Smallcap

Lightweight Image Captioning Prompted with Retrieval Augmentation

Ramos, Rita; Martins, Bruno; Elliott, Desmond; Kementchedjhieva, Yova

Published in:
Proceedings - 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023

DOI:
10.1109/CVPR52729.2023.00278

Publication date:
2023

Document version
Peer reviewed version

Document license:
Unspecified

Citation for published version (APA):
https://doi.org/10.1109/CVPR52729.2023.00278
**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. 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 pretrained 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-
properties: 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).

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.

**2.2. 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

---

2The nocaps results shown in the figure include only models that follow the challenge guidelines, by training on the COCO dataset only.
a prompt. SMALLCAP combines powerful pre-trained uni-modal 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 $d_{\text{model}} = 768$ hidden dimensions and $h = 12$ cross-attention heads, $d$ defaults to 64 ($d_{\text{model}}/h$), 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 \{caption\}_1\ldots\{caption\}_k. This image shows ...\textsuperscript{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 ($\theta$) 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$_{\text{Base}}$, as available

---

\textsuperscript{3} See Appendix C for more information on the prompt template.
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, \( \text{SALLCAP} \), 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.\(^6\)

### 4.2. 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, \( \text{SALLCAP} \) 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, \( \text{SALLCAP}_{d=8, \text{Base}} \) still yields competitive performance, and even with just 1.8M trainable parameters, \( \text{SALLCAP}_{d=4, \text{Base}} \) is better than the substantially larger models ClipCap and I-Tuning.\(^\text{Base}\). We also experiment with Medium and Large GPT-2 decoders (\( \text{SALLCAP}_{\text{Medium}} \) and \( \text{SALLCAP}_{\text{Large}} \) 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, \( \text{SALLCAP} \) shows competitive performance on COCO, the dataset it was trained on. In contrast to previous lightweight-training models, \( \text{SALLCAP} \) further has the ability to generalize and transfer out-of-domain without retraining, as shown in subsequent experiments.

### Table 1. Results on the COCO test set with cross-entropy training.

| 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 | \(|\theta|\) | B@4 | M  | CIDEr | S  |
|------------------------------|-------------|-----|----|-------|----|
| 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 |
| \( \text{SALLCAP} \) [7]      | 7  | 37.0 | 27.9 | 119.7 | 21.3 |
| \( \text{SALLCAP}_{d=16, \text{Large}} \) | 47 | 37.2 | 28.3 | 121.8 | 21.5 |
| \( \text{SALLCAP}_{d=16, \text{Med}} \) | 22 | 36.5 | 28.1 | 120.7 | 21.6 |
| \( \text{SALLCAP}_{d=8, \text{Base}} \) | 3.6 | 36.7 | 27.8 | 119.1 | 21.1 |
| \( \text{SALLCAP}_{d=4, \text{Base}} \) | 1.8 | 36.0 | 27.4 | 117.4 | 21.0 |

<table>
<thead>
<tr>
<th>Model</th>
<th>In</th>
<th>Near</th>
<th>Out</th>
<th>Entire</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSCAR(_{\text{Large}} ) [6]</td>
<td>84.8</td>
<td>82.1</td>
<td>73.8</td>
<td>80.9</td>
</tr>
<tr>
<td>CaMEL* [9]</td>
<td><strong>88.1</strong></td>
<td>79.1</td>
<td>54.6</td>
<td>75.9</td>
</tr>
<tr>
<td>ClipCap* [7]</td>
<td>74.5</td>
<td>65.6</td>
<td>47.1</td>
<td>63.4</td>
</tr>
<tr>
<td>( \text{SALLCAP} )</td>
<td>83.3</td>
<td>77.1</td>
<td>65.0</td>
<td>75.8</td>
</tr>
<tr>
<td>( \text{SALLCAP}_{+W+H} )</td>
<td>87.9</td>
<td><strong>84.6</strong></td>
<td><strong>84.4</strong></td>
<td><strong>85.0</strong></td>
</tr>
</tbody>
</table>

### Table 2. CIDEr results on the nocaps test set.

\(^4\)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.
\(^6\)https://github.com/tylin/coco-caption
\(^7\)See Appendix E for more results regarding scaling the decoder.
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: Flickr30k [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).

5.2.1. Datastore augmentation

<table>
<thead>
<tr>
<th>Datastore augmentation</th>
<th>F30K</th>
<th>VW</th>
<th>MV</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-domain</td>
<td>55.4</td>
<td>47.7</td>
<td>29.2</td>
</tr>
<tr>
<td>In-domain + Web</td>
<td>58.6</td>
<td>48.0</td>
<td>29.8</td>
</tr>
<tr>
<td>In-domain + Human-labeled</td>
<td>57.6</td>
<td>47.5</td>
<td>30.9</td>
</tr>
<tr>
<td>In-domain + W + H</td>
<td>57.9</td>
<td>48.0</td>
<td>30.7</td>
</tr>
</tbody>
</table>

*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]). 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
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.

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.
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 been 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.

### 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.
language models and is therefore model agnostic.\textsuperscript{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 PID-DAC programme (UIDB/50021/2020).

\footnotesize
\textsuperscript{16}See Appendix E for OPT results without retrieval.
\textsuperscript{17}Due to our academic computing budget, we only repeat the experiments from Table 1. Future work can experiment further in this direction.
References


