Towards Multimodal Vision-Language Models Generating Non-generic Text

Wes Robbins

Montana State University
Bozeman, Montana 59717
wesley.robbins10@gmail.com

Abstract

While text generated by current vision-language models may be accurate and syntactically correct, it is often general. Recent work has used optical character recognition to supplement visual information with text extracted from an image. In many cases, using text in the image improves the specificity and usefulness of generated text. We contend that vision-language models can benefit from additional information extracted from an image. We modify previous multimodal frameworks to accept relevant information from a number of auxiliary classifiers. In particular, we focus on person names as an additional set of tokens and create a novel image-caption dataset to facilitate captioning with person names. The dataset, Politicians and Athletes in Captions (PAC), consists of captioned images of well-known people in context. By fine-tuning pretrained models with this dataset, we demonstrate a model that can naturally integrate facial recognition tokens into generated text by training on limited data.

Introduction

Vision-language models combine deep learning techniques from computer vision and natural language processing to assimilate visual and textual understanding. Such models demonstrate both visual and linguistic knowledge by performing tasks such as vision question answering (VQA) and image captioning. There are many applications of these tasks, including aiding the visually impaired by providing scene information and screen reading (Morris et al. 2018).

The text generated by recent vision-language models are general and overlook content that allow for richer text generation with improved contextualization. For example, they ignore clearly visible text or the presence of well-known individuals.

To improve specificity in generated text, recent work has used optical character recognition (OCR) as an additional modality to incorporate text that appears in images, significantly enhancing generated text (Hu et al. 2020).

Specific information that exists in human-level description may also come from additional sources besides OCR.

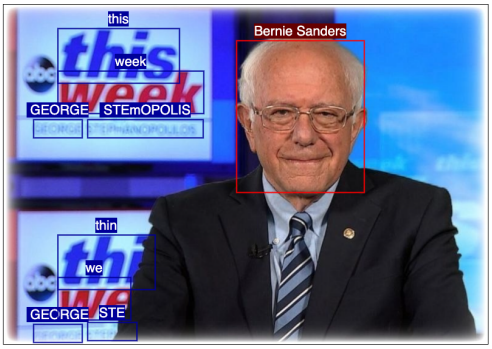

Figure 1: Our captioning model accepts tokens from upstream classifiers, learns representations for tokens, and uses each token appropriately. By using the facial recognition token 'Bernie Sanders', our model's caption is more informative than previous work which just uses OCR.[1]

Without proper nouns or other specific vocabulary, the generated text is at the risk of being awkwardly general, demonstrating a lack of shared knowledge that is expected in society. For example in Figure 1, arguably the most relevant content in the image is the presence of a well-known political figure. Consequently, a reasonable description of the image should include the name, which is 'Bernie Sanders' in this case, instead of a the generic 'a man'. This is notably absent in the caption from the previous model.

The significance of our work is that we contribute a novel method for integrating tokens from upstream classifiers into image captions. Additionally, we create a new dataset to facilitate using person names in image captions. By training on our new dataset along with other vision-language datasets and using our proposed method, we demonstrate a model that can generate informative captions. This work can be used for applications that aid the visually impaired by generating informative captions for screen reading or scene analysis.

[1]The previous model in Figure 1 is M4C Captioner (Sidorov et al. 2020) with weights from the M4C repository.

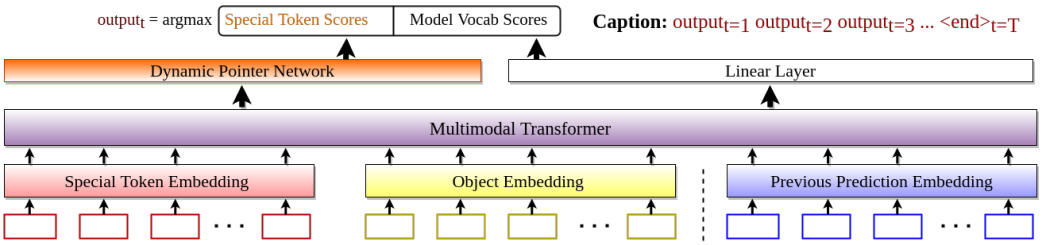

Figure 2: The architecture of the M4C + Special Tokens model.

Approach

We propose a novel method for integrating specific terms into image captions which we call the *special token approach*. A *special token* is a placeholder for extracted relevant information that is identified in an image by upstream sources. Tokens from upstream classifiers are *special* in that they often are named entities, offering unique descriptors for generic objects. For example in Figure 1, 'Bernie Sanders' is not a new object, but rather a special descriptor for an already recognized generic object (i.e. man).

In our approach, there are two modalities that hold information about an image. The first modality corresponds to generic visual features (yellow box in Figure 2) which are responsible for informing the model of general context (all vision-language models have a visual modality). The second modality, special tokens (red box in Figure 2), is responsible for informing the model of specific terms that are relevant to the image. A transformer is used to combine modalities and iteratively select words to generate a caption.

The embeddings for the first modality are visual features calculated from an object detector. The embeddings from the special token modality are calculated from visual feature vectors (Faster-RCNN and a bounding box), textual features (fasttext and PHOC), and a source feature (one-hot encoding). Special tokens are made available for direct copy into generated text which allows zero-shot inclusion of words not seen prior. We adopt token copying mechanisms from the M4C model (Hu et al. 2020), and we name our captioning model M4C+Special Tokens (M4C+ST).

The key hypothesis of our work is that a model can learn to differentiate tokens from separate upstream classifiers. The model can learn to use each token type appropriately in generated text. For example, a caption for the image in Figure 1 should neither say "A screen that says Bernie Sanders" nor should it say " 'this week' standing in front of a screen."!

Dataset

To facilitate this work, we have created the Politicians and Athletes in Captions (PAC) dataset. PAC is an image-caption dataset consisting of images of well-known individuals in context. PAC has 1,572 images and three captions per image. The non-generic terms emphasized in PAC are person names and OCR tokens. High performance on the PAC dataset requires overcoming three primary technical challenges: 1) correctly identifying people in a variety of settings, 2) reasoning about the effect of the *presence* of the individual, and 3) natural integration of a name into a generated caption.

	PAC Test Set Metrics				
Model	B-4	M	R	S	C
M4C	2.1	6.4	14.3	24.6	4.3
M4C+ST	9.1	14.8	30.4	102.6	18.7

Table 1: Between 112-334% performance increases on PAC with proposed approach—scored with five common metrics: BLEU-4, METEOR, ROUGUE, CIEDEr, and SPICE.

Results

We train our model on PAC along with TextCaps, a large scale image-caption dataset (Sidorov et al. 2020). By training the M4C+ST model on these datasets, we find that our model effectively learns to integrate tokens from both OCR and facial recognition tokens. Quantitatively, M4C+ST vastly outperforms vanilla M4C on PAC on five different metrics (see Table 1). Qualitatively, we see that M4C+ST captions are often more informative by using both person names and OCR. Figure 1 is one such example. Additionally, we used a t-SNE projection to visualize the embedding space of the special tokens. In the 2D projection, we find that facial recogonition tokens and OCR tokens are found in distinct sections of the embedding space—confirming that M4C+ST learns distinct representations for OCR and facial recognition tokens.

Future Work

We introduce the special token approach as an adaptable way to introduce non-generic information to a vision-language model. Improvements to the proposed method could include use of more external sources or integration of open-domain knowledge. Further progression in this direction could result in captions that are truly interesting, vivid, and useful.

References

Hu, R.; Singh, A.; Darrell, T.; and Rohrbach, M. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*, 9992–10002.

Morris, M. R.; Johnson, J.; Bennett, C. L.; and Cutrell, E. 2018. Rich Representations of Visual Content for Screen Reader Users. *ACM SIGCHI*, 1–11.

Sidorov, O.; Hu, R.; Rohrbach, M.; and Singh, A. 2020. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In *LNCS*, 742–758.