Retrieval-augmented Image Captioning

Rita Ramos[†] **Desmond Elliott**[⋆] **Bruno Martins**[†]
[†]INESC-ID, Instituto Superior Técnico, University of Lisbon
[⋆]Department of Computer Science, University of Copenhagen
`ritaparadaramos@tecnico.ulisboa.pt`

Abstract

Inspired by retrieval-augmented language generation and pretrained Vision and Language (V&L) encoders, we present a new approach to image captioning that generates sentences given the input image and a set of captions retrieved from a datastore, as opposed to the image alone. The encoder in our model jointly processes the image and retrieved captions using a pretrained V&L BERT, while the decoder attends to the multimodal encoder representations, benefiting from the extra textual evidence from the retrieved captions. Experimental results on the COCO dataset show that image captioning can be effectively formulated from this new perspective. Our model, named EXTRA, benefits from using captions retrieved from the training dataset, and it can also benefit from using an external dataset without the need for retraining. Ablation studies show that retrieving a sufficient number of captions (e.g., k=5) can improve captioning quality. Our work contributes towards using pretrained V&L encoders for generative tasks, instead of standard classification tasks.

1 Introduction

Image captioning is the task of automatically generating a short textual description for a given image. The standard approach involves the use of encoder-decoder neural models, combining a visual encoder with a language generation decoder (see Hossain et al. (2019) for a survey). In early studies, the encoder was typically a Convolutional Neural Network model (CNN) pretrained on the ImageNet classification dataset (Russakovsky et al., 2015) or a pretrained Faster-RCNN object detector (Ren et al., 2015), whereas the decoder was commonly an LSTM (Hochreiter and Schmidhuber, 1997) together with an attention mechanism (Bahdanau et al., 2014). More recently, Transformer based models have been achieving state-of-the-art results on a variety of language processing (Vaswani et al.,

2017; Devlin et al., 2018; Radford et al., 2019) and computer vision tasks (Dosovitskiy et al., 2020). Accordingly, state-of-the art image captioning models have replaced the conventional CNN-LSTM approach with encoder-decoder Transformers (Liu et al., 2021). Still, in both cases, the encoder only attains visual representations, whereas richer features could be captured from image–text interactions if the encoder had access to useful textual context related to the input image (e.g., sentences associated to similar images).

In this paper, we present a new type of image captioning model that uses a pretrained V&L BERT (Tan and Bansal, 2019; Li et al., 2019; Bugliarello et al., 2020, *inter-alia*) to encode both the input image and captions retrieved from similar images. This model generates captions conditioned on representations that consider linguistic information beyond the image alone. Moreover, specifically using the retrieved captions as textual contexts rather than other alternatives (e.g., image tags or object names) can aid guiding the language generation process, since the model is now provided with well-formed sentences that are semantically similar to what the predicted caption should resemble.

In experiments on the COCO dataset (Chen et al., 2015), the proposed model is competitive against state of the art methods. In a series of ablation experiments, we find that the model improves when encoding multiple retrieved captions, and that it could reach better performance if it was able to retrieve better captions from the datastore. In experiments on the smaller Flickr30K dataset, we show that allowing the model to retrieve captions from the larger COCO dataset can improve performance without needing to retrain the model.

We hope that our work inspires the adoption of pretrained V&L encoders for a broader range of generative multimodal tasks. There have been several recent studies proposing V&L BERTs to learn generic multi-modal representations with

large amounts of paired image and text data, which can then be fine-tuned to downstream tasks. However, these pretrained models have mostly been applied to classification tasks and have seen limited use for image captioning, a task which typically only considers single-input images, as opposed to image-text pairs, as proposed in this work.

2 Model

We present a model that captions images, given both the image and a set of k captions retrieved from similar images using a retrieval system. This approach belongs to the class of retrieval-augmented language generation models (Weston et al., 2018; Izacard and Grave, 2020). In our model, the image and the retrieved captions are jointly encoded using a pretrained V&L encoder to capture cross-modal representations in the combined input data. We denote our model as EX-TRA: **E**ncoder with **C**ross-modal representations **T**hrough **R**etrieval **A**ugmentation. It consists of three components, namely an encoder, a retrieval system, and a decoder.

2.1 Encoder

The encoder in EXTRA is LXMERT[1] (Tan and Bansal, 2019), a pretrained vision-and-language Transformer that jointly encodes a visual input **V** and a linguistic input **L**. The visual input is represented as $N{=}36$ regions-of-interest V=$\{v_1, ..., v_N\}$ extracted from the image using the Faster-RCNN object detector, pretrained (Anderson et al., 2018) on the Visual Genome dataset (Krishna et al., 2016). A sentence in the linguistic input is tokenized into M sub-words using the BERT tokenizer (Devlin et al., 2018), starting with a special classification token CLS and ending with a special delimiter token SEP. We extended LXMERT to encode k sentences by concatenating the tokenized sentences into a single input, each separated by the delimiter token: L=$\{$CLS$, w_1^{L_1} \dots w_M^{L_1},$ SEP$\dots w_1^{L_k} \dots w_M^{L_k},$ SEP$\}$. The sentences are obtained from a datastore via a retrieval system, as explained in Section 2.2.

The encoder produces a sequence of cross-modal representations of image and the text, which are the inputs to the decoder, described in Section 2.3.

2.2 Image–Text Retrieval and Datastore

The retrieval system builds on the Facebook AI Similarity Search (FAISS) nearest-neighbour search library (Johnson et al., 2017). FAISS allows for the indexing of high-dimensional vectors, i.e., a datastore **D**, and it offers the ability to quickly search through the datastore given a similarity measure **S**, e.g., Euclidean distance or cosine similarity.

Given an input image **V**, the retrieval system finds L, the set of k captions retrieved from the datastore, which EXTRA encodes together with the image. The datastore consists of captions associated with images in a dataset[2]. Each caption in the datastore, and the query input image, are represented using vectors extracted from CLIP (Radford et al., 2021), allowing image–text search by projecting images and text to a shared latent space. Using FAISS, the input image can then be compared against the vectors[3] from **D** to search over the corresponding k nearest-neighbours captions.

2.3 Decoder

The decoder is a conditional auto-regressive language model based on GPT-2 (Radford et al., 2019) with additional cross-attention layers to the encoder. The Transformer layers in the decoder already contain a masked multi-head self-attention sublayer, which self-attends to the previous words. We add cross-attention layers (Vaswani et al., 2017) subsequent to the masked self-attention sublayers, so the decoder can attend to the encoder outputs.

The decoder predicts a caption $y_1 \dots y_M$ token-by-token, conditioned on the previous tokens and the outputs of the V&L encoder. The model's parameters θ are trained by minimizing the sum of the negative log-likelihood of predicting the ground truth token at each time-step, using the standard cross-entropy loss:

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

We can also fine-tune the model with Self-Critical Sequence Training (Rennie et al., 2017).

3 Experimental Protocol

3.1 Datasets and Metrics

We evaluate our model on the COCO dataset (Chen et al., 2015), using the standard Karpathy splits of

---

[1]The exploration of other encoders is left for future work.

[2]This can either be the training set or an external dataset.
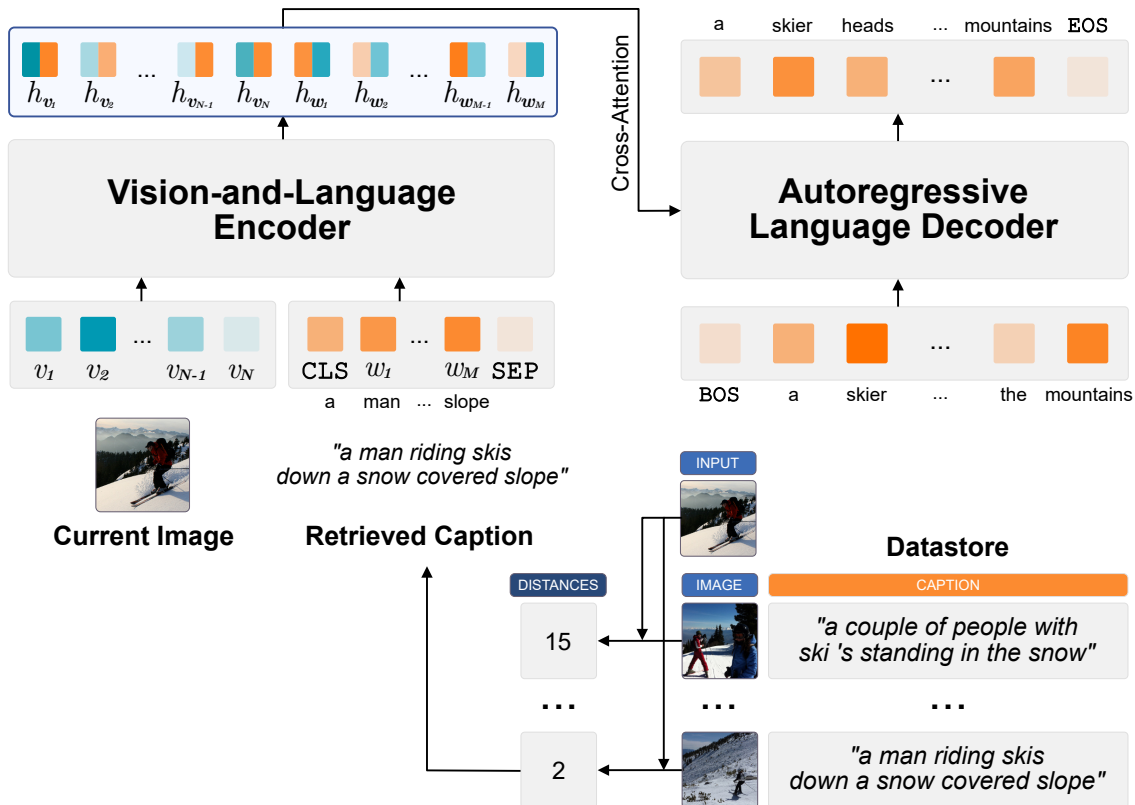[3]This comparison can be pre-computed for efficiency.

Figure 1: Illustration of the EXTRA model. Given an input image, EXTRA retrieves captions from a datastore and encodes both the input image and the retrieved captions using a pretrained vision-and-language encoder. The decoder attends over both the visual and linguistic outputs, improving the quality of the generated caption.

113287 images for training, 5000 for validation, and 5000 for testing, with 5 captions per image.

Standard metrics were used to evaluate caption generation, namely BLEU-4 (B4) (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016), using the MS COCO caption evaluation package[4].

3.2 Implementation and Training Details

The implementation[5] of EXTRA uses the Hugging-Face Transformers library (Wolf et al., 2020). The encoder is LXMERT (Tan and Bansal, 2019), a 14-layer V&L model pretrained on 9 million image–sentence pairs across a variety of datasets and tasks. Following Liu et al. (2021), the decoder is a 4-layer randomly initialized GPT-2-style Transformer network with 12 attention heads and additional cross-attention layers. The retrieval systems uses FAISS with a flat index (*IndexFlatIP*) without any training. The corresponding datastore **D** consists of

all the captions associated to the 113287 images in the COCO training set. For caption retrieval, the captions in the datastore and the input image (i.e., the query) are both represented with features extracted from the CLIP-ResNet50×4 pretrained model. Using the cosine similarity for comparison, a total of $k = 5$ captions are retrieved to be jointly encoded with the input image by EXTRA. Notice that CLIP-ResNet50×4 features are only used for retrieval, while the EXTRA encoder, i.e. the pretrained LXMERT, requires Faster-RCNN features, and thus it cannot use CLIP visual features.

EXTRA is trained in two stages using a single NVIDIA V100S 32GB GPU. In the first stage, EXTRA is trained end-to-end with the cross-entropy loss, using a batch size of 64 and the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of $3e - 5$. The encoder is trained with a linear warmup for the first epoch to prevent gradients from the randomly initialized decoder from harming the pretrained encoder. The model was trained with early stopping: training ends if there is no improvement after 5 consecutive epochs on the val-

	Cross-Entropy Optimization				CIDEr Optimization			
	B4	METEOR	CIDEr	SPICE	B4	METEOR	CIDEr	SPICE
Encoder-Decoder models								
Up-Down	36.2	27.0	113.5	20.3	36.3	27.7	120.1	21.4
CaMEL$_{\text{Faster R-CNN}}$	36.1	28.0	114.8	20.8	-	-	-	-
GCN-LSTM	36.8	27.9	116.3	20.9	38.2	28.5	127.6	22.0
VL-T5	34.5	28.7	116.5	21.9	-	-	-	-
AoANet	37.2	28.4	119.8	21.3	38.9	29.2	129.8	22.4
CPTR	-	-	-	-	40.0	29.1	129.4	-
EXTRA ($k = 5$)	37.5	28.5	120.9	21.7	36.4	28.2	131.1	21.3
CaMEL$_{\text{CLIP-RN50}\times16}$	38.8	29.4	125.0	22.2	41.3	30.2	140.6	23.9
V&L BERT models								
VLP	36.5	28.4	116.9	21.2	39.5	29.3	129.3	23.2
OSCAR$_B$	36.5	30.3	123.7	23.1	40.5	29.7	137.6	22.8
VinVL$_B$	38.2	30.3	129.3	23.6	40.9	30.9	140.4	25.1

Table 1: Results on the Karpathy COCO test split. **EXTRA** ($k = 5$) is competitive against encoder-decoder models. We present results with cross-entropy training and after Self-Critical Sequence Training using the CIDEr metric.

idation set over the BLEU-4 metric. In the second stage, EXTRA is fine-tuned with Self-Critical Sequence Training (Rennie et al., 2017) with CIDEr optimization and greedy search decoding as a baseline, using a batch size of 55, a learning rate of 3e-5, and a frozen encoder. Captions are decoded using beam search with a beam size of 3.

4 Results

Table 1 shows the performance of EXTRA compared to strong encoder-decoder models. We compare against the widely-used Up-Down (Anderson et al., 2018) and AoANet models (Huang et al., 2019), both using a Faster-RCNN image encoder; the GCN-LSTM model (Yao et al., 2018) with a Graph Convolutional Network (GCN) encoder; the CPTR model (Liu et al., 2021) employing a ViT Transformer encoder (Dosovitskiy et al., 2020); the VL-T5 Transformer model (Cho et al., 2021) with a vision and language encoder; and the recent CaMEL model (Barraco et al., 2022) with the CLIP-RN50×16 encoder. Our model is also compared with state-of-art models that do not use the encoder-decoder paradigm but instead unify the Transformer encoder and decoder into a single model, namely the VLP (Zhou et al., 2020), OSCAR-base (Li et al., 2020), and the VinVL-base (Zhang et al., 2021) models. We note that these are general purpose V&L models, not specifically designed for image captioning.

Overall, EXTRA is competitive to state-of-the art captioning models. It outperforms captioning models with vision encoders, and VL-T5, which, like EXTRA, uses a V&L encoder, but with object tags as linguistic inputs rather than retrieved captions. Although EXTRA does not outperform the state of the art captioning model, CaMEL, that uses a dual decoder, it outperforms the variant of CaMEL that uses the same Faster-RCNN features. EXTRA also competes with general purpose V&L BERT models. Notice that our approach can be adapted to other V&L encoders besides LXMERT (e.g., OSCAR, VinVL, etc.), or to more powerful decoders (e.g., as in CaMEL). Likewise, other models could benefit from retrieval-augmentation with captions.

4.1 Ablation Studies

We conducted a series of ablation studies in the Karpathy COCO validation split to better understand what contributes to the success of EXTRA.

Varying the Number of Retrieved Captions: We start by studying the importance of training with multiple retrieved captions, training with k=1 and k=3 captions to explore the effect of retrieving fewer captions. Table 2 reports the result of this experiment, showing that performance degrades when retrieving less captions.

Encoding Irrelevant Captions: We also studied the performance of EXTRA when it encodes textual

	B4	CIDEr
$k = 1$	36.7	118.0
$k = 3$	37.4	119.1
$k = 5$	38.3	121.2

Table 2: The effect of training and evaluating using different numbers of retrieved captions. Performance reported after training with cross-entropy optimization.

input that is not expected to be useful. We conduct two experiments where EXTRA is trained with textual input that is either an empty caption or a randomly chosen caption.

- **Empty Caption**: encode the image with an empty sentence: L={CLS, SEP};

- **Random Caption**: encode the image with a random caption from the datastore.

Table 3 shows the result of this experiment. EXTRA outperforms both variants, further showing that the generation process is improved by encoding the image together with relevant textual context from nearest-neighbour captions. Although having an inferior performance, both models reach reasonable results compared to other models in the literature (see Table 1), showing that LXMERT can be used as a strong encoder for image captioning without providing relevant input image-text pairs.

	B4	CIDEr
Empty caption	37.8	119.1
Random caption	37.1	117.7
EXTRA	38.3	121.2

Table 3: The effect of training and evaluating with captions that are not expected to be useful.

Encoding Irrelevant Images: We tested ablating the visual input (i.e., setting the visual features to zero). Training on "blacked out" input images achieves 102.1 in CIDEr, which is substantially lower than training with the actual input images, as seen in Table 4. This further shows that EXTRA uses the visual input, and does not just rely on the retrieved information.

Changing the Retrieval System and Datastore: We then studied the effect of changing the retrieval system and the representations in the datastore. Recall that EXTRA relies on captions obtained by

	B4	CIDEr
Blacked out image	32.1	102.1
EXTRA	38.3	121.2

Table 4: The effect of training and evaluating with "blacked out" input images.

Image–Text retrieval, where the datastore contains the captions from the COCO training set, represented as vectors extracted from CLIP. We conducted experiments with Image–Image and Image–Text retrieval to understand which performs better:

- **Image–Image Retrieval**: the datastore consists of all the images in the training data. The representation of the input image is compared against those in the datastore to find the k nearest-neighbour images, and, subsequently, to obtain the k captions associated to those images. Specifically, one reference caption is retrieved from each of the top-k nearest-neighbour images.

- **Image–Text Retrieval**: the datastore consists of all the captions associated to the images in the training data. The representation of the input image is compared against the captions to directly find the top-k captions.

For Image–Image retrieval, the input image and the images in **D** are represented with *Faster R-CNN* features, after global average pooling the embeddings of the 36 region-of-interest vectors. For Image–Text Retrieval, the input image and the caption vectors should already belong to a shared semantic space. We use the pretrained CLIP model because it satisfies this criteria and thus allows for direct image–text comparison. We considered two variants of CLIP based on their visual backbone: *ViT* or *ResNet50x4*[6]. The results of this experiment are reported in Table 5.

EXTRA performs worse when it uses Image–Image retrieval in comparison to retrieving captions directly with Image–Text retrieval. The best performance is obtained with the ResNet-variant of the CLIP encoder. We also assess the performance of directly using *only* one of the retrieved captions, with the results shown in Figure 2. In this figure, we can visualize the expected CIDEr score

---

[6]Regarding the comparison measure **S**, the Euclidean distance and cosine similarity were used respectively for Image-to-Image retrieval and Image-to-Text retrieval.

	B4	CIDEr
Image–Image (Faster R-CNN)	36.8	117.1
Image–Text (CLIP ViT)	38.1	120.3
Image–Text (CLIP ResNet)	38.3	121.2

Table 5: The effect of training and evaluating EXTRA with different retrieval systems. $k = 5$ in both settings.

	B4	CIDEr
$k = 5$ retrieved captions	38.3	121.2
$k = 4$ and 1 reference	39.0	123.0
$k = 0$ and 5 references	40.9	129.5

Table 6: Simulation of an oracle experiment, where EX-TRA can "retrieve" reference captions of an image instead of retrieving all 5 captions from the datastore.

of the first retrieved captions and observe that some of them do not sufficiently describe the image, or are mismatches, with a CIDEr of zero. We also observe that the CIDEr score can change significantly depending on the retrieval system. A larger number of mismatch captions are retrieved with Image-to-Image retrieval. This suggests that the retrieval system and the datastore can largely impact a retrieval-augmented image captioning model, hence they should be carefully considered.
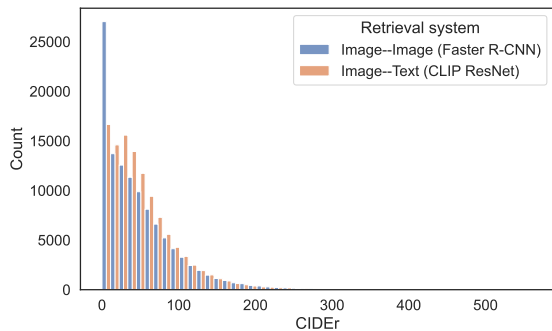

Figure 2: Histogram of the CIDEr scores for the nearest caption ($k = 1$) retrieved with Image-Image and Image-Text retrieval. This shows the evaluations scores of using only the retrieved captions.

Oracle Performance: Given that the retrieval system and datastore affect the performance of EX-TRA, we also study whether EXTRA could continue to improve if it could retrieve *better* captions. After training EXTRA with the $k = 5$ retrieved captions, we simulate an oracle retrieval system during inference, by allowing the actual reference captions to be encoded by EXTRA. Table 6 reports on experiments in the validation data with respect to replacing one of the k retrieved captions with one of the reference captions, as well as replacing all with the 5 references associated to the input images. These experiments bring a 1.8 and 8.3 point increase in CIDEr score, respectively, showing the potential for EXTRA to improve by retrieving captions that better match the input image.

5 Discussion

5.1 Vision First and Language Later

How does EXTRA use the encoded image and retrieved captions? We quantify this by estimating the behaviour of the cross-modal attention heads at each layer in the decoder. Specifically, we compute the average of the cross-modal attention across either the number of image regions or the sub-words in the encoder, at each time-step of generating a caption and across each of the 12 attention heads.

Figure 3 shows that across the layers, the decoder's attention shifts to the textual outputs. In Layer 1, the model attends both to the visual and textual representations, but the model hardly pays attention to the visual outputs by Layer 4, relying more on the textual information from the retrieved captions. This behaviour further shows that the semantics of the nearest captions can aid guiding the language generation process. We performed an identical calculation for the variants of EXTRA that encoded an empty or a random caption, finding in this case the opposite behaviour: the model learned to ignore the textual embeddings provided by the encoder (see Appendix A).
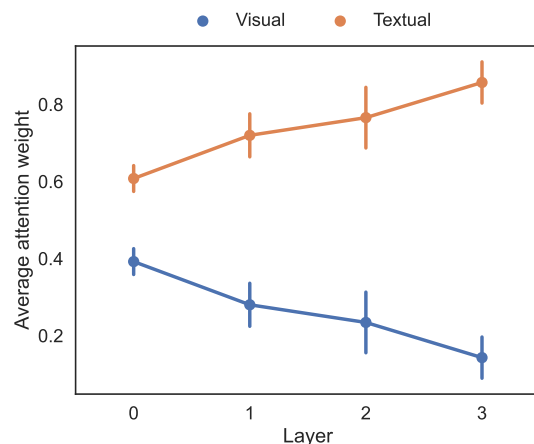

Figure 3: The average cross-attention from the decoder to the outputs of the encoder in respect to the visual V and textual L outputs. Values from COCO validation.

5.2 Retrieve Enough Captions to Overcome Retrieval Mistakes

We note that training with an empty set of captions was better than encoding a single $k = 1$ and $k = 3$ retrieved captions, observing Tables 2 and Table 3. Thus, retrieval augmentation aids to improve caption quality when a sufficient number ($k = 5$) is considered. This further shows that retrieving enough captions can be crucial for success. For this, we hypothesise that retrieving more captions makes the model more robust in the presence of mismatches from certain captions, as shown for instance in the second example in Figure 4.

5.3 Hot-swapping the Datastore

Besides taking advantage of similar training examples, we study whether EXTRA works with external image–caption collections without needing to retrain the model. For this experiment, EXTRA was first trained and evaluated in a small dataset, and then the retrieval datastore was augmented with a larger dataset. The considered datasets were Flickr30k and COCO, respectively. While Flickr30k only contains 30k images, COCO contains 113K, each paired with five sentences. Table 7 reports the results of these experiment. EXTRA got a better performance considering a larger external dataset than just using the current training set, showing the potential for EXTRA to adapt the retrieval datastore.

Retrieval Datastore	B4	CIDEr
Flickr30k	28.8	59.6
+ COCO	29.5	59.9

Table 7: Performance of EXTRA on the Flickr30k validation set. The model is trained on the Flickr30K dataset with the Flickr30K datastore. The datastore for inference is either the Flickr30K training set or combined with the COCO training set.

5.4 Qualitative Examples

Figure 4 shows examples of captions generated by EXTRA, given the input image and the $k = 5$ retrieved captions. EXTRA benefits from textual evidence from nearest-neighbour captions, even though sometimes the retrieved information can be misleading, as depicted in the last example. More examples are provided in Appendix B.

6 Related Work

Image Captioning: The task of image captioning is usually addressed by one of these three main approaches: templates, retrieval, and encoder-decoder methods. Early approaches involved template-based methods that consisted of filling blanks of predefined captions through object detection (Farhadi et al., 2010; Kulkarni et al., 2013; Elliott and de Vries, 2015). Retrieval-based methods instead search over a dataset for the most similar image and fetch the corresponding caption (Hodosh et al., 2013; Ordonez et al., 2011). Currently, the most common approach is the encoder-decoder framework (Xu et al., 2015; Hossain et al., 2019). The encoder typically used a pretrained CNN (Vinyals et al., 2016) or a Faster R-CNN (Anderson et al., 2018), encoding the image into a grid of image features or object proposal image regions. The decoder was usually a LSTM with an attention mechanism (Xu et al., 2015) to dynamically focus on different parts of the encoded image during the prediction of each word.

Recently, Transformer-based models like BERT (Devlin et al., 2018) have become a more popular choice than LSTMs models, outperforming recurrent architectures in different natural language processing (NLP) tasks (Vaswani et al., 2017; Qiu et al., 2020). Transformers can capture long-range dependencies with self-attention layers and they can process each word of a sentence in parallel, reducing training time. After the successful application in NLP, vision Transformers like ViT (Dosovitskiy et al., 2020) are also starting to become the model of choice in the field of computer vision in place of CNNs. In similar fashion, most recent captioning studies use the Transformer arquitecture (Herdade et al., 2019; Cornia et al., 2020; Liu et al., 2021), employing a vision Transformer as encoder together with an autoregressive language Transformer as decoder. Similarly to these models, this work proposes a encoder-decoder Transformer model for the task of image captioning. However, unlike them, the proposed model incorporates a pretrained V&L BERT to exploit cross-modal representations, encoding images along with textual context. Also differently from previous work, this approach explores retrieval-augmented generation, i.e., combining neural encoder-decoder methods with traditional retrieval-based methods.
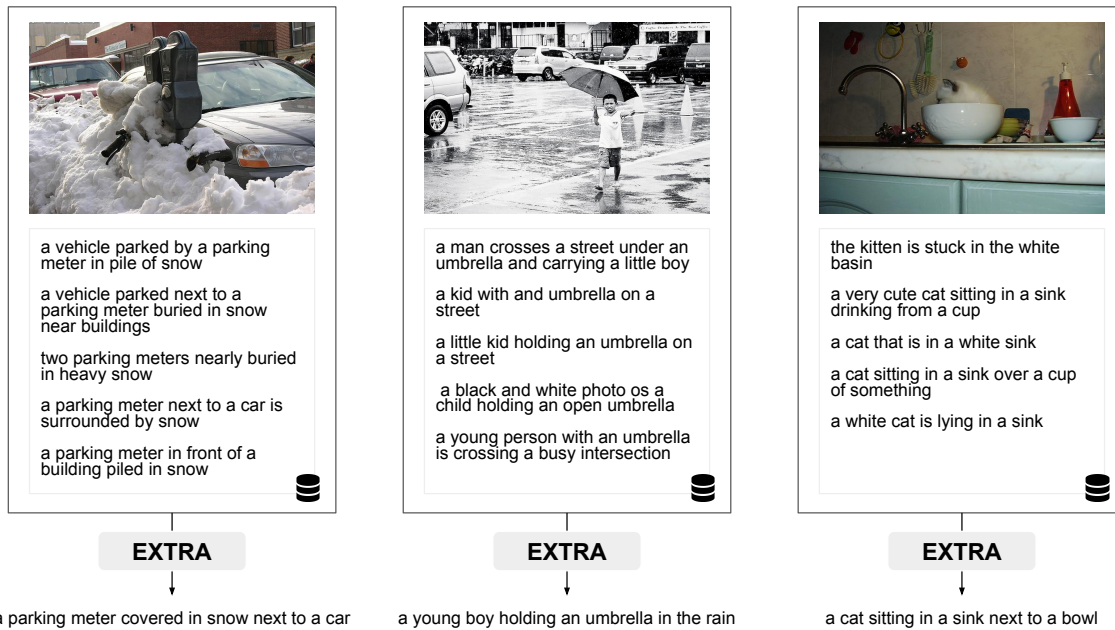
a vehicle parked by a parking meter in pile of snow

a vehicle parked next to a parking meter buried in snow near buildings

two parking meters nearly buried in heavy snow

a parking meter next to a car is surrounded by snow

a parking meter in front of a building piled in snow

a man crosses a street under an umbrella and carrying a little boy

a kid with and umbrella on a street

a little kid holding an umbrella on a street

a black and white photo os a child holding an open umbrella

a young person with an umbrella is crossing a busy intersection

the kitten is stuck in the white basin

a very cute cat sitting in a sink drinking from a cup

a cat that is in a white sink

a cat sitting in a sink over a cup of something

a white cat is lying in a sink

EXTRA → a parking meter covered in snow next to a car

EXTRA → a young boy holding an umbrella in the rain

EXTRA → a cat sitting in a sink next to a bowl

Figure 4: Examples of captions generated by EXTRA conditioned on the input image and retrieved captions.

V&L BERTs: Previous studies have proposed pretrained Vision and Language (V&L) BERTs to learn generic cross-modal representations of images and text, that can later be used for a variety of downstream V&L tasks (Bugliarello et al., 2020). Examples include LXMERT (Tan and Bansal, 2019), VL-BERT (Su et al., 2019), Visual BERT (Li et al., 2019), OSCAR (Li et al., 2020), or UNITER (Chen et al., 2020), which were applied to VQA and other V&L classification tasks. Given that these models are encoder-only Transformers, only few of them have been applied to generation tasks such as image captioning. In such cases, the generation is made from left to right by encoding the input image and using the textual input elements with uni-directional attention masks, i.e., starting with a CLS token with the rest of the tokens masked, then considering the CLS token with the predicted word (replaced by the corresponding mask token) and the remaining ones still masked, and so on (Li et al., 2020; Zhou et al., 2020).

The use of pretrained V&L BERTs, as encoders in the standard encoder-decoder captioning framework, remains largely unexplored. The task of image captioning typically just considers single-input images, and not image-text pairs to be encoded. In our work, a pretrained V&L encoder is used with a decoder for image captioning, by leveraging not just the images as input but also retrieved captions.

Besides pretrained V&L encoders, pretrained

V&L encoder-decoder models have recently been proposed to tackle classification and generation tasks, such as VL-T5 (Cho et al., 2021). Their captioning approach is similar to the present paper, but VL-T5 uses object tags as textual inputs, whereas EXTRA is conditioned on retrieved captions.

Retrieval-augmented Generation: The proposed approach is also similar to some studies on language generation that predict the output conditioned on retrieved examples (Weston et al., 2018; Gu et al., 2018; Khandelwal et al., 2019; Lewis et al., 2020). For instance, this work relates to Weston et al. (2018), in which a sequence-to-sequence LSTM model, for dialog generation, encodes the current input concatenated with the nearest retrieved response. Similarly, Izacard and Grave (2020) used an encoder-decoder Transformer conditioned on retrieved passages for open domain question answering. Retrieval-augmented generation is gaining traction in NLP but has only been explored for image captioning by few studies (Wang et al., 2020; Fei, 2021; Ramos et al., 2021; Sarto et al., 2022; Ramos et al., 2022). Concurrent work proposed Transformer-based captioning models augmented with retrieval as well (Sarto et al., 2022; Ramos et al., 2022). However, differently from these previous studies, we encode the retrieved captions by exploiting cross-modal representations with a V&L encoder.

7 Conclusions

We propose EXTRA, a retrieval-augmented image captioning model that improves performance by exploiting cross-modal representations of the input image together with captions retrieved from a datastore. EXTRA make uses of a pretrained V&L BERT, instead of an image-only encoder, combined with a language decoder. To generate a caption, the decoder attends to the cross-modal encoder features, containing information from image regions and also textual evidence from the retrieved captions. Image captioning is therefore addressed as language generation conditioned on vision and language inputs, instead of vision only. To evaluate this model, EXTRA was assessed against strong encoder-decoder models in the area, and ablation studies were also conducted. The experiments conducted on the COCO dataset confirmed the effectiveness of the proposed captioning approach.

For future work, we plan to explore the utility of EXTRA in out-of-domain and in few-shot learning settings, since the retrieval component can be easily modified to include external datastores, without the need to retrain the whole model. We also plan to explore how this approach can be adapted to other powerful vision and language encoders besides LXMERT. Finally, we will explore methods that allow us to jointly train the retrieval mechanism with the full model in order to retrieve captions that are more similar to the input image.

Limitations

Previous work has shown that generative models suffer from biases inherent to the data they are trained on (Weidinger et al., 2021; Thoppilan et al., 2022). Likewise, our EXTRA model can suffer from biases present in the COCO image captioning dataset (Chen et al., 2015). Particularly, it has been shown that there is significant gender imbalance in COCO, and that captioning models can exhibit gender bias amplification (e.g., they are likely to generate the word "woman" in kitchen scenarios, and the word "man" in snowboarding scenes) (Hendricks et al., 2018; Zhao et al., 2017).

However, differently from most captioning models, EXTRA is a retrieval-augmented captioning model, and thus it has the potential to make predictions beyond the training data, by relying on information from an external datastore. Still, the datastore knowledge might also have inherent bias, as mentioned by previous studies on retrieval-augmented generation (Lewis et al., 2020). In the paper, we show examples of such limitations wherein mismatched retrieved captions can bias the model towards incorrect predictions (see the results and appendix sections).

As a way to mitigate these limitations, we recommend analyzing the corresponding nearest captions when using EXTRA, since the retrieved captions can give useful insight of the bias involved in the generation process. EXTRA can provide interpretability through textual descriptions, whereas most captioning models only provide explanations as visual attention maps.

EXTRA also has the downside of focusing on an English-centric dataset. Captioning datasets are primarily available in English, and most image captioning models are trained on COCO or other english-centric datasets. To avoid hindered research on image captioning, it is important to consider multilingual captioning datasets that contain both language-diverse captions and geographically-diverse visual concepts (Thapliyal et al., 2022).

References

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

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6077–6086.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Manuele Barraco, Matteo Stefanini, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cuc-

chiara. 2022. Camel: Mean teacher learning for image captioning. *arXiv preprint arXiv:2202.10492*.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2020. Multimodal pretraining unmasked: Unifying the vision and language berts. *arXiv preprint arXiv:2011.15124*.

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

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision*, pages 104–120.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. *arXiv preprint arXiv:2102.02779*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Desmond Elliott and Arjen de Vries. 2015. Describing images using inferred visual dependency representations. In *Proceedings of the Joint Conference of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 42–52.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of the European Conference on Computer Vision*, pages 15–29. Springer.

Zhengcong Fei. 2021. Memory-augmented image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2–9.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *Proceedings of the Conference of the Association for the Advancement of Artificial Intelligence*, pages 5133–5140.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision*, pages 771–787.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. *arXiv preprint arXiv:1906.05963*.

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

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys*, 51(6):1–36.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

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

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Proceedings of the Conference on Neural Information Processing Systems*, 33:9459–9474.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. OSCAR: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision*, pages 121–137.

Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. 2021. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Proceedings of the Conference on Neural Information Processing Systems*, 24:1143–1151.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

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

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

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2022. SmallCap: Lightweight image captioning prompted with retrieval augmentation. *arXiv preprint arXiv:2209.15323*.

Rita Parada Ramos, Patrícia Pereira, Helena Moniz, Joao Paulo Carvalho, and Bruno Martins. 2021. Retrieval augmentation for deep neural networks. *Proceedings of the International Joint Conference on Neural Networks*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Proceedings of the Conference on Neural Information Processing Systems*, 28:91–99.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7008–7024.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

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

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Ashish V Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. *arXiv preprint arXiv:2205.12522*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning

challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663.

Li Wang, Zechen Bai, Yonghua Zhang, and Hongtao Lu. 2020. Show, recall, and tell: Image captioning with recall mechanism. *Proceedings of the Conference of the Association for the Advancement of Artificial Intelligence*, 34(07):12176–12183.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Jason Weston, Emily Dinan, and Alexander H Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. *arXiv preprint arXiv:1808.04776*.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP: System Demonstrations*, pages 38–45.

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

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision*, pages 684–699.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Making visual representations matter in vision-language models. *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the Conference of the Association for the Advancement of Artificial Intelligence*, volume 34, pages 13041–13049.

A   Cross-Attention

In Section 5.1, we quantified how much attention EXTRA pays to the encoded image and retrieved captions. We also quantify this for the two other
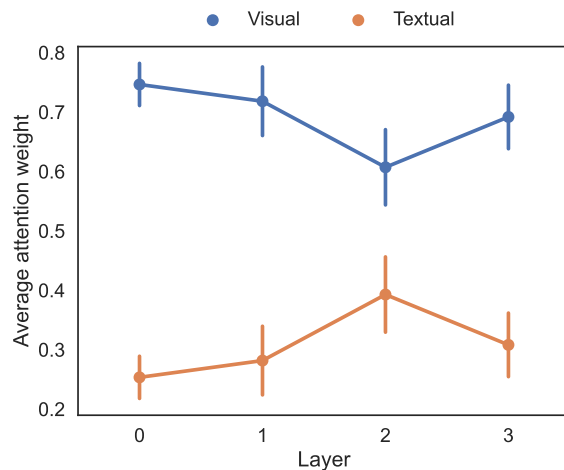

Figure 5: Cross-attention for the variant of EXTRA that that encodes an empty caption.
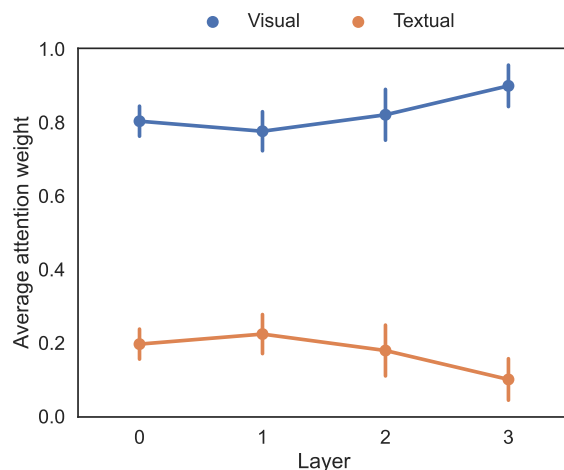
Figure 6: Cross-attention for the variant of EXTRA that that encodes a random caption.

variants of EXTRA which encode irrelevant captions, using either an empty or a random caption. Figures 5 and 6 show the average cross-attention weights from the decoder to the outputs of the encoder in respect to the visual $\mathbf{V}$ and textual $\mathbf{L}$ outputs, respectively for the empty and random caption encoding. Contrary to the findings presented in Section 5.1, regarding the encoding of retrieved captions, in this scenario the two variants pay more attention to the visual outputs instead.

For details on how we calculated the corresponding attention weights, we present the corresponding formula. Specifically, we calculated the average of the cross-modal attention C across either the number of image regions or the sub-words in the encoder at each of T time-step of generating a caption and across each of the $H = 12$ attention heads.

This calculation happens independently for each of the $L = 4$ layers in the decoder:

$$A(C^L, V) = \frac{1}{H} \sum_{j=1}^{H} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{|V|} \alpha_{j,t \to i}^L. \qquad (2)$$

$$A(C^L, T) = 1 - A(C^L, V). \qquad (3)$$

B   More Examples

Figure 7 shows additional examples of the captions generated by EXTRA considering the retrieved captions, against the other two variants: encoding an empty and random caption instead. For the first image, the two variants fail to recognize that the image shows kids playing basketball (perhaps given the small size of the ball), whereas EXTRA was able to identify it by having that information in the retrieved captions. In the second image[7], the two variants produced the error of generating *sandwich* while EXTRA correctly mentioned *hot-dog*, similar to the retrieved captions. EXTRA considers the semantics from the nearest captions retrieved during generation, sometimes even copying an entire sentence, as shown in Figure 9.

Figure 8 shows examples where the retrieved captions mislead the model. We note however that EXTRA is also able to succeed, despite the mismatch from retrieved captions, as seen in Figure 10.

---

[7]The person was blurred for privacy concerns.

a group of children play a game of basketball

a group of young people playing a game of basketball

young children playing a basketball game with the ball flying

a group of people play a game inside on a court

young kids playing a game of basketball on a basketball court

EXTRA

a group of children playing a game of basketball

CLS SEP

"EXTRA"
(empty caption)

a group of people walking down a street

two men are playing wii together in the living room

"EXTRA"
(random caption)

a group of young people walking across a street

a man putting a hot dog in a bun at a restaurant

holding a hot dog in a <unk> bun with a napkin

a person holds a hot dog with onions up to the camera

a person holding a bottle of drink and a hotdog in a napkin at a hot dog stand

a man holds a hotdog near a food stand and parking lot

EXTRA

a man holding a hot dog in his hand

CLS SEP

"EXTRA"
(empty caption)

a man holding a sandwich in his hand

a man playing baseball in the middle of a pitch

"EXTRA"
(random caption)

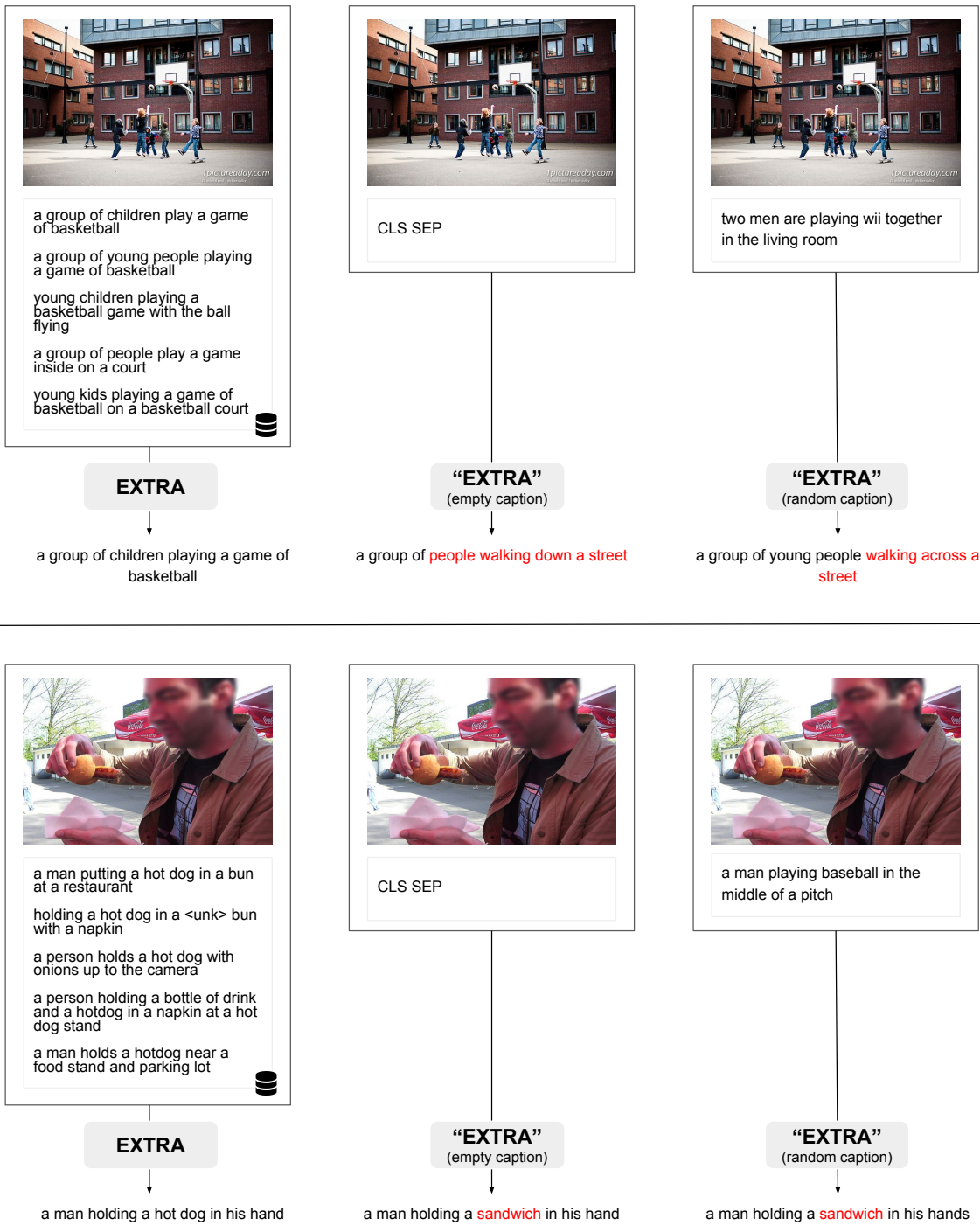a man holding a sandwich in his hands

Figure 7: Examples of generated captions by EXTRA and the other two variants (empty and random caption). Better image captions are obtained from generating with retrieval augmentation.

a dog running on sand with a frisbee in its mouth

a dog with a toy in its mouth while running across a beach

a very big cute dog running on the beach

a dog <unk> running on beach being chased

a dog running on a beach with a toy in its mouth

EXTRA

a dog running on the beach with a ball in its mouth

CLS SEP

"EXTRA"
(empty caption)

a dog running on the beach with a leash

a street sign is hanging on a partially rusty pole

"EXTRA"
(random caption)

a brown and white dog running on sand

a conveyor belt topped of doughnuts inside of a kitchen

a conveyor belt topped with deep fried donuts

gourmet sandwiches of meat and mushrooms on fresh rolls

some doughnuts are being made on a conveyor belt

a hot dog covered in toppings sitting next to beer

EXTRA

a bunch of doughnuts that are on a table

CLS SEP

"EXTRA"
(empty caption)

a couple of plates of food on a table

roadway intersection near large brick building in city

"EXTRA"
(random caption)
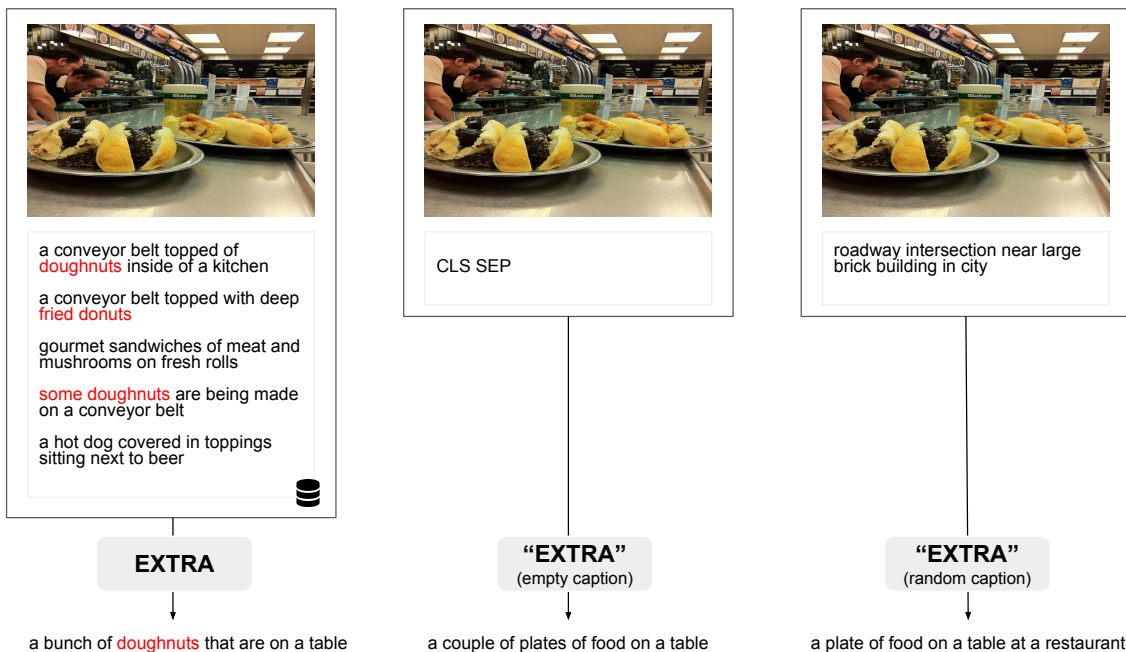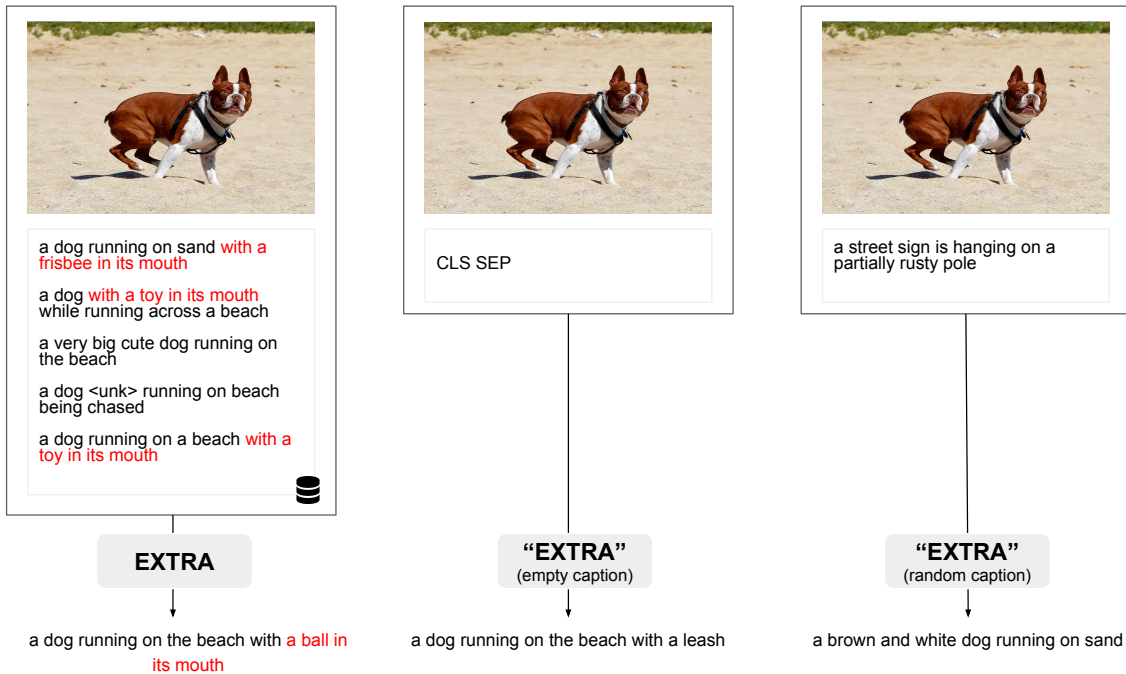
a plate of food on a table at a restaurant

Figure 8: Qualitative results in which the retrieved captions are not that related to the input image.
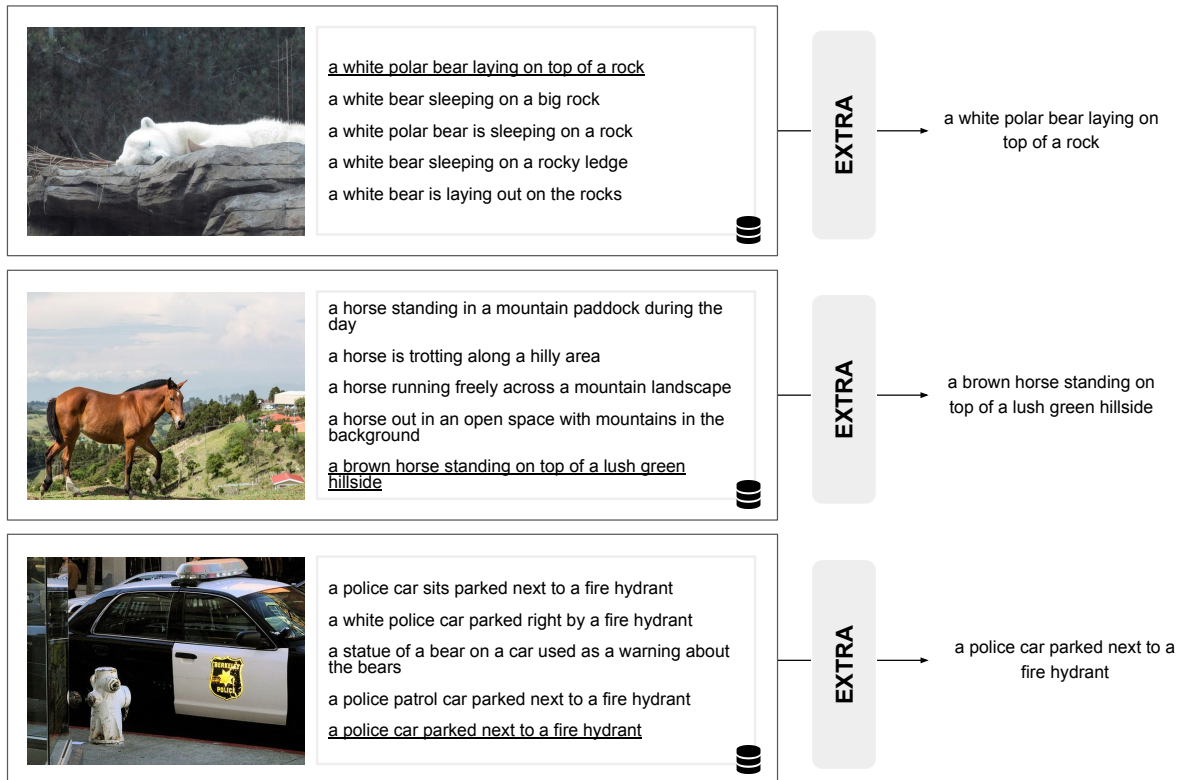
Figure 9: Examples of generated captions for which EXTRA copied from the retrieved captions.
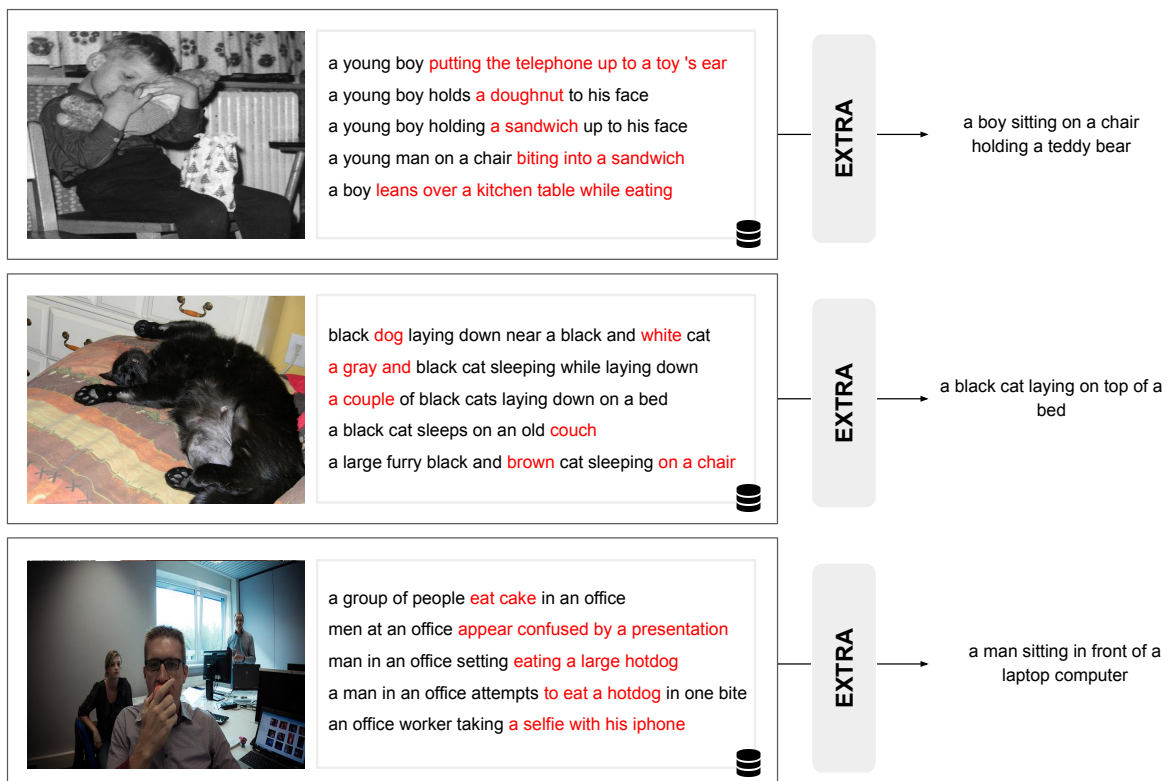

Figure 10: Examples where EXTRA is able to succeeded even with mismatches from retrieved captions.