Object Hallucination in Image Captioning

Anna Rohrbach[*1], **Lisa Anne Hendricks**[*1],
Kaylee Burns[1], **Trevor Darrell**[1], **Kate Saenko**[2]
[1] UC Berkeley, [2] Boston University

Abstract

Despite continuously improving performance, contemporary image captioning models are prone to "hallucinating" objects that are not actually in a scene. One problem is that standard metrics only measure similarity to ground truth captions and may not fully capture image relevance. In this work, we propose a new image relevance metric to evaluate current models with veridical visual labels and assess their rate of object hallucination. We analyze how captioning model architectures and learning objectives contribute to object hallucination, explore when hallucination is likely due to image misclassification or language priors, and assess how well current sentence metrics capture object hallucination. We investigate these questions on the standard image captioning benchmark, MSCOCO, using a diverse set of models. Our analysis yields several interesting findings, including that models which score best on standard sentence metrics do not always have lower hallucination and that models which hallucinate more tend to make errors driven by language priors.

1 Introduction

Image captioning performance has dramatically improved over the past decade. Despite such impressive results, it is unclear to what extent captioning models actually rely on image content: as we show, existing metrics fall short of fully capturing the captions' relevance to the image. In Figure 1 we show an example where a competitive captioning model, Neural Baby Talk (NBT) (Lu et al., 2018), incorrectly generates the object "bench." We refer to this issue as object *hallucination*.

While missing salient objects is also a failure mode, captions are summaries and thus generally not expected to describe all objects in the scene. On the other hand, describing objects that are *not present* in the image has been shown to be less preferable to humans. For example, the LSMDC challenge (Rohrbach et al., 2017a) documents that correctness is more important to human judges than specificity. In another study, (MacLeod et al., 2017) analyzed how visually impaired people react to automatic image captions. They found that people vary in their preference of either coverage or correctness. For many visually impaired who value correctness over coverage, hallucination is an obvious concern.

Besides being poorly received by humans, object hallucination reveals an internal issue of a captioning model, such as not learning a very good representation of the visual scene or overfitting to its loss function.

In this paper we assess the phenomenon of object hallucination in contemporary captioning models, and consider several key questions. The

NBT: A woman talking on a cell phone while sitting on a ***bench***.
CIDEr: **0.87**, METEOR: 0.23, SPICE: **0.22,** CHs: **1.00**, CHi: **0.33**

TopDown: A woman is talking on a cell phone.
CIDEr: 0.54, METEOR: **0.26**, SPICE: 0.13, CHs: **0.00**, CHi: **0.00**

Figure 1: Image captioning models often "hallucinate" objects that may appear in a given context, like e.g. a *bench* here. Moreover, the sentence metrics do not always appropriately penalize such hallucination. Our proposed metrics (CHAIRs and CHAIRi) reflect hallucination. For CHAIR *lower is better*.

* Denotes equal contribution.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4035–4045
Brussels, Belgium, October 31 - November 4, 2018.

Ảo giác đối tượng trong chú thích hình ảnh

Anna Rohrbach 1 ,Lisa Anne Hendricks 1 ,
Kaylee Burns1 ,Trevor Darrell1,Kate Saenko2
[1] Đại học California, Berkeley, [2] Đại học Boston

Tóm tắt

Mặc dù liên tục cải thiện hiệu suất,
các mô hình chú thích hình ảnh đương đại là
dễ bị "ảo giác" về những vật thể không phải là
thực sự trong một cảnh. Một vấn đề là các số liệu chuẩn
chỉ đo lường sự tương đồng với mặt đất
chú thích sự thật và có thể không nắm bắt đầy đủ sự liên
quan của hình ảnh-tuổi. Trong tác phẩm này, chúng tôi đề xuất một
thước đo mức độ liên quan của hình ảnh mới để đánh giá hiện tại
các mô hình với nhãn trực quan chân thực và đánh giá
tỷ lệ ảo giác đối tượng của họ. Chúng tôi phân tích
cách các kiến trúc mô hình chú thích và mục tiêu học
tập góp phần vào ảo giác đối tượng, khám phá khi nào
ảo giác có khả năng xảy ra
để hình ảnh phân loại sai hoặc ngôn ngữ trước đó,
và đánh giá mức độ hiệu quả của các số liệu câu hiện tại
bắt giữ ảo giác đối tượng. Chúng tôi điều tra
những câu hỏi này về chuẩn mực chú thích hình ảnh,
MSCOCO, sử dụng một bộ đa dạng
của các mô hình. Phân tích của chúng tôi đưa ra một số
phát hiện thú vị, bao gồm các mô hình
đạt điểm cao nhất về số liệu câu chuẩn không
luôn có ảo giác thấp hơn và các mô hình tạo ra nhiều
ảo giác hơn có xu hướng mắc lỗi
được thúc đẩy bởi các kiến thức ngôn ngữ trước đó.

1 Giới thiệu

Hiệu suất chú thích hình ảnh đã tăng đáng kể
được cải thiện trong thập kỷ qua. Mặc dù như vậy
kết quả ấn tượng, không rõ ở mức độ nào
các mô hình chú thích thực sự dựa vào nội dung hình ảnh:
như chúng tôi trình bày, các số liệu hiện có không đạt yêu cầu
nắm bắt đầy đủ sự liên quan của chú thích với hình ảnh.
Trong Hình 1, chúng tôi trình bày một ví dụ trong đó
mô hình chú thích cạnh tranh, Neural Baby Talk
(NBT) (Lu et al., 2018), tạo ra không chính xác
đối tượng "băng ghế." Chúng tôi gọi vấn đề này là đối tượng
ảo giác.

Trong khi thiếu các đối tượng nổi bật cũng là một thất bại
chế độ, chú thích là tóm tắt và do đó nói chung

NBT: Một người phụ nữ đang nói chuyện điện thoại di động trong khi ngồi trên băng ghế.
CIDEr: 0,87, METEOR: 0,23, SPICE: 0,22, CHs: 1,00, CHi: 0,33

TopDown: Một người phụ nữ đang nói chuyện trên điện thoại di động.
CIDEr: 0,54, METEOR: 0,26, SPICE: 0,13, CHs: 0,00, CHi: 0,00

Hình 1: Các mô hình chú thích hình ảnh thường "ảo giác"
các đối tượng có thể xuất hiện trong một bối cảnh nhất định, như ví dụ a
ghế ở đây. Hơn nữa, các số liệu của bản án không phải lúc nào
cũng phạt thích đáng đối với ảo giác như vậy.
số liệu đề xuất (CHAIR và CHAIRi) phản ánh ảo giác. Đối với CHAIR
thì thấp hơn là tốt hơn.

không mong đợi mô tả được tất cả các đối tượng trong cảnh.
Mặt khác, mô tả các đối tượng không phải là
có trong hình ảnh đã được chứng minh là ít hơn
thích hợp hơn với con người. Ví dụ, LSMDC
thách thức (Rohrbach và cộng sự, 2017a) chứng minh rằng
sự chính xác quan trọng hơn đối với các thẩm phán con người
hơn là tính đặc hiệu. Trong một nghiên cứu khác, (MacLeod et al.,
2017) đã phân tích cách những người khiếm thị phản ứng với
chú thích hình ảnh tự động. Họ thấy rằng
mọi người có sở thích khác nhau về phạm vi bảo hiểm
hoặc tính chính xác. Đối với nhiều người khiếm thị
giá trị chính xác hơn phạm vi bao phủ, ảo giác là
một mối quan tâm rõ ràng.

Bên cạnh việc không được con người đón nhận, ảo giác
đối tượng còn bộc lộ một vấn đề nội tại của mô hình chú
thích, chẳng hạn như không học được một điều gì đó rất hay.
sự biểu diễn của cảnh tượng trực quan hoặc sự phù hợp quá mức với
hàm mất mát của nó.

Trong bài báo này chúng tôi đánh giá hiện tượng
ảo giác đối tượng trong chú thích đương đại
mô hình và xem xét một số câu hỏi chính.

* Biểu thị sự đóng góp ngang nhau.

Biên bản Hội nghị năm 2018 về Phương pháp thực nghiệm trong Xử lý ngôn ngữ tự nhiên, trang 4035–4045
Brussels, Bỉ, 31 tháng 10 - 4 tháng 11 năm 2018.

first question we aim to answer is: *Which models are more prone to hallucination?* We analyze this question on a diverse set of captioning models, spanning different architectures and learning objectives. To measure object hallucination, we propose a new metric, *CHAIR (Caption Hallucination Assessment with Image Relevance)*, which captures image relevance of the generated captions. Specifically, we consider both ground truth object annotations (MSCOCO Object segmentation (Lin et al., 2014)) and ground truth sentence annotations (MSCOCO Captions (Chen et al., 2015)). Interestingly, we find that models which score best on standard sentence metrics do not always hallucinate less.

The second question we raise is: *What are the likely causes of hallucination?* While hallucination may occur due to a number of reasons, we believe the top factors include visual misclassification and over-reliance on language priors. The latter may result in memorizing which words "go together" regardless of image content, which may lead to poor generalization, once the test distribution is changed. We propose *image and language model consistency* scores to investigate this issue, and find that models which hallucinate more tend to make mistakes consistent with a language model.

Finally, we ask: *How well do the standard metrics capture hallucination?* It is a common practice to rely on automatic sentence metrics, e.g. CIDEr (Vedantam et al., 2015), to evaluate captioning performance during development, and few employ human evaluation to measure the final performance of their models. As we largely rely on these metrics, it is important to understand how well they capture the hallucination phenomenon. In Figure 1 we show how two sentences, from NBT with hallucination and from TopDown model (Anderson et al., 2018) – without, are scored by the standard metrics. As we see, hallucination is not always appropriately penalized. We find that by using additional ground truth data about the image in the form of object labels, our metric CHAIR allows us to catch discrepancies that the standard captioning metrics cannot fully capture. We then investigate ways to assess object hallucination risk with the standard metrics. Finally, we show that CHAIR is complementary to the standard metrics in terms of capturing human preference.

2 Caption Hallucination Assessment

We first introduce our image relevance metric, *CHAIR*, which assesses captions w.r.t. objects that are actually in an image. It is used as a main tool in our evaluation. Next we discuss the notions of *image and language model consistency*, which we use to reason about the causes of hallucination.

2.1 The CHAIR Metric

To measure object hallucination, we propose the *CHAIR (Caption Hallucination Assessment with Image Relevance)* metric, which calculates what proportion of words generated are actually in the image according to the ground truth sentences and object segmentations. This metric has two variants: per-instance, or what fraction of object instances are hallucinated (denoted as CHAIRi), and per-sentence, or what fraction of sentences include a hallucinated object (denoted as CHAIRs):

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}$$

$$\text{CHAIR}_s = \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{ all sentences}\}|}$$

For easier analysis, we restrict our study to the 80 MSCOCO objects which appear in the MSCOCO segmentation challenge. To determine whether a generated sentence contains hallucinated objects, we first tokenize each sentence and then singularize each word. We then use a list of synonyms for MSCOCO objects (based on the list from Lu et al. (2018)) to map words (e.g., "player") to MSCOCO objects (e.g., "person"). Additionally, for sentences which include two word compounds (e.g., "hot dog") we take care that other MSCOCO objects (in this case "dog") are not incorrectly assigned to the list of MSCOCO objects in the sentence. For each ground truth sentence, we determine a list of MSCOCO objects in the same way. The MSCOCO segmentation annotations are used by simply relying on the provided object labels.

We find that considering both sources of annotation is important. For example, MSCOCO contains an object "dining table" annotated with segmentation maps. However, humans refer to many different kinds of objects as "table" (e.g., "coffee table" or "side table"), though these objects are not annotated as they are not specifically "dining table". By using sentence annotations to

Câu hỏi đầu tiên chúng tôi muốn trả lời là: Những mô hình
nào dễ gây ảo giác hơn? Chúng tôi phân tích
câu hỏi này về một tập hợp đa dạng các mô hình chú thích,
trải dài trên các kiến trúc và mục tiêu học tập khác nhau.
Để đo lường ảo giác đối tượng, chúng tôi đề xuất một số
liệu mới, CHAIR (Tiêu đề Ảo giác
Đánh giá mức độ liên quan của hình ảnh), ghi lại mức độ
liên quan của hình ảnh trong các chú thích được tạo.
Cụ thể, chúng tôi xem xét cả đối tượng thực tế
chú thích (Phân đoạn đối tượng MSCOCO (Lin
et al., 2014)) và chú thích câu sự thật cơ bản (Chú thích
MSCOCO (Chen et al., 2015)). Điều thú vị là chúng tôi thấy
rằng các mô hình đạt điểm cao nhất
về các phép đo câu chuẩn không phải lúc nào cũng ít gây ảo
giác hơn.

Câu hỏi thứ hai chúng tôi nêu ra là: Những gì là
nguyên nhân có thể gây ra ảo giác? Mặc dù ảo giác có thể
xảy ra do một số lý do, chúng tôi
tin rằng các yếu tố hàng đầu bao gồm phân loại sai trực
quan và quá phụ thuộc vào các tiên nghiệm ngôn ngữ.
sau này có thể dẫn đến việc ghi nhớ những từ nào "đi
cùng nhau" bất kể nội dung hình ảnh, có thể
dẫn đến sự khái quát kém, một khi sự phân phối thử nghiệm
bị thay đổi. Chúng tôi đề xuất điểm số nhất quán của mô
hình hình ảnh và ngôn ngữ để điều tra điều này
vấn đề, và tìm thấy rằng các mô hình gây ảo giác nhiều hơn
có xu hướng mắc lỗi phù hợp với ngôn ngữ
người mẫu.

Cuối cùng, chúng tôi hỏi: Tiêu chuẩn được thực hiện tốt như thế nào?
số liệu thống kê nắm bắt ảo giác? Đó là một phổ biến
thực hành dựa vào các số liệu câu tự động,
ví dụ CIDEr (Vedantam et al., 2015), để đánh giá
hiệu suất chú thích trong quá trình phát triển và
ít người sử dụng đánh giá của con người để đo lường hiệu
suất cuối cùng của mô hình của họ. Vì chúng tôi phần lớn
dựa vào các số liệu này, điều quan trọng là phải hiểu chúng
nắm bắt hiện tượng ảo giác tốt như thế nào. Trong Hình 1,
chúng tôi trình bày cách hai câu, từ NBT với ảo giác và từ

Mô hình TopDown (Anderson et al., 2018) - không có, được
chấm điểm theo các số liệu chuẩn. Như chúng tôi
thấy đấy, ảo giác không phải lúc nào cũng bị phạt thích
đáng. Chúng tôi thấy rằng bằng cách sử dụng căn cứ bổ sung
dữ liệu sự thật về hình ảnh dưới dạng nhãn đối tượng, số
liệu CHAIR của chúng tôi cho phép chúng tôi nắm bắt những
điểm khác biệt mà số liệu chú thích tiêu chuẩn không thể
nắm bắt hoàn toàn. Sau đó chúng tôi điều tra các cách để đánh giá
nguy cơ ảo giác vật thể với các số liệu chuẩn.
Cuối cùng, chúng tôi chỉ ra rằng CHAIR bổ sung cho
các số liệu chuẩn về việc nắm bắt con người
sự ưa thích.

2 Đánh giá ảo giác chú thích

Đầu tiên chúng tôi giới thiệu số liệu liên quan đến hình ảnh của chúng tôi,
CHAIR, đánh giá các chú thích liên quan đến các đối tượng
thực sự có trong một hình ảnh. Nó được sử dụng như một công cụ chính
trong đánh giá của chúng tôi. Tiếp theo chúng tôi thảo luận về các khái niệm
sự nhất quán của mô hình hình ảnh và ngôn ngữ, mà chúng tôi
dùng để lý giải nguyên nhân gây ra ảo giác.

2.1 Chỉ số CHAIR

Để đo lường ảo giác đối tượng, chúng tôi đề xuất
GHẾ (Chú thích Đánh giá ảo giác với
Số liệu về mức độ liên quan của hình ảnh, tính toán những gì
tỷ lệ các từ được tạo ra thực sự nằm trong
hình ảnh theo các câu sự thật cơ bản và
phân đoạn đối tượng. Số liệu này có hai biến thể: mỗi
trường hợp, hoặc phần nào của các trường hợp đối tượng
được ảo giác (được biểu thị là CHAIRi) và
mỗi câu, hoặc bao nhiêu phần của câu bao gồm
một vật thể ảo giác (được biểu thị là GHẾ):

$$\text{CHAIRi} = \frac{|\{\text{vật thể ảo giác}\}|}{|\{\text{tất cả các đối tượng được đề cập}\}|}$$

$$\text{GHẾ} = \frac{|\{\text{câu có tân ngữ ảo giác}\}|}{|\{\text{ tất cả các câu}\}|}$$

Để phân tích dễ dàng hơn, chúng tôi giới hạn nghiên cứu của mình vào
80 đối tượng MSCOCO xuất hiện trong
Thách thức phân đoạn MSCOCO. Để xác định
cho dù một câu được tạo ra có chứa các đối tượng ảo giác
hay không, trước tiên chúng ta sẽ mã hóa từng câu
và sau đó đơn lẻ hóa từng từ. Sau đó chúng ta sử dụng
danh sách các từ đồng nghĩa cho các đối tượng MSCOCO (dựa trên
trong danh sách từ Lu et al. (2018)) để lập bản đồ các từ
(ví dụ, "player") đến các đối tượng MSCOCO (ví dụ, "person"). Ngoài ra, đối với các câu bao gồm các hợp chất hai
từ (ví dụ, "hot dog"), chúng tôi
hãy cẩn thận rằng các đối tượng MSCOCO khác (trong này
trường hợp "chó") không được gán sai cho
danh sách các đối tượng MSCOCO trong câu. Đối với
mỗi câu sự thật cơ bản, chúng tôi xác định một danh sách
của các đối tượng MSCOCO theo cùng một cách.
Chú thích phân đoạn MSCOCO được sử dụng bởi
chỉ cần dựa vào các nhãn đối tượng được cung cấp.

Chúng tôi thấy rằng việc xem xét cả hai nguồn chú thích
đều quan trọng. Ví dụ, MSCOCO
chứa một đối tượng "bàn ăn" được chú thích bằng
bản đồ phân đoạn. Tuy nhiên, con người tham khảo
nhiều loại đối tượng khác nhau như "bảng" (ví dụ,
"bàn cà phê" hoặc "bàn phụ"), mặc dù các đối tượng này
không được chú thích vì chúng không được chú thích cụ thể
"bàn ăn". Bằng cách sử dụng chú thích câu để

scrape ground truth objects, we account for variation in how human annotators refer to different objects. Inversely, we find that frequently humans will not mention all objects in a scene. Qualitatively, we observe that both annotations are important to capture hallucination. Empirically, we verify that using only segmentation labels or only reference captions leads to higher hallucination (and practically incorrect) rates.

2.2 Image Consistency

We define a notion of *image consistency*, or how consistent errors from the captioning model are with a model which predicts objects based on an image alone. To measure image consistency for a particular generated word, we train an image model and record $P(w|I)$ or the probability of predicting the word given only the image. To score the image consistency of a caption we use the average of $P(w|I)$ for all MSCOCO objects, where higher values mean that errors are *more* consistent with the image model. Our image model is a multi-label classification model with labels corresponding to MSCOCO objects (labels determined the same way as is done for CHAIR) which shares the visual features with the caption models.

2.3 Language Consistency

We also introduce a notion of *language consistency*, i.e. how consistent errors from the captioning model are with a model which predicts words based only on previously generated words. We train an LSTM (Hochreiter and Schmidhuber, 1997) based language model which predicts a word w_t given previous words $w_{0:t-1}$ on MSCOCO data. We report language consistency as $1/R(w_t)$ where $R(w_t)$ is the rank of the predicted word in the language model. Again, for a caption we report average rank across all MSCOCO objects in the sentence and higher language consistency implies that errors are *more* consistent with the language model.

We illustrate image and language consistency in Figure 2, i.e. the hallucination error ("fork") is more consistent with the Language Model predictions than with the Image Model predictions. We use these consistency measures in Section 3.3 to help us investigate the causes of hallucination.

Image Model predictions:
bowl, broccoli, carrot, dining table

Language Model predictions for the last word:
fork, spoon, bowl

Generated caption: A plate of food with broccoli and a ***fork***.

Figure 2: Example of image and language consistency. The hallucination error ("fork") is more consistent with the Language Model.

3 Evaluation

In this section we present the findings of our study, where we aim to answer the questions posed in Section 1: *Which models are more prone to hallucination? What are the likely causes of hallucination? How well do the standard metrics capture hallucination?*

3.1 Baseline Captioning Models

We compare object hallucination across a wide range of models. We define two axes for comparison: model architecture and learning objective.

Model architecture. Regarding model architecture, we consider models both with and without attention mechanisms. In this work, we use "attention" to refer to any mechanism which learns to focus on different image regions, whether image regions be determined by a high level feature map, or by object proposals from a trained detector. All models are end-to-end trainable and use a recurrent neural network (LSTM (Hochreiter and Schmidhuber, 1997) in our case) to output text. For non-attention based methods we consider the **FC model** from Rennie et al. (2017) which incorporates visual information by initializing the LSTM hidden state with high level image features. We also consider **LRCN** (Donahue et al., 2015) which considers visual information at each time step, as opposed to just initializing the LSTM hidden state with extracted features.

For attention based models, we consider **Att2In** (Rennie et al., 2017), which is similar to the original attention based model proposed by (Xu et al., 2015), except the image feature is only input into the cell gate as this was shown to lead to better performance. We then consider the attention model proposed by (Anderson et al., 2018) which proposes a specific "top-down attention" LSTM as well as a "language" LSTM.

đối tượng sự thật cơ bản, chúng tôi tính đến sự khác biệt trong cách chú thích của con người tham chiếu đến các đối tượng khác nhau các đối tượng. Ngược lại, chúng ta thấy rằng con người thường xuyên sẽ không đề cập đến tất cả các đối tượng trong một cảnh. Về mặt định tính, chúng tôi quan sát thấy rằng cả hai chú thích đều quan trọng để nắm bắt ảo giác. Theo kinh nghiệm, chúng tôi xác minh rằng chỉ sử dụng nhãn phân đoạn hoặc chỉ chú thích tham chiếu dẫn đến ảo giác cao hơn (và (thực tế là không chính xác).

2.2 Sự nhất quán của hình ảnh

Chúng tôi định nghĩa một khái niệm về tính nhất quán của hình ảnh hoặc cách lỗi nhất quán từ mô hình chú thích là với một mô hình dự đoán các đối tượng dựa trên hình ảnh một mình. Để đo tính nhất quán của hình ảnh cho một từ cụ thể được tạo ra, chúng tôi đào tạo một hình ảnh mô hình và ghi lại P(w|I) hoặc xác suất dự đoán từ chỉ dựa vào hình ảnh. Để ghi điểm tính nhất quán của hình ảnh trong chú thích chúng tôi sử dụng giá trị trung bình của P(w|I) cho tất cả các đối tượng MSCOCO, trong đó giá trị cao hơn có nghĩa là lỗi nhất quán hơn với mô hình hình ảnh. Mô hình hình ảnh của chúng tôi là mô hình phân loại đa nhãn với các nhãn tương ứng với các đối tượng MSCOCO (nhãn được xác định cùng một cách như được thực hiện đối với CHAIR) chia sẻ các tính năng trực quan với các mô hình chú thích.

2.3 Sự nhất quán của ngôn ngữ

Chúng tôi cũng giới thiệu một khái niệm về tính nhất quán của ngôn ngữ, tức là mức độ nhất quán của các lỗi từ mô hình chú thích với một mô hình dự đoán từ chỉ dựa trên các từ đã tạo trước đó. Chúng tôi đào tạo một mô hình ngôn ngữ dựa trên LSTM (Hochreiter và Schmidhu-ber, 1997) dự đoán một từ wt dựa trên các từ trước đó w0:t 1 trên dữ liệu MSCOCO. Chúng tôi báo cáo tính nhất quán của ngôn ngữ là 1/R(wt) trong đó R(wt) là thứ hạng của từ dự đoán trong mô hình ngôn ngữ. Một lần nữa, đối với một chú thích, chúng tôi báo cáo thứ hạng trung bình trên tất cả Các đối tượng MSCOCO trong câu và tính nhất quán ngôn ngữ cao hơn ngụ ý rằng lỗi nhiều hơn phù hợp với mô hình ngôn ngữ.

Chúng tôi minh họa sự nhất quán của hình ảnh và ngôn ngữ trong Hình 2, tức là lỗi ảo giác ("fork") là phù hợp hơn với các dự đoán của Mô hình Ngôn ngữ so với các dự đoán của Mô hình Hình ảnh. Chúng tôi sử dụng các biện pháp nhất quán này trong Phần 3.3 để giúp chúng tôi điều tra nguyên nhân gây ra ảo giác.

Dự đoán mô hình hình ảnh:
bát, bông cải xanh, cà rốt, bàn ăn

Dự đoán Mô hình ngôn ngữ cho từ cuối cùng:
nĩa, thìa, bát

Tạo chú thích: Một đĩa thức ăn với bông cải xanh và một cái nĩa.

Hình 2: Ví dụ về tính nhất quán của hình ảnh và ngôn ngữ. Lỗi ảo giác ("cái nĩa") phù hợp hơn với Mô hình ngôn ngữ.

3 Đánh giá

Trong phần này chúng tôi trình bày những phát hiện của nghiên cứu của chúng tôi, nơi chúng tôi muốn trả lời những câu hỏi được đặt ra trong Phần 1: Những mô hình nào dễ bị ảo giác hơn? Những nguyên nhân có thể gây ra ảo giác là gì? Các số liệu chuẩn nắm bắt tốt như thế nào ảo giác?

3.1 Mô hình chú thích cơ bản

Chúng tôi so sánh ảo giác đối tượng trên một phạm vi rộng phạm vi mô hình. Chúng tôi xác định hai trục để so sánh: kiến trúc mô hình và mục tiêu học tập.

Kiến trúc mô hình. Về kiến trúc mô hình, chúng tôi xem xét các mô hình có và không có cơ chế chú ý. Trong tác phẩm này, chúng tôi sử dụng "attention" để chỉ bất kỳ cơ chế nào học được để tập trung vào các vùng hình ảnh khác nhau, cho dù các vùng hình ảnh có được xác định bởi một tính năng cấp cao hay không bản đồ, hoặc theo đề xuất đối tượng từ một máy dò được đào tạo. Tất cả các mô hình đều có thể đào tạo từ đầu đến cuối và sử dụng mạng nơ-ron hồi quy (LSTM (Hochre-iter và Schmidhuber, 1997) trong trường hợp của chúng tôi) để đưa ra văn bản. Đối với các phương pháp không dựa trên sự chú ý, chúng tôi xem xét mô hình FC từ Rennie et al. (2017) mà kết hợp thông tin trực quan bằng cách khởi tạo Trạng thái ẩn LSTM với các đặc điểm hình ảnh cấp cao. Chúng tôi cũng xem xét LRCN (Donahue et al., 2015) xem xét thông tin trực quan tại mỗi thời điểm bước, trái ngược với việc chỉ khởi tạo trạng thái ẩn LSTM bằng các tính năng được trích xuất.

Đối với các mô hình dựa trên sự chú ý, chúng tôi xem xét Att2In (Rennie et al., 2017), tương tự như với mô hình chú ý ban đầu được đề xuất bởi (Xu et al., 2015), ngoại trừ tính năng hình ảnh là chỉ nhập vào cổng tế bào như đã trình bày để dẫn đến hiệu suất tốt hơn. Sau đó chúng tôi xem xét mô hình chú ý được đề xuất bởi (Anderson et al., 2018) đề xuất một LSTM "chú ý từ trên xuống" cụ thể cũng như một LSTM "ngôn ngữ".

Generally attention mechanisms operate over high level convolutional layers. The attention mechanism from (Anderson et al., 2018) can be used on such feature maps, but Anderson et al. also consider feature maps corresponding to object proposals from a detection model. We consider both models, denoted as **TopDown** (feature map extracted from high level convolutional layer) and **TopDown-BB** (feature map extracted from object proposals from a detection model). Finally, we consider the recently proposed **Neural Baby Talk (NBT)** model (Lu et al., 2018) which explicitly uses object detections (as opposed to just bounding boxes) for sentence generation.

Learning objective. All of the above models are trained with the standard *cross entropy* (CE) loss as well as the *self-critical* (SC) loss proposed by Rennie et al. (2017) (with an exception of NBT, where only the CE version is included). The SC loss directly optimizes the CIDEr metric with a reinforcement learning technique. We additionally consider a model trained with a *GAN* loss (Shetty et al., 2017) (denoted **GAN**), which applies adversarial training to obtain more diverse and "human-like" captions, and their respective non-GAN baseline with the CE loss.

TopDown deconstruction. To better evaluate how each component of a model might influence hallucination, we "deconstruct" the TopDown model by gradually removing components until it is equivalent to the FC model. The intermediate networks are *NoAttention*, in which the attention mechanism is replaced by mean pooling, *NoConv* in which spatial feature maps are not input into the network (the model is provided with fully connected feature maps), *SingleLayer* in which only one LSTM is included in the model, and finally, instead of inputting visual features at each time step, visual features are used to initialize the LSTM embedding as is done in the FC model. By deconstructing the TopDown model in this way, we ensure that model design choices and hyperparameters do not confound results.

Implementation details. All the baseline models employ features extracted from the fourth layer of ResNet-101 (He et al., 2016), except for the GAN model which employs ResNet-152. Models without attention traditionally use fully connected layers as opposed to convolutional layers. However, as ResNet-101 does not have intermediate fully connected layers, it is standard to average pool convolutional activations and input these features into non-attention based description models. Note that this means the difference between the *NoAttention* and *NoConv* model is that the *NoAttention* model learns a visual embedding of spatial feature maps as opposed to relying on pre-pooled feature maps. All models except for TopDown-BB, NBT, and GAN are implemented in the same open source framework from Luo et al. (2018).[1]

Training/Test splits. We evaluate the captioning models on two MSCOCO splits. First, we consider the split from Karpathy et al. (Karpathy and Fei-Fei, 2015), specifically in that case the models are trained on the respective Karpathy Training set, tuned on Karpathy Validation set and the reported numbers are on the Karpathy Test set. We also consider the *Robust* split, introduced in (Lu et al., 2018), which provides a compositional split for MSCOCO. Specifically, it is ensured that the object pairs present in the training, validation and test captions do not overlap. In this case the captioning models are trained on the Robust Training set, tuned on the Robust Validation set and the reported numbers are on the Robust Test set.

3.2 Which Models Are More Prone To Hallucination?

We first present how well competitive models perform on our proposed CHAIR metric (Table 1). We report CHAIR at sentence-level and at instance-level (CHs and CHi in the table). In general, we see that models which perform better on standard evaluation metrics, perform better on CHAIR, though this is not always true. In particular, models which optimize for CIDEr frequently hallucinate more. Out of all generated captions on the Karpathy Test set, anywhere between 7.4% and 17.7% include a hallucinated object. When shifting to more difficult training scenarios in which new combinations of objects are seen at test time, hallucination consistently increases (Table 2).

Karpathy Test set. Table 1 presents object hallucination on the Karpathy Test set. All sentences are generated using beam search and a beam size of 5. We note a few important trends. First, models with attention tend to perform better on the CHAIR metric than models without attention. As we explore later, this is likely because they have

[1] https://github.com/ruotianluo/self-critical.pytorch

Nói chung các cơ chế chú ý hoạt động ở mức cao
lớp tích chập cấp độ. Cơ chế chú ý từ (Anderson và cộng sự,
2018) có thể được sử dụng trên
các bản đồ đặc điểm như vậy, nhưng Anderson et al. cũng xem
xét các bản đồ đặc điểm tương ứng với các đề xuất đối tượng
từ một mô hình phát hiện. Chúng tôi xem xét cả hai
mô hình, được biểu thị là TopDown (bản đồ đặc điểm được
trích xuất từ lớp tích chập cấp cao) và
TopDown-BB (bản đồ đặc điểm được trích xuất từ đối tượng
đề xuất từ một mô hình phát hiện). Cuối cùng, chúng tôi
hãy xem xét Neural Baby Talk được đề xuất gần đây
(NBT) mô hình (Lu et al., 2018) mà rõ ràng
sử dụng tính năng phát hiện đối tượng (khác với chỉ sử dụng
hộp giới hạn) để tạo câu.

Mục tiêu học tập. Tất cả các mô hình trên
được đào tạo với entropy chéo chuẩn (CE)
mất mát cũng như mất mát tự phê bình (SC) do Rennie et al.
(2017) đề xuất (trừ một ngoại lệ
của NBT, trong đó chỉ có phiên bản CE được bao gồm).
Sự mất mát SC trực tiếp tối ưu hóa số liệu CIDEr
với một kỹ thuật học tăng cường. Chúng tôi cũng xem xét một
mô hình được đào tạo với GAN
mất mát (Shetty et al., 2017) (được biểu thị bằng GAN), trong đó
áp dụng đào tạo đối kháng để có được sự đa dạng hơn
và chú thích "giống con người", và chú thích tương ứng của chúng
đường cơ sở không phải GAN với mất mát CE.

Phân tích TopDown. Để đánh giá tốt hơn
cách mỗi thành phần của một mô hình có thể ảnh hưởng đến ảo
giác, chúng tôi "phân tích" mô hình từ trên xuống bằng cách
loại bỏ dần các thành phần
cho đến khi nó tương đương với mô hình FC. Các mạng trung
gian là NoAttention, trong đó cơ chế chú ý được thay thế
bằng nhóm trung bình, No-Conv trong đó bản đồ đặc điểm không
gian không được nhập
vào mạng (mô hình được cung cấp đầy đủ
bản đồ tính năng được kết nối), SingleLayer trong đó
chỉ có một LSTM được đưa vào mô hình và cuối cùng, thay vì
nhập các tính năng trực quan tại mỗi
bước thời gian, các tính năng trực quan được sử dụng để khởi tạo
Nhúng LSTM như được thực hiện trong mô hình FC. Bằng cách
phân tích mô hình TopDown theo cách này,
chúng tôi đảm bảo rằng các lựa chọn thiết kế mô hình và
siêu tham số không làm sai lệch kết quả.

Chi tiết triển khai. Tất cả các mô hình cơ sở đều sử
dụng các tính năng được trích xuất từ lớp thứ tư
của ResNet-101 (He et al., 2016), ngoại trừ
Mô hình GAN sử dụng ResNet-152. Các mô hình không được chú
ý thường sử dụng các lớp được kết nối hoàn toàn thay vì các
lớp tích chập.
Tuy nhiên, vì ResNet-101 không có các lớp trung gian được
kết nối đầy đủ nên nó là chuẩn mực để trung bình
nhóm các kích hoạt tích chập và đưa các tính năng này vào
các mô hình mô tả không dựa trên sự chú ý.
Lưu ý rằng điều này có nghĩa là sự khác biệt giữa
Mô hình NoAttention và NoConv là NoAt-
mô hình căng thẳng học cách nhúng trực quan của không gian
bản đồ đặc điểm trái ngược với việc dựa vào các bản đồ được gộp trước
bản đồ tính năng. Tất cả các mô hình ngoại trừ TopDown-BB,
NBT và GAN đều được triển khai trong cùng một
khung nguồn mở từ Luo et al. (2018).1

Phân chia Đào tạo/Kiểm tra. Chúng tôi đánh giá phụ đề
mô hình trên hai phân chia MSCOCO. Đầu tiên, chúng tôi xem
xét phân chia từ Karpathy et al. (Karpathy và
Fei-Fei, 2015), cụ thể trong trường hợp đó các mô hình được
đào tạo trên Đào tạo Karpathy tương ứng
được thiết lập, điều chỉnh trên bộ Xác thực Karpathy và các
số được báo cáo nằm trên bộ Kiểm tra Karpathy. Chúng tôi
cũng xem xét sự chia tách Robust, được giới thiệu trong (Lu
et al., 2018), cung cấp sự phân chia thành phần
đối với MSCOCO. Cụ thể, nó được đảm bảo rằng
cặp đối tượng có trong quá trình đào tạo, xác thực và
chú thích thử nghiệm không chồng chéo. Trong trường hợp
này, các mô hình chú thích được đào tạo trên Robust Training
được thiết lập, điều chỉnh trên bộ Xác thực mạnh mẽ và các
số được báo cáo nằm trên bộ Kiểm tra mạnh mẽ.

3.2 Những mô hình nào dễ bị hơn
Ảo giác?

Đầu tiên chúng tôi trình bày các mô hình cạnh tranh tốt như thế nào
thực hiện trên số liệu CHAIR được đề xuất của chúng tôi
(Bảng 1). Chúng tôi báo cáo CHAIR ở cấp độ câu và
ở cấp độ trường hợp (CH và CHi trong bảng). Trong
Nhìn chung, chúng ta thấy rằng các mô hình hoạt động tốt
hơn trên các số liệu đánh giá tiêu chuẩn, hoạt động tốt hơn
trên CHAIR, mặc dù điều này không phải lúc nào cũng đúng. Trong
cụ thể, các mô hình tối ưu hóa cho CIDEr thường xuyên gây
ảo giác nhiều hơn. Trong số tất cả các mô hình được tạo ra
chú thích trên bộ Kiểm tra Karpathy, bất kỳ nơi nào trong
khoảng từ 7,4% đến 17,7% đều bao gồm một vật thể ảo giác. Khi
chuyển sang các kịch bản đào tạo khó hơn trong đó các kết hợp
vật thể mới được
được thấy tại thời điểm thử nghiệm, ảo giác liên tục tăng
lên (Bảng 2).

Bộ kiểm tra Karpathy. Bảng 1 trình bày sự phân tách đối
tượng trên bộ kiểm tra Karpathy. Tất cả các câu
được tạo ra bằng cách sử dụng tìm kiếm chùm tia và kích thước chùm tia
của 5. Chúng tôi lưu ý một vài xu hướng quan trọng. Đầu
tiên, các mô hình được chú ý có xu hướng hoạt động tốt hơn trên
CHAIR đo lường hơn các mô hình không có sự chú ý. Như
chúng ta khám phá sau, điều này có thể là do họ có

1https://github.com/ruotianluo/tự phê bình.pytorch

Model	Att.	Cross Entropy S	M	C	CHs	CHi	Self Critical S	M	C	CHs	CHi
LRCN*		17.0	23.9	90.8	17.7	12.6	16.9	23.5	93.0	17.7	12.9
FC*		17.9	24.9	95.8	15.4	10.9	18.4	25.0	103.9	14.4	10.1
Att2In*	✓	18.9	25.8	102.0	10.8	7.8	19.0	25.7	106.7	12.2	8.4
TopDown*	✓	19.9	26.7	107.6	8.4	6.0	20.4	27.0	117.2	13.6	8.8
TopDown-BB †	✓	20.4	27.1	113.7	8.3	5.9	21.4	27.7	120.6	10.4	6.9
NBT †	✓	19.4	26.2	105.1	7.4	5.4	-	-	-	-	
		Cross Entropy					GAN				
GAN ‡		18.7	25.7	100.4	10.6	7.6	16.6	22.7	79.3	8.2	6.5

Table 1: Hallucination analysis on the Karpathy Test set: Spice (S), CIDEr (C) and METEOR (M) scores across different image captioning models as well as CHAIRs (sentence level, CHs) and CHAIRi (instance level, CHi). All models are generated with beam search (beam size=5). * are trained/evaluated within the same implementation (Luo et al., 2018), † are trained/evaluated with implementation publicly released with corresponding papers, and ‡ sentences obtained directly from the author. For discussion see Section 3.2.

a better understanding of the image. In particular, methods that incorporate bounding box attention (as opposed to relying on coarse feature maps), consistently have lower hallucination as measured by our CHAIR metric. Note that the NBT model does not perform as well on standard captioning metrics as the TopDown-BB model but has lower hallucination. This is perhaps because bounding box proposals come from the MSCOCO detection task and are thus "in-domain" as opposed to the TopDown-BB model which relies on proposals learned from the Visual Genome (Krishna et al., 2017) dataset. Second, frequently training models with the self-critical loss actually increases the amount of hallucination. One hypothesis is that CIDEr does not penalize object hallucination sufficiently, leading to both increased CIDEr and increased hallucination. Finally, the LRCN model has a higher hallucination rate than the FC model, indicating that inputting the visual features only at the first step, instead of at every step, leads to more image relevant captions.

We also consider a GAN based model (Shetty et al., 2017) in our analysis. We include a baseline model (trained with CE) as well as a model trained with the GAN loss.[2] Unlike other models, the GAN model uses a stronger visual network (ResNet-152) which could explain the lower hallucination rate for both the baseline and the GAN model. Interestingly, when comparing the baseline and the GAN model (both trained with ResNet-152), standard metrics decrease substantially, even though human evaluations from (Shetty et al., 2017) demonstrate that sentences are of comparable quality. On the other hand, hallucination

[2]Sentences were procured directly from the authors.

	Att	S	M	C	CHs	CHi
FC*		15.5	22.7	76.2	21.3	15.3
Att2In*	✓	16.9	24.0	85.8	14.1	10.1
TopDown*	✓	17.7	24.7	89.8	11.3	7.9
NBT †	✓	18.1	24.8	94.5	6.8	4.6

Table 2: Hallucination Analysis on the Robust Test set: Spice (S), CIDEr (C) and METEOR (M) scores across different image captioning models as well as CHAIRs (sentence level, CHs) and CHAIRi (instance level, CHi). * are trained/evaluated within the same implementation (Luo et al., 2018), † are trained/evaluated with implementation publicly released with corresponding papers. All models trained with cross-entropy loss. See Section 3.2.

decreases, implying that the GAN loss actually helps decrease hallucination. Unlike the self critical loss, the GAN loss encourages sentences to be human-like as opposed to optimizing a metric. Human-like sentences are not likely to hallucinate objects, and a hallucinated object is likely a strong signal to the discriminator that a sentence is generated, and is not from a human.

We also assess the effect of beam size on CHAIR. We find that generally beam search decreases hallucination. We use beam size of 5, and for all models trained with cross entropy, it outperforms lower beam sizes on CHAIR. However, when training models with the self-critical loss, beam size sometimes leads to worse performance on CHAIR. For example, on the Att2In model trained with SC loss, a beam size of 5 leads to 12.8 on CHAIRs and 8.7 on CHAIRi, while a beam size of 1 leads to 10.8 on CHAIRs and 8.1 on CHAIRi.

Robust Test set. Next we review the hallucination behavior on the Robust Test set (Table 2). For almost all models the hallucination increases on the Robust split (e.g. for TopDown from 8.4% to 11.4% of sentences), indicating that the issue of

Người mẫu	Chú thích	Entropy chéo SMC	CHs	CHi	Tự phê bình SMC	CHs	CHi
LRCN*		17,0	23,9	90,8 17,7 12,6	16,9	23,5	93,0 17,7 12,9
FC*		17,9	24,9	95,8 15,4 10,9	18,4	25,0	103,9 14,4 10,1
Att2In*		18,9	25,8	102,0 10,8 19,9	7,8	19,0	25,7 106,7 12,2 6,0 20,4 8.4
Trên xuống*		26,7	107,6	8,4	27,0	117,2	13,6 8.8
TopDown-BB †		20,4	27,1	113,7 8.3	5,9	21,4	27,7 120,6 10,4 6.9
Không có dữ liệu !		19,4	26,2	105,1 7.4	5.4	-	- - -
		Entropy chéo			GAN		
GAN ‡		18,7	25,7	100,4 10,6	7,6	16,6	22,7 79,3 8.2 6,5

Bảng 1: Phân tích ảo giác trên bộ Kiểm tra Karpathy: Điểm Spice (S), CIDEr (C) và METEOR (M) trên các hình ảnh khác nhau các mô hình chú thích cũng như CHAIR (mức câu, CH) và CHAIRi (mức thể hiện, CHi). Tất cả các mô hình được tạo ra với tìm kiếm chùm tia (kích thước chùm tia = 5). * được đào tạo/đánh giá trong cùng một triển khai (Luo et al., 2018), được đào tạo/đánh giá † với việc triển khai được công bố công khai cùng với các giấy tờ tương ứng, và ‡ câu được lấy trực tiếp từ tác giả. Đối với thảo luận xem Mục 3.2.

hiểu rõ hơn về hình ảnh. Đặc biệt,
phương pháp kết hợp sự chú ý của hộp giới hạn
(trái ngược với việc dựa vào bản đồ đặc điểm thô),
luôn có ảo giác thấp hơn khi đo
theo số liệu CHAIR của chúng tôi. Lưu ý rằng mô hình NBT
không hoạt động tốt trên phụ đề chuẩn
số liệu như mô hình TopDown-BB nhưng có thấp hơn
ảo giác. Có lẽ là do sự ràng buộc
các đề xuất hộp xuất phát từ nhiệm vụ phát hiện MSCOCO
và do đó là "trong miền" trái ngược với
mô hình TopDown-BB dựa trên các đề xuất
được học từ Bộ gen thị giác (Krishna et al.,
2017) tập dữ liệu. Thứ hai, thường xuyên đào tạo các mô
hình với tổn thất tự phê bình thực sự làm tăng
lượng ảo giác. Một giả thuyết là
CIDEr không trừng phạt ảo giác đối tượng một cách đầy
đủ, dẫn đến cả việc tăng CIDEr và tăng ảo giác. Cuối
cùng, mô hình LRCN
có tỷ lệ ảo giác cao hơn mô hình FC,
chỉ ra rằng việc nhập các tính năng trực quan chỉ ở
bước đầu tiên, thay vì ở mọi bước, dẫn đến nhiều hơn
chú thích có liên quan đến hình ảnh.

Chúng tôi cũng xem xét một mô hình dựa trên GAN (Shetty et al., 2017) trong phân tích của chúng tôi. Chúng tôi bao gồm một mô hình cơ sở (được đào tạo với CE) cũng như một mô hình được đào tạo với GAN loss.2 Không giống như các mô hình khác, mô hình GAN sử dụng mạng lưới thị giác mạnh hơn (ResNet-152) có thể giải thích tỷ lệ halucination thấp hơn cho cả đường cơ sở và GAN
mô hình. Điều thú vị là khi so sánh đường cơ sở
và mô hình GAN (cả hai đều được đào tạo bằng ResNet-152),
các số liệu chuẩn giảm đáng kể, thậm chí
mặc dù đánh giá của con người từ (Shetty et al.,
2017) chứng minh rằng các câu có chất lượng tương đương.
Mặt khác, ảo giác

[2] Các câu được lấy trực tiếp từ tác giả.

	Att	SMC	CHs	CHi
FC*		15,5	22,7	76,2 21,3 15,3
Att2In*		16,9	24,0	85,8 14,1 10,1
Trên xuống*		17,7	24,7	89,8 11,3 7,9
Không có dữ liệu !		18,1	24,8	94,5 6,8 4,6

Bảng 2: Phân tích ảo giác trên bộ thử nghiệm mạnh mẽ: Điểm Spice (S), CIDEr (C) và METEOR (M) trên các mô hình chú thích hình ảnh khác nhau cũng như CHAIR (mức câu, CH) và CHAIRi (mức thể hiện, CHi). * được đào tạo/đánh giá trong là cùng một triển khai (Luo et al., 2018), † được đào tạo/đánh giá với việc triển khai công khai được phát hành với các bài báo tương ứng. Tất cả các mô hình được đào tạo với mất entropy chéo. Xem Phần 3.2.

giảm, ngụ ý rằng sự mất mát GAN thực sự
giúp giảm ảo giác. Không giống như sự mất mát tự phê
bình, sự mất mát GAN khuyến khích các câu
giống con người thay vì tối ưu hóa một số liệu.
Những câu nói giống con người không có khả năng gây ảo giác
các đối tượng, và một đối tượng ảo giác có khả năng là một đối tượng mạnh mẽ
báo hiệu cho bộ phân biệt biết rằng câu được tạo ra chứ
không phải từ con người.

Chúng tôi cũng đánh giá tác động của kích thước chùm tia trên GHẾ. Chúng tôi thấy rằng tìm kiếm chùm tia nói chung làm giảm ảo giác. Chúng tôi sử dụng kích thước chùm tia là 5 và đối với tất cả các mô hình được đào tạo với entropy chéo, nó hoạt động tốt hơn các kích thước chùm tia thấp hơn trên CHAIR. Tuy nhiên, khi đào tạo các mô hình với sự mất mát tự phê bình, kích thước chùm tia đôi khi dẫn đến hiệu suất kém hơn trên CHAIR. Ví dụ, trên mô hình Att2In được đào tạo với mất SC, kích thước chùm tia là 5 dẫn đến 12,8 trên CHAIR và 8,7 trên CHAIRi, trong khi kích thước chùm tia của 1 dẫn đến 10,8 trên CHAIR và 8,1 trên CHAIRi.

Bộ kiểm tra mạnh mẽ. Tiếp theo chúng tôi xem xét hành vi ảo giác trên bộ kiểm tra mạnh mẽ (Bảng 2). Đối với hầu như tất cả các mô hình ảo giác tăng lên
sự phân chia mạnh mẽ (ví dụ đối với TopDown từ 8,4% đến 11,4% số câu), cho thấy vấn đề của

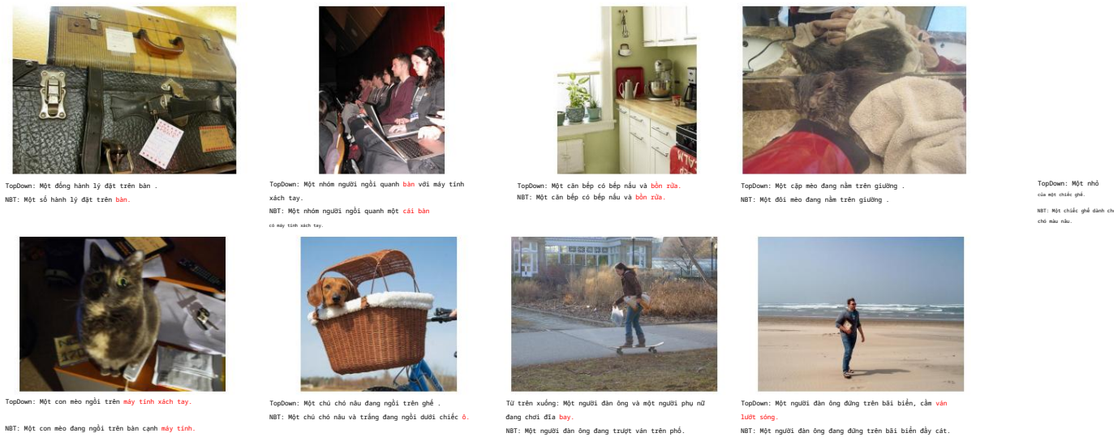
Figure 3: Examples of object hallucination from two state-of-the-art captioning models, TopDown and NBT, see Section 3.2.

hallucination is more critical in scenarios where test examples can not be assumed to have the same distribution as train examples. We again note that attention is helpful for decreasing hallucination. We note that the NBT model actually has lower hallucination scores on the robust split. This is in part because when generating sentences we use the detector outputs provided by Lu et al. (2018). Separate detectors on the Karpathy test and robust split are not available and the detector has access to images in the robust split during training. Consequently, the comparison between NBT and other models is not completely fair, but we include the number for completeness.

In addition to the Robust Test set, we also consider a set of MSCOCO in which certain objects are held out, which we call the *Novel Object split* (Hendricks et al., 2016). We train on the training set outlined in (Hendricks et al., 2016) and test on the Karpathy test split, which includes objects unseen during training. Similarly to the Robust Test set, we see hallucination increase substantially on this split. For example, for the TopDown model hallucination increases from 8.4% to 12.1% for CHAIRs and 6.0% to 9.1% for CHAIRi.

We find no obvious correlation between the average length of the generated captions and the hallucination rate. Moreover, vocabulary size does not correlate with hallucination either, i.e. models with *more diverse* descriptions may actually *hallucinate less*. We notice that hallucinated objects tend to be mentioned towards *the end of the sentence* (on average at position 6, with average sentence length 9), suggesting that some of the preceding words may have triggered hallucination. We investigate this below.

Which objects are hallucinated and in what context? Here we analyze which MSCOCO objects tend to be hallucinated more often and what are the common preceding words and image context. Across all models the super-category *Furniture* is hallucinated most often, accounting for $20-50\%$ of all hallucinated objects. Other common super-categories are *Outdoor objects*, *Sports* and *Kitchenware*. On the Robust Test set, *Animals* are often hallucinated. The *dining table* is the most frequently hallucinated object across all models (with an exception of GAN, where *person* is the most hallucinated object). We find that often words like "sitting" and "top" precede the "dining table" hallucination, implying the two common scenarios: a person "sitting at the table" and an object "sitting on top of the table" (Figure 3, row 1, examples 1, 2). Similar observations can be made for other objects, e.g. word "kitchen" often precedes "sink" hallucination (Figure 3, row 1, example 3) and "laying" precedes "bed" (Figure 3, row 1, example 4). At the same time, if we look at which objects are actually present in the image (based on MSCOCO object annotations), we can similarly identify that presence of a "cat" co-occurs with hallucinating a "laptop" (Figure 3, row 2, example 1), a "dog" – with a "chair" (Figure 3, row 2, example 2) etc. In most cases we observe that the hallucinated objects appear in the relevant scenes (e.g. "surfboard" on a beach), but



Hình 3: Ví dụ về ảo giác đối tượng từ hai mô hình chú thích hiện đại, TopDown và NBT, xem Mục 3.2.

ảo giác quan trọng hơn trong những tình huống mà
các ví dụ kiểm tra không thể được cho là có cùng
phân phối như ví dụ về tàu hỏa. Chúng tôi một lần nữa lưu ý rằng
sự chú ý có tác dụng làm giảm ảo giác.
Chúng tôi lưu ý rằng mô hình NBT thực sự có giá trị thấp hơn
ảo giác ghi điểm trên sự chia tách mạnh mẽ. Đây là
một phần vì khi tạo câu chúng ta sử dụng
đầu ra của máy dò được cung cấp bởi Lu et al. (2018).
Các máy dò riêng biệt trên thử nghiệm Karpathy và mạnh mẽ
tách không có sẵn và máy dò có quyền truy cập
đến hình ảnh trong sự phân chia mạnh mẽ trong quá trình
đào tạo. Do đó, sự so sánh giữa NBT và các
mô hình không hoàn toàn công bằng, nhưng chúng tôi bao gồm
số để hoàn thiện.

Ngoài bộ Kiểm tra mạnh mẽ, chúng tôi cũng xem xét một bộ
MSCOCO trong đó một số đối tượng nhất định được đưa ra, mà
chúng tôi gọi là Phân tách đối tượng mới (Hendricks và cộng
sự, 2016). Chúng tôi đào tạo trên
bộ đào tạo được nêu trong (Hendricks et al., 2016) và
kiểm tra trên phân chia kiểm tra Karpathy, bao gồm các đối
tượng không nhìn thấy trong quá trình đào tạo. Tương tự như
bộ Kiểm tra Ro-bust, chúng ta thấy ảo giác tăng đáng kể trên
phân chia này. Ví dụ, đối với mô hình Top-Down, ảo giác tăng
từ 8,4% lên
12,1% cho CHAIR và 6,0% đến 9,1% cho CHAIRi.

Chúng tôi không tìm thấy mối tương quan rõ ràng nào giữa
độ dài trung bình của các chú thích được tạo ra và tỷ lệ
ảo giác. Hơn nữa, kích thước từ vựng
cũng không tương quan với ảo giác, tức là các mô hình có
nhiều mô tả đa dạng hơn thực sự có thể
ít gây ảo giác hơn. Chúng tôi nhận thấy rằng các đối tượng
gây ảo giác có xu hướng được đề cập đến vào cuối
câu (trung bình ở vị trí 6, với trung bình
độ dài câu 9), cho thấy rằng một số
những từ đứng trước có thể gây ra ảo giác.
Chúng tôi sẽ điều tra vấn đề này bên dưới.

Những vật thể nào bị ảo giác và trong cái gì
bối cảnh? Ở đây chúng tôi phân tích những vật thể MSCOCO
nào có xu hướng bị ảo giác thường xuyên hơn và những gì
là những từ ngữ và hình ảnh chung trước đó. Trong tất cả
các mô hình, siêu thể loại Fur-niture thường bị ảo giác
nhất, chiếm
20 50% của tất cả các đối tượng ảo giác. Các siêu thể
loại phổ biến khác là Đối tượng ngoài trời, Thể thao
và Đồ dùng nhà bếp. Trong bộ Kiểm tra mạnh mẽ, Động vật
thường bị ảo giác. Bàn ăn là
vật thể gây ảo giác thường xuyên nhất trên tất cả
mô hình (ngoại trừ GAN, nơi người
là đối tượng gây ảo giác nhiều nhất). Chúng tôi thấy rằng thường
những từ như "ngồi" và "trên cùng" xuất hiện trước ảo giác
"bàn ăn", ám chỉ hai kịch bản phổ biến: một người "ngồi ở
bàn" và
một vật thể "ngồi trên đầu bàn" (Hình 3,
hàng 1, ví dụ 1, 2). Những quan sát tương tự có thể
được tạo ra cho các đối tượng khác, ví dụ từ "bếp" thường
đứng trước "bồn rửa" ảo giác (Hình 3, hàng
1, ví dụ 3) và "laying" đứng trước "bed" (Hình 3, hàng 1,
ví dụ 4). Đồng thời, nếu chúng ta
hãy xem những vật thể nào thực sự có mặt trong
hình ảnh (dựa trên chú thích đối tượng MSCOCO),
chúng ta cũng có thể xác định sự hiện diện của một "con mèo"
xảy ra đồng thời với ảo giác về một "máy tính xách tay" (Hình 3,
hàng 2, ví dụ 1), một "con chó" – với một "chiếc ghế" (Hình
3, hàng 2, ví dụ 2) v.v. Trong hầu hết các trường hợp, chúng tôi
quan sát rằng các vật thể ảo giác xuất hiện trong
cảnh có liên quan (ví dụ "ván lướt sóng" trên bãi biển), nhưng

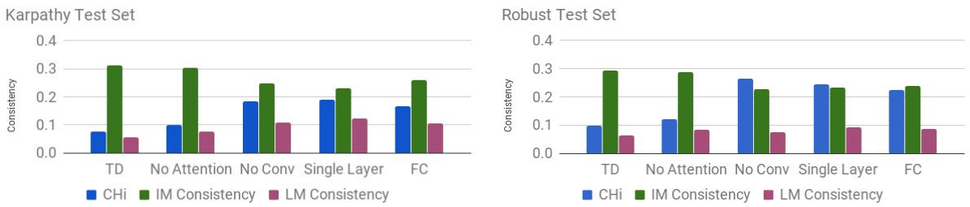


Figure 4: Image and Language model consistency (IM, LM) and CHAIRi (instance-level, CHi) on deconstructed TopDown models. Images with less hallucination tend to make errors consistent with the image model, whereas models with more hallucination tend to make errors consistent with the language model, see Section 3.3.

there are cases where objects are hallucinated out of context (e.g. "bed" in the bathroom, Figure 3, row 1, example 4).

3.3 What Are The Likely Causes Of Hallucination?

In this section we investigate the likely causes of object hallucination. We have earlier described how we deconstruct the TopDown model to enable a controlled experimental setup. We rely on the deconstructed TopDown models to analyze the impact of model components on hallucination.

First, we summarize the hallucination analysis on the deconstructed TopDown models (Table 3). Interestingly, the *NoAttention* model does not do substantially worse than the full model (w.r.t. sentence metrics and CHAIR). However, removing Conv input (*NoConv* model) and relying only on FC features, decreases the performance dramatically. This suggests that much of the gain in attention based models is primarily due to *access to feature maps with spatial locality*, not the actual attention mechanism. Also, similar to LRCN vs. FC in Table 1, initializing the LSTM hidden state with image features, as opposed to inputting image features at each time step, leads to lower hallucination (*Single Layer* vs. *FC*). This is somewhat surprising, as a model which has access to image information at each time step should be less likely to "forget" image content and hallucinate objects. However, it is possible that models which include image inputs at each time step with no access to spatial features overfit to the visual features.

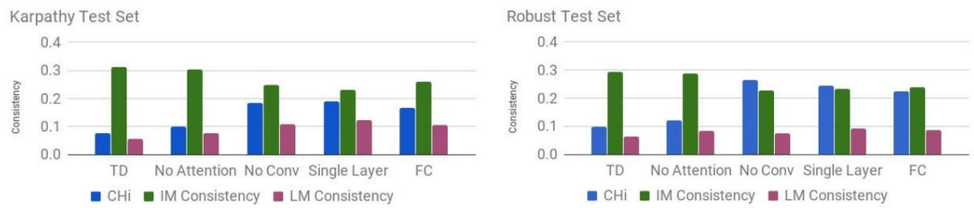
Now we investigate what causes hallucination using the deconstructed TopDown models and the *image consistency* and *language consistency* scores, introduced in Sections 2.2 and 2.3 which capture how consistent the hallucinations errors are with image- / language-only models.

Karpathy Split	METEOR	CIDEr	SPICE	CHs	CHi
TD	26.10	103.40	19.50	10.80	7.40
No Attention	25.60	99.70	18.80	14.20	9.40
No Conv	22.90	81.30	15.70	25.70	17.70
Single Layer	22.70	80.20	15.50	25.60	18.00
FC	23.30	85.10	16.40	23.60	15.70

Table 3: Hallucination analysis on deconstructed TopDown models with sentence metrics, CHAIRs (sentence level, CHs) and CHAIRi (instance level, CHi). See Section 3.3.

Figure 4 shows the CHAIR metric, image consistency and language consistency for the deconstructed TopDown models on the Karpathy Test set (left) and the Robust Test set (right). We note that models with *less* hallucination tend to make errors consistent with the image model, whereas models with *more* hallucination tend to make errors consistent with the language model. This implies that models with less hallucination are better at integrating knowledge from an image into the sentence generation process. When looking at the Robust Test set, Figure 4 (right), which is more challenging, as we have shown earlier, we see that image consistency *decreases* when comparing to the same models on the Karpathy split, whereas language consistency is similar across all models trained on the Robust split. This is perhaps because the Robust split contains novel compositions of objects at test time, and all of the models are heavily biased by language.

Finally, we measure image and language consistency during training for the FC model and note that at the beginning of training errors are more consistent with the language model, whereas towards the end of training, errors are more consistent with the image model. This suggests that models first learn to produce fluent language before learning to incorporate visual information.



Hình 4: Sự nhất quán của mô hình Ngôn ngữ và Hình ảnh (IM, LM) và CHAIRi (mức độ thể hiện, CHi) trên TopDown được giải cấu trúc mô hình. Hình ảnh có ít ảo giác hơn có xu hướng tạo ra lỗi phù hợp với mô hình hình ảnh, trong khi các mô hình có nhiều ảo giác có xu hướng tạo ra lỗi phù hợp với mô hình ngôn ngữ, xem Mục 3.3.

có những trường hợp các vật thể bị ảo giác
của ngữ cảnh (ví dụ "giường" trong phòng tắm, Hình 3,
hàng 1, ví dụ 4).

3.3 Những Nguyên Nhân Có Thể Xảy Ra Là Gì?
Ảo giác?

Trong phần này chúng tôi điều tra các nguyên nhân có thể xảy ra
ảo giác đối tượng. Chúng tôi đã mô tả trước đó
cách chúng tôi phân tích mô hình TopDown để cho phép thiết
lập thử nghiệm được kiểm soát. Chúng tôi dựa vào
các mô hình TopDown được phân tích để phân tích
tác động của các thành phần mô hình lên ảo giác.

Đầu tiên, chúng tôi tóm tắt phân tích ảo giác
trên các mô hình TopDown được phân tích (Bảng 3).
Điều thú vị là mô hình NoAttention không làm được điều đó
tệ hơn đáng kể so với mô hình đầy đủ (về số liệu câu và
CHAIR). Tuy nhiên, việc loại bỏ
Đầu vào Conv (mô hình NoConv) và chỉ dựa vào
Các tính năng FC làm giảm hiệu suất đáng kể. Điều này cho
thấy rằng phần lớn lợi ích trong các mô hình dựa trên sự
chú ý chủ yếu là do khả năng truy cập vào
bản đồ đặc điểm với vị trí không gian, không phải thực tế
cơ chế chú ý. Cũng tương tự như LRCN so với
FC trong Bảng 1, khởi tạo trạng thái ẩn LSTM
với các tính năng hình ảnh, trái ngược với việc nhập hình ảnh
các tính năng tại mỗi bước thời gian, dẫn đến ảo giác thấp
hơn (Lớp đơn so với FC). Điều này có phần
thật đáng ngạc nhiên, như một mô hình có thể truy cập vào hình ảnh
thông tin tại mỗi bước thời gian sẽ ít có khả năng xảy ra hơn
để "quên" nội dung hình ảnh và tạo ra ảo giác về các vật thể.
Tuy nhiên, có thể các mô hình bao gồm
đầu vào hình ảnh tại mỗi bước thời gian không có quyền truy cập vào
các đặc điểm không gian quá phù hợp với các đặc điểm trực quan.

Bây giờ chúng ta hãy tìm hiểu nguyên nhân gây ra ảo giác
sử dụng các mô hình TopDown đã được giải cấu trúc và
sự nhất quán của hình ảnh và sự nhất quán của ngôn ngữ
điểm số, được giới thiệu trong Mục 2.2 và 2.3
nắm bắt được mức độ nhất quán của các lỗi ảo giác
là những mô hình chỉ có hình ảnh/ngôn ngữ.

Karpathy Split	METEOR	CIDEr	SPICE	CHi
T.Đ	26,10	103,40	19,50	10,80 7,40
Không chú ý	25,60	99,70	18,80	14.20 9.40
Không có chuyển đổi	22,90	81,30	15,70	25,70 17,70
Lớp đơn	22,70	80,20	15,50	25,60 18,00
FC	23,30	85,10	16,40	23,60 15,70

Bảng 3: Phân tích ảo giác trên TopDown đã giải cấu trúc mô hình với số liệu câu, CHAIR (mức câu, CH) và CHAIRi (mức độ thể hiện, CHi). Xem Phần 3.3.

Hình 4 cho thấy số liệu CHAIR, tính nhất quán của hình
ảnh và tính nhất quán của ngôn ngữ đối với các mô hình
TopDown được phân tích trên Bài kiểm tra Karpathy
bộ (trái) và bộ Kiểm tra mạnh mẽ (phải). Chúng tôi lưu ý
rằng các mô hình có ít ảo giác hơn có xu hướng tạo ra
lỗi phù hợp với mô hình hình ảnh, trong khi
các mô hình có nhiều ảo giác hơn có xu hướng tạo ra er-
rors phù hợp với mô hình ngôn ngữ. Điều này ngụ ý rằng các
mô hình có ít ảo giác hơn sẽ tốt hơn trong việc tích hợp
kiến thức từ hình ảnh vào
quá trình tạo câu. Khi nhìn
tại bộ Kiểm tra mạnh mẽ, Hình 4 (bên phải), là
thách thức hơn, như chúng tôi đã trình bày trước đó, chúng tôi
thấy rằng tính nhất quán của hình ảnh giảm đi khi so sánh
với các mô hình tương tự trên phân tách Karpathy,
trong khi tính nhất quán của ngôn ngữ là tương tự nhau trên tất cả
các mô hình được đào tạo trên Robust split. Đây có lẽ là
bởi vì Robust split chứa các thành phần mới của các đối
tượng tại thời điểm thử nghiệm và tất cả các mô hình
có sự thiên vị rất lớn về ngôn ngữ.

Cuối cùng, chúng tôi đo lường sự nhất quán của hình ảnh
và ngôn ngữ trong quá trình đào tạo cho mô hình FC và lưu ý
rằng khi bắt đầu đào tạo, lỗi thường nhiều hơn
phù hợp với mô hình ngôn ngữ, trong khi về cuối quá trình
đào tạo, lỗi phù hợp hơn với mô hình hình ảnh. Điều này cho
thấy rằng
Các mô hình đầu tiên học cách tạo ra ngôn ngữ lưu loát trước
khi học cách kết hợp thông tin trực quan.

TD: A cat is sitting on a bed in a room.
S: 12.1 M: 23.8 C: 69.7
TD Restrict: A bed with a blanket and a pillow on it.
S: 23.5 M: 25.4 C: 52.5

TD: A cat laying on the ground with a frisbee.
S: 8.0 M: 13.1 C: 37.0
TD Restrict: A black and white animal laying on the ground.
S: 7.7 M: 15.9 C: 17.4

Figure 5: Examples of how TopDown (TD) sentences change when we enforce that objects cannot be hallucinated: SPICE (S), Meteor (M), CIDEr (C), see Section 3.4.

3.4 How Well Do The Standard Metrics Capture Hallucination?

In this section we analyze how well SPICE (Anderson et al., 2016), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015) capture hallucination. All three metrics do penalize sentences for mentioning incorrect words, either via an F score (METEOR and SPICE) or cosine distance (CIDEr). However, if a caption mentions enough words correctly, it can have a high METEOR, SPICE, or CIDEr score while still hallucinating specific objects.

Our first analysis tool is the TD-Restrict model. This is a modification of the TopDown model, where we enforce that MSCOCO objects which are not present in an image are *not generated* in the caption. We determine which words refer to objects absent in an image following our approach in Section 2.1. We then set the log probability for such words to a very low value. We generate sentences with the TopDown and TD-Restrict model with beam search of size 1, meaning all words produced by both models are the same, until the TopDown model produces a hallucinated word.

We compare which scores are assigned to such captions in Figure 5. TD-Restrict generates captions that do not contain hallucinated objects, while TD hallucinates a "cat" in both cases. In Figure 5 (left) we see that CIDEr scores the more correct caption much lower. In Figure 5 (right), the TopDown model incorrectly calls the animal a "cat." Interestingly, it then correctly identifies the "frisbee," which the TD-Restrict model fails to mention, leading to lower SPICE and CIDEr.

In Table 4 we compute Pearson correlation coefficient between individual sentence scores and

	CIDEr	METEOR	SPICE
FC	0.197	0.198	0.266
Att2In	0.177	0.178	0.246
TopDown	0.135	0.140	0.172

Table 4: Pearson correlation coefficients between 1-CHs and CIDEr, METEOR, and SPICE scores, see Section 3.4.

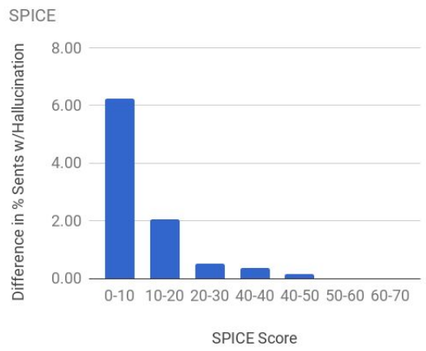


Figure 6: Difference in percentage of sentences with *no* hallucination for TopDown and FC models when SPICE scores fall into specific ranges. For sentences with low SPICE scores, the hallucination is generally larger for the FC model, even though the SPICE scores are similar, see Section 3.4.

the *absence* of hallucination, i.e. 1−CHAIRs; we find that SPICE consistently correlates higher with 1−CHAIRs. E.g., for the FC model the correlation for SPICE is 0.27, while for METEOR and CIDEr – around 0.2.

We further analyze the metrics in terms of their predictiveness of hallucination risk. Predictiveness means that a certain score should imply a certain percentage of hallucination. Here we show the results for SPICE and the captioning models FC and TopDown. For each model and a score interval (e.g. $10-20$) we compute the percentage of captions *without* hallucination (1−CHAIRs). We plot the difference between the percentages from both models (TopDown - FC) in Figure 6. Comparing the models, we note that even when scores are similar (e.g., all sentences with SPICE score in the range of $10-20$), the TopDown model has fewer sentences with hallucinated objects. We see similar trends across other metrics. Consequently, object hallucination can *not* be always predicted based on the traditional sentence metrics.

Is CHAIR complementary to standard metrics? In order to measure usefulness of our proposed metrics, we have conducted the following

TD: Một con mèo đang ngồi trên giường trong phòng.
S: 12,1 Nam: 23,8 NỮ: 69,7
TD Restrict: Một chiếc giường có chăn và gối trên đó.
S: 23,5 M: 25,4 C: 52,5

TD: Một con mèo nằm trên mặt đất với một chiếc đĩa bay.
S: 8.0 Nam: 13.1 NỮ: 37.0
TD Restrict: Đen và trắng động vật nằm trên mặt đất.
S: 7,7 M: 15,9 C: 17,4

Hình 5: Ví dụ về cách câu TopDown (TD) thay đổi khi chúng ta thực thi rằng các vật thể không thể bị ảo giác: SPICE (S), Meteor (M), CIDEr (C), xem Mục 3.4.

3.4 Các số liệu chuẩn tốt như thế nào Bắt ảo giác?

Trong phần này, chúng tôi phân tích mức độ hiệu quả của SPICE (An-derson và cộng sự, 2016), METEOR (Banerjee và Lavie, 2005) và CIDEr (Vedantam và cộng sự, 2015) bắt ảo giác. Cả ba số liệu đều phạt các câu vì đề cập đến các từ không chính xác, thông qua điểm F (METEOR và SPICE) hoặc khoảng cách co-sin (CIDEr). Tuy nhiên, nếu chú thích đề cập đến đủ số từ chính xác, nó có thể có

Điểm METEOR, SPICE hoặc CIDEr vẫn có thể làm ảo giác các vật thể cụ thể.

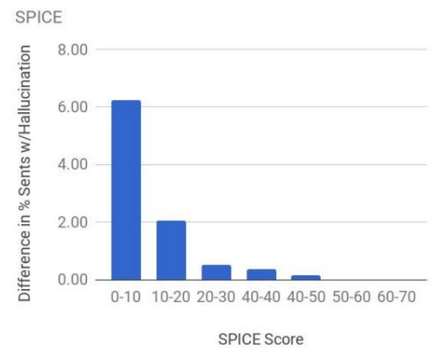
Công cụ phân tích đầu tiên của chúng tôi là mô hình TD-Restrict. Đây là một sửa đổi của mô hình TopDown, nơi chúng tôi thực thi các đối tượng MSCOCO không có trong hình ảnh không được tạo ra trong chú thích. Chúng tôi xác định những từ nào đề cập đến các vật thể không có trong hình ảnh theo cách tiếp cận của chúng tôi trong Phần 2.1. Sau đó chúng tôi đặt xác suất logarit cho những từ như vậy xuống một giá trị rất thấp. Chúng tôi tạo ra các câu với mô hình TopDown và TD-Restrict với tìm kiếm chùm tia có kích thước 1, nghĩa là tất cả các từ do cả hai mô hình tạo ra đều giống nhau, cho đến khi mô hình Top-Down tạo ra một từ ảo giác.

Chúng tôi so sánh những điểm số nào được gán cho chú thích trong Hình 5. TD-Restrict tạo ra các chú thích không chứa các đối tượng ảo giác, trong khi TD ảo giác một "con mèo" trong cả hai trường hợp. Trong Hình 5 (bên trái) chúng ta thấy rằng CIDEr ghi được nhiều điểm hơn chú thích đúng thấp hơn nhiều. Trong Hình 5 (bên phải), mô hình TopDown gọi sai động vật một "con mèo". Điều thú vị là sau đó nó xác định chính xác "Frisbee", mà mô hình TD-Restrict không làm được đề cập đến, dẫn đến lượng SPICE và CIDEr thấp hơn.

Trong Bảng 4, chúng tôi tính toán hệ số tương quan Pearson giữa điểm số câu riêng lẻ và

	CIDer METEOR GIA VỊ		
FC 0,197 Att2In 0,177		0,198	0,266
TopDown 0,135		0,178	0,246
		0,140	0,172

Bảng 4: Hệ số tương quan Pearson giữa 1-CH và Điểm CIDEr, METEOR và SPICE, xem Mục 3.4.



Hình 6: Sự khác biệt về tỷ lệ phần trăm các câu không có hiện tượng ảo giác đối với các mô hình TopDown và FC khi chấm điểm SPICE rơi vào các phạm vi cụ thể. Đối với các câu có SPICE thấp điểm, ảo giác thường lớn hơn đối với mô hình FC, mặc dù điểm SPICE tương tự nhau, hãy xem Mục 3.4.

sự vắng mặt của ảo giác, tức là 1 CHAIR; chúng tôi thấy rằng SPICE luôn có mối tương quan cao hơn với 1 CHAIRs. Ví dụ, đối với mô hình FC, mối tương quan đối với SPICE là 0,27, trong khi đối với METEOR và CIDEr – khoảng 0,2.

Chúng tôi tiếp tục phân tích các số liệu theo khía cạnh của chúng khả năng dự đoán rủi ro ảo giác. Khả năng dự đoán có nghĩa là một số điểm nhất định sẽ ngụ ý một tỷ lệ ảo giác nhất định. Ở đây chúng tôi trình bày kết quả cho SPICE và các mô hình chú thích FC và TopDown. Đối với mỗi mô hình và một điểm số trong khoảng thời gian (ví dụ 10 20), chúng tôi tính toán phần trăm chú thích không có ảo giác (1 CHAIR). Chúng tôi vẽ biểu đồ sự khác biệt giữa các phần trăm từ cả hai mô hình (TopDown - FC) trong Hình 6. Khi so sánh các mô hình, chúng tôi lưu ý rằng ngay cả khi điểm số tương tự nhau (ví dụ, tất cả các câu có điểm SPICE trong phạm vi 10 20), mô hình TopDown có ít câu hơn với các đối tượng ảo giác. Chúng ta thấy xu hướng tương tự trên các số liệu khác. Do đó, ảo giác đối tượng không phải lúc nào cũng có thể dự đoán được dựa trên các phép đo câu truyền thống.

CHAIR có bổ sung cho các số liệu chuẩn không? Để đo lường tính hữu ích của các số liệu được đề xuất của chúng tôi, chúng tôi đã tiến hành các bước sau

	Metric	Metric +(1-CHs)	Metric +(1-CHi)
METEOR	0.269	0.299	0.304
CIDEr	0.282	0.321	0.322
SPICE	0.248	0.277	0.281

Table 5: Pearson correlation coefficients between individual/combined metrics and human scores. See Section 3.4.

human evaluation (via the Amazon Mechanical Turk). We have randomly selected 500 test images and respective captions from 5 models: non-GAN baseline, GAN, NBT, TopDown and TopDown - Self Critical. The AMT workers were asked to score the presented captions w.r.t. the given image based on their preference. They could score each caption from 5 (very good) to 1 (very bad). We did not use ranking, i.e. different captions could get the same score; each image was scored by three annotators, and the average score is used as the final human score. For each image we consider the 5 captions from all models and their corresponding sentence scores (METEOR, CIDEr, SPICE). We then compute Pearson correlation between the human scores and sentence scores; we also consider a simple combination of sentence metrics and 1-CHAIRs or 1-CHAIRi by summation. The final correlation is computed by averaging across all 500 images. The results are presented in Table 5. Our findings indicate that a simple combination of CHAIRs or CHAIRi with the sentence metrics leads to an increased correlation with the human scores, showing the usefulness and complementarity of our proposed metrics.

Does hallucination impact generation of other words? Hallucinating objects impacts sentence quality not only because an object is predicted incorrectly, but also because the hallucinated word impacts generation of other words in the sentence. Comparing the sentences generated by Top-Down and TD-Restrict allows us to analyze this phenomenon. We find that after the hallucinated word is generated, the following words in the sentence are different 47.3% of the time. This implies that hallucination impacts sentence quality beyond simply naming an incorrect object. We observe that one hallucination may lead to another, e.g. hallucinating a “cat” leading to hallucinating a “chair”, hallucinating a “dog” – to a “frisbee”.

4 Discussion

In this work we closely analyze hallucination in object captioning models. Our work is similar to other works which attempt to characterize flaws of different evaluation metrics (Kilickaya et al., 2016), though we focus specifically on hallucination. Likewise, our work is related to other work which aims to build better evaluation tools ((Vedantam et al., 2015), (Anderson et al., 2016), (Cui et al., 2018)). However, we focus on carefully quantifying and characterizing one important type of error: object hallucination.

A significant number of objects are hallucinated in current captioning models (between 5.5% and 13.1% of MSCOCO objects). Furthermore, hallucination does not always agree with the output of standard captioning metrics. For instance, the popular self critical loss increases CIDEr score, but also the amount of hallucination. Additionally, we find that given two sentences with similar CIDEr, SPICE, or METEOR scores from two different models, the number of hallucinated objects might be quite different. This is especially apparent when standard metrics assign a low score to a generated sentence. Thus, for challenging caption tasks on which standard metrics are currently poor (e.g., the LSMDC dataset (Rohrbach et al., 2017b)), the CHAIR metric might be helpful to tease apart the most favorable model. Our results indicate that CHAIR complements the standard sentence metrics in capturing human preference.

Additionally, attention lowers hallucination, but it appears that much of the gain from attention models is due to access to the underlying convolutional features as opposed the attention mechanism itself. Furthermore, we see that models with stronger *image consistency* frequently hallucinate fewer objects, suggesting that strong visual processing is important for avoiding hallucination.

Based on our results, we argue that the design and training of captioning models should be guided not only by cross-entropy loss or standard sentence metrics, but also by image relevance. Our CHAIR metric gives a way to evaluate the phenomenon of hallucination, but other image relevance metrics e.g. those that incorporate missed salient objects, should also be investigated. We believe that incorporating visual information in the form of ground truth objects in a scene (as opposed to only reference captions) helps us better understand the performance of captioning models.

	Hệ mét	Hệ mét +(1-CH)	Hệ mét +(1-CHi)
METEOR 0.269	CIDEr	0,299	0,304
0.282 SPICE 0.248		0,321	0,322
		0,277	0,281

Bảng 5: Hệ số tương quan Pearson giữa các số liệu cá nhân/kết hợp và điểm số của con người. Xem Mục 3.4.

đánh giá của con người (thông qua Amazon Mechanical (Thổ Nhĩ Kỳ). Chúng tôi đã chọn ngẫu nhiên 500 hình ảnh thử nghiệm và chú thích tương ứng từ 5 mô hình: không phải GAN đường cơ sở, GAN, NBT, TopDown và TopDown - Tự phê bình. Các công nhân AMT được yêu cầu chấm điểm các chú thích được trình bày liên quan đến hình ảnh đã cho dựa trên sở thích của họ. Họ có thể chấm điểm từng chú thích từ 5 (rất tốt) đến 1 (rất tệ). Chúng tôi đã làm không sử dụng thứ hạng, tức là các chú thích khác nhau có thể nhận được cùng một điểm; mỗi hình ảnh được chấm điểm bởi ba người chú thích, và điểm trung bình được sử dụng làm điểm cuối cùng của con người. Đối với mỗi hình ảnh, chúng tôi xem xét 5 chú thích từ tất cả các mô hình và điểm câu tương ứng của chúng (METEOR, CIDEr, SPICE).
Sau đó chúng tôi tính toán tương quan Pearson giữa điểm số của con người và điểm số câu; chúng tôi cũng xem xét sự kết hợp đơn giản của các số liệu câu và 1-CHAIRs hoặc 1-CHAIRi bằng cách tổng hợp. tương quan cuối cùng được tính bằng cách tính trung bình trên tất cả 500 hình ảnh. Kết quả được trình bày trong Bảng 5. Phát hiện của chúng tôi chỉ ra rằng sự kết hợp đơn giản của CHAIR hoặc CHAIRi với câu số liệu dẫn đến mối tương quan gia tăng với điểm số của con người, cho thấy tính hữu ích và tính bổ sung của các số liệu chúng tôi đề xuất.

Ảo giác có ảnh hưởng đến việc tạo ra những thứ khác không? từ ngữ? Các vật thể gây ảo giác tác động đến câu chất lượng không chỉ vì một đối tượng được dự đoán không chính xác, mà còn vì từ ảo giác tác động đến việc tạo ra các từ khác trong câu. So sánh các câu được tạo ra bởi Top-Down và TD-Restrict cho phép chúng ta phân tích điều này hiện tượng. Chúng tôi thấy rằng sau khi ảo giác từ được tạo ra, các từ sau trong câu khác nhau 47,3% thời gian. Điều này ngụ ý rằng ảo giác ảnh hưởng đến chất lượng câu ngoài việc chỉ đơn giản là đặt tên cho một vật thể không chính xác. Chúng tôi quan sát thấy rằng một ảo giác có thể dẫn đến một ảo giác khác, ví dụ như ảo giác về một "con mèo" dẫn đến ảo giác một “chiếc ghế”, gây ảo giác về một “con chó” – cho đến một “chiếc đĩa bay”.

4 Thảo luận

Trong công trình này chúng tôi phân tích chặt chẽ ảo giác trong mô hình chú thích đối tượng. Công việc của chúng tôi tương tự như các tác phẩm khác cố gắng mô tả các khuyết điểm của các số liệu đánh giá khác nhau (Kilickaya et al., 2016), mặc dù chúng tôi tập trung cụ thể vào ảo giác. Tương tự như vậy, công việc của chúng tôi liên quan đến các công việc nhằm mục đích xây dựng các công cụ đánh giá tốt hơn ((Vedantam và cộng sự, 2015), (Anderson và cộng sự, 2016), (Cui et al., 2018)). Tuy nhiên, chúng tôi tập trung vào việc cẩn thận định lượng và mô tả một loại quan trọng của lỗi: ảo giác đối tượng.

Một số lượng đáng kể các đối tượng bị ảo giác trong các mô hình chú thích hiện tại (giữa 5,5% và 13,1% MSCOCO phản đối). Hơn nữa, hal-lucination không phải lúc nào cũng đồng ý với đầu ra của các số liệu chú thích tiêu chuẩn. Ví dụ, mất mát tự phê bình phổ biến làm tăng điểm CIDEr, nhưng cũng là lượng ảo giác. Ngoài ra, chúng ta thấy rằng cho hai câu có nội dung tương tự Điểm số CIDEr, SPICE hoặc METEOR từ hai mô hình khác nhau, số lượng vật thể ảo giác có thể khá khác biệt. Điều này đặc biệt rõ ràng khi các số liệu chuẩn chỉ định điểm thấp cho một câu được tạo ra. Do đó, đối với các nhiệm vụ chú thích đầy thách thức mà các số liệu chuẩn hiện đang kém (ví dụ: bộ dữ liệu LSMDC (Rohrbach et al., 2017b)), số liệu CHAIR có thể hữu ích để tách rời mô hình thuận lợi nhất. Kết quả của chúng tôi chỉ ra rằng CHAIR bổ sung cho tiêu chuẩn thước đo câu trong việc nắm bắt sở thích của con người.

Ngoài ra, sự chú ý làm giảm ảo giác, nhưng có vẻ như phần lớn lợi ích thu được từ sự chú ý các mô hình là do truy cập vào các tính năng tích chập cơ bản trái ngược với cơ chế chú ý. Hơn nữa, chúng ta thấy rằng các mô hình với hình ảnh mạnh hơn nhất quán thường xuyên ảo giác ít vật thể hơn, cho thấy rằng xử lý thị giác mạnh là quan trọng để tránh ảo giác.

Dựa trên kết quả của chúng tôi, chúng tôi lập luận rằng việc thiết kế và đào tạo các mô hình chú thích nên được không chỉ được hướng dẫn bởi mất mát entropy chéo hoặc tiêu chuẩn số liệu câu, mà còn theo mức độ liên quan của hình ảnh. Của chúng tôi Chỉ số CHAIR cung cấp một cách để đánh giá hiện tượng ảo giác, nhưng các chỉ số liên quan đến hình ảnh khác, ví dụ như những chỉ số kết hợp bị mất các đối tượng nổi bật, cũng nên được điều tra. Chúng tôi tin rằng việc kết hợp thông tin trực quan vào hình dạng của các đối tượng thực tế trong một cảnh (trái ngược với (chỉ tham chiếu đến phụ đề) giúp chúng ta hiểu rõ hơn về hiệu suất của các mô hình phụ đề.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and vqa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *CVPR*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634.

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

Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2016. Re-evaluating automatic metrics for image captioning. In *European Chapter of the Association for Computational Linguistics*.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind peoples experiences with computer-generated captions of social media images. In *Proceedings of the 2017 SIGCHI Conference on Human Factors in Computing Systems*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Anna Rohrbach, Makarand Tapaswi, Atousa Torabi, Tegan Maharaj, Marcus Rohrbach, Sanja Fidler Christopher Pal, and Bernt Schiele. 2017a. The Joint Video and Language Understanding Workshop: MovieQA and The Large Scale Movie Description Challenge (LSMDC). `https://sites.google.com/site/describingmovies/lsmdc-2017`.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017b. Movie description. *International Journal of Computer Vision*, 123(1):94–120.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE*

Tài liệu tham khảo

Peter Anderson, Basura Fernando, Mark Johnson và Stephen Gould. 2016. Spice: Đánh giá chú thích hình ảnh mệnh đề ngữ nghĩa. Ở châu Â u Hội nghị về Thị giác máy tính, trang 382–398. Mùa xuân.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould và Lei Zhang. 2018. Sự chú ý từ dưới lên và từ trên xuống cho chú thích hình ảnh và vqa. Trong Biên bản của Hội nghị IEEE về Tầm nhìn máy tính và Mẫu Sự công nhận.

Satanjeev Banerjee và Alon Lavie. 2005. Sao băng: Một tự động đo lường để đánh giá mt với cải tiến tương quan với phán đoán của con người. Trong Biên bản của Hội thảo ACL về các Biện pháp Đánh giá Nội tại và Ngoại tại cho Dịch máy và/hoặc Tóm tắt, trang 65–72.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakr-ishna Vedantam, Saurabh Gupta, Piotr Dollar và C Lawrence Zitnick. 2015. Chú thích coco của Microsoft: Máy chủ thu thập và đánh giá dữ liệu. Bản in trước arXiv arXiv:1504.00325.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, và Serge Belongie. 2018. Học cách đánh giá chú thích hình ảnh. Trong CVPR.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadar-rama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko và Trevor Darrell. 2015. Dài hạn mạng lưới tích chập tuần hoàn để nhận dạng và mô tả trực quan. Trong Biên bản báo cáo của IEEE Hội nghị về Thị giác máy tính và Nhận dạng mẫu, trang 2625–2634.

Kaiming He, Xiangyu Zhang, Shaoqing Ren và Jian CN. 2016. Học tập dư thừa sâu cho nhận dạng hình ảnh. Trong Biên bản báo cáo của hội nghị IEEE về tầm nhìn máy tính và nhận dạng mẫu, trang 770–778.

Lisa Anne Hendricks, Subhashini Venugopalan, Mar-cus Rohrbach, Raymond Mooney, Kate Saenko, và Trevor Darrell. 2016. Chú thích sáng tác sâu sắc: Mô tả các danh mục đối tượng mới lạ mà không dữ liệu đào tạo được ghép nối. Trong Biên bản báo cáo của IEEE Hội nghị về Thị giác máy tính và Nhận dạng mẫu, trang 1–10.

Sepp Hochreiter và Jurgen Schmidhuber. 1997. Bộ nhớ ngắn hạn dài hạn. Tính toán thần kinh, 9(8):1735–1780.

Andrej Karpathy và Li Fei-Fei. 2015. Căn chỉnh ngữ nghĩa thị giác sâu để tạo ra các mô tả hình ảnh. Trong Biên bản báo cáo của hội nghị IEEE về tầm nhìn máy tính và nhận dạng mẫu, các trang 3128–3137.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, và Erkut Erdem. 2016. Đánh giá lại tự động số liệu cho chú thích hình ảnh. Trong Chương Châu Â u của Hiệp hội Ngôn ngữ học tính toán.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin John-son, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Bộ gen trực quan: Kết nối ngôn ngữ và tầm nhìn sử dụng chú thích hình ảnh dày đặc do cộng đồng đóng góp. Tạp chí quốc tế về thị giác máy tính, 123(1):32–73.

Lin Tsung Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar và C Lawrence Zitnick. 2014. Microsoft dừa: Các đối tượng chung trong ngữ cảnh. Trong hội nghị châu Â u về tầm nhìn máy tính, trang 740–755. Springer.

Jiasen Lu, Jianwei Yang, Dhruv Batra và Devi Parikh. 2018. Trò chuyện của trẻ sơ sinh thần kinh. Trong Biên bản báo cáo Hội nghị IEEE về Thị giác máy tính và Nhận dạng mẫu.

Ruotian Luo, Brian Price, Scott Cohen và Gregory Shakhnarovich. 2018. Mục tiêu phân biệt đối xử cho đào tạo chú thích mô tả. Trong Biên bản của Hội nghị IEEE về Tầm nhìn máy tính và Mẫu Sự công nhận.

Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris và Edward Cutrell. 2017. Hiểu biết những trải nghiệm của người mù với máy tính tạo ra chú thích của hình ảnh phương tiện truyền thông xã hội. Trong Biên bản Hội nghị SIGCHI năm 2017 về các yếu tố con người trong Hệ thống máy tính.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross và Vaibhava Goel. 2017. Tự phê bình đào tạo trình tự cho chú thích hình ảnh. Trong Biên bản báo cáo của Hội nghị IEEE về Tầm nhìn máy tính và Nhận dạng mẫu.

Anna Rohrbach, Makarand Tapaswi, Atousa Torabi, Tegan Maharaj, Marcus Rohrbach, Sanja Fi-dler Christopher Pal và Bernt Schiele. 2017a. Hội thảo chung về hiểu biết ngôn ngữ và video: MovieQA và The Large Thử thách mô tả phim theo tỷ lệ (LSMDC). https://sites.google.com/site/ mô tảphim/lsmdc-2017.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville và Bernt Schiele. 2017b. Phim mô tả. Tạp chí quốc tế về thị giác máy tính, 123(1):94–120.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hen-dricks, Mario Fritz và Bernt Schiele. 2017. Nói cùng một ngôn ngữ: So sánh máy móc với con người chú thích bằng cách đào tạo đối nghịch. Trong Biên bản Hội nghị quốc tế IEEE về thị giác máy tính (ICCV).

Ramakrishna Vedantam, C Lawrence Zitnick và Devi Parikh. 2015. Cider: Đánh giá mô tả hình ảnh dựa trên sự đồng thuận. Trong Biên bản báo cáo của IEEE

Conference on Computer Vision and Pattern Recognition, pages 4566–4575.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Hội nghị về Thị giác máy tính và Nhận dạng mẫu, trang 4566–4575.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Cho Kyunghyun, Aaron Courville, Ruslan Salakhudinov, Zemel giàu có, và Yoshua Bengio. 2015. Hiển thị, tham dự và kể lại: Tạo chú thích hình ảnh thần kinh với sự chú ý trực quan. Trong Hội nghị quốc tế về máy móc Học tập, trang 2048–2057.

4045