SMALLCAP: Lightweight Image Captioning Prompted with Retrieval Augmentation

Rita Ramos† Bruno Martins† Desmond Elliott*,‡ Yova Kementchedjhieva*
†INESC-ID, Instituto Superior Técnico, University of Lisbon
*Department of Computer Science, University of Copenhagen
‡Pioneer Center for AI
ritaparadaramos@tecnico.ulisboa.pt

Abstract

Recent advances in image captioning have focused on scaling the data and model size, substantially increasing the cost of pre-training and finetuning. As an alternative to large models, we present SMALLCAP*, which generates a caption conditioned on an input image and related captions retrieved from a datastore. Our model is lightweight and fast to train, as the only learned parameters are in newly introduced cross-attention layers between a pre-trained CLIP encoder and GPT-2 decoder.* SMALLCAP *can transfer to new domains without additional finetuning and can exploit large-scale data in a training-free fashion since the contents of the datastore can be readily replaced. Our experiments show that* SMALLCAP*, trained only on COCO, has competitive performance on this benchmark, and also transfers to other domains without retraining, solely through retrieval from target-domain data. Further improvement is achieved through the training-free exploitation of diverse human-labeled and web data, which proves to be effective for a range of domains, including the* nocaps *benchmark, designed to test generalization to unseen visual concepts.*[1]

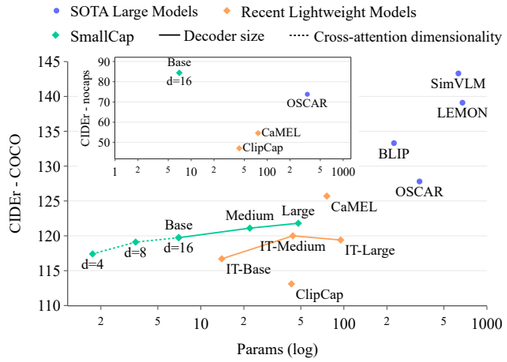


Figure 1. SMALLCAP's performance on the COCO dataset and on the out-of-domain split of the nocaps dataset, compared to other approaches in terms of number of trainable parameters. We can control the number of trainable parameters through the dimensionality of the cross-attention ($d = d_v = d_k$) and the size of the decoder. SMALLCAP is competitive to other lightweight models on COCO, and outperforms much larger models on nocaps.

1. Introduction

The state-of-the-art in image captioning is defined by increasingly large-scale models trained on increasingly large-scale datasets [11, 18, 39, 42]. Scaling up leads to higher computational demands for model pre-training and finetuning on downstream tasks. This becomes especially relevant when numerous model versions may be needed for different visual domains [1] and end-users in practical applications, e.g. image captioning for the visually impaired [10].

Some efforts have been made recently to reduce the cost of model training, e.g., ClipCap [25] and I-Tuning [22]. These models use an off-the-shelf pre-trained vision encoder and language decoder. The parameters of these pre-trained components are frozen and only a mapping between the two is trained for the task of image captioning. This results in a highly reduced number of trainable parameters (∼43M in each case) and faster training time. While these models operate on a much more manageable scale from a research perspective, they can still be unsuitable for the aforementioned practical applications, as both models require separate training for every use-case.

This work presents SMALLCAP, an image captioning model, prompted with captions retrieved from an external datastore of text, based on the input image. This formulation of image captioning enables a range of desirable prop-

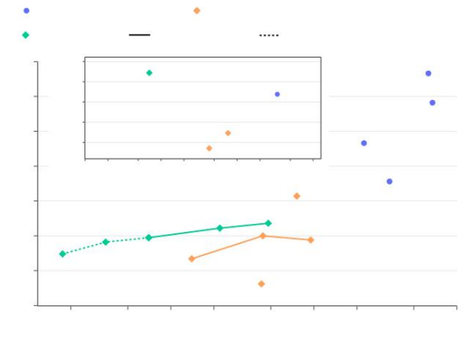
[1]Code: https://github.com/RitaRamo/smallcap.

SMALLCAP: Chú thích hình ảnh nhẹ được nhắc nhở với Tăng cường Truy xuất

Rita Ramos† Bruno Martins† Desmond Elliott?,‡ Yova Kementchedjhieva?
† INESC-ID, Instituto Superior Tecnico, Đại học Lisbon ?
Khoa Khoa học máy tính, Đại học Copenhagen
‡ Trung tâm tiên phong về AI
ritaparadaramos@tecnico.ulisboa.pt

Tóm tắt

Những tiến bộ gần đây trong chú thích hình ảnh đã tập trung vào mở rộng dữ liệu và kích thước mô hình, tăng đáng kể chi phí đào tạo trước và tinh chỉnh. Như một giải pháp thay thế đối với các mô hình lớn, chúng tôi trình bày SMALLCAP, tạo ra một chú thích có điều kiện trên hình ảnh đầu vào và chú thích liên quan được lấy từ kho dữ liệu. Mô hình của chúng tôi nhẹ và nhanh chóng để đào tạo, vì các tham số học được duy nhất nằm trong các lớp chú ý chéo mới được giới thiệu giữa một CLIP được đào tạo trước bộ mã hóa và bộ giải mã GPT-2. SMALLCAP có thể chuyển đến các miền mới mà không cần tinh chỉnh thêm và có thể khai thác dữ liệu quy mô lớn theo cách không cần đào tạo vì nội dung của kho dữ liệu có thể dễ dàng được thay thế. Các thí nghiệm của chúng tôi cho thấy SMALLCAP, chỉ được đào tạo trên COCO, có hiệu suất cạnh tranh trên chuẩn mực này và cũng chuyển sang các miền khác mà không cần đào tạo lại, chỉ thông qua việc truy xuất lại từ dữ liệu miền mục tiêu. Cải tiến hơn nữa là đạt được thông qua việc khai thác không cần đào tạo của nhiều dữ liệu web và được gắn nhãn của con người, chứng minh là có hiệu quả cho một loạt các miền, bao gồm chuẩn mực nocaps, được thiết kế để kiểm tra khả năng khái quát hóa các khái niệm trực quan chưa được biết đến.1



Hình 1. Hiệu suất của SMALLCAP trên tập dữ liệu COCO và về sự phân chia ngoài miền của tập dữ liệu nocaps, so với các cách tiếp cận khác về số lượng các tham số có thể đào tạo được. Chúng tôi có thể kiểm soát số lượng các tham số có thể đào tạo thông qua tính đa chiều của sự chú ý chéo (d = dv = dk) và kích thước của bộ giải mã. SMALLCAP có khả năng cạnh tranh với các mô hình nhẹ khác trên COCO và hoạt động tốt hơn nhiều so với các mô hình lớn hơn trên nocaps.

1. Giới thiệu

Tình trạng hiện đại trong chú thích hình ảnh được xác định bởi các mô hình ngày càng lớn được đào tạo trên các tập dữ liệu ngày càng lớn [11, 18, 39, 42]. Việc mở rộng quy mô dẫn đến cao hơn nhu cầu tính toán cho việc đào tạo trước mô hình và tinh chỉnh các tác vụ hạ lưu. Điều này trở nên đặc biệt có liên quan khi nhiều phiên bản mô hình có thể cần thiết cho các mục đích khác nhau lĩnh vực trực quan [1] và người dùng cuối trong các ứng dụng thực tế, ví dụ chú thích hình ảnh cho người khiếm thị [10].

Một số nỗ lực đã được thực hiện gần đây để giảm chi phí của đào tạo mô hình, ví dụ, ClipCap [25] và I-Tuning [22]. Các mô hình này sử dụng bộ mã hóa thị giác được đào tạo sẵn và bộ giải mã ngôn ngữ. Các tham số của các thành phần được đào tạo sẵn này bị đóng băng và chỉ có một ánh xạ giữa hai người được đào tạo cho nhiệm vụ chú thích hình ảnh. Điều này dẫn đến số lượng các tham số có thể đào tạo được giảm đáng kể (43M trong mỗi trường hợp) và thời gian đào tạo nhanh hơn. Trong khi những các mô hình hoạt động ở quy mô dễ quản lý hơn nhiều theo quan điểm nghiên cứu, chúng vẫn có thể không phù hợp với các ứng dụng thực tế đã đề cập ở trên, vì cả hai mô hình đều yêu cầu đào tạo riêng cho từng trường hợp sử dụng.

Tác phẩm này trình bày SMALLCAP, một chú thích hình ảnh mô hình, được nhắc nhở với các chú thích được lấy từ bên ngoài kho dữ liệu văn bản, dựa trên hình ảnh đầu vào. Công thức chú thích hình ảnh này cho phép một loạt các prop- mong muốn

1Mã: https://github.com/RitaRamo/smallcap.

erties: lightweight training, training-free domain transfer, and exploitation of large data in a training-free fashion.

SMALLCAP is both light to train and highly effective (see Figure 1).[2] It uses a pre-trained CLIP vision encoder [29] and GPT-2 language model [31], which are frozen and linked through new cross-attention layers amounting to 7 million trainable parameters. Through retrieval, the model leverages external data and therefore has to store less information within its weights (as demonstrated in Figure 6). Trained on the common COCO benchmark [7], SMALLCAP performs on par with other lightweight-training models, despite an 83% reduction in number of trainable parameters.

SMALLCAP can also leverage data in a training-free manner. Once the model is trained, we can replace the datastore with either (i) captions from a new domain or (ii) a large and diverse collection of captions. In the first case, which presents an alternative to finetuning, SMALLCAP gains access to the style and concepts that characterize the new domain and can generate captions accordingly. In the second case, which presents an alternative to generalized pre-training, SMALLCAP gains access to general knowledge that it can apply to any domain. Our experiments show that SMALLCAP effectively leverages new knowledge accessed through a retrieval-based prompt, improving its performance on different datasets. This includes the challenging VizWiz dataset, where images are captioned for the visually impaired [10], and the `nocaps` challenge dataset with rarely-seen and unseen visual concepts [1].

SMALLCAP competes with other lightweight-training models on in-domain evaluations and outperforms them by a large margin out-of-domain. It overcomes a key limitation of previous models, which require explicit finetuning to adapt to new domains, and in this way attests to the potential of retrieval augmentation for multimodal tasks.

2. Related Work

2.1. Image Captioning Models

Current approaches to image captioning employ encoder-decoder methods, where an input image is passed to a visual encoder and a caption is generated by an autoregressive language decoder [4, 46]. The state-of-the-art is currently held by general purpose vision-and-language (V&L) models [11, 18, 19, 42]. These large-scale models are pre-trained on large amounts of image-text pairs to learn generic multimodal features, after which they can be finetuned to a downstream task such as image captioning, with a separately-optimized model needed for each image captioning dataset. As such, these models require excessive resources for training and deployment.

[2]The `nocaps` results shown in the figure include only models that follow the challenge guidelines, by training on the COCO dataset only.

2.2. Freezing Image Captioning Models

Components of the image captioning model can be initialized with pre-trained weights, frozen in part or completely [2], as a way to prevent catastrophic forgetting [24], i.e. to maintain good generalization. As frozen model parameters require no gradient updates, training becomes faster and occupies less GPU memory. ClipCap and I-Tuning [22, 25] are two lightweight-training image captioning models which use a pre-trained vision encoder, CLIP [29], and language decoder, GPT-2 [31], as frozen model components. To map between these two independently trained components, ClipCap employs prefix-tuning, mapping a fixed-length CLIP embedding of the image into the GPT-2 language space. I-Tuning extracts *visual memory embeddings* from CLIP and uses those to adjust the output hidden states of GPT-2. In SMALLCAP, we also use CLIP and GPT-2, instead connected through a set of trainable cross-attention layers. The novelty here is that SMALLCAP uses retrieval augmentation to maintain performance while substantially reducing the number of trainable parameters.

2.3. Retrieval-Augmented Generation

Retrieval-augmented language generation consists of conditioning generation on additional information that is retrieved from an external datastore [16]. Retrieval augmentation has been gaining traction in other tasks [12, 17], but remains largely unexplored in image captioning. Some relevant works in image captioning include [32–34, 44, 49]. Closest to our work, Sarto *et al*. [34] and Ramos *et al*. [32] recently proposed retrieval-augmented transformer-based captioning models that perform cross-attention over the encoded retrieved captions. Our work differs from previous work in two main ways. We employ a simple prompt-based conditioning method, wherein retrieved captions are used as a prompt to a generative language model. Moreover, we are the first to leverage retrieval augmentation for training-free domain transfer and generalization in image captioning.

2.4. Prompting Text Generation

Prompts have become a common way to pass additional instructions and task demonstrations to a pre-trained language model [30]. In vision-and-language learning, prompts have been used to instruct a model to perform one of multiple tasks it was trained for [18], or to apply it to a new task in a zero-shot fashion [13, 36]. We use prompts with a task demonstration tailored to the specific input image, as a means towards retrieval augmentation.

3. Proposed Approach

3.1. Model

SMALLCAP is a lightweight-training image captioning model augmented with retrieved captions through the use of

erties: đào tạo nhẹ, chuyển miền không cần đào tạo,
và khai thác dữ liệu lớn theo cách không cần đào tạo.

SMALLCAP vừa nhẹ để tập luyện vừa có hiệu quả cao
(xem Hình 1).2 Nó sử dụng bộ mã hóa tầm nhìn CLIP đư ợc đào tạo trư ớc
[29] và mô hình ngôn ngữ GPT-2 [31], đư ợc đóng băng và
đư ợc liên kết thông qua các lớp chú ý chéo mới lên tới 7
triệu tham số có thể đào tạo đư ợc. Thông qua việc truy xuất, mô hình
tận dụng dữ liệu bên ngoài và do đó phải lư u trữ ít thông tin hơn trong
trọng số của nó (như minh họa trong Hình 6).
Đư ợc đào tạo trên chuẩn COCO chung [7], SMALLCAP
có hiệu suất ngang bằng với các mô hình luyện tập nhẹ khác, mặc dù số lư ợng
các thông số có thể luyện tập đư ợc giảm 83%.

SMALLCAP cũng có thể tận dụng dữ liệu trong quá trình đào tạo miễn phí
cách. Khi mô hình đư ợc đào tạo, chúng ta có thể thay thế
kho dữ liệu với (i) chú thích từ một miền mới hoặc
(ii) một bộ sư u tập lớn và đa dạng các chú thích. Trong lần đầu tiên
trư ờng hợp, trong đó trình bày một giải pháp thay thế cho việc tinh chỉnh,
SMALL- CAP có quyền truy cập vào phong cách và các khái niệm đặc trư ng cho
miền mới và có thể tạo chú thích theo đó. Trong trư ờng hợp thứ hai, trong
đó trình bày một giải pháp thay thế cho
đào tạo trư ớc tổng quát, SMALLCAP có đư ợc quyền truy cập vào kiến thức
chung mà nó có thể áp dụng cho bất kỳ miền nào. Các thí nghiệm của chúng
tôi cho thấy SMALLCAP tận dụng hiệu quả các kiến thức mới
kiến thức đư ợc truy cập thông qua lời nhắc dựa trên truy xuất, cải thiện
hiệu suất của nó trên các tập dữ liệu khác nhau. Điều này bao gồm
bộ dữ liệu VizWiz đầy thử thách, trong đó hình ảnh đư ợc chú thích
dành cho ngư ời khiếm thị [10] và thử thách nocaps
tập dữ liệu với các khái niệm trực quan hiếm thấy và chư a từng thấy [1].

SMALLCAP cạnh tranh với các chư ơng trình đào tạo nhẹ khác
các mô hình đánh giá trong miền và vư ợt trội hơn chúng bằng
một biên độ lớn ngoài phạm vi. Nó khắc phục đư ợc một hạn chế quan trọng của
các mô hình trư ớc đó, đòi hỏi phải tinh chỉnh rõ ràng
để thích ứng với các lĩnh vực mới và theo cách này chứng minh tiềm năng tăng
cư ờng khả năng truy xuất cho các nhiệm vụ đa phư ơng thức.

2. Công trình liên quan

2.1. Mô hình chú thích hình ảnh

Các phư ơng pháp tiếp cận hiện tại để chú thích hình ảnh sử dụng
phư ơng pháp mã hóa-giải mã, trong đó hình ảnh đầu vào đư ợc truyền qua
đến một bộ mã hóa trực quan và một chú thích đư ợc tạo ra bởi một bộ giải
mã ngôn ngữ tự động hồi quy [4, 46]. Tình trạng nghệ thuật
hiện đang đư ợc tổ chức bởi tầm nhìn và ngôn ngữ mục đích chung
(V&L) mô hình [11, 18, 19, 42]. Các mô hình quy mô lớn này
đư ợc đào tạo trư ớc trên một lư ợng lớn cặp hình ảnh-văn bản để
tìm hiểu các tính năng đa phư ơng thức chung, sau đó chúng có thể đư ợc
đư ợc tinh chỉnh cho một nhiệm vụ hạ nguồn như chú thích hình ảnh,
với một mô hình đư ợc tối ư u hóa riêng biệt cần thiết cho mỗi hình ảnh
tập dữ liệu chú thích. Do đó, các mô hình này đòi hỏi quá nhiều
nguồn lực cho đào tạo và triển khai.

2Kết quả nocaps đư ợc hiển thị trong hình chỉ bao gồm các mô hình sau
hư ớng dẫn thử thách thấp, chỉ bằng cách đào tạo trên tập dữ liệu COCO.

2.2. Đóng băng các mô hình chú thích hình ảnh

Các thành phần của mô hình chú thích hình ảnh có thể đư ợc khởi tạo bằng
trọng số đư ợc đào tạo trư ớc, đóng băng một phần hoặc toàn bộ [2], như một
cách để ngăn ngừa tình trạng quên thảm khốc [24],
tức là để duy trì sự khái quát tốt. Như mô hình đóng băng
các tham số không yêu cầu cập nhật gradient, đào tạo trở thành
nhanh hơn và chiếm ít bộ nhớ GPU hơn. ClipCap và I-Tuning [22, 25] là hai
mô hình chú thích hình ảnh đào tạo nhẹ sử dụng bộ mã hóa thị giác đư ợc đào
tạo trư ớc,
CLIP [29] và bộ giải mã ngôn ngữ, GPT-2 [31], như đã đóng băng
thành phần mô hình. Để ánh xạ giữa hai thành phần đư ợc đào tạo độc lập này,
ClipCap sử dụng điều chỉnh tiền tố,
ánh xạ một đoạn nhúng CLIP có độ dài cố định của hình ảnh vào
không gian ngôn ngữ GPT-2. I-Tuning trích xuất bộ nhớ hình ảnh
nhúng từ CLIP và sử dụng chúng để điều chỉnh đầu ra
trạng thái ẩn của GPT-2. Trong SMALLCAP, chúng tôi cũng sử dụng CLIP
và GPT-2, thay vào đó đư ợc kết nối thông qua một tập hợp các
lớp chú ý chéo. Điểm mới lạ ở đây là SMALLCAP
sử dụng tăng cư ờng truy xuất để duy trì hiệu suất trong khi
giảm đáng kể số lư ợng các tham số có thể đào tạo đư ợc.

2.3. Thế hệ tăng cư ờng truy xuất

Thế hệ ngôn ngữ tăng cư ờng truy xuất bao gồm
tạo điều kiện trên thông tin bổ sung đư ợc lấy lại từ kho dữ liệu bên ngoài
[16]. Việc tăng cư ờng truy xuất đã và đang đư ợc chú ý trong các nhiệm vụ
khác [12, 17], như ng
vẫn chư a đư ợc khám phá nhiều trong chú thích hình ảnh. Một số tác phẩm có
liên quan trong chú thích hình ảnh bao gồm [32–34, 44, 49].
Gần nhất với công trình của chúng tôi là Sarto et al. [34] và Ramos et al. [32]
gần đây đề xuất tăng cư ờng thu hồi dựa trên máy biến áp
các mô hình chú thích thực hiện sự chú ý chéo qua các chú thích đã đư ợc mã
hóa. Công việc của chúng tôi khác với các công trình trư ớc đây
làm việc theo hai cách chính. Chúng tôi sử dụng một lời nhắc đơn giản dựa trên
phư ơng pháp điều kiện hóa, trong đó các chú thích đư ợc lấy ra đư ợc sử dụng như
một lời nhắc đến một mô hình ngôn ngữ tạo sinh. Hơn nữa, chúng tôi
đầu tiên tận dụng sự gia tăng khả năng truy xuất để đào tạo miễn phí
chuyển miền và khái quát hóa trong chú thích hình ảnh.

2.4. Nhắc nhở tạo văn bản

Lời nhắc đã trở thành một cách phổ biến để truyền đạt các hư ớng dẫn bổ
sung và trình diễn nhiệm vụ cho một ngư ời đư ợc đào tạo trư ớc
mô hình ngôn ngữ [30]. Trong việc học ngôn ngữ và thị giác,
lời nhắc đã đư ợc sử dụng để hư ớng dẫn một mô hình thực hiện một
của nhiều nhiệm vụ mà nó đư ợc đào tạo cho [18], hoặc để áp dụng nó vào một
nhiệm vụ mới theo cách không có cú đánh nào [13, 36]. Chúng tôi sử dụng lời nhắc
với một bản trình bày nhiệm vụ đư ợc thiết kế riêng cho hình ảnh đầu vào cụ thể,
như một phư ơng tiện hư ớng tới việc tăng cư ờng khả năng truy xuất.

3. Phư ơng pháp tiếp cận đư ợc đề xuất

3.1. Mô hình

SMALLCAP là một chú thích hình ảnh đào tạo nhẹ
mô hình đư ợc tăng cư ờng với các chú thích đư ợc lấy thông qua việc sử dụng

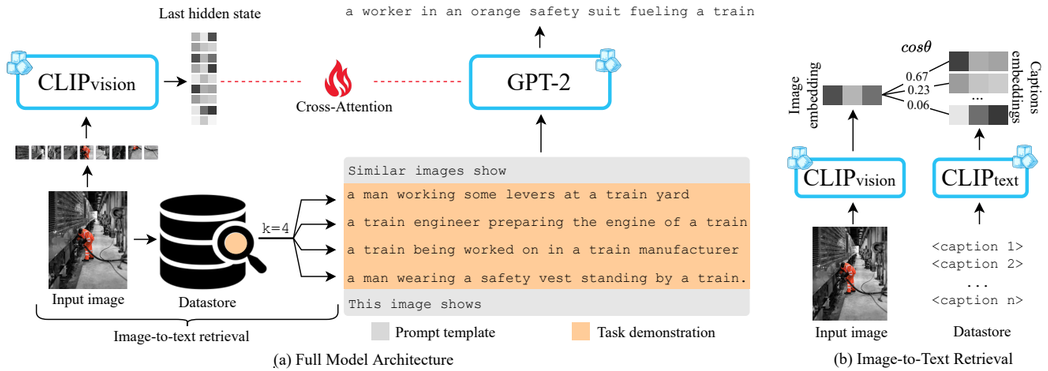


Figure 2. The SMALLCAP approach to image captioning. (a) SMALLCAP generates a caption conditioned on the encoded input image, as well as a set of k retrieved captions which are used as a task demonstration, input to the decoder as a prompt. (b) The k captions are retrieved from a datastore of N captions via image-to-text retrieval.

a prompt. SMALLCAP combines powerful pre-trained unimodal models in an encoder-decoder architecture, as shown in Figure 2 (a). As encoder we use CLIP [29], which produces a sequence of patch embeddings. As decoder we use GPT-2 [31]. These two models operate in different vector spaces, so we connect them with multi-head cross-attention, through which each layer of the decoder attends to the encoder outputs [37]. In order to reduce the compute requirements for training and to preserve their generalization capabilities, we freeze the encoder and decoder and only train the randomly-initialized cross-attention layers between them. We further control the number of trainable parameters through the dimensionality of the projection matrices in the cross-attention layers, which we denote as d. For GPT-2, a model with $d_{model} = 768$ hidden dimensions and $h = 12$ cross-attention heads, d defaults to 64 (d_{model}/h), as per Vaswani *et al.* [37], but can be arbitrarily set to any value (see Appendix A for more details).

Similarly to retrieval-augmented models for other tasks [12, 16, 17, 40], SMALLCAP does not need to store all necessary information within its parameters, because it has access to external knowledge from a datastore of text.

3.2. Prompting with Retrieved Captions

Instead of the image-to-image retrieval methods used in recent work [34], which are limited to image captioning data in the datastore, we employ image-to-text retrieval, as shown in Figure 2 (b). In this way, SMALLCAP can make use of a datastore containing any type of text that is considered useful for describing images, be that image captions, video captions, audio captions, etc. Here, we exploit the full CLIP model, with its vision and text encoders, which map the two modalities into a shared vector space. We encode an input image and the contents of the datastore, and use nearest neighbor search based on cosine similarity to retrieve the k text items from the datastore most similar to the image. The retrieved text is used to fill the slots in a fixed prompt template of the following form: `Similar images show` $\{\texttt{caption}_1\}\ldots\{\texttt{caption}_k\}$. `This image shows ___`.[3] The last sentence of the prompt is similar to the simple, fixed prompts used in other studies [18], but here this cue is preceded by a demonstration of the captioning task, tailored to the input image. The decoder receives this prompt as input tokens and then generates a caption conditioned on the image features $\mathbf{V}$ and the task demonstration $\mathbf{X}$. The weights in the cross-attention layers (θ) are trained by minimizing the cross-entropy loss of predicting the M tokens in the reference $y_1, \ldots, y_M$:

$$L_\theta = -\sum_{i=1}^{M} \log P_\theta(y_i | y_{<i}, \mathbf{X}, \mathbf{V}; \theta). \quad (1)$$

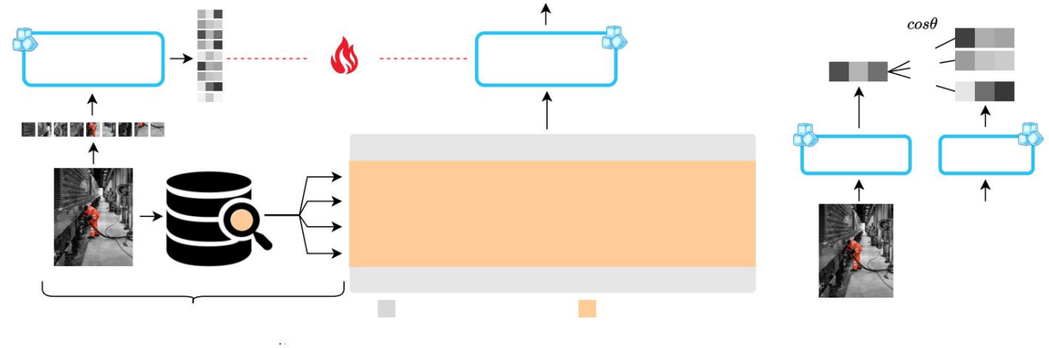
The datastore used to train SMALLCAP can change from training to inference, depending on the application. For example, additional data can be added to enable better generalization, or the datastore can be entirely swapped for new data at inference time to enable domain transfer without the need for retraining, as shown in Section 5.

4. Main Experiments

4.1. Experimental Setup

SMALLCAP's encoder and decoder are initialized respectively from CLIP-ViT-B/32 and GPT-2_{Base}, as available

[3] See Appendix C for more information on the prompt template.



Hình 2. Phương pháp SMALLCAP để chú thích hình ảnh. (a) SMALLCAP tạo ra một chú thích có điều kiện trên hình ảnh đầu vào được mã hóa, cũng như một tập hợp k chú thích đã lấy được được sử dụng như một bản trình diễn nhiệm vụ, đầu vào cho bộ giải mã như một lời nhắc. (b) k chú thích là được lấy từ kho dữ liệu gồm N chú thích thông qua việc truy xuất hình ảnh thành văn bản.

một lời nhắc. SMALLCAP kết hợp các mô hình đơn phương thức được đào tạo trước mạnh mẽ trong kiến trúc mã hóa-giải mã, như được hiển thị trong Hình 2 (a). Với tư cách là bộ mã hóa, chúng tôi sử dụng CLIP [29], tạo ra một chuỗi nhúng bản vá. Với tư cách là bộ giải mã, chúng tôi sử dụng GPT-2 [31]. Hai mô hình này hoạt động khác nhau không gian vectơ, vì vậy chúng tôi kết nối chúng với sự chú ý chéo nhiều đầu, thông qua đó mỗi lớp của bộ giải mã tham dự đến đầu ra của bộ mã hóa [37]. Để giảm yêu cầu tính toán cho việc đào tạo và bảo toàn khả năng khái quát hóa của chúng, chúng tôi đóng băng bộ mã hóa và bộ giải mã và chỉ đào tạo các lớp chú ý chéo được khởi tạo ngẫu nhiên giữa chúng. Chúng tôi tiếp tục kiểm soát số lượng tham số có thể đào tạo thông qua tính đa chiều của các ma trận chiếu trong các lớp chú ý chéo, mà chúng tôi ký hiệu là d. Đối với GPT-2, một mô hình với dmodel = 768 chiều ẩn và h = 12 đầu chú ý chéo, d mặc định là 64 (dmodel/h), theo Vaswani et al. [37], như ng có thể được thiết lập tùy ý thành bất kỳ giá trị (xem Phụ lục A để biết thêm chi tiết).

Tương tự như các mô hình tăng cường truy xuất cho các mô hình khác nhiệm vụ [12, 16, 17, 40], SMALLCAP không cần phải lưu trữ tất cả thông tin cần thiết trong các tham số của nó, bởi vì nó có quyền truy cập vào kiến thức bên ngoài từ kho dữ liệu văn bản.

3.2. Nhắc nhở với Phụ đề đã Lấy

Thay vì các phương pháp truy xuất hình ảnh sang hình ảnh được sử dụng trong công trình gần đây [34], được giới hạn trong việc chú thích hình ảnh dữ liệu trong kho dữ liệu, chúng tôi sử dụng phương pháp truy xuất hình ảnh thành văn bản, như được hiển thị trong Hình 2 (b). Theo cách này, SMALLCAP có thể tạo ra sử dụng kho dữ liệu chứa bất kỳ loại văn bản nào được coi là hữu ích để mô tả hình ảnh, có thể là chú thích hình ảnh, phụ đề video, phụ đề âm thanh, v.v. Ở đây, chúng tôi khai thác mô hình CLIP đầy đủ, với tầm nhìn và bộ mã hóa văn bản, ánh xạ hai phương thức vào một không gian vectơ được chia sẻ. Chúng tôi encode mã hóa hình ảnh đầu vào và nội dung của kho dữ liệu, và sử dụng tìm kiếm hàng xóm gần nhất dựa trên độ tương đồng cosin với lấy k mục văn bản từ kho dữ liệu giống nhất với hình ảnh. Văn bản được lấy lại được sử dụng để điền vào các ô trong mẫu nhắc nhở cố định có dạng sau: Tương tự hình ảnh hiển thị {caption1}...{captionk}. Điều này hình ảnh hiển thị . ___[3] Câu cuối cùng của lời nhắc là tương tự như các lời nhắc đơn giản, cố định được sử dụng trong các nghiên cứu khác [18], như ng ở đây lời nhắc này được đưa ra trước bằng một cuộc trình diễn nhiệm vụ chú thích, được điều chỉnh theo hình ảnh đầu vào. Bộ giải mã nhận lời nhắc này như là mã thông báo đầu vào và sau đó tạo ra một chú thích có điều kiện trên các tính năng hình ảnh V và nhiệm vụ trình diễn X. Các trọng số trong các lớp chú ý chéo ( ) được đào tạo bằng cách giảm thiểu tổn thất entropy chéo khi dự đoán M mã thông báo trong tham chiếu y1,...,yM:

$$L = \sum_{\text{tôi}=1}^{\text{Tôi}} \log P\ (\text{yi}|\text{y<i, X, V; }). \quad (1)$$

Kho dữ liệu được sử dụng để đào tạo SMALLCAP có thể thay đổi từ đào tạo để suy luận, tùy thuộc vào ứng dụng. Ví dụ, dữ liệu bổ sung có thể được thêm vào để cho phép tổng quát hóa tốt hơn hoặc kho dữ liệu có thể được hoán đổi hoàn toàn cho dữ liệu mới dữ liệu tại thời điểm suy luận để cho phép chuyển miền mà không cần nhu cầu đào tạo lại, như thể hiện ở Mục 5.

4. Thí nghiệm chính

4.1. Thiết lập thử nghiệm

Bộ mã hóa và giải mã của SMALLCAP được khởi tạo từ ứng từ CLIP-ViT-B/32 và GPT-2Base, tùy theo khả dụng

3Xem Phụ lục C để biết thêm thông tin về mẫu lời nhắc.

Model	$\|\theta\|$	B@4	M	CIDEr	S
Large Models with V&L pre-training					
LEMON$_{\text{Huge}}$ [11]	675	**41.5**	30.8	139.1	24.1
SimVLM$_{\text{Huge}}$ [42]	632	40.6	**33.7**	**143.3**	**25.4**
OSCAR$_{\text{Large}}$ [19]	338	37.4	30.7	127.8	23.5
BLIP$_{\text{CapFilt-L}}$ [18]	224	39.7	-	133.3	-
Lightweight-training models					
I-Tuning$_{\text{Large}}$ [22]	95	34.8	29.3	119.4	**22.4**
CaMEL [5]	76	**39.1**	**29.4**	**125.7**	22.2
I-Tuning$_{\text{Medium}}$ [22]	44	35.5	28.8	120.0	22.0
ClipCap [25]	43	33.5	27.5	113.1	21.1
I-Tuning$_{\text{Base}}$ [22]	14	34.8	28.3	116.7	21.8
SMALLCAP	**7**	37.0	27.9	119.7	21.3
SMALLCAP$_{\text{d=16, Large}}$	47	37.2	28.3	121.8	21.5
SMALLCAP$_{\text{d=16, Med}}$	22	36.5	28.1	120.7	21.6
SMALLCAP$_{\text{d=8, Base}}$	3.6	36.7	27.8	119.1	21.1
SMALLCAP$_{\text{d=4, Base}}$	1.8	36.0	27.4	117.4	21.0

Table 1. Results on the COCO test set with cross-entropy training. $|\theta|$: number of trainable parameters in the model (in millions).

on HuggingFace [43]. The encoder and decoder are not updated and only the cross-attention layers between them are trained. A 12-head cross-attention layer is added to each of the 12 layers of GPT-2. To achieve a low number of trainable parameters, we vary the dimensionality of the projection matrices in the cross-attention layers, d, by scaling from the default size of 64 down to 16, 8 and 4, which results in model variants with 7M, 3.6M and 1.8M trainable parameters, respectively. Our main model, SMALLCAP, has 7M trainable parameters and a total of 218M parameters (including the frozen CLIP encoder and GPT-2 decoder).

The cross-attention layers are trained on the COCO dataset [7] using the standard Karpathy splits [15]. The models are trained to minimize the cross-entropy loss using an AdamW optimizer [21] with an initial learning rate of 1e-4 and a batch size of 64. Training runs for 10 epochs and we use the epoch checkpoint with the best CIDEr score on the validation set. Training takes up to 8 hours on a single NVIDIA A100 GPU, using 16 GB of the available memory.

During training, the model is prompted with a set of $k = 4$ captions per image, retrieved from a datastore of the training captions from COCO. Retrieval is based on CLIP-ResNet-50x64[4] representations of input images and captions in the datastore, the latter being precomputed offline and indexed with FAISS [14] for efficient nearest neighbor searching.[5] During inference, the model generates a caption using beam search decoding with a beam size of 3. Inference, including retrieval and prompting, takes 0.22 seconds on average across 1,000 randomly sampled images, compared to 0.19 seconds without retrieval. For more details on design choices and hyperparameters, see Appendix B.

For evaluation, we compute the standard metrics: BLEU-4 (B@4) [27], METEOR (M) [8], CIDEr [38], and SPICE (S) [3], using the COCO evaluation package.[6]

[4]Downloaded from https://github.com/openai/CLIP

[5]We use an inner product index (`IndexFlatIP`) without any training and normalize the representations to search based on cosine similarity.

Model	In	Near	Out	Entire
OSCAR$_{\text{Large}}^{\diamond}$	84.8	82.1	73.8	80.9
CaMEL*	**88.1**	79.1	54.6	75.9
ClipCap*	74.5	65.6	47.1	63.4
SMALLCAP	83.3	77.1	65.0	75.8
SMALLCAP$_{\text{+W+H}}$	87.9	**84.6**	**84.4**	**85.0**

Table 2. CIDEr results on the `nocaps` test set. $\diamond$: Results copied from the respective publications. $\star$: Results computed by us. +W+H: datastore with additional Web and Human-labeled data.

4.2. Benchmark Results

Here, we report results on COCO [7], as well as on `nocaps` [1], a challenge dataset for evaluating the generalization capabilities of models trained on COCO.

COCO: In Table 1 we benchmark our approach on the COCO dataset. In the top half of the table, we acknowledge the strong performance of large-scale pre-trained models, ranging in size from 224M to 675M trainable parameters. We also note that these models are pre-trained on 4M–1.8B image-caption pairs, i.e., much more than the COCO data.

In the lower half of the table we see how our approach compares to other lightweight-training models. With only 7M parameters, SMALLCAP performs better or on par with ClipCap and I-Tuning. In this in-domain setting, it is only outperformed by CaMEL, which is trained end-to-end with eleven times as many trainable parameters. Reducing the number of trainable parameters to 3.6M, SMALLCAP$_{\text{d=8, Base}}$ still yields competitive performance, and even with just 1.8M trainable parameters, SMALLCAP$_{\text{d=4, Base}}$ is better than the substantially larger models ClipCap and I-Tuning$_{\text{Base}}$. We also experiment with Medium and Large GPT-2 decoders (SMALLCAP$_{\text{Medium}}$ and SMALLCAP$_{\text{Large}}$ in Table 1), and find that performance scales: by one CIDEr point from Base to Medium and by another point from Medium to Large.[7] Despite its small size, SMALLCAP shows competitive performance on COCO, the dataset it was trained on. In contrast to previous lightweight-training models, SMALLCAP further has the ability to generalize and transfer out-of-domain without retraining, as shown in subsequent experiments.

[6]https://github.com/tylin/coco-caption

[7]See Appendix E for more results regarding scaling the decoder.

Ngư ời mẫu	| | B@4 M CIDEr S
Các mô hình lớn với V&L đư ợc đào tạo trư ớc	
CHÌM LÒNG [11]	675 41,5 30,8 139,1 24,1
SimVLMHuge [42]	632 40,6 33,7 143,3 25,4
OSCAR Lớn [19]	338 37,4 30,7 127,8 23,5
BLIPCapFilt-L [18]	224 39,7 133,3 -
Các mô hình đào tạo nhẹ	
I-Tuning Lớn [22]	95 34,8 29,3 119,4 22,4
CaMEL [5]	76 39,1 29,4 125,7 22,2
Trung bình [22]	44 35,5 28,8 120,0 22,0
ClipCap [25]	43 33,5 27,5 113,1 21,1
Cơ sở điều chỉnh I [22]	14 34,8 28,3 116,7 21,8
VỐN NHỎ 7 37.0 27.9 119.7 21.3	
SMALLCAPd=16, Lớn	47 37,2 28,3 121,8 21,5
SMALLCAPd=16, Trung bình	22 36,5 28,1 120,7 21,6
SMALLCAPd=8, Cơ sở	3,6 36,7 27,8 119,1 21.1
SMALLCAPd=4, Cơ sở	1,8 36,0 27,4 117,4 21,0

Bảng 1. Kết quả trên bộ kiểm tra COCO với đào tạo entropy chéo.
| |: số lư ợng tham số có thể đào tạo đư ợc trong mô hình (tính bằng triệu).

trên HuggingFace [43]. Bộ mã hóa và bộ giải mã không đư ợc cập nhật và chỉ có các lớp chú ý chéo giữa chúng đư ợc đư ợc đào tạo. Một lớp chú ý chéo 12 đầu đư ợc thêm vào mỗi của 12 lớp GPT-2. Để đạt đư ợc số lư ợng thấp các tham số có thể đào tạo, chúng tôi thay đổi chiều của các ma trận chiếu trong các lớp chú ý chéo, d, bằng cách chia tỷ lệ từ kích thư ớc mặc định là 64 xuống còn 16, 8 và 4, dẫn đến các biến thể mô hình có thể đào tạo đư ợc 7M, 3,6M và 1,8M tham số, tư ơng ứng. Mô hình chính của chúng tôi, SMALLCAP, có 7 triệu tham số có thể đào tạo và tổng cộng 218 triệu tham số (bao gồm bộ mã hóa CLIP bị đóng băng và bộ giải mã GPT-2).

Các lớp chú ý chéo đư ợc đào tạo trên COCO tập dữ liệu [7] sử dụng các phân chia Karpathy chuẩn [15]. các mô hình đư ợc đào tạo để giảm thiểu tổn thất entropy chéo bằng cách sử dụng một trình tối ư u hóa AdamW [21] với tốc độ học ban đầu là 1e-4 và kích thư ớc lô là 64. Chạy đào tạo trong 10 kỷ nguyên và chúng tôi sử dụng điểm kiểm tra kỷ nguyên với điểm CIDEr tốt nhất trên bộ xác thực. Đào tạo mất tới 8 giờ trên một GPU NVIDIA A100, sử dụng 16 GB bộ nhớ khả dụng.

Trong quá trình đào tạo, mô hình đư ợc nhắc nhở bằng một tập hợp k = 4 chú thích cho mỗi hình ảnh, đư ợc lấy từ kho dữ liệu của chú thích đào tạo từ COCO. Việc truy xuất dựa trên các biểu diễn CLIP-ResNet-50x644 của hình ảnh đầu vào và chú thích trong kho dữ liệu, sau này đư ợc tính toán trư ớc ngoại tuyến và đư ợc lập chỉ mục với FAISS [14] để có hàng xóm gần nhất hiệu quả tìm kiếm.5 Trong quá trình suy luận, mô hình tạo ra một chú thích sử dụng giải mã tìm kiếm chùm tia với kích thư ớc chùm tia là 3. Suy luận, bao gồm truy xuất và nhắc nhở, mất 0,22 giây trung bình trên 1.000 hình ảnh đư ợc lấy mẫu ngẫu nhiên, so với 0,19 giây mà không cần truy xuất. Để biết thêm chi tiết về lựa chọn thiết kế và siêu tham số, xem Phụ lục B.

4Tải xuống từ https://github.com/openai/CLIP 5Chúng tôi sử dụng chỉ số tích trong (IndexFlatIP) mà không cần bất kỳ đào tạo nào và chuẩn hóa các biểu diễn để tìm kiếm dựa trên độ tư ơng đồng cosin.

Ngư ời mẫu	TRONG Gần Ra Toàn Bộ
OSCAR Lớn	84,8 82,1 73,8 80,9
CaMEL? 88,1 79,1 54,6 75,9	
ClipCap? 74,5 65,6 47,1 63,4	
VỐN NHỎ 83,3 77,1 65,0 75,8	
CHỮ HOA NHỎ+RỘNG+CAO 87,9 84,6 84,4 85,0	

Bảng 2. Kết quả CIDEr trên bộ thử nghiệm nocaps. : Kết quả đã sao chép từ các ấn phẩm tư ơng ứng. ?: Kết quả do chúng tôi tính toán. +W+H: kho dữ liệu có thêm dữ liệu đư ợc gắn nhãn bởi Web và Con ngư ời.

Để đánh giá, chúng tôi tính toán các số liệu tiêu chuẩn: BLEU-4 (B@4) [27], METEOR (M) [8], CIDEr [38], và SPICE (S) [3], sử dụng gói đánh giá COCO.6

4.2. Kết quả chuẩn

Ở đây, chúng tôi báo cáo kết quả về COCO [7], cũng như về nocaps [1], một tập dữ liệu thử thách để đánh giá khả năng tổng quát hóa của các mô hình đư ợc đào tạo trên COCO.

COCO: Trong Bảng 1, chúng tôi đánh giá chuẩn mực cách tiếp cận của mình trên Bộ dữ liệu COCO. Trong nửa trên của bảng, chúng tôi thừa nhận hiệu suất mạnh mẽ của các mô hình đư ợc đào tạo trư ớc quy mô lớn, có kích thư ớc từ 224M đến 675M các thông số có thể đào tạo đư ợc. Chúng tôi cũng lư u ý rằng các mô hình này đư ợc đào tạo trư ớc trên 4M-1.8B cặp hình ảnh-chú thích, nghĩa là nhiều hơn nhiều so với dữ liệu COCO.

Ở nửa dư ới của bảng, chúng ta thấy cách tiếp cận của chúng tôi so sánh với các mô hình đào tạo nhẹ khác. Chỉ với 7M tham số, SMALLCAP hoạt động tốt hơn hoặc ngang bằng với ClipCap và I-Tuning. Trong cài đặt trong miền này, nó là chỉ bị CaMEL đánh bại, đư ợc đào tạo từ đầu đến cuối với số lư ợng tham số có thể đào tạo nhiều gấp mư ời một lần. Giảm số lư ợng tham số có thể đào tạo xuống còn 3,6M, SMALL- CAPd=8, Cơ sở vẫn mang lại hiệu suất cạnh tranh và thậm chí chỉ với 1,8 triệu tham số có thể đào tạo, SMALLCAPd=4, Cơ sở là tốt hơn nhiều so với các mô hình lớn hơn đáng kể là ClipCap và I-TuningBase. Chúng tôi cũng thử nghiệm với Medium và Large Bộ giải mã GPT-2 (SMALLCAPMedium và SMALLCAPLarge trong Bảng 1) và thấy rằng hiệu suất đư ợc điều chỉnh: theo một CIDEr điểm từ Cơ sở đến Trung bình và bằng một điểm khác từ Trung bình đến Lớn.7 Mặc dù có kích thư ớc nhỏ, SMALLCAP cho thấy hiệu suất cạnh tranh trên COCO, tập dữ liệu nó đã đư ợc đào tạo. Trái ngư ợc với đào tạo nhẹ trư ớc đây mô hình, SMALLCAP còn có khả năng khái quát hóa và chuyển ra khỏi miền mà không cần đào tạo lại, như đư ợc hiển thị trong các thí nghiệm tiếp theo.

6https://github.com/tylin/coco-caption

7Xem Phụ lục E để biết thêm kết quả về việc mở rộng bộ giải mã.

nocaps: Results on the `nocaps` test set are reported in Table 2.[8,9,10] SMALLCAP clearly outperforms other lightweight methods *Out*-of-domain and achieves competitive performance *In*-domain and *Near*-domain. The model's strong generalization capabilities point to it being less prone to over-fitting as it does not need to memorize its training data, available also through retrieval. Our model can further improve when additional data is placed in the datastore, as seen in SMALLCAP$_{+W+H}$. In this variant, described in more detail in Section 5.2, the COCO datastore is augmented with diverse web (W) and human-labeled (H) data. SMALLCAP$_{+W+H}$ shows impressive generalization capabilities, outperforming the much larger OSCAR$_{\text{Large}}$ by over 10 points in the *Out*-of-domain setting. Next, we further explore SMALLCAP's ability for training-free transfer to new domains on diverse datasets.

5. Training-Free Use of Data

In this section, we study SMALLCAP's ability to leverage new data in its datastore in a training-free manner, i.e. all experiments presented here constitute changes made to the datastore at inference time, while the model, trained on COCO, remains fixed. The focus is on out-of-domain performance as measured on a diverse set of captioning datasets: Flick30k [47], VizWiz [10] and MSR-VTT [45]. The latter is in fact a video captioning dataset, which we adapt by converting video clips into an image of four 4 frames, sampled at 0, 25, 50 and 100% of the clip duration (see the MSR-VTT example in Figure 5). We start by exploring different configurations of the datastore, with the results in Table 3 reported on validation data.

5.1. In-domain Data

In the top of Table 3, we show how SMALLCAP performs when its datastore is populated with the training data associated with each respective dataset (*In-domain*). In comparison to using COCO captions in the datastore (*COCO*), the model performance substantially increases for all three datasets. This shows that SMALLCAP adapts to the retrieved information to achieve domain transfer. The improvement is most notable for VizWiz, likely because the nature of this dataset is very distinct from COCO, and thus there is a larger domain gap to be closed.

5.2. Augmenting the Datastore

In Table 3 (*Datastore augmentation*), we augment the in-domain datastore with additional large-scale data in an effort to improve generalization. We experiment with diverse web data (which is large-scale but automatically labeled) and human-labeled data (smaller-scale but clean).[11]

SMALLCAP datastore	F30K	VW	MV
COCO	52.2	34.5	23.3
In-domain	55.4	47.7	29.2
Datastore augmentation			
In-domain + **W**eb	**58.6**	**48.0**	29.8
In-domain + **H**uman-labeled	57.6	47.5	**30.9**
In-domain + **W** + **H**	57.9	**48.0**	30.7
Domain-agnostic			
Web	58.4	42.4	27.6
Human-labeled	56.6	36.4	29.0
Web + Human-labeled	57.8	42.2	29.9

Table 3. Exploration of the training-free use of data. Validation performance of SMALLCAP measured in CIDEr score, with different contents of the datastore, without any finetuning on Flickr30k (F30K), VizWiz (VW), and MSR-VTT (MV). The best number per section is underlined; the best number overall is in bold.

+ Web Data: We first consider large-scale data from the web, expanding the datastore with text from three web datasets [18] (Conceptual Captions [35], Conceptual 12M [6], and SBU captions [26]).[12] The results with *In-domain + Web* in Table 3 show that performance improves for all three datasets. We can see a bigger improvement on Flickr30K and MSR-VTT when using a large and diverse datastore compared to just using in-domain data. Improvement on VizWiz, on the other hand, remains low, in line with the earlier observation that this dataset has a distinct distribution that is not easily matched by other data.

+ Human-labeled Data: We also consider smaller-scale but clean human-labeled data. As discussed in Section 3.2, the datastore can contain any type of text that can be useful to describe images, thus not being constrained by the assumption of image-caption pairs. As such, we consider text not only from image captions (COCO [7], Flickr30k [47], VizWiz [10]), but also from video captions (MSR-VTT [45], VATEX [41], TGIF [20]), audio captions (Clotho [9]), and localized narratives (LN ADE20k, LN COCO, LN Flickr30k, LN OpenImages [28]).

As seen in *In-domain + Human-labeled*, adding human-labeled data to the datastore leads to an improvement over using in-domain data only for Flickr30k and MSR-VTT but not for VizWiz. In comparison to *In-domain + Web*, this

[8] OSCAR$_{\text{Large}}$ results with COCO-only training. CaMEL results with CLIP-ResNet-50×16, $\lambda_k d$ = 0.1, no mesh connectivity, and a cross-entropy objective (checkpoint obtained through personal communication).

[9] We only include results from models which follow the `nocaps` guidelines to not train on image-caption pairs beyond COCO [1]. As we use only captions for retrieval, our method is also in line with these guidelines.

[10] We also include results on the validation set in Appendix D.

[11] Data size and further details can be found in Appendix F.

[12] We use a trained FAISS index (`IndexIVFFlat`) for faster search.

nocaps: Kết quả trên bộ kiểm tra nocaps đư ợc báo cáo trong Bảng 2. 8,9,10. SMALLCAP rõ ràng vư ợt trội hơn các phư ơng pháp nhẹ Ngoài miền và đạt đư ợc hiệu suất cạnh tranh Trong miền và Gần miền. Mô hình khả năng khái quát hóa mạnh mẽ chỉ ra rằng nó ít có khả năng xảy ra hơn quá phù hợp vì nó không cần phải ghi nhớ quá trình đào tạo của nó dữ liệu, cũng có sẵn thông qua việc truy xuất. Mô hình của chúng tôi có thể cải thiện hơn nữa khi dữ liệu bổ sung đư ợc đặt trong kho dữ liệu, như đư ợc thấy trong SMALLCAP+W+H. Trong biến thể này, mô tả chi tiết hơn trong Phần 5.2, kho dữ liệu COCO đư ợc bổ sung bằng dữ liệu web (W) và dữ liệu do con ngư ời gắn nhãn (H) đa dạng. SMALLCAP+W+H cho thấy khả năng khái quát hóa ấn tư ợng, vư ợt trội hơn OSCARLarge lớn hơn nhiều , hơn 10 điểm trong thiết lập Ngoài miền. Tiếp theo, chúng tôi tiếp tục khám phá khả năng chuyển giao không cần đào tạo của SMALLCAP sang miền trên nhiều tập dữ liệu khác nhau.

5. Sử dụng dữ liệu không cần đào tạo

Trong phần này, chúng tôi nghiên cứu khả năng của SMALLCAP trong việc tận dụng dữ liệu mới trong kho dữ liệu của mình theo cách không cần đào tạo, tức là tất cả các thí nghiệm đư ợc trình bày ở đây đều cấu thành những thay đổi đư ợc thực hiện đối với kho dữ liệu tại thời điểm suy luận, trong khi mô hình đư ợc đào tạo trên COCO, vẫn cố định. Trọng tâm là ngoài miền hiệu suất đư ợc đo lư ờng trên một tập hợp đa dạng các phụ đề bộ dữ liệu: Flick30k [47], VizWiz [10] và MSR-VTT [45]. Sau này thực chất là một tập dữ liệu phụ đề video, mà chúng tôi thích ứng bằng cách chuyển đổi các đoạn video clip thành hình ảnh bốn 4 khung hình, đư ợc lấy mẫu ở 0, 25, 50 và 100% thời lư ợng clip (xem ví dụ MSR-VTT trong Hình 5). Chúng tôi bắt đầu bằng khám phá các cấu hình khác nhau của kho dữ liệu, với kết quả trong Bảng 3 đư ợc báo cáo trên dữ liệu xác thực.

5.1. Dữ liệu trong miền

Ở đầu Bảng 3, chúng tôi trình bày cách SMALLCAP hoạt động khi kho dữ liệu của nó đư ợc điền dữ liệu đào tạo liên kết với từng tập dữ liệu tư ơng ứng (Trong miền). So với việc sử dụng chú thích COCO trong kho dữ liệu (COCO), hiệu suất mô hình tăng đáng kể cho cả ba bộ dữ liệu. Điều này cho thấy SMALLCAP thích ứng với thông tin đư ợc lấy lại để đạt đư ợc chuyển miền. Sự cải thiện đáng chú ý nhất đối với VizWiz, có thể là do bản chất của tập dữ liệu này rất khác biệt so với COCO, và do đó vẫn còn một khoảng cách lớn hơn cần phải đư ợc thu hẹp.

5.2. Mở rộng kho dữ liệu

Trong Bảng 3 (Tăng cư ờng kho dữ liệu), chúng tôi tăng cư ờng kho dữ liệu trong miền bằng dữ liệu quy mô lớn bổ sung theo hiệu ứng pháo đài để cải thiện khái quát hóa. Chúng tôi thử nghiệm với nhiều dữ liệu web (có quy mô lớn như ng đư ợc gắn nhãn tự động) và dữ liệu đư ợc gắn nhãn của con ngư ời (quy mô nhỏ hơn như ng sạch hơn).11

+ Dữ liệu Web: Đầu tiên chúng ta xem xét dữ liệu quy mô lớn từ web, mở rộng kho dữ liệu với văn bản từ ba trang web bộ dữ liệu [18] (Chú thích khái niệm [35], Khái niệm 12M [6], và chú thích SBU [26]).12 Kết quả với In-domain + Web trong Bảng 3 cho thấy hiệu suất đư ợc cải thiện đối với cả ba tập dữ liệu. Chúng ta có thể thấy sự cải thiện lớn hơn về Flickr30K và MSR-VTT khi sử dụng một lư ợng lớn và đa dạng kho dữ liệu so với việc chỉ sử dụng dữ liệu trong miền. Mặt khác, sự cải thiện trên VizWiz vẫn còn thấp, phù hợp với quan sát trư ớc đó rằng tập dữ liệu này có một sự khác biệt phân phối không dễ dàng khớp với dữ liệu khác.

+ Dữ liệu đư ợc gắn nhãn của con ngư ời: Chúng tôi cũng xem xét dữ liệu quy mô nhỏ hơn như ng dữ liệu đư ợc gắn nhãn của con ngư ời sạch sẽ. Như đã thảo luận trong Phần 3.2, kho dữ liệu có thể chứa bất kỳ loại văn bản nào có thể hữu ích để mô tả hình ảnh, do đó không bị hạn chế bởi giả định của cặp hình ảnh-chú thích. Như vậy, chúng tôi xem xét văn bản không chỉ từ chú thích hình ảnh (COCO [7], Flickr30k [47], VizWiz [10]), mà còn từ phụ đề video (MSR-VTT [45], VATEX [41], TGIF [20]), phụ đề âm thanh (Clotho [9]), và các tư ơng thuật cục bộ (LN ADE20k, LN COCO, LN Flickr30k, LN OpenImages [28]).

Như đã thấy trong In-domain + Human-labeled, việc thêm dữ liệu đư ợc gắn nhãn của con ngư ời vào kho dữ liệu dẫn đến cải thiện so với chỉ sử dụng dữ liệu trong miền cho Flickr30k và MSR-VTT như ng không dành cho VizWiz. So với In-domain + Web, điều này

Kho dữ liệu SMALLCAP	Xe F30K	VW	
COCO	52,2	34,5	23,3
Trong miền	55,4	47,7	29,2
Tăng cư ờng kho dữ liệu			
Trong miền + Web	58,6	48,0	29,8
Trong miền + Đư ợc gắn nhãn bởi con ngư ời	57,6	47,5	30,9
Trong miền + W + H	57,9	48,0	30,7
Không phụ thuộc vào miền			
Trang web	58,4	42,4	27,6
Đư ợc gắn nhãn bởi con ngư ời	56,6	36,4	29,0
Web + Nhãn của con ngư ời	57,8	42,2	29,9

Bảng 3. Khám phá việc sử dụng dữ liệu không cần đào tạo. Xác thực hiệu suất của SMALLCAP đư ợc đo bằng điểm CIDEr, với các nội dung khác nhau của kho dữ liệu, mà không có bất kỳ tinh chỉnh nào trên Flickr30k (F30K), VizWiz (VW) và MSR-VTT (MV). Số tốt nhất mỗi phần đư ợc gạch chân; số tốt nhất đư ợc in đậm.

8OSCARKết quả lớn với đào tạo chỉ COCO. Kết quả CaMEL với CLIP-ResNet-50×16, kd = 0,1, không có kết nối lư ới và mục tiêu entropy chéo (điểm kiểm tra thu đư ợc thông qua giao tiếp cá nhân).

9Chúng tôi chỉ bao gồm kết quả từ các mô hình tuân theo hư ớng dẫn nocaps để không đào tạo trên các cặp chú thích hình ảnh ngoài COCO [1]. Vì chúng tôi chỉ sử dụng chú thích để truy xuất, phư ơng pháp của chúng tôi cũng phù hợp với những hư ớng dẫn này.

10Chúng tôi cũng bao gồm kết quả về bộ xác thực trong Phụ lục D.

11Kích thư ớc dữ liệu và thông tin chi tiết hơn có thể đư ợc tìm thấy trong Phụ lục F.

12Chúng tôi sử dụng chỉ mục FAISS đã đư ợc đào tạo (IndexIVFFlat) để tìm kiếm nhanh hơn.

	Flickr30K	VizWiz	MSR-VTT
ClipCap	41.2	28.3	12.5
CaMEL	55.2	37.6	20.7
SMALLCAP	**60.6**	**55.0**	**28.4**
Pre-training & finetuning			
SOTA	79.6 [23]	120.8 [39]	75.9 [39]

Table 4. Out-of-domain performance without additional training, measured in CIDEr score on the test data. Flickr30K and VizWiz results with *In-domain* + *Web*, and MSR-VTT result with *In-domain* + *Human-labeled*. We include SOTA results from large-scale pre-trained models, finetuned on the respective datasets.

improvement is smaller for Flickr30k, but larger for MSR-VTT. Although smaller than web data, human-labeled data benefits MSR-VTT more, because it contains text from different tasks, including video captioning.

+ Web + Human-labeled Data: Seeing that SMALLCAP can benefit both from Web and from Human-labeled data as augmentations over in-domain data alone, we also consider a combination of the two, to determine whether their contributions are complementary or overlapping. The results for *In-domain* + *W* + *H* in Table 3 show that combining the two sources of data is not beneficial for any of the three datasets.

5.3. Domain-agnostic Datastore

In this section, we study whether SMALLCAP could still perform well without access to in-domain data and report results under the heading *Domain-agnostic* in Table 3. We find that the patterns observed above with in-domain data largely hold without it as well. With the large and diverse *Web* datastore, SMALLCAP performs close to or even better than with *In-domain* data. *Human-labeled* data is again seen to benefit MSR-VTT the most, the optimal configuration for this dataset being *Web* + *Human-labeled*.

From the exploration presented above, we conclude that SMALLCAP's image captioning capabilities can transfer with access to web data in addition to or in place of in-domain data. The model can also leverage human-labeled data beyond image-captioning pairs in solving tasks other than image captioning, such as video captioning.

5.4. Results with the Best Configuration

Having explored different datastore configurations for each of the three datasets, we use the best configuration for each to compare zero-shot performance against ClipCap and CaMEL, both models also trained only on COCO. In Table 4 we show test set performance (in CIDEr score) with a datastore consisting of *In-domain* + *Web* for Flickr30k and VizWiz, and *In-domain* + *Human-labeled* for MSR-VTT. SMALLCAP outperforms both ClipCap and CaMEL by a large margin on all three datasets. In comparison to CaMEL, the stronger baseline of the two, we see a 5.4 point improvement on Flickr30k, a noteworthy 17.4 point improvement on VizWiz and an increase of 7.7 points on MSR-VTT. The large improvement on VizWiz demonstrates SMALLCAP's ability to transfer to domains very distinct from the training data, i.e., COCO. The improvement on MSR-VTT, on the other hand, shows our approach has potential not only for other domains but for other tasks as well. These results show that while other lightweight-training models lack out-of-domain generalization without finetuning, our model can transfer across domains by only swapping the datastore contents. In the bottom of the table, we provide state-of-the-art results for context, which were achieved by large-scale pre-trained V&L models, finetuned specifically on the respective datasets.

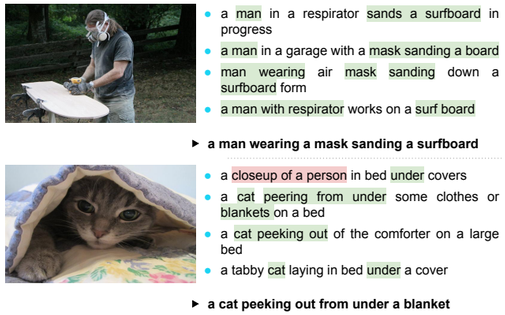


Figure 3. Examples generated by SMALLCAP, together with the retrieved predictions from the COCO datastore. • denotes the retrieved captions, highlighted as green or red to indicate correct and mismatch captions, respectively. ▶ denotes the generated caption.

6. Discussion

6.1. Qualitative Examples

Figure 3 shows examples of the retrieved and generated captions for two images from the COCO dataset. In first example, we observe that the retrieved captions are highly relevant to the input image and the generated captions are semantically similar to them. As seen in the second example, SMALLCAP can also be robust to misleading information from retrieval. Figure 5 shows examples of captions generated for Flickr30k, VizWiz, and MSR-VTT, with a datastore populated with COCO or with in-domain data. These qualitative results show how SMALLCAP adapts to new domains: with the help of the retrieved captions, it correctly refers to the concepts *tutu*, the *Swanson* brand name, and *Pokemon*. The first two concepts are not present in the COCO training data at all, while the last is seen just six times.

	Flickr30K	VizWiz	MSR-VTT
ClipCap 41.2 CaMEL		28,3	12,5
55.2 SMALLCAP 60.6		37,6	20,7
		55,0	28,4
Đào tạo trư ớc và tinh chỉnh			
SOTA	79.6 [23]	120,8 [39]	75,9 [39]

Bảng 4. Hiệu suất ngoài miền mà không cần đào tạo bổ sung, đư ợc đo bằng điểm CIDEr trên dữ liệu thử nghiệm. Flickr30K và VizWiz kết quả với In-domain + Web và kết quả MSR-VTT với In-domain + Human-labeled. Chúng tôi bao gồm kết quả SOTA từ các mô hình đư ợc đào tạo trư ớc quy mô lớn, đư ợc tinh chỉnh trên các tập dữ liệu tư ơng ứng.

cải thiện nhỏ hơn đối với Flickr30k, như ng lớn hơn đối với MSR-VTT. Mặc dù nhỏ hơn dữ liệu web, dữ liệu đư ợc gắn nhãn của con ngư ời mang lại nhiều lợi ích hơn cho MSR-VTT vì nó chứa văn bản từ nhiều tác vụ khác nhau, bao gồm cả phụ đề video.

+ Web + Dữ liệu đư ợc gắn nhãn của con ngư ời: Nhìn thấy SMALLCAP có thể hư ởng lợi từ cả Web và dữ liệu đư ợc gắn nhãn của con ngư ời tăng cư ờng trên dữ liệu trong miền một mình, chúng tôi cũng xem xét sự kết hợp của cả hai, để xác định xem những đóng góp của họ có bổ sung hay chồng chéo nhau không. Kết quả cho Trong miền + W + H trong Bảng 3 cho thấy việc kết hợp hai nguồn dữ liệu không có lợi cho bất kỳ bộ dữ liệu nào trong ba bộ dữ liệu.

5.3. Kho dữ liệu không phụ thuộc vào miền

Trong phần này, chúng tôi nghiên cứu liệu SMALLCAP vẫn có thể thực hiện tốt mà không cần truy cập vào dữ liệu và báo cáo trong miền kết quả dư ới tiêu đề Không phụ thuộc vào miền trong Bảng 3. Chúng tôi tìm thấy các mẫu đư ợc quan sát ở trên với dữ liệu trong miền phần lớn giữ mà không có nó cũng như vậy. Với sự lớn và đa dạng Kho dữ liệu web, SMALLCAP hoạt động gần bằng hoặc thậm chí tốt hơn so với dữ liệu trong miền. Dữ liệu đư ợc gắn nhãn của con ngư ời một lần nữa đư ợc coi là có lợi nhất cho MSR-VTT, cấu hình tối ư u cho tập dữ liệu này là Web + Nhãn con ngư ời.

Từ việc khám phá đư ợc trình bày ở trên, chúng tôi kết luận rằng Khả năng chú thích hình ảnh của SMALLCAP có thể chuyển giao với quyền truy cập vào dữ liệu web ngoài hoặc thay thế dữ liệu trong miền. Mô hình cũng có thể tận dụng nhãn của con ngư ời dữ liệu vư ợt ra ngoài các cặp chú thích hình ảnh trong việc giải quyết các nhiệm vụ khác hơn là chú thích hình ảnh, chẳng hạn như chú thích video.

5.4. Kết quả với cấu hình tốt nhất

Đã khám phá các cấu hình kho dữ liệu khác nhau cho mỗi một trong ba tập dữ liệu, chúng tôi sử dụng cấu hình tốt nhất để so sánh hiệu suất không bắn với ClipCap và CaMEL, cả hai mô hình cũng chỉ đư ợc đào tạo trên COCO. Trong Bảng 4 chúng tôi hiển thị hiệu suất của bộ kiểm tra (theo điểm CIDEr) với kho dữ liệu bao gồm In-domain + Web cho Flickr30k và VizWiz, và Trong miền + Đư ợc gắn nhãn bởi con ngư ời cho MSR-VTT. SMALLCAP vư ợt trội hơn cả ClipCap và CaMEL với biên độ lớn trên cả ba tập dữ liệu. Khi so sánh đối với CaMEL, đư ờng cơ sở mạnh hơn trong hai đư ờng cơ sở, chúng ta thấy một Cải thiện 5,4 điểm trên Flickr30k, đáng chú ý là 17,4 cải thiện điểm trên VizWiz và tăng 7,7 điểm trên MSR-VTT. Sự cải tiến lớn trên VizWiz chứng minh khả năng của SMALLCAP trong việc chuyển đến các miền rất khác biệt so với dữ liệu đào tạo, tức là COCO. Sự cải tiến Mặt khác, trên MSR-VTT, cho thấy cách tiếp cận của chúng tôi đã tiềm năng không chỉ cho các lĩnh vực khác mà còn cho các nhiệm vụ khác cũng như vậy. Những kết quả này cho thấy rằng trong khi các mô hình đào tạo nhẹ khác thiếu khái quát hóa ngoài miền mà không có tinh chỉnh, mô hình của chúng tôi có thể chuyển qua các miền chỉ bằng hoán đổi nội dung kho dữ liệu. Ở cuối bảng, chúng tôi cung cấp kết quả hiện đại nhất cho bối cảnh, đó là đạt đư ợc bằng các mô hình V&L đư ợc đào tạo trư ớc quy mô lớn, đư ợc tinh chỉnh cụ thể trên các tập dữ liệu tư ơng ứng.

Hình 3. Các ví dụ đư ợc tạo ra bởi SMALLCAP, cùng với đã lấy lại các dự đoán từ kho dữ liệu COCO. • biểu thị các chú thích đã lấy lại, đư ợc tô sáng màu xanh lá cây hoặc màu đỏ để chỉ ra các chú thích chính xác và chú thích không khớp, tư ơng ứng. I biểu thị chú thích đã tạo.

6. Thảo luận

6.1. Ví dụ định tính

Hình 3 cho thấy các ví dụ về dữ liệu đư ợc lấy và tạo ra chú thích cho hai hình ảnh từ tập dữ liệu COCO. Trong ví dụ đầu tiên, chúng tôi quan sát thấy rằng các chú thích đư ợc lấy ra có liên quan cao đến hình ảnh đầu vào và các chú thích đư ợc tạo ra có ngữ nghĩa tư ơng tự với chúng. Như đã thấy trong ví dụ thứ hai, SMALLCAP cũng có thể mạnh mẽ chống lại thông tin sai lệch từ việc truy xuất. Hình 5 hiển thị các ví dụ về chú thích đư ợc tạo cho Flickr30k, VizWiz và MSR-VTT, với kho dữ liệu đư ợc điền bằng COCO hoặc bằng dữ liệu trong miền. Những kết quả định tính này cho thấy SMALLCAP thích ứng với các miền mới như thế nào: với sự trợ giúp của các chú thích đã lấy lại, nó đề cập chính xác đến các khái niệm tutu, tên thư ơng hiệu Swanson và Pokemon. Hai khái niệm đầu tiên không có trong chư ơng trình đào tạo COCO dữ liệu nào cả, trong khi dữ liệu cuối cùng chỉ đư ợc nhìn thấy sáu lần.

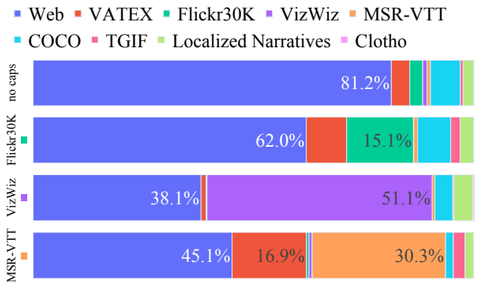


Figure 4. Percentage of the retrieved captions that come from each data source, when testing the model on the different benchmarks of `nocaps` , Flickr30k, VizWiz and MSR-VTT.

6.2. Analysis of the Retrieved Captions

In Section 5.3, we demonstrated the ability of SMALLCAP to exploit large data in a training-free fashion. Here, we inspect the distribution of retrieved captions in the *In-domain + Web + Human-labeled* setting, in order to understand the individual impact of each dataset. As can be seen in Figure 4, most text is retrieved from web data, especially in the presence of unseen visual concepts, as is the case for `nocaps`. Besides web data, the model tends to retrieve text from the corresponding dataset or from a similar domain; for instance, MSR-VTT retrieval also relies on other video datasets. Due to its unique distribution, VizWiz stands out as the case with the highest rate of in-domain retrieval.

Seeing that text from all types of human-labeled data is retrieved, we measure the actual impact of each type on performance. In Table 5, we report performance on Flickr30k, VizWiz, and MSR-VTT, with an in-domain datastore augmented with either Image captions, Video captions, localized Narratives, or Audio captions. We see that SMALLCAP can indeed benefit from data beyond image captions. For instance, video captions help not only for MSR-VTT, but also for Flickr30k and VizWiz. Flickr30k benefits the most from localized narratives since this dataset contains narratives for the Flickr30k images. Audio captions are beneficial for both Flickr30k and MSR-VTT. Considering the distinct nature of the audio and visual modalities, this finding demonstrates the potential of leveraging data which has previously seen limited application to image captioning.

6.3. The Impact of Retrieval

In Figure 6, we show validation performance with 1.8, 3.6, 7, 14 and 28 million trainable parameters with and without retrieval augmentation.[13] For variants with retrieval augmentation, performance is stable across the range of model sizes considered. Reducing the number of trainable parameters by a factor of four, from 28M to 7M, leads to a slight drop of 0.6 CIDEr points. This indicates that SMALLCAP has a close-to-optimal size to performance trade-off.

Next, we ablate the retrieval augmentation to quantify its impact. We train models without retrieval augmentation, prompting them with just the phrase `This image shows`. As seen in Figure 6, without the aid of retrieved captions, there is a notable drop in performance compared to results with retrieval. Moreover, model performance degrades at a higher rate: while performance at the two extremes of model sizes differs by just 1.7 CIDEr points with retrieval, without it the difference is 4.3 points.[14]

In order to confirm that SMALLCAP is not simply paraphrasing the retrieved captions without attending to the visual input, we experiment with ablating the visual modality. For this, we train a model on "blank" input images, setting the visual features from the encoder to zero. This yields a much lower CIDEr score of 90.1 on the validation set, showing that SMALLCAP indeed uses the visual input.

	Flickr30K	VizWiz	MSR-VTT
In-domain	52.2	47.7	29.2
+ Image	56.7	47.8	29.8
+ Video	57.0	47.8	31.1
+ Narratives	57.1	47.2	28.7
+ Audio	55.4	47.7	29.4

Table 5. SMALLCAP performance with retrieval from the different sources of the Human-labeled data. The model can benefit from having access to text that is not only from image captioning tasks, but also from other tasks such as audio captioning.

Decoder	$\|\theta\|$	B@4	M	CIDEr	S
GPT2-Base$_{d=16}$	7	37.0	27.9	119.7	21.3
OPT-125M$_{d=16}$	7	37.6	28.4	122.0	21.7
GPT2-Medium$_{d=16}$	22	36.5	28.1	120.7	21.6
OPT-350M$_{d=16}$	22	37.5	28.7	122.7	22.0

Table 6. Results with different decoders on the COCO test set.

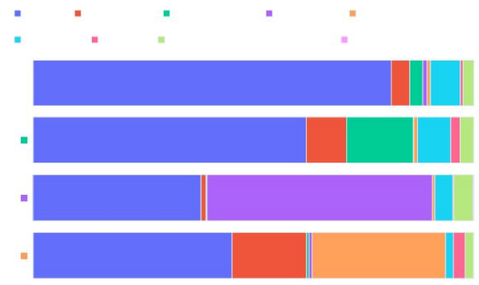
6.4. Alternative Decoders

At the request of the anonymous reviewers, we include additional experiments with some more recent language models: OPT-125M and OPT-350M [48], equivalent in size to GPT2-Base and GPT2-Medium.[15] The results in Table 6 show that our approach performs well with these stronger

[13] The model sizes correspond to $d = 4, 8, 16, 32$ and 64.

[14] See Appendix H for qualitative examples with and without retrieval.

[15] There is no OPT variant equivalent in size to GPT2-Large.



Hình 4. Tỷ lệ phần trăm phụ đề đư ợc lấy từ mỗi nguồn dữ liệu, khi thử nghiệm mô hình trên các chuẩn mực khác nhau của nocaps, Flickr30k, VizWiz và MSR-VTT.

	Flickr30K	VizWiz	MSR-VTT
Trong miền	52,2	47,7	29.2
+ Hình ảnh	56,7	47,8	29,8
+ Video	57.0	47,8	31.1
+ Tư ờng thuật	57,1	47,2	28,7
+ Âm thanh	55,4	47,7	29,4

Bảng 5. Hiệu suất SMALLCAP với việc truy xuất từ các nguồn dữ liệu đư ợc gắn nhãn của con ngư ời. Mô hình có thể đư ợc hư ởng lợi từ có quyền truy cập vào văn bản không chỉ từ các tác vụ chú thích hình ảnh, mà còn từ các nhiệm vụ khác như chú thích âm thanh.

Bộ giải mã	|	B@4	M	CIDEr	S
Dựa trên GPT2=16	7	37.0	27.9	119.7	21.3
TÙY CHỌN-125Md=16	7	37,6	28,4	122,0	21,7
GPT2-Trung bình=16	22	36,5	28,1	120,7	21,6
OPT-350Md=16	22	37,5	28,7	122,7	22,0

Bảng 6. Kết quả với các bộ giải mã khác nhau trên bộ thử nghiệm COCO.

6.2. Phân tích các chú thích đã thu thập

Trong Phần 5.3, chúng tôi đã chứng minh khả năng của SMALL- CAP trong việc khai thác dữ liệu lớn theo cách không cần đào tạo. Ở đây, chúng tôi kiểm tra sự phân phối của các chú thích đã lấy đư ợc trong bối cảnh Trong miền + Web + Đư ợc gắn nhãn bởi con ngư ời, để hiểu đư ợc tác động riêng lẻ của từng tập dữ liệu. Như có thể thấy trong Hình 4, hầu hết văn bản đư ợc lấy từ dữ liệu web, đặc biệt là trong sự hiện diện của các khái niệm trực quan vô hình, như trư ờng hợp của nocaps. Bên cạnh dữ liệu web, mô hình có xu hư ớng lấy văn bản từ tập dữ liệu tư ơng ứng hoặc từ một miền tư ơng tự; ví dụ, việc truy xuất MSR-VTT cũng dựa vào các video khác bộ dữ liệu. Do sự phân phối độc đáo của nó, VizWiz nổi bật như trư ờng hợp có tỷ lệ truy xuất trong miền cao nhất.

Nhìn thấy văn bản từ tất cả các loại dữ liệu đư ợc gắn nhãn của con ngư ời là đã thu thập, chúng tôi đo lư ờng tác động thực tế của từng loại đối với hiệu suất. Trong Bảng 5, chúng tôi báo cáo hiệu suất trên Flickr30k, VizWiz và MSR-VTT, với kho dữ liệu trong miền đư ợc tăng cư ờng bằng chú thích Hình ảnh, chú thích Video, Tư ờng thuật cục bộ hoặc chú thích Âm thanh. Chúng tôi thấy rằng SMALLCAP thực sự có thể hư ởng lợi từ dữ liệu ngoài chú thích hình ảnh. Ví dụ, chú thích video không chỉ giúp ích cho MSR-VTT mà còn cho Flickr30k và VizWiz. Flickr30k đư ợc hư ởng lợi nhiều nhất từ các câu chuyện đư ợc bản địa hóa vì tập dữ liệu này chứa các câu chuyện cho hình ảnh Flickr30k. Phụ đề âm thanh có lợi cho cả hai Flickr30k và MSR-VTT. Xem xét bản chất riêng biệt của phư ơng thức âm thanh và hình ảnh, phát hiện này chứng minh tiềm năng của việc tận dụng dữ liệu đã từng thấy trư ớc đây ứng dụng hạn chế trong việc chú thích hình ảnh.

6.3. Tác động của việc truy xuất

Trong Hình 6, chúng tôi hiển thị hiệu suất xác thực với 1.8, 3,6, 7, 14 và 28 triệu tham số có thể đào tạo với và không có sự gia tăng truy xuất.13 Đối với các biến thể có sự truy xuất tăng cư ờng, hiệu suất ổn định trên phạm vi kích thư ớc mô hình đư ợc xem xét. Giảm số lư ợng có thể đào tạo các thông số theo hệ số bốn, từ 28M đến 7M, dẫn đến giảm nhẹ 0,6 điểm CIDEr. Điều này cho thấy SMALL- CAP có sự đánh đổi giữa kích thư ớc và hiệu suất gần như tối ư u.

Tiếp theo, chúng tôi loại bỏ sự gia tăng truy xuất để định lư ợng tác động của nó. Chúng tôi đào tạo các mô hình mà không cần tăng cư ờng truy xuất, chỉ nhắc nhở chúng bằng cụm từ Hình ảnh này hiển thị. Như đư ợc thấy trong Hình 6, không có sự trợ giúp của lấy lại chú thích, có một sự sụt giảm đáng kể về hiệu suất so với để có kết quả với việc truy xuất. Hơn nữa, hiệu suất mô hình giảm ở mức cao hơn: trong khi hiệu suất ở hai cực trị của kích thư ớc mô hình chỉ khác nhau 1,7 điểm CIDEr với lấy lại, nếu không có nó thì sự khác biệt là 4,3 điểm.14

Để xác nhận rằng SMALLCAP không chỉ diễn đạt lại các chú thích đã lấy đư ợc mà không chú ý đến đầu vào trực quan, chúng tôi thử nghiệm bằng cách loại bỏ phư ơng thức trực quan. Đối với điều này, chúng tôi đào tạo một mô hình trên hình ảnh đầu vào "trống", thiết lập các tính năng trực quan từ bộ mã hóa đến số không. Điều này mang lại điểm CIDEr thấp hơn nhiều là 90,1 trên bộ xác thực, cho thấy SMALLCAP thực sự sử dụng đầu vào trực quan.

6.4. Bộ giải mã thay thế

Theo yêu cầu của ngư ời đánh giá ẩn danh, chúng tôi bao gồm các thí nghiệm bổ sung với một số ngôn ngữ gần đây hơn các mô hình: OPT-125M và OPT-350M [48], có kích thư ớc tư ơng đư ơng đến GPT2-Base và GPT2-Medium.15 Kết quả trong Bảng 6 cho thấy cách tiếp cận của chúng tôi hoạt động tốt với những điều mạnh mẽ hơn này

13Kích thư ớc mô hình tư ơng ứng với d = 4, 8, 16, 32 và 64.

14Xem Phụ lục H để biết các ví dụ định tính có và không có truy xuất.

15Không có biến thể OPT nào có kích thư ớc tư ơng đư ơng với GPT2-Large.

Figure 5. Examples of captions generated for Flickr30k, VizWiz and MSR-VTT, with retrieval either from COCO or in-domain data. The captions use words retrieved from the in-domain datastores which were rarely seen in the COCO training data (tutu, swanson, pokemon).

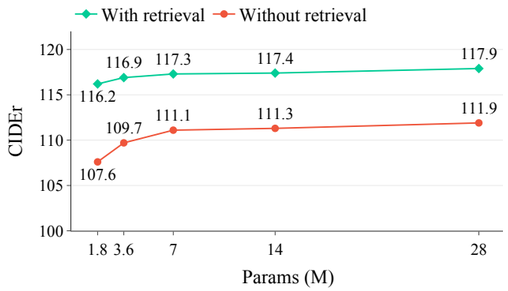


Figure 6. CIDEr scores on the COCO validation set, with and without retrieval, across different cross-attention sizes.

language models and is therefore model agnostic.[16,17]

7. Conclusion

In this paper, we propose SMALLCAP, an image captioning model augmented with retrieval, which is light to train and can be transferred across domains without retraining. Results on the COCO dataset show that SMALLCAP is competitive to other lightweight-training models despite having substantially less trainable parameters, instead leveraging non-parametric information from a datastore of text. Out-of-domain evaluations show that SMALLCAP can also perform training-free domain transfer when given access to a datastore with target-domain data. Our model further benefits from diverse web and human-labeled data in addition to or in place of target-domain data. We find that SMALLCAP benefits not just from access to image captions, but also to video and audio captions (resources neglected in image captioning work in the past).

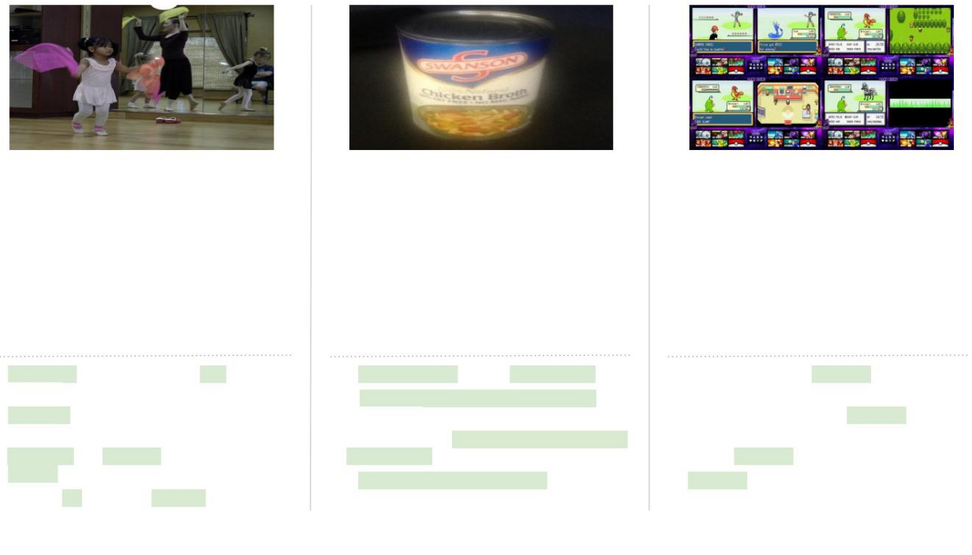
SMALLCAP's small size and impressive performance in out-of-domain settings attest to the potential of retrieval augmentation as an alternative to the expensive training found in large pre-trained vision-and-language models and the costly finetuning that even previous lightweight-training models require in order to adapt to different image captioning datasets. Future work can apply our retrieval augmentation approach to a wider range of multimodal tasks, and further explore the scalability of the data used for retrieval.

Acknowledgements

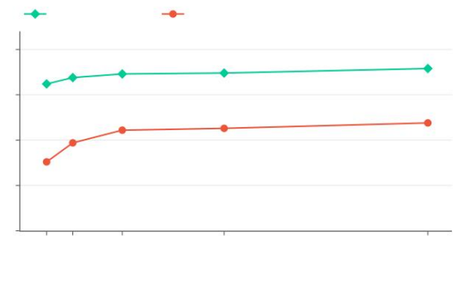
This research was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055, through Fundação para a Ciência e Tecnologia (FCT) with the Ph.D. scholarship 2020.06106.BD, and through the INESC-ID multi-annual funding from the PIDDAC programme (UIDB/50021/2020).

[16] See Appendix E for OPT results without retrieval.

[17] Due to our academic computing budget, we only repeat the experiments from Table 1. Future work can experiment further in this direction.



Hình 5. Ví dụ về chú thích đư ợc tạo cho Flickr30k, VizWiz và MSR-VTT, với việc truy xuất từ COCO hoặc dữ liệu trong miền. chú thích sử dụng các từ đư ợc lấy từ kho dữ liệu trong miền, hiếm khi thấy trong dữ liệu đào tạo COCO (tutu, swanson, pokemon).



Hình 6. Điểm CIDEr trên bộ xác thực COCO, với và không cần truy xuất, trên nhiều kích thư ớc chú ý chéo khác nhau.

có ít tham số có thể đào tạo hơn đáng kể, thay vào đó là sử dụng thông tin phi tham số để đào tạo từ kho dữ liệu văn bản. Đánh giá ngoài miền cho thấy SMALLCAP cũng có thể thực hiện chuyển miền không cần đào tạo khi đư ợc cấp quyền truy cập một kho dữ liệu với dữ liệu miền mục tiêu. Mô hình của chúng tôi đư ợc hư ởng lợi nhiều hơn từ dữ liệu web và dữ liệu đư ợc gắn nhãn của con ngư ời đến hoặc thay thế dữ liệu miền mục tiêu. Chúng tôi thấy rằng SMALL - CAP không chỉ đư ợc hư ởng lợi từ việc truy cập vào chú thích hình ảnh mà còn cũng như phụ đề video và âm thanh (những nguồn lực bị bỏ qua trong công việc chú thích hình ảnh trư ớc đây).

Kích thư ớc nhỏ và hiệu suất ấn tư ợng của SMALLCAP trong thiết lập ngoài miền chứng minh tiềm năng của việc truy xuất tăng cư ờng như một giải pháp thay thế cho việc đào tạo tốn kém đư ợc tìm thấy trong các mô hình ngôn ngữ và thị giác đư ợc đào tạo trư ớc lớn và sự tinh chỉnh tốn kém mà ngay cả những ngư ời tập luyện nhẹ trư ớc đó các mô hình yêu cầu để thích ứng với các tập dữ liệu chú thích hình ảnh khác nhau. Công việc trong tư ơng lai có thể áp dụng phư ơng pháp tăng cư ờng truy xuất của chúng tôi vào nhiều nhiệm vụ đa phư ơng thức hơn và tiếp tục khám phá khả năng mở rộng của dữ liệu đư ợc sử dụng để truy xuất.

Lời cảm ơn

Nghiên cứu này đư ợc hỗ trợ bởi Kế hoạch phục hồi và phục hồi của Bồ Đào Nha thông qua dự án C645008882- 00000055, thông qua Fundac,ao para a Ci encia e Tecnolo- gia (FCT) với học bổng Tiến sĩ 2020.06106.BD, và thông qua nguồn tài trợ đa niên của INESC-ID từ chư ơng trình PID-DAC (UIDB/50021/2020).

16Xem Phụ lục E để biết kết quả OPT mà không cần truy xuất.

17Do ngân sách tính toán học thuật của chúng tôi, chúng tôi chỉ lặp lại các thí nghiệm từ Bảng 1. Các công trình trong tư ơng lai có thể thử nghiệm sâu hơn theo hư ớng này.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 1, 2, 4, 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. 4

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[5] Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. CaMEL: Mean Teacher Learning for Image Captioning. In *International Conference on Pattern Recognition*, 2022. 4

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 5

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 4, 5, 12

[8] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 4

[9] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020. 5, 12

[10] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pages 417–434. Springer, 2020. 1, 2, 5, 12

[11] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022. 1, 2, 4

[12] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022. 2, 3

[13] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*, 2021. 2, 11

[14] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. 4

[15] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 4

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. 2, 3

[17] Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*, 2022. 2, 3

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1, 2, 3, 4, 5, 12

[19] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2, 4, 11

[20] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016. 5, 12

[21] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 4

[22] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. I-tuning: Tuning language models with image for caption generation. *arXiv preprint arXiv:2202.06574*, 2022. 1, 2, 4

[23] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. Vc-gpt: Visual conditioned gpt for end-to-end generative vision-and-language pre-training. *arXiv preprint arXiv:2201.12723*, 2022. 6

[24] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989. 2

[25] Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning, 2021. 1, 2, 4

[26] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 5

Tài liệu tham khảo

[1] Agrawal khắc nghiệt, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Ste-fan Lee và Peter Anderson. nocaps: chú thích đối tượng mới lạ ở quy mô lớn. Trong Hội nghị quốc tế IEEE/CVF năm 2019 về Computer Vision (ICCV). IEEE, tháng 10 năm 2019. 1, 2, 4, 5

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: một mô hình ngôn ngữ trực quan cho việc học ít cảnh. Bản in trư ớc arXiv arXiv:2204.14198, 2022. 2

[3] Peter Anderson, Basura Fernando, Mark Johnson và Stephen Gould. Spice: Đánh giá chú thích hình ảnh mệnh đề ngữ nghĩa. Trong hội nghị châu Âu về tầm nhìn máy tính, trang 382–398. Springer, 2016. 4

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould và Lei Zhang. Sự chú ý từ dư ới lên và từ trên xuống cho chú thích hình ảnh và trả lời câu hỏi trực quan. Trong Biên bản báo cáo của hội nghị IEEE về tầm nhìn máy tính và nhận dạng mẫu, các trang 6077–6086, 2018. 2

[5] Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi và Rita Cucchiara. Con lạc đà: Giáo viên trung bình học tập cho chú thích hình ảnh. Trong Hội nghị quốc tế về nhận dạng mẫu, 2022. 4

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding và Radu Soricut. Khái niệm 12m: Đẩy mạnh việc đào tạo trư ớc hình ảnh-văn bản quy mô web để nhận dạng các khái niệm trực quan đuôi dài. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Tầm nhìn máy tính và Nhận dạng mẫu, trang 3558–3568, 2021. 5

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedan-tam, Saurabh Gupta, Piotr Dollar và C Lawrence Zitnick. Tiêu đề Microsoft coco: Thu thập và đánh giá dữ liệu máy chủ. bản in trư ớc arXiv arXiv:1504.00325, 2015. 2, 4, 5, 12

[8] Michael Denkowski và Alon Lavie. Meteor universal: Đánh giá bản dịch theo ngôn ngữ cụ thể cho bất kỳ ngôn ngữ đích nào. Trong Biên bản hội thảo lần thứ chín về máy thống kê bản dịch, trang 376–380, 2014. 4

[9] Konstantinos Drossos, Samuel Lipping và Tuomas Virta-nen. Clotho: Một tập dữ liệu phụ đề âm thanh. Trong ICASSP 2020-Hội nghị quốc tế IEEE năm 2020 về âm học, bài phát biểu và Xử lý tín hiệu (ICASSP), trang 736–740. IEEE, 2020. 5, 12

[10] Danna Gurari, Yinan Zhao, Meng Zhang và Nilavra Bhat-tacharya. Chú thích hình ảnh đư ợc chụp bởi ngư ời mù. Trong Hội nghị Châu Âu về Tầm nhìn Máy tính, trang 417–434. Springer, 2020. 1, 2, 5, 12

[11] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Lư u Tử Thành, Lộ Ngọc Mao và Vư ơng Lệ Quyên. Mở rộng quy mô đào tạo trư ớc ngôn ngữ thị giác cho chú thích hình ảnh. Trong Biên bản báo cáo của Hội nghị IEEE/CVF về Thị giác máy tính và Nhận dạng Mẫu, trang 17980–17989, 2022. 1, 2, 4

[12] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hos-seini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Ar-mand Joulin, Sebastian Riedel và Edouard Grave. Học tập ít lần với các mô hình ngôn ngữ tăng cư ờng truy xuất. bản in trư ớc arXiv arXiv:2208.03299, 2022. 2, 3

[13] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, và Xiang Ren. Một lời nhắc tốt đáng giá hàng triệu tham số? Học tập dựa trên lời nhắc ít tài nguyên cho các mô hình ngôn ngữ thị giác. Bản in trư ớc arXiv arXiv:2110.08484, 2021. 2, 11

[14] Jeff Johnson, Matthijs Douze và Herve J egou. Tìm kiếm sự tư ơng đồng ở quy mô tỷ với gpus. Bản in trư ớc arXiv arXiv:1702.08734, 2017. 4

[15] Andrej Karpathy và Li Fei-Fei. Sự liên kết ngữ nghĩa thị giác sâu sắc để tạo ra các mô tả hình ảnh. Trong Biên bản hội nghị IEEE về tầm nhìn máy tính và nhận dạng mẫu, trang 3128–3137, 2015. 4

[16] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rockt aschel,Sebastian Riedel và Douwe Kiela. Tạo ra tăng cư ờng truy xuất cho các nhiệm vụ nlp chuyên sâu về kiến thức. Trong H. Larochelle, M. Ranzato, R. Hadsell, MF Balcan và H. Lin, biên tập viên, Những tiến bộ trong hệ thống xử lý thông tin thần kinh, tập 33, trang 9459–9474. Curran Associates, Inc., 2020. 2, 3

[17] Huayang Li, Yixuan Su, Đặng Thái, Yan Wang và Lemao Liu. Một cuộc khảo sát về việc tạo văn bản đư ợc tăng cư ờng truy xuất. arXiv bản in trư ớc arXiv:2202.01110, 2022. 2, 3

[18] Junnan Li, Dongxu Li, Caiming Xiong và Steven Hoi. Blip: Khởi động quá trình đào tạo trư ớc ngôn ngữ-hình ảnh để hiểu và tạo ra ngôn ngữ-thị giác thống nhất. arXiv bản in trư ớc arXiv:2201.12086, 2022. 1, 2, 3, 4, 5, 12

[19] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Đối tư ợng-ngữ nghĩa đư ợc căn chỉnh trư ớc khi đào tạo cho nhiệm vụ ngôn ngữ thị giác. Trong Hội nghị máy tính châu Âu Tầm nhìn, trang 121–137. Springer, 2020. 2, 4, 11

[20] Vận Thành Lý, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes và Jiebo Luo. Tgif: Một bộ dữ liệu và chuẩn mực mới về mô tả ảnh gif động. Trong Biên bản Hội nghị IEEE về Tầm nhìn Máy tính và Nhận dạng Mẫu, trang 4641–4650, 2016. 5, 12

[21] Ilya Loshchilov và Frank Hutter. Sửa đổi quy định giảm cân ularization trong adam. 2018. 4

[22] Ziyang Luo, Yadong Xi, Rongsheng Zhang và Jing Ma. I-tuning: Điều chỉnh mô hình ngôn ngữ bằng hình ảnh cho chú thích thế hệ. bản in trư ớc arXiv arXiv:2202.06574, 2022. 1, 2, 4

[23] Ziyang Luo, Yadong Xi, Rongsheng Zhang và Jing Ma. Vc-gpt: Gpt có điều kiện thị giác để đào tạo trư ớc ngôn ngữ và thị giác tạo ra đầu cuối. Bản in trư ớc arXiv arXiv:2201.12723, 2022. 6

[24] Michael McCloskey và Neal J Cohen. Sự can thiệp thảm khốc trong mạng lư ới kết nối: Học tập tuần tự vấn đề. Trong Tâm lý học về học tập và động lực, tập 24, trang 109–165. Elsevier, 1989. 2

[25] Ron Mokady, Amir Hertz và Amit H. Bermano. Clipcap: Tiền tố clip cho chú thích hình ảnh, 2021. 1, 2, 4

[26] Vicente Ordonez, Girish Kulkarni, và Tamara Berg. Im2text: Mô tả hình ảnh bằng 1 triệu bức ảnh có chú thích. Tiến bộ trong hệ thống xử lý thông tin thần kinh, 24, 2011. 5

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 4

[28] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *European conference on computer vision*, pages 647–664. Springer, 2020. 5, 12

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2, 3

[32] Rita Ramos, Desmond Elliott, and Bruno Martins. Retrieval-augmented image captioning. *arXiv preprint arXiv:2302.08268*, 2023. 2

[33] Rita Parada Ramos, Patrícia Pereira, Helena Moniz, Joao Paulo Carvalho, and Bruno Martins. Retrieval augmentation for deep neural networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 2

[34] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Retrieval-augmented transformer for image captioning. *arXiv preprint arXiv:2207.13162*, 2022. 2, 3

[35] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5

[36] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 200–212. Curran Associates, Inc., 2021. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 3, 11

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 4

[39] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 1, 6

[40] Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu, and Michael Zeng. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3170–3179, Dublin, Ireland, May 2022. Association for Computational Linguistics. 3

[41] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 5, 12

[42] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1, 2, 4

[43] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020. 4

[44] Chunpu Xu, Wei Zhao, Min Yang, Xiang Ao, Wangrong Cheng, and Jinwen Tian. A unified generation-retrieval framework for image captioning. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2019. 2

[45] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 5, 12

[46] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[47] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5, 12

[48] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 7, 12

[49] Shanshan Zhao, Lixiang Li, Haipeng Peng, Zihang Yang, and Jiaxuan Zhang. Image caption generation via unified retrieval and generation-based method. *Applied Sciences*, 10(18), 2020. 2

[27] Kishore Papineni, Salim Roukos, Todd Ward và Wei-Jing Zhu. Bleu: một phư ơng pháp đánh giá tự động máy móc bản dịch. Trong Biên bản cuộc họp thư ờng niên lần thứ 40 của Hiệp hội Ngôn ngữ học tính toán, trang 311–318, 2002. 4

[28] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut và Vittorio Ferrari. Kết nối tầm nhìn và ngôn ngữ với các câu chuyện địa phư ơng. Trong hội nghị châu Âu về tầm nhìn máy tính, trang 647–664. Springer, 2020. 5, 12

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Học các mô hình trực quan có thể chuyển giao từ tầm nhìn siêu ngôn ngữ tự nhiên. Trong Hội nghị quốc tế về học máy, trang 8748–8763. PMLR, 2021. 2, 3

[30] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei và Ilya Sutskever. Các mô hình ngôn ngữ là những ngư ời học đa nhiệm không có giám sát. 2019. 2

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Các mô hình ngôn ngữ là những ngư ời học đa nhiệm không đư ợc giám sát. Blog OpenAI, 1(8):9, 2019. 2, 3

[32] Rita Ramos, Desmond Elliott và Bruno Martins. Chú thích hình ảnh đư ợc tăng cư ờng truy xuất. Bản in trư ớc arXiv arXiv:2302.08268, 2023. 2

[33] Rita Parada Ramos, Patr'ıcia Pereira, Helena Moniz, Joao Paulo Carvalho và Bruno Martins. Tăng cư ờng truy xuất cho mạng nơ-ron sâu. Năm 2021, International Joint Hội nghị về Mạng nơ-ron (IJCNN), trang 1–8. IEEE, 2021. 2

[34] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, và Rita Cúc-chiara. Máy biến áp tăng cư ờng truy xuất để tạo chú thích cho hình ảnh. bản in trư ớc arXiv arXiv:2207.13162, 2022. 2, 3

[35] Piyush Sharma, Nan Ding, Sebastian Goodman và Radu Soricut. Chú thích khái niệm: Một tập dữ liệu văn bản thay thế hình ảnh đư ợc làm sạch, có siêu ẩn danh, để tạo chú thích hình ảnh tự động. Trong Biên bản báo cáo của Cuộc họp thư ờng niên lần thứ 56 của Hiệp hội Ngôn ngữ học tính toán (Tập 1: Bài báo dài), trang 2556–2565, 2018. 5

[36] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Ali Eslami, Oriol Vinyals và Felix Hill. Học tập đa phư ơng thức với mô hình ngôn ngữ đông lạnh. Trong M. Ranzato, A. Beygelzimer, Y. Dauphin, PS Liang và J. Wortman Vaughan, biên tập viên, Những tiến bộ trong Hệ thống xử lý thông tin thần kinh, tập 34, trang 200–212. Curran Associates, Inc., 2021. 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-reit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser và Il-lia Polosukhin. Sự chú ý là tất cả những gì bạn cần. Trong I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vish-wanathan và R. Garnett, biên tập viên, Những tiến bộ trong hệ thống xử lý thông tin thần kinh, tập 30. Curran Associates, Inc., 2017. 3, 11

[38] Ramakrishna Vedantam, C Lawrence Zitnick và Devi Parikh. Cider: Đánh giá mô tả hình ảnh dựa trên sự đồng thuận. Trong Biên bản báo cáo của hội nghị IEEE về máy tính tầm nhìn và nhận dạng mẫu, trang 4566–4575, 2015. 4

[39] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zi Cheng Liu, Ce Liu và Lijuan Wang. Git: Một công cụ chuyển đổi hình ảnh thành văn bản phục vụ cho mục đích thị giác và ngôn ngữ. bản in trư ớc arXiv arXiv:2205.14100, 2022. 1, 6

[40] Shuohang Wang, Yichong Xu, Yuwei Fang, Yang Liu, Siqi Sun, Ruochen Xu, Chenguang Zhu và Michael Zeng. Dữ liệu đào tạo có giá trị hơn bạn nghĩ: Một và phư ơng pháp hiệu quả bằng cách lấy từ dữ liệu đào tạo. Trong Biên bản cuộc họp thư ờng niên lần thứ 60 của Hiệp hội cho Ngôn ngữ học tính toán (Tập 1: Bài báo dài), trang 3170–3179, Dublin, Ireland, tháng 5 năm 2022. Hiệp hội cho Ngôn ngữ học tính toán. 3

[41] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang và William Yang Wang. Vatex: Một bộ dữ liệu đa ngôn ngữ chất lư ợng cao, quy mô lớn dành cho nghiên cứu video và ngôn ngữ. Trong Biên bản Hội nghị quốc tế IEEE/CVF về Tầm nhìn máy tính, trang 4581–4591, 2019. 5, 12

[42] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov và Yuan Cao. Simvlm: Ngôn ngữ hình ảnh đơn giản mô hình đào tạo trư ớc với sự giám sát yếu. bản in trư ớc arXiv arXiv:2108.10904, 2021. 1, 2, 4

[43] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. Transformers: Xử lý ngôn ngữ tự nhiên hiện đại. Trong Biên bản Hội nghị về Phư ơng pháp Thực nghiệm trong Xử lý Ngôn ngữ Tự nhiên: Biểu diễn Hệ thống, các trang 38–45, 2020. 4

[44] Chunpu Xu, Wei Zhao, Min Yang, Xiang Ao, Wangrong Cheng, và Jinwen Tian. Một thế hệ thống nhất-thu hồi khuôn khổ cho chú thích hình ảnh. Biên bản báo cáo của ngày 28 Hội nghị quốc tế ACM về quản lý thông tin và kiến thức, 2019. 2

[45] Jun Xu, Tao Mei, Ting Yao và Yong Rui. Msr-vtt: Một bộ dữ liệu mô tả video để kết nối video và ngôn ngữ. Trong Biên bản hội nghị IEEE về tầm nhìn máy tính và nhận dạng mẫu, trang 5288–5296, 2016. 5, 12

[46] Kelvin Xu, Jimmy Ba, Ryan Kiros, Cho Kyunghyun, Aaron Courville, Ruslan Salakhudinov, Rich Zemel và Yoshua Bengio. Hiển thị, tham dự và kể: Tạo chú thích hình ảnh thần kinh với sự chú ý trực quan. Trong hội nghị quốc tế về học máy, trang 2048–2057. PMLR, 2015. 2

[47] Peter Young, Alice Lai, Micah Hodosh và Julia Hocken-maier. Từ mô tả hình ảnh đến biểu thị trực quan: Mới số liệu tư ơng tự cho suy luận ngữ nghĩa trên các mô tả sự kiện. Giao dịch của Hiệp hội tính toán Ngôn ngữ học, 2:67–78, 2014. 5, 12

[48] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Mở các mô hình ngôn ngữ chuyển đổi đư ợc đào tạo trư ớc. arXiv preprint arXiv:2205.01068, 2022. 7, 12

[49] Shanshan Zhao, Lixiang Li, Haipeng Peng, Zihang Yang, và Jiaxuan Zhang. Tạo chú thích hình ảnh thông qua thống nhất phư ơng pháp truy xuất và dựa trên thế hệ. Khoa học ứng dụng, 10(18), 2020. 2