluating the role of bleu in machine translation research. In *In EACL*, pages 249–256, 2006. 1, 2
- [5] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *ArXiv e-prints*, Apr. 2015. 2, 8
- [6] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. *CoRR*, abs/1411.5654, 2014. 2, 3
- [7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1
- [8] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, 2014. 6
- [9] P. K. Dokania, A. Behl, C. V. Jawahar, and P. M. Kumar. Learning to rank using high-order information. *ECCV*, 2014. 1
- [10] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389, 2014. 2, 3
- [11] D. Elliott and F. Keller. Image description using visual dependency representations. In *EMNLP*, pages 1292–1302. ACL, 2013. 1, 2, 3
- [12] D. Elliott and F. Keller. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland, June 2014. Association for Computational Linguistics. 3, 5, 7
- [13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. <http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html>. 1
- [14] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, 2010. 2, 5, 6, 7, 12, 16
- [15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 1
- [16] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In D. A. Forsyth, P. H. S. Torr, and A. Zisserman, editors, *ECCV (1)*, volume 5302 of *Lecture Notes in Computer Science*, pages 16–29. Springer, 2008. 2
- [17] A. Gupta, Y. Verma, and C. Jawahar. Choosing linguistics over vision to describe images. 2012. 2
- [18] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res. (JAIR)*, 47:853–899, 2013. 1, 2, 3, 5, 7
- [19] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306, 2014. 2, 3
- [20] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *CoRR*, 2014. 2
- [21] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014. 2, 3
- [22] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*, 2011. 1, 2, 5, 6, 7, 11, 12, 16
- [23] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by betweenclass attribute transfer. In *In CVPR*, 2009. 1, 2
- [24] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 220–228, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. 2
- [25] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3, 4
- [26] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 1
- [27] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *CoRR*, abs/1410.1090, 2014. 2, 3
- [28] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. 8th Int'l Conf. Computer Vision*, volume 2, pages 416–423, July 2001. 1
- [29] M. Mitchell, X. Han, and J. Hayes. Midge: Generating descriptions of images. In *Proceedings of the Seventh International Natural Language Generation Conference*, INLG '12, pages 131–133, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. 1, 2, 5, 6, 7, 12, 16
- Tài liệu tham khảo
- [1] S. Banerjee và A. Lavie. Thiên thạch: Một phép đo tự động cho đánh giá mt với mối tương quan dựa trên các phân tích của con người. trang 65–72, 2005. 2, 5, 6
- [2] AC Berg, TL Berg, HD III, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos và K. Yamaguchi. Hiểu và dự đoán tầm quan trọng trong hình ảnh. Trong *CVPR*. IEEE, 2012. 2
- [3] R. Bogacz, E. Brown, J. Moehlis, P. Holmes và JD Cohen. Vật lý của quá trình ra quyết định tối ưu: một hình thức phân tích các mô hình hiệu suất trong các nhiệm vụ lựa chọn bắt buộc thay thế hai. *Psychol Rev*, 113(4):700–765, tháng 10 năm 2006. 1, 4
- [4] C. Callison-burch và M. Osborne. Đánh giá lại vai trò của bleu trong nghiên cứu dịch máy. Trong *EACL*, các trang 249–256, 2006. 1, 2
- [5] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar và CL Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *ArXiv e-prints*, tháng 4 năm 2015. 2, 8
- [6] X. Chen và CL Zitnick. Học cách biểu diễn trực quan toàn hoàn để tạo chú thích hình ảnh. *CoRR*, abs/1411.5654, 2014. 2, 3
- [7] J. Đặng, W. Dong, R. Socher, L.-J. Li, K. Li và L. Fei-Fei. ImageNet: Cơ sở dữ liệu hình ảnh phân cấp quy mô lớn. Trong *CVPR09*, 2009. 1
- [8] M. Denkowski và A. Lavie. Thiên thạch phổ quát: Ngôn ngữ đánh giá bản dịch cụ thể cho bất kỳ ngôn ngữ đích nào. Trong *Biên bản Hội thảo EACL 2014 về Biên dịch máy thông kê*, 2014. 6
- [9] PK Dokania, A. Behl, CV Jawahar và Thủ tư ống Kumar. Học cách xếp hạng bằng cách sử dụng thông tin bậc cao. *ECCV*, 2014. 1
- [10] J. Donahue, LA Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, và T. Darrell. Mạng tích chập toàn hoàn dài hạn để nhận dạng và mô tả trực quan. *CoRR*, abs/1411.4389, 2014. 2, 3
- [11] D. Elliott và F. Keller. Mô tả hình ảnh sử dụng biểu diễn phụ thuộc trực quan. Trong *EMNLP*, trang 1292–1302. ACL, 2013. 1, 2, 3
- [12] D. Elliott và F. Keller. So sánh đánh giá tự động biện pháp mô tả hình ảnh. Trong *Biên bản báo cáo lần thứ 52: Cuộc họp thường niên của Hiệp hội Ngôn ngữ học tính toán (Tập 2: Bài báo ngắn)*, trang 452–457, Baltimore, Maryland, tháng 6 năm 2014. Hiệp hội Ngôn ngữ học tính toán
- [13] M. Everingham, L. Van Gool, CKI Williams, J. Winn, và A. Zisserman. Các lớp đối tượng trực quan PASCAL Kết quả Thủ thách 2010 (VOC2010). <http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html>. 1
- [14] A. Farhadi, M. Hejrati, MA Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier và D. Forsyth. Mỗi bức ảnh đều kể một câu chuyện: Tạo câu từ hình ảnh. Trong *Biên bản báo cáo của Hội nghị châu Âu lần thứ 11 về Thị giác máy tính: Phần IV, ECCV'10*, 2010. 2, 5, 6, 7, 12, 16
- [15] PF Felzenszwalb, RB Girshick, D. McAllester và D. Raamanan. Phát hiện đối tượng với phản hồi được đào tạo phân biệt dựa trên các mô hình. Giao dịch IEEE về Phân tích Mẫu và Trí tuệ máy móc, 32(9):1627–1645, 2010. 1
- [16] A. Gupta và LS Davis. Ngoài danh từ: Khai thác giới từ và tính từ so sánh để học các phân loại trực quan. Trong DA Forsyth, PHS Torr và A. Zisserman, biên tập viên, *ECCV (1)*, tập 5302 của *Ghi chú bài giảng về Khoa học máy tính*, trang 16–29. Springer, 2008. 2
- [17] A. Gupta, Y. Verma, và C. Jawahar. Lựa chọn ngôn ngữ học trên tầm nhìn để mô tả hình ảnh. 2012. 2
- [18] M. Hodosh, P. Young, và J. Hockenmaier. Khung hình ảnh mô tả như một nhiệm vụ xếp hạng: Dữ liệu, mô hình và đánh giá số liệu. *J. Artif. Intell. Res. (JAIR)*, 47:853–899, 2013. 1, 2, 3, 5, 7
- [19] A. Karpathy và L. Fei-Fei. Ngôn ngữ thị giác sâu sắc căn chỉnh để tạo mô tả hình ảnh. *CoRR*, abs/1412.2306, 2014. 2, 3
- [20] A. Karpathy, A. Joulin và L. Fei-Fei. Nhúng đoạn sâu cho ánh xạ câu hình ảnh hai chiều. *CoRR*, 2014. 2
- [21] R. Kiros, R. Salakhutdinov và RS Zemel. Thông nhất nhúng ngữ nghĩa thị giác với các mô hình ngôn ngữ thần kinh đa phương thức. *CoRR*, abs/1411.2539, 2014. 2, 3
- [22] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, AC Berg, và TL Berg. Trò chuyện với trẻ con: Hiểu và tạo ra mô tả hình ảnh. Trong *Biên bản báo cáo CVPR lần thứ 24*, 2011. 1, 2, 5, 6, 7, 11, 12, 16
- [23] CH Lampert, H. Nickisch, và S. Harmeling. Học cách phát hiện các lớp đối tượng không nhìn thấy bằng cách chuyển giao thuộc tính betweenclass. Trong *CVPR*, 2009. 1, 2
- [24] S. Li, G. Kulkarni, TL Berg, AC Berg và Y. Choi. Soạn thảo các mô tả hình ảnh đơn giản bằng cách sử dụng n-gram quy mô web. Trong *Biên bản Hội nghị lần thứ muối lăm về Học ngôn ngữ tự nhiên tính toán*, CoNLL '11, trang 220–228, Stroudsburg, PA, Hoa Kỳ, 2011. Hiệp hội Ngôn ngữ học tính toán. 2
- [25] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar và CL Zitnick. Microsoft COCO: Đối tượng chung trong ngữ cảnh. Trong *ECCV*, 2014. 3, 4
- [26] S. Maji, L. Bourdev, và J. Malik. Nhận dạng hành động từ một biểu diễn phân tán của từ thẻ và hình dáng. Trong *IEEE Hội nghị quốc tế về thị giác máy tính và mẫu* Sự công nhận (CVPR), 2011. 1
- [27] J. Mao, W. Xu, Y. Yang, J. Wang, và AL Yuille. Giải thích hình ảnh với mạng lưới nơ-ron hồi quy đa phương thức. *CoRR*, abs/1410.1090, 2014. 2, 3
- [28] D. Martin, C. Fowlkes, D. Tal, và J. Malik. Một cơ sở dữ liệu của hình ảnh tự nhiên phân đoạn của con người và ứng dụng của nó vào đánh giá các thuật toán phân đoạn và do lường các số liệu thống kê sinh thái. Trong *Proc. 8th Int'l Conf. Computer Vision*, tập 2, trang 416–423, tháng 7 năm 2001. 1
- [29] M. Mitchell, X. Han, và J. Hayes. Midge: Tạo mô tả hình ảnh. Trong *Biên bản Hội nghị quốc tế lần thứ bảy về thẻ hệ ngôn ngữ tự nhiên, INLG '12*, trang 131–133, Stroudsburg, PA, Hoa Kỳ, 2012. Hiệp hội cho Ngôn ngữ học tính toán. 1, 2, 5, 6, 7, 12, 16

- [30] H. Müller, P. Clough, T. Deselaers, and B. Caputo. *Image-CLEF: Experimental Evaluation in Visual Information Retrieval*. Springer Publishing Company, Incorporated, 1st edition, 2010. 3
- [31] A. Nenkova and R. J. Passonneau. Evaluating content selection in summarization: The pyramid method. In *HLT-NAACL*, pages 145–152, 2004. 3
- [32] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011. 2, 3, 4, 11
- [33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. 1, 2, 5, 12
- [34] D. Parikh and K. Grauman. Relative Attributes. In *ICCV*, 2011. 2
- [35] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT '10*, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 2, 3, 4
- [36] S. Robertson. Understanding inverse document frequency: On theoretical arguments for idf. *Journal of Documentation*, 60:2004, 2004. 4
- [37] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *IEEE International Conference on Computer Vision (ICCV)*, December 2013. 1, 2, 5, 6, 8, 12, 16
- [38] M. A. Sadeghi and A. Farhadi. Recognition using visual phrases. 2011. 2
- [39] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Comput. Vision*, 2002. 1
- [40] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference on*, pages 1–8, June 2008. 1
- [41] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai. Adaptively learning the crowd kernel. In *In ICML11*, 2011. 1
- [42] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014. 2, 3
- [43] Y. Yang, C. L. Teo, H. D. III, and Y. Aloimonos. Corpus-guided sentence generation of natural images. In *EMNLP. ACL*, 2011. 1, 2
- [44] M. Yatskar, M. Galley, L. Vanderwende, and L. Zettlemoyer. See no evil, say no evil: Description generation from densely labeled images. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, page 110120, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University. 1, 2
- [45] C. yew Lin. Rouge: a package for automatic evaluation of summaries. pages 25–26, 2004. 1, 3, 5
- [46] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013. 2, 3, 4
- [30] H. Müller, P. Clough, T. Deselaers, và B. Caputo. *Đánh giá thử nghiệm trong việc truy xuất thông tin trực quan*. Springer Publishing Company, Incorporated, ấn bản lần thứ 1, 2010. 3
- [46] CL Zitnick và D. Parikh. Đưa ngữ nghĩa vào trọng tâm bằng cách sử dụng trừu tượng trực quan. Trong *CVPR*, 2013. 2, 3, 4
- [31] A. Nenkova và RJ Passonneau. *Đánh giá lựa chọn nội dung trong tóm tắt: Phương pháp kim tự tháp*. Trong *HLT-NAACL*, trang 145–152, 2004. 3
- [32] V. Ordonez, G. Kulkarni và TL Berg. *Im2text: Mô tả hình ảnh bằng 1 triệu bức ảnh có chủ thích*. Trong *Hệ thống xử lý thông tin thần kinh (NIPS)*, 2011. 2, 3, 4, 11 [33] K. Papineni, S. Roukos, T. Ward và W.-J. Zhu. Bleu: Một phương pháp đánh giá tự động bản dịch máy.
- Trong Biên bản báo cáo của Hội nghị thư ờng niên lần thứ 40 về Hiệp hội Ngôn ngữ học tính toán, ACL '02, trang 311-318, Stroudsburg, PA, Hoa Kỳ, 2002. Hiệp hội Ngôn ngữ học tính toán. 1, 2, 5, 12 [34] D. Parikh và K. Grauman. Thuộc tính tương đối. Trong *ICCV*, 2011. 2
- [35] C. Rashtchian, P. Young, M. Hodosh và J. Hockenmaier. Thu thập chủ thích hình ảnh bằng cách sử dụng Amazon Mechanical Turk. Trong Biên bản báo cáo của Hội thảo NAACL HLT 2010 về Tạo dữ liệu ngôn ngữ và giọng nói bằng Amazon Mechanical Turk, CSLDAMT '10, Stroudsburg, PA, Hoa Kỳ, 2010. Hiệp hội Ngôn ngữ học tính toán. 2, 3, 4 [36] S. Robertson. Hiểu tần suất tài liệu nghịch đảo: Về các lập luận lý thuyết cho idf. *Tạp chí Tài liệu*, 60:2004, 2004. 4
- [37] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal và B. Schiele. Biên dịch nội dung video thành mô tả ngôn ngữ tự nhiên. Trong *Hội nghị quốc tế về tầm nhìn máy tính của IEEE (ICCV)*, tháng 12 năm 2013. 1, 2, 5, 6, 8, 12, 16
- [38] MA Sadeghi và A. Farhadi. Nhận dạng bằng cách sử dụng hình ảnh cụm từ. 2011. 2
- [39] D. Scharstein và R. Szeliski. Phân loại và đánh giá các thuật toán tương ứng âm thanh nỗi hai khung dày đặc. *Int. J. Comput. Vision*, 2002. 1
- [40] A. Sorokin và D. Forsyth. Chủ thích dữ liệu tiện ích với amazon mechanical turk. Trong *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE Computer Society Conference* về, trang 1–8, tháng 6 năm 2008. 1 [41] O. Tamuz, C. Liu, S. Belongie, O. Shamir và AT Kalai. Học tập thích ứng hạt nhân đám đông. Trong *ICML11*, 2011. 1
- [42] O. Vinyals, A. Toshev, S. Bengio, và D. Erhan. *Hiển thị và kể: Một trình tạo chủ thích hình ảnh thần kinh*. *CoRR*, abs/1411.4555, 2014. 2, 3
- [43] Y. Yang, CL Teo, HD III, và Y. Aloimonos. *Tạo câu theo hướng dẫn của các hình ảnh tự nhiên*. Trong *EMNLP. ACL*, 2011. 1, 2
- [44] M. Yatskar, M. Galley, L. Vanderwende và L. Zettlemoyer. Không nhìn thấy điều xấu, không nói điều xấu: Tạo mô tả từ hình ảnh được gắn nhãn dày đặc. Trong *Biên bản báo cáo của Hội nghị chung lần thứ ba về Ngữ nghĩa từ vựng và tính toán (*SEM 2014)*, trang 110120, Dublin, Ireland, tháng 8 năm 2014. Hiệp hội Ngôn ngữ học tính toán và Đại học thành phố Dublin. 1, 2 [45] C. yew Lin. Rouge: một gói để đánh giá tự động các bản tóm tắt. trang 25–26, 2004. 1, 3, 5

Appendix Overview

List of items:

I Comparison of metrics on triplet annotations to pairwise annotations: Compares the accuracy of CIDEr on triplet annotation to existing choices of metrics on pairwise annotations

II Ranking of reference sentences for various automated metrics: Qualitative examples of the kind of sentences preferred by each metric

III Comparison of rankings from CIDEr and CIDEr-D: Establishes that both CIDEr and CIDEr-D are similar qualitatively, in terms of how they rank reference sentences

IV Difference between human-like and what humans like: Shows examples of differences between pairwise and triplet annotations. Pairwise annotations often favor longer sentences

V Sentence collection interface for PASCAL-50S and ABSTRACT-50S: Shows a snapshot of the interface used to collect our datasets, and explains the instructions

VI Equations for BLEU, ROUGE, and METEOR: Formulates some existing metrics in terms of the notation used in the rest of the paper

VII Qualitative examples of outputs of image description methods evaluated in the paper: Gives a sense for the kind of outputs produced by each of the image description methods evaluated in the paper

VIII Performance of different versions of metrics on consensus: Benchmarks the performance of different versions of metrics discussed in the paper at matching human consensus

Appendix I : Comparison to Pairwise Annotations

We consider some alternate annotation modalities and compare the performance of present metrics on them with that of CIDEr on consensus. The first such modality is a pairwise interface described as follows. Subjects on Amazon Mechanical Turk (AMT) are shown just the two candidate sentences (B and C) with the image (instead of sentence A), and asked to pick the *better* description out of the two. 11 such human judgments are collected for each such pair. These annotations are collected for the same PASCAL-50S candidate sentences as those used for the triplet experiments in the paper. We compare accuracy on *consensus* for CIDEr to accuracy of other metrics on picking the *better* candidate sentence. We find that ROUGE_L

at 5 sentences performs at 75.6% whereas the BLEU₄ version performs at 74.75%. ROUGE₁ and BLEU₁ perform at 73.15% and 73.4% respectively at 5 sentences. With METEOR at 5 sentences, the performance is at 79.5%. In contrast, CIDEr at 48 sentences reaches an accuracy of 84% on consensus. Thus the consensus-based protocol comprising of our proposed metric, dataset and human annotation modality provides more accurate automated evaluation.

Appendix II : Ranking of Sentences

We now show a ranking of the 48 sentences collected for a particular image as per the CIDEr, BLEU₁, BLEU₁ without Brevity Penalty and ROUGE₁ scores (Fig. 5). Each reference sentence is considered in turn as a candidate and scored with the remaining (47) reference sentences using the corresponding metric. Note how the top-ranked CIDEr sentences show high consensus. The top-ranked ROUGE sentences are typically more detailed, whereas the top ranked BLEU sentences are not as consistent as those with CIDEr. If BLEU was used without the brevity penalty, as some previous works have [22, 32] one would see that really short sentences get high scores. Intuitively, we can see that the ranking produced by CIDEr is more meaningful.

Appendix III : Difference between Human-like and What Humans Like

In our experiments, we found that there can often be a difference in the sentence that is rated as “better” (measured via pairwise annotation) by subjects *versus* the kind of sentences written by subjects when asked to describe the image (measured via consensus annotation). We refer to this distinction as human-like vs what humans like. Some qualitative examples are shown in Fig. 7. Candidate sentences shown in bold are those that the consensus-based measure picks and those shown in thin font are those picked by the pairwise evaluation based on “better”. Reference sentences rated similar to the winning candidate sentence using the triplet annotation are shown in bold.

Appendix IV : Ranking of sentences - CIDEr and CIDEr-D

As we report in Sec. 8, we find that CIDEr and CIDEr-D agree with a high correlation (Spearman’s $\rho=0.94$) on ranking of sentences. We now compare CIDEr₁ and CIDEr-D₁ rankings, since results are easier to interpret for the unigram case. An example of ranking can be found in Fig. 6. Notice that the rankings of CIDEr and CIDEr-D are very similar qualitatively. However, the formulation of CIDEr-D avoids gaming effects as explained in Sec. 8.

Appendix V : Sentence Collection Interface

The sentence collection interface for both ABSTRACT-50S and PASCAL-50S is shown in Fig. 8. Stringent rejection criteria were specified (Fig. 9).

Phụ lục Tổng quan

Danh sách các mục:

Tôi So sánh các số liệu trên chú thích bộ ba với
chú thích từng cặp: So sánh độ chính xác của
CIDEr về chú thích bộ ba cho các lựa chọn số liệu hiện có trên
chú thích từng cặp

II Xếp hạng các câu tham chiếu cho các số liệu tự động khác nhau:
Các ví dụ định tính về loại
câu được ưa thích bởi mỗi số liệu

III So sánh thứ hạng từ CIDEr và CIDEr-D:
Xác định rằng cả CIDEr và CIDEr-D đều tương tự về mặt chất
lượng, xét về cách chúng xếp hạng tham chiếu
câu

IV Sự khác nhau giữa giống người và con người
thích: Hiển thị các ví dụ về sự khác biệt giữa từng cặp
và chú thích bộ ba. Chú thích từng cặp thường ưu tiên các câu
dài hơn

V Giao diện thu thập câu cho PASCAL-50S và
TÓM TẮT-50S: Hiển thị ảnh chụp nhanh của giao diện
được sử dụng để thu thập các tập dữ liệu của chúng tôi và giải thích các
hướng dẫn

VI Các chương trình cho BLEU, ROUGE và METEOR: Công thức hóa một
số số liệu hiện có theo ký hiệu
được sử dụng trong phần còn lại của bài báo

VII Các ví dụ định tính về kết quả đầu ra của các phương pháp mô
tả hình ảnh được đánh giá trong bài báo: Cung cấp một ý nghĩa
cho loại đầu ra được tạo ra bởi mỗi hình ảnh
phương pháp mô tả được đánh giá trong bài báo

VIII Hiệu suất của các phiên bản số liệu khác nhau trên
sự đồng thuận: Đánh giá hiệu suất của các
phiên bản số liệu được thảo luận trong bài báo tại matching
sự đồng thuận của con người

Phụ lục I: So sánh với chú thích theo cặp

Chúng tôi xem xét một số phương thức chú thích thay thế và
so sánh hiệu suất của các số liệu hiện tại trên chúng với
của CIDEr về sự đồng thuận. Phương thức đầu tiên như vậy là
giao diện từng cặp được mô tả như sau. Các đối tượng trên Amazon
Mechanical Turk (AMT) chỉ được hiển thị hai câu có thể (B và C) có
hình ảnh (thay vì câu A) và được yêu cầu chọn mô tả tốt hơn trong
số

hai. 11 phán đoán của con người như vậy được thu thập cho mỗi
cặp như vậy. Những chủ thích này được thu thập cho cùng một
Câu ứng viên PASCAL-50S như những câu được sử dụng cho
thí nghiệm bộ ba trong bài báo. Chúng tôi so sánh độ chính xác trên
sự đồng thuận cho CIDEr về độ chính xác của các số liệu khác khi
chọn câu ứng viên tốt hơn. Chúng tôi thấy rằng ROUGE

ở 5 câu thực hiện ở mức 75,6% trong phiên bản BLEU4 thực hiện ở
mức 74,75%. ROUGE1 và BLEU1 thực hiện ở mức
73,15% và 73,4% tương ứng ở 5 câu. Với ME-TEOR ở 5 câu, hiệu suất
là 79,5%. Ngược lại, CIDEr ở 48 câu đạt độ chính xác là 84%

về sự đồng thuận. Do đó, giao thức dựa trên sự đồng thuận bao gồm số
liệu, tập dữ liệu và chú thích của con người được đề xuất của chúng tôi
phương thức cung cấp đánh giá tự động chính xác hơn.

Phụ lục II: Xếp hạng các câu

Bây giờ chúng tôi hiển thị bảng xếp hạng của 48 câu đã thu thập
đối với một hình ảnh cụ thể theo CIDEr, BLEU1, BLEU1
không có Hình phạt ngắn gọn và điểm ROUGE1 (Hình 5). Mỗi
câu tham chiếu được xem xét lần lượt như một ứng cử viên
và ghi điểm với các câu tham chiếu còn lại (47) bằng cách sử dụng
số liệu tương ứng. Lưu ý cách xếp hạng cao nhất
các câu của CIDEr cho thấy sự đồng thuận cao. Các câu được xếp hạng cao nhất
Câu ROUGE thường chi tiết hơn, trong khi
các câu BLEU được xếp hạng cao nhất không quán như những câu đó
với CIDEr. Nếu BLEU được sử dụng mà không có hình phạt ngắn gọn,
như một số tác phẩm trước đây [22, 32] người ta sẽ thấy rằng
các câu thực sự ngắn đạt điểm cao. Theo trực giác, chúng ta có thể thấy
rằng thứ hạng do CIDEr đưa ra có ý nghĩa hơn.

Phụ lục III: Sự khác biệt giữa giống người và Con người thích gì

Trong các thí nghiệm của chúng tôi, chúng tôi thấy rằng thường có thể có một
sự khác biệt trong câu được đánh giá là “tốt hơn” (được do lưỡng
thông qua chú thích từng cặp) theo chủ thể so với loại câu do chủ
thể viết khi được yêu cầu mô tả hình ảnh (được do thông qua chú
thích đồng thuận). Chúng tôi thanh khảo điều này
sự phân biệt giữa giống con người và những gì con người thích. Một
số ví dụ định tính được thể hiện trong Hình 7. Câu ứng viên
được in đậm là những biện pháp dựa trên sự đồng thuận
những cái được chọn và những cái được hiển thị bằng phông chữ mờ là những cái được chọn bởi
đánh giá từng cặp dựa trên “tốt hơn”. Câu tham khảo
được đánh giá tương tự như câu ứng cử viên chiến thắng bằng cách sử dụng
chú thích bộ ba được hiển thị bằng chữ in đậm.

Phụ lục IV: Xếp hạng câu - CIDEr và Rouge-Tao-D

Như chúng tôi báo cáo trong Mục 8, chúng tôi thấy rằng CIDEr và CIDEr-D
đồng ý với mối tương quan cao (Spearman’s $\rho=0.94$) về thứ hạng của
các câu. Bây giờ chúng ta so sánh CIDEr và CIDEr-D₁
xếp hạng, vì kết quả dễ diễn giải hơn đối với unigram
trưởng hợp. Một ví dụ về xếp hạng có thể được tìm thấy trong Hình 6. Lưu ý
rằng thứ hạng của CIDEr và CIDEr-D rất giống nhau
về mặt chất lượng. Tuy nhiên, công thức của CIDEr-D tránh
tác động của trò chơi như đã giải thích trong Mục 8.

Phụ lục V: Giao diện thu thập câu

Giao diện thu thập câu cho cả ABSTRACT-50S và PASCAL-50S được hiển
 thị trong Hình 8. Tiêu chí từ chối nghiêm ngặt đã được chỉ định (Hình 9).



Figure 8: Interface used for collecting image descriptions

This is a whacy image.

Reasons for rejection:

- (1) Does not actually describe what is going on in the scene.
- (2) Too short.

A man is next to a woman, the man is on the couch, the woman is on the couch, a dog is on the man

Reasons for rejection:

- (1) Too literal.
- (2) Others probably would not describe this scene the same way.

Figure 9: An illustration of our rejection criteria with examples shown to subjects on Amazon Mechanical Turk (AMT)

Appendix VI : Image Description Method Outputs

In the paper, we compared the relative performance of five image description methods: Midge [29], Babytalk [22], Story [14], and two versions of Translating Video Content to Natural Language Descriptions [37] (Video and Video+). Here, we show a sample image with the descriptions generated by the five methods compared in the paper (Fig. 10). We can see that Midge [29] and Babytalk [22] produce the better descriptions on this image, consistent with our finding in the paper.

Appendix VII : Other Metrics

Our goal is to automatically evaluate for an image I_i how well a candidate sentence c_i matches the consensus of a set of image descriptions $S_i = \{s_{i1}, \dots, s_{im}\}$. The sentences are represented using sets of n -grams, where an n -gram $\omega_k \in \Omega$ is a set of one or more ordered words. In this paper we explore n -grams with one to four words. Each word in an n -gram is modified to its stemming or root form.

That is, “fishes”, “fishing” and “fished” all get reduced to “fish”. The number of times an n -gram ω_k occurs in a sentence s_{ij} is denoted $h_k(s_{ij})$ or $h_k(c_i)$ for the candidate sentence $c_i \in C$.

BLEU

BLEU [33] is a popular machine translation metric that analyzes the co-occurrences of n -grams between the candidate and reference sentences. As we explain in Sec.6, we compute the sentence level BLEU scores between a candidate sentence and a set of reference sentences. The BLEU score is computed as follows:

$$P_n(c_i, S_i) = \frac{\sum_k \min(h_k(c_i), \max_j h_k(s_{ij}))}{\sum_k h_k(c_i)}, \quad (6)$$

where k indexes the set of possible n -grams of length n . The clipped precision metric limits the number of times an n -gram may be counted to the maximum number of times it is observed in a single reference sentence. Note that P_n is a precision score and it favors short sentences. So a brevity penalty is also used:

$$b(C, S) = \begin{cases} 1 & \text{if } l_C > l_S \\ e^{1-l_S/l_C} & \text{if } l_C \leq l_S \end{cases}, \quad (7)$$

where l_C is the total length of candidate sentences c_i 's and l_S is the length of the corpus-level effective reference length. When there are multiple references for a candidate sentence, we choose to use the *closest* reference length for the brevity penalty.

The overall BLEU score is computed using a weighted geometric mean of the individual n -gram precision:

$$\text{BLEU}_N(c_i, S_i) = b(c_i, S_i) \exp \left(\sum_{n=1}^N w_n \log P_n(c_i, S_i) \right), \quad (8)$$

where $N = 1, 2, 3, 4$ and w_n is typically held constant for all n .

BLEU has shown good performance for corpus-level comparisons over which a high number of n -gram matches exist. However, at a sentence-level the n -gram matches for higher n rarely occur. As a result, BLEU performs poorly when comparing individual sentences.

ROUGE

ROUGE is a set of evaluation metrics designed to evaluate text summarization algorithms.

1. ROUGE_N: The first ROUGE metric computes a simple n -gram recall over all reference summaries given a



Hình 8: Giao diện được sử dụng để thu thập mô tả hình ảnh

This is a whacy image.

Reasons for rejection:

- (1) Does not actually describe what is going on in the scene.
- (2) Too short.

A man is next to a woman, the man is on the couch, the woman is on the couch, a dog is on the man

Reasons for rejection:

- (1) Too literal.
- (2) Others probably would not describe this scene the same way.

Hình 9: Minh họa về tiêu chí từ chối của chúng tôi với các ví dụ được hiển thị cho các đối tượng trên Amazon Mechanical Turk (AMT)

Phụ lục VI: Đầu ra của phương pháp mô tả hình ảnh

Trong bài báo, chúng tôi đã so sánh hiệu suất tương đối của năm phương pháp mô tả hình ảnh: Midge [29], Babytalk [22], Story [14] và hai phiên bản của Dịch nội dung video sang mô tả ngôn ngữ tự nhiên [37] (Video và Video+).

Ở đây, chúng tôi trình bày một hình ảnh mẫu với các mô tả được tạo ra bởi năm phương pháp được so sánh trong bài báo (Hình 10).

Chúng ta có thể thấy rằng Midge [29] và Babytalk [22] đưa ra những mô tả tốt hơn về hình ảnh này, phù hợp với phát hiện của chúng tôi trong bài báo.

Phụ lục VII: Các số liệu khác

Mục tiêu của chúng tôi là tự động đánh giá cho một hình ảnh. Để đạt được độ phù hợp của một câu ứng viên với sự đồng thuận của một tập hợp các mô tả hình ảnh $S_i = \{s_{i1}, \dots, s_{im}\}$. Các câu được biểu diễn bằng các tập hợp n -gram, trong đó n -gram $\omega_k \in \Omega$ là một tập hợp gồm một hoặc nhiều từ được sắp xếp.

Trong bài báo này, chúng tôi khám phá n -gram với một điều kiện: Mỗi từ trong n -gram được sửa đổi thành dạng gốc hoặc dạng gốc của nó.

Nghĩa là, “fishes”, “fishing” và “fished” đều được rút gọn thành “fish”. Số lần n -gram ω_k xuất hiện trong câu s_{ij} được ký hiệu là $h_k(s_{ij})$ hoặc $h_k(c_i)$ đối với câu ứng viên $c_i \in C$.

Màu xanh

BLEU [33] là một phép đo dịch máy phổ biến phân tích sự đồng xuất hiện của n -gram giữa câu ứng viên và câu tham chiếu. Như chúng tôi giải thích trong Mục 6, chúng tôi tính toán điểm BLEU cấp câu giữa một câu ứng viên và một tập hợp các câu tham chiếu. Điểm BLEU được tính như sau:

$$P_n(c_i, S_i) = \frac{\min(h_k(c_i), \max_j h_k(s_{ij}))}{\sum_k h_k(c_i)}, \quad (6)$$

trong đó k chỉ mục tập hợp các n -gram có thể có độ dài n . Thước đo độ chính xác bị cắt giới hạn số lần n -gram có thể được đến số lần tối đa được quan sát thấy trong một câu tham chiếu duy nhất. Lưu ý rằng P_n là điểm chính xác và nó ưu tiên các câu ngắn. Vì vậy, hình phạt ngắn gọn cũng được sử dụng:

$$b(C, S) = \begin{cases} 1 & \text{if } l_C > l_S \\ e^{1-l_S/l_C} & \text{if } l_C \leq l_S \end{cases}, \quad (7)$$

trong đó l_C là tổng độ dài của các câu ứng viên c_i và l_S là độ dài của độ dài tham chiếu hiệu quả ở cấp độ ngũ liệu. Khi có nhiều tham chiếu cho một câu ứng viên, chúng tôi chọn sử dụng độ dài tham chiếu gần nhất cho hình phạt ngắn gọn.

Điểm BLEU tổng thể được tính bằng cách sử dụng một trọng số giá trị trung bình hình học của độ chính xác n -gram riêng lẻ:

$$\text{BLEU}_N(c_i, S_i) = b(c_i, S_i) \exp \left(\sum_{n=1}^N w_n \log P_n(c_i, S_i) \right), \quad (8)$$

trong đó $N = 1, 2, 3, 4$ và w_n thường giữ không đổi với mọi n .

BLEU đã cho thấy hiệu suất tốt đối với các phép so sánh cấp ngũ liệu trong đó có số lượng lớn các phép khớp n -gram. Tuy nhiên, ở cấp câu, các phép khớp n -gram cho n cao hơn hiếm khi xảy ra. Do đó, BLEU hoạt động kém khi so sánh các câu riêng lẻ.

MÀU ĐỎ

ROUGE là một tập hợp các số liệu đánh giá được thiết kế để đánh giá các thuật toán tóm tắt văn bản.

1. ROUGEN : Chỉ số ROUGE đầu tiên tính toán một phép thu hồi n -gram đơn giản trên tất cả các bản tóm tắt tham chiếu đưa ra

candidate sentence:

$$ROUGE_N(c_i, S_i) = \frac{\sum_j \sum_k \min(h_k(c_i), h_k(s_{ij}))}{\sum_j \sum_k h_k(s_{ij})} \quad (9)$$

2. ROUGE_L: ROUGE_L uses a measure based on the Longest Common Subsequence (LCS). An LCS is a set words shared by two sentences which occur in the same order. However, unlike *n*-grams there may be words in between the words that create the LCS. Given the length $l(c_i, s_{ij})$ of the LCS between a pair of sentences, ROUGE_L is found by computing an F-measure:

$$R_l = \max_j \frac{l(c_i, s_{ij})}{|s_{ij}|} \quad (10)$$

$$P_l = \max_j \frac{l(c_i, s_{ij})}{|c_i|} \quad (11)$$

$$ROUGE_L(c_i, S_i) = \frac{(1 + \beta^2)R_l P_l}{R_l + \beta^2 P_l} \quad (12)$$

R_l and P_l are recall and precision of LCS. β is usually set to favor *recall* ($\beta = 2$). Since *n*-grams are implicit in this measure due to the use of the LCS, they need not be specified.

3. ROUGE_S: The final ROUGE metric uses skip bi-grams instead of the LCS or *n*-grams. Skip bi-grams are pairs of ordered words in a sentence. However, similar to the LCS, words may be skipped between pairs of words. Thus, a sentence with 4 words would have $C_2^4 = 6$ skip bi-grams. Precision and recall are again incorporated to compute an F-measure score. If $f_k(s_{ij})$ is the skip bi-gram count for sentence s_{ij} , ROUGE_S is computed as:

$$R_s = \max_j \frac{\sum_k \min(f_k(c_i), f_k(s_{ij}))}{\sum_k f_k(s_{ij})} \quad (13)$$

$$P_s = \max_j \frac{\sum_k \min(f_k(c_i), f_k(s_{ij}))}{\sum_k f_k(c_i)} \quad (14)$$

$$ROUGE_S(c_i, S_i) = \frac{(1 + \beta^2)R_s P_s}{R_s + \beta^2 P_s} \quad (15)$$

Skip bi-grams are capable of capturing long range sentence structure. In practice, skip bi-grams are computed so that the component words occur at a distance of at most 4 from each other.

METEOR

METEOR is calculated by generating an alignment between the words in the candidate and reference sentences, with an aim of 1:1 correspondence. This alignment is computed while minimizing the number of chunks, ch , of contiguous and identically ordered tokens in the sentence pair. The alignment is based on exact token matching, followed by WordNet synonyms and then stemmed tokens. Given a set of alignments, m , the METEOR score is the harmonic mean of precision and recall between the best scoring reference and candidate:

$$Pen = \gamma \left(\frac{ch}{m} \right)^{\theta} \quad (16)$$

$$F_{mean} = \frac{P_m R_m}{\alpha P_m + (1 - \alpha) R_m} \quad (17)$$

$$P_m = \frac{|m|}{\sum_k h_k(c_i)} \quad (18)$$

$$R_m = \frac{|m|}{\sum_k h_k(s_{ij})} \quad (19)$$

$$METEOR = (1 - Pen)F_{mean} \quad (20)$$

Thus, the final METEOR score includes a penalty based on chunkiness of resolved matches and a harmonic mean term that gives the quality of the resolved matches.

Appendix VIII : Detailed Evaluation

We now show the results for different versions of each metric in the family of BLEU and ROUGE metrics, along with some variations of CIDEr. We use only one (latest) version of METEOR, thus it is not a part of this evaluation. The versions of CIDEr shown here are as follows. **CIDEr exp** refers to an exponential combination of scores obtained by varying *n*-gram counts w_n instead of taking a mean, which we describe in Sec. 4. **CIDEr max** refers to taking a max across scores with different reference sentences, instead of the mean we discuss in the paper. **CIDEr no idf** version sets uniform IDF weights in CIDEr. The rest of the versions of other metrics are explained in the previous section. The results on PASCAL-50S are shown in Fig. 11 and ABSTRACT-50S are shown in Fig. 12. We find that removing the IDF weights in the **CIDEr no idf** version hurts performance significantly. **CIDEr max** and **CIDEr exp** perform slightly worse than CIDEr. The best performing version of each of these metrics was discussed in Sec. 7.

câu ứng cử viên:

$$ROUGEN(c_i, S_i) = \frac{\sum_{j,k} \min(h_k(c_i), h_k(s_{ij}))}{\sum_k h_k(s_{ij})} \quad (9)$$

2. ROUGEL: ROUGEL sử dụng một phép đo dựa trên Longest Common Subsequence (LCS). LCS là một tập hợp các từ được chia sẻ bởi hai câu xuất hiện theo cùng một thứ tự. Tuy nhiên, không giống như *n*-gram, có thể có các từ ở giữa các từ tạo nên LCS.

Với độ dài $l(c_i, s_{ij})$ của LCS giữa một cặp câu, ROUGEL được tìm thấy bằng cách tính toán F-do lường:

$$R_l = \max_j \frac{l(c_i, s_{ij})}{|s_{ij}|} \quad (10)$$

$$P_l = \text{tối đa}_{i,j} \frac{l(c_i, s_{ij})}{|c_i|} \quad (11)$$

$$ROUGEL(c_i, S_i) = \frac{(1 + \beta^2)R_l P_l}{R_l + \beta^2 P_l} \quad (12)$$

R_l và P_l là độ thu hồi và độ chính xác của LCS. β thường được thiết lập để ưu tiên thu hồi ($\beta = 2$). Vì *n*-gram được ngầm định trong phép đo này do sử dụng LCS nên chúng không cần phải được chỉ định.

3. ROUGES: Thủ tục ROUGE cuối cùng sử dụng skip bi-gram thay vì LCS hoặc *n*-gram. Skip bi-gram là cặp từ được sắp xếp trong một câu. Tuy nhiên, tự ngưng như LCS, các từ có thể bị bỏ qua giữa các cặp từ. Do đó, một câu có 4 từ sẽ có $C = 6$ skip bi-gram. Độ chính xác và độ thu hồi lại được kết hợp lại để tính điểm F-measure. $\frac{4}{2}$

Nếu $f_k(s_{ij})$ là số lưỡng bi-gram bỏ qua cho câu s_{ij} , ROUGES được tính như sau:

$$R_s = j \text{ tối đa} \frac{\min(f_k(c_i), f_k(s_{ij}))}{f_k(s_{ij})} \quad (13)$$

$$P_s = j \text{ tối đa} \frac{\min(f_k(c_i), f_k(s_{ij}))}{f_k(c_i)} \quad (14)$$

$$ROUGES(c_i, S_i) = \frac{(1 + \beta^2)R_s P_s}{R_s + \beta^2 P_s} \quad (15)$$

Skip bi-gram có khả năng nắm bắt cấu trúc câu dài. Trong thực tế, skip bi-gram được tính toán sao cho các từ thành phần xuất hiện ở khoảng cách tối đa là 4 từ với nhau.

SAO TINH

METEOR được tính toán bằng cách tạo ra sự liên kết giữa các từ trong câu ứng viên và câu tham chiếu, với mục tiêu là sự tương ứng 1:1. Sự liên kết này được tính toán trong khi giảm thiểu số lượng các khối, ch , của các mã thông báo liên tiếp và có thứ tự giống nhau trong cặp câu. Sự liên kết dựa trên sự khớp mã thông báo chính xác, tiếp theo là các từ đồng nghĩa của WordNet và sau đó là các mã thông báo có gốc. Với một tập hợp các sự liên kết, m , điểm METEOR là giá trị trung bình hài hòa của độ chính xác và khả năng thu hồi giữa tham chiếu có điểm cao nhất và ứng viên:

$$Pen = \gamma \frac{ch}{m} \quad (16)$$

$$F_{mean} = \frac{\chiều_giờ_chiều}{\alpha P_m + (1 - \alpha) R_m |m|} \quad (17)$$

$$\chiều = \frac{\chiều_giờ_chiều}{m} \quad (18)$$

$$R_m = \frac{|m|}{\sum h_k(s_{ij})} \quad (19)$$

$$METEOR = (1 - Pen)F_{mean} \quad (20)$$

Do đó, điểm METEOR cuối cùng bao gồm một hình phạt dựa trên độ chi tiết của các trận đấu đã giải quyết và một thuật ngữ trung bình hài hòa cho biết chất lượng của các trận đấu đã giải quyết.

Phụ lục VIII: Đánh giá chi tiết

Bây giờ chúng tôi sẽ trình bày kết quả cho các phiên bản khác nhau của từng số liệu trong họ số liệu BLEU và ROUGE, cùng với một số biến thể của CIDEr. Chúng tôi chỉ sử dụng một phiên bản (mới nhất) của METEOR, do đó nó không phải là một phần của đánh giá này. Các phiên bản của CIDEr được hiển thị ở đây như sau.

CIDEr exp đề cập đến sự kết hợp theo cấp số nhân của các điểm thu được bằng cách thay đổi số lưỡng n -gram w_n thay vì lấy giá trị trung bình, mà chúng tôi sẽ mô tả trong Phần 4. CIDEr max đề cập đến việc lấy giá trị tối đa trên các điểm có các câu tham chiếu khác nhau, thay vì lấy giá trị trung bình mà chúng tôi sẽ thảo luận trong bài báo. Phiên bản CIDEr không có idf đặt trọng số IDF thống nhất trong CIDEr. Các phiên bản còn lại của các số liệu khác được giải thích trong phần trước. Kết quả trên PASCAL-50S được hiển thị trong Hình 11 và ABSTRACT-50S được hiển thị trong Hình 12. Chúng tôi thấy rằng việc loại bỏ trọng số IDF trong phiên bản CIDEr không có idf làm giảm đáng kể hiệu suất. CIDEr max và CIDEr exp hoạt động kém hơn một chút so với CIDEr. Phiên bản hoạt động tốt nhất của mỗi số liệu này đã được thảo luận trong Phần 7.



CIDER	ROUGE	BLEU w/o BP	BLEU
<p>[1] A man is fishing in a canoe on a lake. [2] A man fishing in a canoe on a lake [3] A man in canoe fishing on a lake [4] A man in his canoe fishing on the lake. [5] A man fishes in a canoe in an empty lake [6] A man fishing off of his canoe on a lake. [7] A person in a canoe fishes on a lake. [8] a person fishes while sitting in a canoe on a lake [9] The man is fishing on a canoe [10] A man in a canoe is fishing. [11] A man fishing in a canoe. [12] man fishing in a canoe [13] Someone is fishing from a canoe on a lake. [14] a man fishing out of a canoe [15] A man in a canoe is fishing on a still lake. [16] A man is fishing on a lake. [17] A person is fishing from a canoe. [18] A man is fishing in a boat in lake [19] A man is on a boat fishing on the lake. [20] A man is fishing in a small boat on a lake. [21] One man fishes in a small boat on the lake. [22] A man is fishing from a boat on a lake. [23] A man in a canoe fishing in a calm lake. [24] A man is fishing alone in a canoe. [25] a man fishing in the middle of a lake in a boat [26] A person in a canoe fishing on a lake surrounded by hills. [27] a person fishing on a lake [28] A person is fishing in a boat on a lake. [29] Man in a boat fishing. [30] A man fishing alone on the lake. [31] A man fishes from his small boat. [32] A man is fishing from his canoe on quiet water. [33] A man is fishing in the river. [34] A man on a canoe fishing near a landmass. [35] A man is fishing alone on a small boat. [36] A lone man sits in a boat and fishes. [37] A guy is canoeing and fishing the middle of a tranquil and calm lake. [38] A man is out fishing from a canoe on a tranquil morning. [39] A man fishing in a kayak. [40] A man is fishing in the sea by a forest. [41] There is a man in the canoe. [42] A person is fishing in the water all by themselves. [43] A person is sitting in a boat on a lake. [44] A small boat in the middle of the lake. [45] A lone fisherman sits in his canoe on a river. [46] A lone fisherman sits in a canoe with a pole in the water. [47] A man is rowing a man in a river. [48] A lone fisherman in a rowboat on an empty lake.</p>	<p>[1] A man is fishing in a canoe on a lake. [2] A man in a canoe is fishing on a still lake. [3] A man is fishing in a small boat on a lake. [4] A man fishing in a canoe [5] A man in canoe fishing on a lake [6] [6] a man fishing in the middle of a lake in a boat [7] A man in a boat [8] A man is fishing in a canoe on a lake. [9] Man in a boat fishing. [10] a man fishing out of a canoe [11] a person fishing on a lake [12] A man in his canoe fishing on the lake. [13] A man is fishing in a boat in lake [14] A man is on a boat fishing on the lake. [15] A person in a canoe fishes on a lake. [16] A man is fishing in a boat on a lake. [17] A man in a canoe fishing on the lake. [18] A man fishing off of his canoe on a lake. [19] A man is in a canoe is fishing alone in a canoe. [20] A man in a small boat on the lake. [21] A man is fishing alone in a canoe. [22] A man is fishing in a boat on a lake. [23] A man is fishing alone in a canoe. [24] A man fishing alone on the lake. [25] A person is fishing in a boat on a lake. [26] A man in a canoe is fishing on a still lake. [27] a person fishes while sitting in a canoe on a lake [28] a man fishing in the middle of a lake in a boat [29] Someone is fishing from a canoe on a lake. [30] There is a man in the canoe. [31] A man in a canoe fishing in a calm lake. [32] A man fishes from his small boat. [33] A man fishing in a kayak. [34] A lone man sits in a boat and fishes. [35] A person is fishing alone on a small boat. [36] A man on a canoe fishing near a landmass. [37] A person in a canoe fishing on a lake surrounded by hills. [38] A man is fishing from his canoe on quiet water. [39] A man is fishing in the sea by a forest. [40] A person is sitting in a boat on a lake. [41] A man is out fishing from a canoe on a tranquil morning. [42] A guy is canoeing and fishing the middle of a tranquil and calm lake. [43] A small boat in the middle of the lake. [44] A person is fishing in the water all by themselves. [45] A lone fisherman sits in his canoe on a river. [46] A lone fisherman sits in a canoe with a pole in the water. [47] A man is rowing a man in a river. [48] A lone fisherman in a rowboat on an empty lake.</p>	<p>[1] A man is fishing on a lake. [2] A man fishing in a canoe. [3] man fishing in a canoe [4] A man in a canoe is fishing. [5] The man is fishing on a canoe [6] A man fishing in a canoe on a lake [7] A man in canoe fishing on a lake [8] A man is fishing in a canoe on a lake. [9] Man in a boat fishing. [10] a man fishing out of a canoe [11] a person fishing on a lake [12] A man in his canoe fishing on the lake. [13] A man is fishing in a boat in lake [14] A man is on a boat fishing on the lake. [15] A person in a canoe fishes on a lake. [16] A man is fishing in a small boat on a lake. [17] One man fishes in a small boat on the lake. [18] A man fishes in a canoe in an empty lake [19] A man is fishing from a boat on a lake. [20] A man is fishing in the river. [21] A person is fishing from a canoe. [22] A man fishing off of his canoe on a lake. [23] A man is fishing alone in a canoe. [24] A man fishing alone on the lake. [25] A person is fishing in a boat on a lake. [26] A man in a canoe is fishing on a still lake. [27] a person fishes while sitting in a canoe on a lake [28] a man fishing in the middle of a lake in a boat [29] Someone is fishing from a canoe on a lake. [30] There is a man in the canoe. [31] A man in a canoe fishing in a calm lake. [32] A man fishes from his small boat. [33] A man fishing in a kayak. [34] A lone man sits in a boat and fishes. [35] A person is fishing alone on a small boat. [36] A man on a canoe fishing near a landmass. [37] A person in a canoe fishing on a lake surrounded by hills. [38] A man is fishing from his canoe on quiet water. [39] A man is fishing in the sea by a forest. [40] A person is sitting in a boat on a lake. [41] A man is out fishing from a canoe on a tranquil morning. [42] A guy is canoeing and fishing the middle of a tranquil and calm lake. [43] A small boat in the middle of the lake. [44] A person is fishing in the water all by themselves. [45] A lone fisherman sits in his canoe on a river. [46] A lone fisherman sits in a canoe with a pole in the water. [47] A man is rowing a man in a river. [48] A lone fisherman in a rowboat on an empty lake.</p>	<p>[1] Man in a boat fishing. [2] a man fishing out of a canoe [3] A person is fishing in a boat on a lake. [4] A man is fishing in the river. [5] A man fishing in a canoe on a lake [6] A man fishes in a canoe in an empty lake [7] A man in a canoe is fishing. [8] A person in a canoe fishes on a lake. [9] A man in his canoe fishing on the lake. [10] a man fishing in the middle of a lake in a boat [11] A man in canoe fishing on a lake [12] A man is fishing in a small boat on a lake. [13] A man fishing alone on the lake. [14] A man is fishing on a lake. [15] A man is fishing from a boat on a lake. [16] a person fishing on a lake [17] A man is fishing alone on a small boat. [18] A person is sitting in a boat on a lake. [19] A man fishes from his small boat. [20] A man is fishing in a boat in lake [21] A man is fishing alone in a canoe. [22] A man is fishing in a canoe on a lake [23] The man is fishing on a canoe [24] A man is on a boat fishing on the lake. [25] One man fishes in a small boat on the lake. [26] A man in a canoe fishing in a calm lake. [27] A lone man sits in a boat and fishes. [28] A man fishing in a canoe. [29] A lone fisherman sits in his canoe on a river. [30] A person is fishing from a canoe. [31] man fishing in a canoe [32] A man is out fishing from a canoe on a tranquil morning. [33] a person fishes while sitting in a canoe on a lake [34] A man in a canoe is fishing on a still lake. [35] A lone fisherman in a rowboat on an empty lake. [36] A man is fishing from his canoe on quiet water. [37] A man fishing off of his canoe on a lake. [38] Someone is fishing from a canoe on a lake. [39] A small boat in the middle of the lake. [40] A guy is canoeing and fishing the middle of a tranquil and calm lake. [41] There is a man in the canoe. [42] A lone fisherman sits in a canoe with a pole in the water. [43] A person in a canoe fishing on a lake surrounded by hills. [44] A man fishing in a kayak. [45] A man is fishing in the sea by a forest. [46] A person is fishing in the water all by themselves. [47] A man is rowing a man in a river. [48] A lone fisherman in a rowboat on an empty lake.</p>



Figure 5: Ranking of 48 sentences, from highest score to lowest score, as predicted by each metric. Notice how CIDEr captures how most humans tend to describe an image (consensus) better, whereas ROUGE scores invariably favor longer, detailed sentences (less salient) and BLEU scores favor shorter sentences (lacking coverage) when used without Brevity Penalty. ROUGE₁ and BLEU₁ versions of ROUGE and BLEU are used.



CIDEr	CIDEr-D
[1] A man is fishing in a canoe on a lake. [2] A man fishing in a canoe on a lake [3] A man in canoe fishing on a lake [4] A man in his canoe fishing on the lake. [5] A man fishes in a canoe on a lake. [6] A man fishing off of his canoe on a lake. [7] A person in a canoe fishes on a lake. [8] a person fishes while sitting in a canoe on a lake [9] The man is fishing on a canoe [10] A man in a canoe is fishing. [11] A man fishing in a canoe. [12] man fishing in a canoe [13] Someone is fishing from a canoe on a lake. [14] a man fishing out of a canoe [15] A man in a canoe is fishing on a still lake. [16] A man is fishing on a lake. [17] A person is fishing from a canoe. [18] A man is fishing in a boat in lake [19] A man is on a boat fishing on the lake. [20] A man is fishing in a small boat on a lake. [21] One man fishes in a small boat on the lake. [22] A man is fishing from a boat on a lake. [23] A man in a canoe fishing in a calm lake. [24] A man is fishing alone in a canoe. [25] a man fishing in the middle of a lake in a boat [26] A person in a canoe fishing on a lake surrounded by hills. [27] a person fishing on a lake [28] A person is fishing in a boat on a lake. [29] Man in a boat fishing. [30] A man fishing alone on the lake. [31] A man fishes from his small boat. [32] A man is fishing from his canoe on quiet water. [33] A man is fishing in the river. [34] A man on a canoe fishing near a landmass. [35] A man is fishing alone on a small boat. [36] A lone man sits in a boat and fishes. [37] A guy is canoeing and fishing the middle of a tranquil and calm lake. [38] A man is out fishing from a canoe on a tranquil morning. [39] A man fishing in a kayak. [40] There is a man in the canoe. [41] A guy is canoeing and fishing the middle of a tranquil and calm lake. [42] A person is fishing in the water all by themselves. [43] A person is sitting in a boat on a lake. [44] A small boat in the middle of the lake. [45] A lone fisherman sits in his canoe on a river. [46] A lone fisherman sits in a canoe with a pole in the water. [47] A man is rowing a man in a river. [48] A lone fisherman in a rowboat on an empty lake.	[1] A man fishing in a canoe on a lake [2] A man in canoe fishing on a lake [3] A man in his canoe fishing on the lake. [4] A man fishes in a canoe on a lake. [5] A person in a canoe fishes on a lake. [6] a person fishes while sitting in a canoe on a lake [7] The man is fishing on a canoe [8] A man in a canoe is fishing. [9] A man fishing in a canoe. [10] man fishing in a canoe [11] Someone is fishing from a canoe on a lake. [12] a man fishing out of a canoe [13] A man in a canoe is fishing on a still lake. [14] A man is fishing on a lake. [15] A person is fishing from a canoe. [16] A man is fishing in a boat in lake [17] A man is on a boat fishing on the lake. [18] A man is fishing in a small boat on a lake. [19] One man fishes in a small boat on the lake. [20] A man is fishing from a boat on a lake. [21] A man in a canoe fishing in a calm lake. [22] A man is fishing alone in a canoe. [23] [24] A man is fishing in a small boat on a lake. [25] a man fishing in the middle of a lake in a boat [26] A person in a canoe fishing on a lake surrounded by hills. [27] a person fishing on a lake [28] A person is fishing in a boat on a lake. [29] Man in a boat fishing. [30] A man fishing alone on the lake. [31] A man fishes from his small boat. [32] A man is fishing from his canoe on quiet water. [33] A man is fishing in the river. [34] A man on a canoe fishing near a landmass. [35] A man is fishing alone on a small boat. [36] A lone man sits in a boat and fishes. [37] A guy is canoeing and fishing the middle of a tranquil and calm lake. [38] A man is out fishing from a canoe on a tranquil morning. [39] A man fishing in a kayak. [40] There is a man in the canoe. [41] A guy is canoeing and fishing the middle of a tranquil and calm lake. [42] A person is fishing in the water all by themselves. [43] A person is sitting in a boat on a lake. [44] A small boat in the middle of the lake. [45] A lone fisherman sits in his canoe on a river. [46] A lone fisherman sits in a canoe with a pole in the water. [47] A man is rowing a man in a river. [48] A lone fisherman in a rowboat on an empty lake.

Figure 6: Ranking of 48 sentences, from highest score to lowest score, as predicted by CIDEr₁ and CIDEr-D₁. Notice that the rankings are mostly similar qualitatively. CIDEr-D is more robust to gaming effects than CIDEr.

Ru câu táo	Ru câu táo-D
[1] Amanis câu cá trên hồ canoe. [2] Aman câu cá trên hồ canoe [3] Aman câu cá trên hồ canoe [4] Aman câu cá trên hồ canoe . [5] Aman câu cá trên hồ canoe. [6] Aman câu cá ngoài hồ canoe. [7] Aman câu cá trên hồ canoe. [8] Aman câu cá trong hồ canoe [9] Theman câu cá trên hồ canoe [10] Amaninacanoeisfishing. [11] Amanfishinginacanoe. [12] manfishinginacanoe [13] Có người đang câu cá từ acanoe trên hồ. [14] amanfishingoutofacanoe [15] Amaninacanoes đang câu cá trên hồ. [16] AmanfishingonalaKE. [17] Amanisfishingfromacanoe. [18] Amanisfishinginaboatinlake [19] Amanisboatfishingonthelake. [20] Amanisfishinginasmallboatonthelake. [21] Onemanfishesinasmallboatonthelake. [22] Amanisfishingfromaboatonthe lake. [23] Amanina câu cá bằng xuồng ở một hồ nước tĩnh lặng. [24] Amanina câu cá một mình bằng xuồng. [25] aman câu cá giữa hồ, một chiếc thuyền [26] aman câu cá bằng xuồng ở một hồ được bao quanh bởi những ngọn đồi. [27] aman câu cá trên hồ [28] aman câu cá trên thuyền, một hồ. [29] aman câu cá bằng thuyền. [30] aman câu cá một mình trên hồ. [31] aman câu cá từ chiếc thuyền nhỏ của mình. [32] aman câu cá từ chiếc xuồng của mình trên vùng nước tĩnh lặng. [33] aman câu cá trên sông. [34] aman câu cá bằng xuồng gần vùng đất thấp. [35] aman câu cá một mình trên một chiếc thuyền nhỏ. [36] aman ngồi một mình trên thuyền và câu cá. [37] aguy câu cá bằng xuồng và câu cá giữa hồ nước tĩnh lặng. [38] Aman đang câu cá ngoài khơi từ acanoe vào một buổi sáng yên tĩnh. [39] Aman đang câu cá trên thuyền. [40] Aman đang câu cá trên biển bên một khu rừng. [41] Có một người đàn ông trên xuồng. [42] Một người đang câu cá trên mặt nước một mình. [43] Một người đàn ông đang lái một chiếc thuyền trên hồ. [44] Một chiếc thuyền nhỏ giữa hồ. [45] Một người đánh cá đơn độc ngồi trên một chiếc thuyền trên sông. [46] Một người đánh cá đơn độc ngồi trên một chiếc thuyền độc mộc trên sông. [47] Aman đang chèo thuyền trên sông. [48] Một người đánh cá đơn độc trên một chiếc thuyền chèo trên mặt hồ nước trống rỗng. Một người đang tự mình câu cá trên mặt nước. [43] Một người đang lái thuyền trên hồ. [44] Một chiếc thuyền giữa hồ. [45] Một người đánh cá đơn độc ngồi trên một chiếc thuyền độc mộc trên sông. [46] Một người đánh cá đơn độc ngồi trên một chiếc thuyền độc mộc với một cây sào trên mặt nước. [47] Một người đang lái thuyền chèo trên mặt hồ nước trống rỗng. [48] Một người đang chèo thuyền trên sông.	[1] Aman đang câu cá trên hồ canoe [2] Aman đang câu cá trên hồ canoe [3] Aman đang câu cá trên hồ canoe . [4] Aman đang câu cá trên hồ canoe . [5] Aman đang câu cá trên hồ canoe . [6] Aman đang câu cá trên hồ canoe . [7] Aman đang câu cá ngoài hồ canoe của mình. [8] Ai đó đang câu cá trên hồ canoe. [9] Theman đang câu cá trong khi câu cá trên hồ canoe [10] Aman đang câu cá ngoài hồ canoe [11] Aman đang câu cá trên thuyền trong hồ [12] Aman đang câu cá trên thuyền trên hồ. [13] Aman đang câu cá trên thuyền trong hồ [14] Aman đang câu cá trên thuyền trên hồ. [15] Một người đang câu cá trên thuyền nhỏ trên hồ. [16] Aman đang câu cá trên thuyền trên hồ. [17] Aman đang câu cá trên hồ M11. [18] AmanisfishingonalaKE. [19] Amanfishinginacanoe. [20] Amanis câu cá một mình trên xuồng. [21] Amanina câu cá trên xuồng ở hồ bình tĩnh. [22] Amanis câu cá từ xuồng. [23] Aman đang câu cá trên một chiếc thuyền nhỏ trên hồ. [24] Aman đang câu cá trên một chiếc thuyền trên hồ. [25] một người đang câu cá trên một chiếc xuồng [26] một người đang câu cá giữa hồ trên một chiếc thuyền [27] một người đang câu cá một mình trên hồ. [28] một người đang câu cá trên hồ [29] một người đang câu cá trên một chiếc xuồng được bao quanh bởi những ngọn đồi. [30] một người đang câu cá từ chiếc thuyền nhỏ của mình. [31] Aman đang câu cá từ chiếc xuồng của mình trên vùng nước yên tĩnh. [32] Aman đang câu cá một mình trên một chiếc thuyền nhỏ. [33] Aman đang câu cá bằng xuồng gần một vùng đất thấp. [34] Một mình ngồi trên một chiếc thuyền và câu cá. [35] Aman đang câu cá trên sông. [36] Man đang câu cá bằng thuyền. [37] Aman đang câu cá trên biển gần một khu rừng. [38] Aman đang câu cá ngoài khơi từ một con sông vào một buổi sáng yên tĩnh. [39] Aman đang câu cá trên thuyền. [40] Có một người đàn ông trên xuồng. [41] Aman đang câu cá bằng xuồng và câu cá giữa hồ yên tĩnh và tĩnh lặng. [42] Một người đang tự mình câu cá trên mặt nước. [43] Một người đang lái thuyền trên hồ. [44] Một chiếc thuyền giữa hồ. [45] Một người đánh cá đơn độc ngồi trên một chiếc thuyền độc mộc trên sông. [46] Một người đánh cá đơn độc ngồi trên một chiếc thuyền độc mộc với một cây sào trên mặt nước. [47] Một người đang lái thuyền chèo trên mặt hồ nước trống rỗng. [48] Một người đang chèo thuyền trên sông.

Hình 6: Xếp hạng 48 câu, từ điểm cao nhất đến thấp nhất, theo dự đoán của CIDEr₁ và CIDEr-D₁. Lưu ý rằng xếp hạng về mặt chất lượng chủ yếu tương tự nhau. CIDEr-D mạnh mẽ hơn đối với các hiệu ứng chơi game so với CIDEr.

Reference Sentences	Candidate Sentences
<p>A baby girl laughs at the camera</p> <p>A woman is getting a baby girl to smile for the camera.</p> <p>A mom is smiling with a baby.</p> <p>A woman sits down next to a baby sitting on the table.</p> <p>A woman smiles at a baby who is sitting on a table.</p> <p>A woman sits with a baby at a table.</p> <p>A baby girl is sitting on a table and smiling.</p> <p>A baby is sitting on the counter smiling while her mom looks on.</p> <p>A woman in spongebob scrub is smiling at a baby in a blue dress.</p> <p>A baby is sitting on a table with her blond mom smiling at her.</p>	<p>[1] A woman with a smiling baby sitting on the table.</p> <p>[2] A tiny blond child in a blue dress sits on a table near her mother.</p>
<p>Multiple cows graze in the open field of grass.</p> <p>Black cows graze in the pasture.</p> <p>Black cows graze in a green pasture.</p> <p>Cows are grazing in a grassy field.</p> <p>Black cows are eating a lot of grass.</p> <p>A herd of cows eats grass.</p> <p>Black cows are grazing in a field.</p> <p>Several black cows wander in a green pasture.</p> <p>Cattle graze in a green pasture near a tall tree.</p> <p>Black cows are grazing in a field in front of a tree.</p>	<p>[1] A number of black cows grazing in front of a large tree.</p> <p>[2] Black cows graze on green grass.</p>
<p>A dog sitting idly on a floral pattern chair.</p> <p>A little dog sits on a flower cushion.</p> <p>A dog relax on a flower patterned chair outside.</p> <p>A dog with bell collar sits on a flower pillow.</p> <p>A dog lying on a flower patterned chair.</p> <p>A dog sitting on a floral chair.</p> <p>A brown and white dog is lying on a floral print chair.</p> <p>A dog is lying on a flower couch.</p> <p>A small dog lying on a flowery cushion stares at the camera.</p> <p>A dog with a bell collar sits on the chair</p>	<p>[1] Brown and white dog with a bell on black collar.</p> <p>[2] A small orange and white dog with a collar and a bell relaxing on a flower print pillow.</p>

Figure 7: Reference sentences shown in **bold** are those which are rated as more similar to the winning candidate sentence, also shown in **bold**, via the triplet interface. The candidate sentence not shown in bold is the one picked by the pairwise interface, which captures “better”. This illustrates the difference between human-like *versus* what humans like.

Câu tham khảo	Ứng cử viên
<p>Một cô bé cười khúc khích trước ống kính</p> <p>Một người phụ nữ thật tuyệt vời khi mỉm cười trước ống kính.</p> <p>Tôi nhớ em bé.</p> <p>Awomansitsdownnexttobabysis(nonatable).</p> <p>Một người phụ nữ ngồi với một đứa bé bất động.</p> <p>Một cô bé dễ thương và tươi cười.</p> <p>Ababyissi (trên quây mỉm cười trong khi mẹ cô nhìn về phía trước).</p> <p>Một người phụ nữ mặc váy xanh.</p> <p>Em bé đang ngồi cùng bàn với bà mẹ tóc vàng và mỉm cười.</p>	<p>[1] Awomanwithasmilingbabysi(ngonatable).</p> <p>[2] A: đứa trẻ tóc vàng mặc váy xanh ngồi trên bàn gần mẹ.</p>
<p>Dàn bò MulEplecow gồm có trên cánh đồng trống.</p> <p>Dàn bò đen gặm cỏ trên đồng cỏ.</p> <p>Đồng cỏ Blackcows và đồng cỏ xanh.</p> <p>Bò đang gặm cỏ trên đồng cỏ.</p> <p>BlackcowsareeaEngalotofgrass.</p> <p>Cô chăn bò.</p> <p>Bò đen đang gặm cỏ trên cánh đồng.</p> <p>Một số con bò đen, con lách bạch, đồng cỏ xanh.</p> <p>Cây-legraeinagreenpasturegầnraalltree.</p> <p>Những con bò đen đang gặm cỏ trên cánh đồng trước một cái cây.</p>	<p>[1] Một số con bò đen gặm cỏ trước một cái cây lớn.</p> <p>[2] Đồng cỏ bò đen trên cỏ xanh.</p>
<p>Ghế bành hoa Adogsi.</p> <p>Alioledogsitsonflowercushion.</p> <p>Ghế Adogrelaxonaflowerpa@erned bên ngoài.</p> <p>Một chú chó đeo vòng cổ hình chuông và một chiếc gối hình hoa.</p> <p>Một chiếc ghế tựa hình hoa.</p> <p>Ghế Adogsi (ngonafloral chair).</p> <p>Một chú chó màu nâu và trắng đang nằm trên chiếc ghế in hoa.</p> <p>Adogsi đang nằm trên chiếc ghế dài đầy hoa.</p> <p>Một chú chó nhỏ nằm trên chiếc đệm hoa nhìn chằm chằm vào máy ảnh.</p> <p>Một con chó đeo vòng cổ hình quả chuông ngồi trên ghế</p>	<p>[1] Con chó màu nâu và trắng có vòng cổ màu đen.</p> <p>[2] Một chú chó nhỏ màu cam và trắng có vòng cổ và chuông đang thư giãn trên chiếc gối in hoa.</p>

Hình 7: Các câu tham chiếu được in đậm là những câu được đánh giá là giống với câu ứng cử viên chiến thắng hơn, cũng được hiển thị bằng chữ in đậm, thông qua giao diện bộ ba. Câu ứng viên không được hiển thị bằng chữ in đậm là câu được chọn bởi cặp Giao diện, nắm bắt “tốt hơn”. Điều này minh họa sự khác biệt giữa giao diện giống con người với giao diện mà con người thích.

Midge	Babytalk	Story	Video	Video+
This is a picture of one person, one sofa and one dog. The person is against the brown sofa. The dog is near the person, and beside the brown sofa.	China doll in a leather recliner. people posing in a restaurant	a man at a table at a restaurant		

Figure 10: Descriptions produced by Midge [29], Babytalk [22], Story [14], Video [37] and Video+ [37] for an image. Note that since Story is a retrieval based approach, we consider the top-ranked output to show here.

ruồi muỗi	Nói chuyện với em bé	Câu chuyện	Bảng hình	Video+
Đây là một bức ảnh của một người, một chiếc ghế sofa và một chó. Ngươi đó dựa vào ghế sofa màu nâu. Chó ở gần ngươi đó và bên cạnh ghế sofa màu nâu.	Búp bê Trung Quốc ngồi trên ghế bành bằng da.	một người đàn ông ở một bàn tại một nhà hàng		

Hình 10: Mô tả được tạo ra bởi Midge [29], Babytalk [22], Story [14], Video [37] và Video+ [37] cho một hình ảnh. Lưu ý vì Story là phương pháp tiếp cận dựa trên truy xuất nên chúng tôi sẽ xem xét kết quả được xếp hạng cao nhất để hiển thị ở đây.

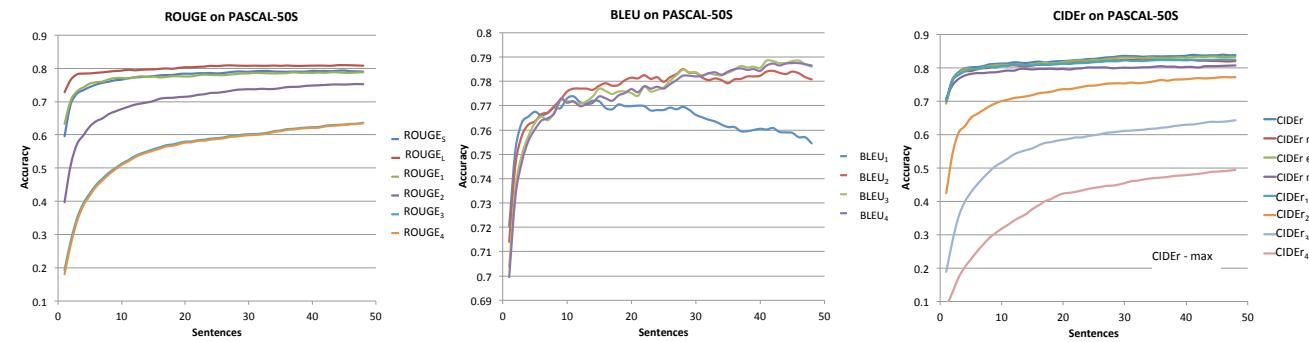
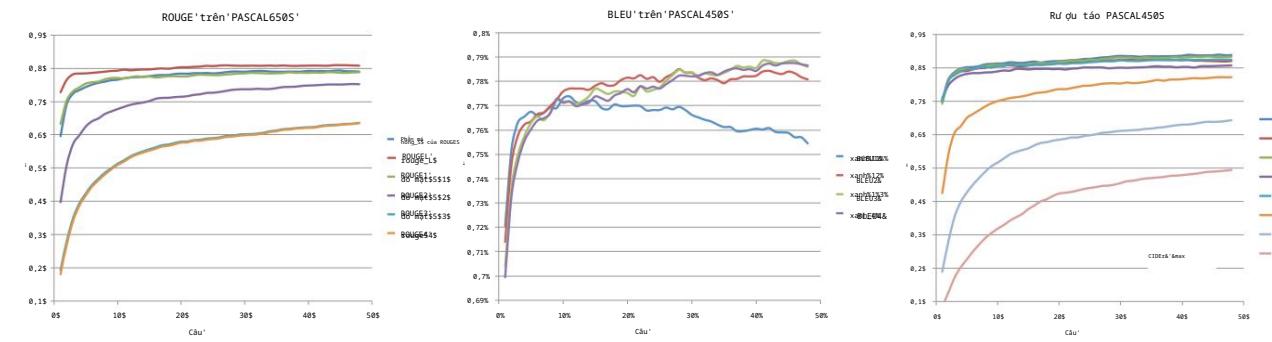


Figure 11: Performance of different versions of metrics on PASCAL-50S



Hình 11: Hiệu suất của các phiên bản số liệu khác nhau trên PASCAL-50S

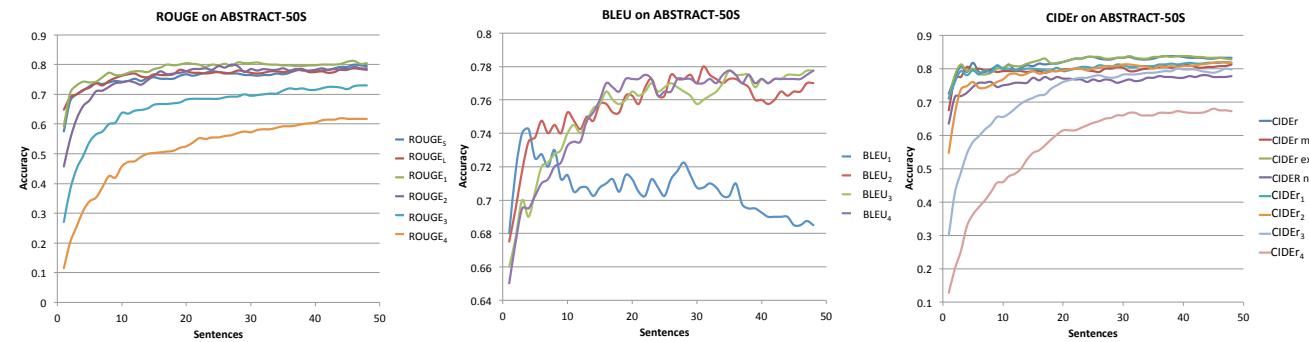
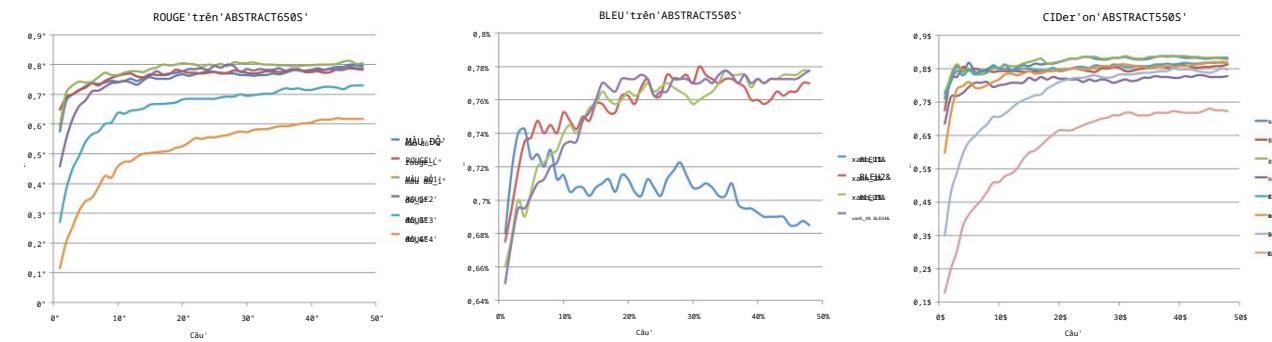


Figure 12: Performance of different versions of metrics on ABSTRACT-50S



Hình 12: Hiệu suất của các phiên bản số liệu khác nhau trên ABSTRACT-50S