Towards Multimodal Vision-Language Models Generating Non-generic Text

Wes Robbins† , Zanyar Zohourianshahzadi‡, and Jugal Kalita‡

†Department of Computer Science, Montana State University
‡Department of Computer Science, University of Colorado, Colorado Springs

Background

Vision-Language models combine imformation from the visual and linguistic domains in order to perform tasks such as image captioning model or visual question answering. These tasks necessitate the understanding of important sematic information from an image and the ability to generate relevant text. These are important taks with many applications, including aiding visually impaired persons with screen reading [3]. Modern approaches are able to generate text that is both accurate syntactically correct after training on large-scale datasets. However, generated text is often overly general and ignore scene specific information such as named entities.

Non-generic terms offer high-value semantics—a known person or location can implicitly evoke rich context. To a basketball fan, 'Lebron James' (see Figure 1) can imply many concepts: 'LA Lakers', 'shooting guard', and '4x NBA Champion'. The generic alternative—'player'—does not do the same. Thus, utilizing non-generic terms increases the quality and usefulness of the generated text for applications such as screen reading.

Approach

In this work, we propose a novel method for integrating non-generic terms into image captions. We leverage auxilary classifers to identify non-generic terms. Specifically, we focus on optical character recognition (OCR) and facial recognition. We inegrate discovered non-generic tokens by 1) introducing an input modality specific to non-generic tokens and 2) appending non-generic tokens to the model vocabulary at inference time. This approach allows *learned* representations of non-generic tokens during training, and inference-time zero-shot adaptation to unseen non-generic terms (i.e a new person name not in the training set). Throughout our work, we use the term *special token* as an absraction for all token generated from auxilary classifiers.

In our approach, there are two modalities that hold information about an image: generic visual features (yellow box in Figure 3) and special tokens (red box in Figure 2). A transformer is used to combine modalities and iteratively select words to generate an image caption.

The input embeddings to the special token modality are calculated by combining visual feature vectors x^v (Faster-RCNN [4] and a bounding box), text embeddings x^t, and a source feature x^s (one-hot encoding). The embeddings x_i^{spec} ($i \in 1..N$ tokens) are calculated as follows:

$$x_i^{spec} = LN(W_1 v_i^v)) + LN(W_2 x_i^t) + LN(W_3 x_i^s). \quad (1)$$

where LN is layer normalization and $W_{1..3}$ are learned weight matrices.

All model inputs (see Figure 2) are combined with stacked multi-headed attention layers. A final bilinear calculates scores for model vocabulary words z^m and special tokens z^{st}. A caption is generated iteratively where the word selected at each time step is $y_t = argmax([z^s t; z^m])$.

The training loss is decoding binary cross entropy loss $\mathcal{L}_{dbce}$ such that the model is supervised at each decoding step t with binary cross entropy $\mathcal{L}_{bce}$.

$$\mathcal{L}_{dbce} = \sum_{t=1}^{T_{end}} \frac{\mathcal{L}_{bce}(t)}{T_{end}} \quad (2)$$

where T_{end} is the number of decoding steps before <*end*> is predicted from the vocabulary.

PAC Dataset

A difficulty in training a captioning model to use named entities is that many image-caption datasets do not include named entities. To overcome this, we collect a new dataset which we title *Politicians and Athletes in Captions (PAC)*. PAC consists 1,572 images of diverse of scenes of famous persons from around the world. The dataset consists of CC liscensed images, and three groud truth captions were labeled per image with amazon mechanical turk.

Results

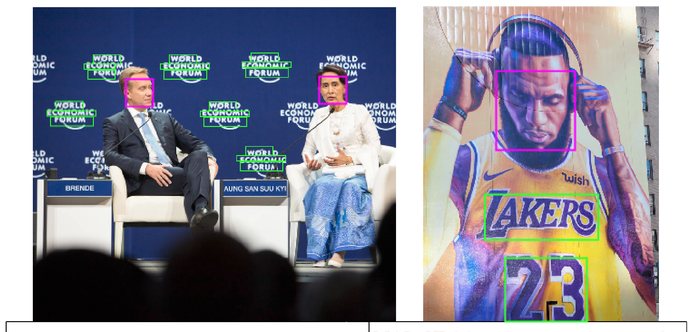
The baseline model for our work is M4C [2] which introduces a copying mechanism for captioning with OCR tokens. Our work extends this model to use special tokens, and we call our model M4C + Special Tokens (M4C+ST). We train both models on PAC (described above) and TextCaps, a large scale image-caption dataset [5]. 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 as can be seen in Table 1.

PAC Test Set Metrics

Model	BLEU-4	METEOR	ROUGUE	SPICE	CIDEr
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 caption metrics which capture how close the generated caption is to the set of ground truth captions.

Qualitatively, we observe that our model uses person names and OCR tokens appropriately throughout the captions. Figure 1 demonstrates M4C+ST appropriately switching between model vocabulary, face tokens, and OCR tokens during caption generation. In comparison, the M4C model refers to people generically (i.e 'man', 'player'), resulting in less informative captions.



M4C+ST: aung san suu kyi is at the world economic forum.	**M4C+ST:** lebron james wearing a purple jersey with the number 23
M4C: a photo of a man in front of a screen that says world world world economic forum.	**M4C:** a close up a soccer player wearing a purple shirt with the word "lakers" on it

Fig. 1: Comparison of our model (M4C+ST) against the baseline on examples from the PAC dataset. Arcface is used for facial recognition and Google Cloud OCR is used for OCR detection [1].

Model Architecture

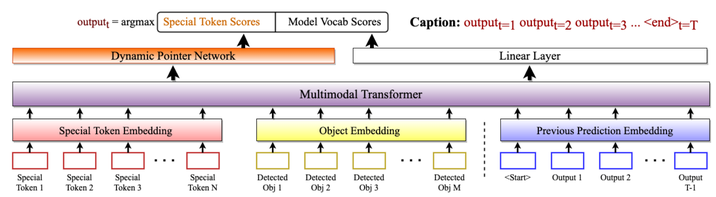


Fig. 2: Architecture diagram of proposed approach. Our primary contribution is introducing the *special token* modality, shown on bottom-left in red.

Conclusion

Text generated by vision-language models often lacks specific terms that would be present in human level descriptions or answers. We introduce the special token approach as an adaptable way to introduce non-generic information to a vision-language model. Our results suggest that our proposed method can generate descriptive captions, which could be used to improve screen reading applications used by those who are visually impaired.

Possible improvements to the proposed method include use of more external sources or integration of open-domain knowledge with special tokens. An additional path of work is adopting the propsed approach to visual question answering or visual dialogue. Further progression in this direction could result in text that is truly interesting, vivid, and useful.

Acknowledgements

The work reported in this paper is supported by the National Science Foundation under Grant No. 2050919. Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Jiankang Deng et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[2] Ronghang Hu et al. "Iterative answer prediction with pointer-augmented multimodal transformers for textvqa". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 9992–10002.

[3] Meredith Ringel Morris et al. "Rich Representations of Visual Content for Screen Reader Users". In: *ACM SIGCHI* (2018), pp. 1–11.

[4] Shaoqing Ren et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 91–99.

[5] Oleksii Sidorov et al. "Textcaps: a dataset for image captioning with reading comprehension". In: *European Conference on Computer Vision*. Springer. 2020, pp. 742–758.