Technological University Dublin
ARROW@TU Dublin

Conference papers

School of Computer Sciences

2023-02-19

Applying Positional Encoding to Enhance Vision-Language Transformers

Xuehao Liu
Technological University Dublin, xuehao.liu@tudublin.ie

Sarah Jane Delany
Technological University Dublin, sarahjane.delany@tudublin.ie

Susan McKeever
Technological University Dublin, susan.mckeever@tudublin.ie

Follow this and additional works at: https://arrow.tudublin.ie/scschcomcon

 Part of the Computer Sciences Commons

Recommended Citation

Liu, X., Delany, S. J., & McKeever, S. (2023). Applying Positional Encoding to Enhance Vision-Language Transformers. Technological University Dublin. DOI: 10.21427/DQ99-6T76

This Presentation is brought to you for free and open access by the School of Computer Sciences at ARROW@TU Dublin. It has been accepted for inclusion in Conference papers by an authorized administrator of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie, gerard.connolly@tudublin.ie.

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 4.0 License

Funder: Science Foundation Ireland

Applying Positional Encoding to Enhance Vision-Language Transformers

Xuehao Liu, Sarah Jane Delany and Susan McKeever
School of Computer Science, Technological University Dublin, Ireland
xuehao.liu@tudublin.ie
VISAPP 2023

Introduction

- Positional encoding is used in both natural language and computer vision transformers. It provides information on sequence order and relative position of input tokens (such as of words in a sentence) for higher performance.
- Unlike the pure language and vision transformers, vision-language transformers do not currently exploit positional encoding schemes to enrich input information. We show that capturing location information of visual features can help vision-language transformers improve their performance.
- We take Oscar[1], one of the state-of-the-art (SOTA) vision-language transformers as an example transformer for implanting positional encoding. We use image captioning as a downstream task to test performance.
- We added two types of positional encoding into Oscar: DETR[2] and iRPE[3].
- We used the off-the-shelf Oscar model with the same BERT self-attention backbone.

Method

- The image captioning performance of two positional encodings will determine if a 2d positional encoding applied to visual features will help vision-language transformers.
- We follow the same training process as Oscar: pre-train the model and fine-tune it to adapt to image captioning.
- We measured image captioning performance using the same metrics as originally used to evaluate Oscar.
- All of the models are pre-trained for 5 epochs with a batch size of 256. Image captioning pre-training and finetuning are then conducted for 10 epochs.

Results

Image captioning performance on the COCO validation set. The percentage in parentheses is the improvement compared to the Oscar (baseline) performance.

	Bleu4	Metor	CIDEr	Spice	Rouge_L
Oscar (baseline)	0.277	0.249	99.6	0.184	0.528
Oscar+DETR	0.318 (14.8%)	0.257 (3%)	109.6 (10%)	0.189 (2.7%)	0.546 (3.4%)
Oscar+iRPE (Q)	0.316 (14.0%)	0.252 (1.2%)	103.9 (4.3%)	0.188 (2.1%)	0.547 (3.5%)
Oscar+iRPE (QK)	**0.344** (24.1%)	0.265 (6.4%)	112.4 (12.8%)	0.197 (7%)	0.560 (6%)
Oscar+iRPE (QKV)	0.342 (23.4%)	**0.269** (8%)	**114.2** (14.6%)	**0.200** (8.6%)	**0.564** (6.8%)

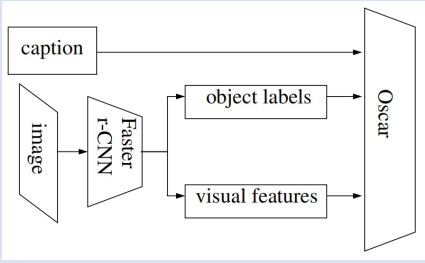

Fig 1: An overview of Oscar

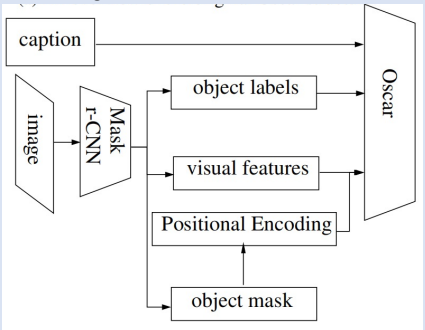

Fig 2: An overview of Oscar with positional encoding

Contribution

- We introduced positional encoding to vision-language pre-training transformers. We built the visual feature vectors with two kinds of positional encoding.
- With positional encoding, we demonstrate that with the same amount of training data, Oscar reaches a better image captioning performance compared to the original model. The Bleu4 score increased by 24.1%. The CIDEr score increased by 14.6%.
- The improvement of Oscar indicates that adding positional encoding into the vision-language transformers can enhance the performance of vision-language downstream tasks.

Discussion

- Future work can examine whether increasing to the level of epoch training (hundreds) and enlarged training sets used in Oscar's original implementation impacts on the positional encoding results.
- Other visual language tasks can be examined with positional encoding to investigate the impact.
- We implemented two common positional encoding schemes, but there are abundant choices for further examination of performance impact.

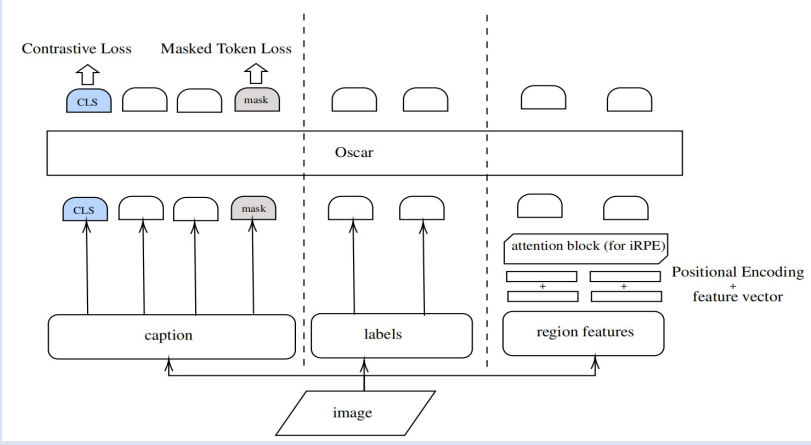

Fig 3: Illustration of Oscar with DETR/iRPE positional encoding

Acknowledgment

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183.
For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Li, Xiujun, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks." In Eur Conf on Computer Vision, pp. 121-137. Springer,, 2020.
2. Carion, Nicolas, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. "End-to-end object detection with transformers." In Eur Conf on Computer Vision, pp. 213-229. Springer, 2020.
3. Wu, Kan, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. "Rethinking and improving relative position encoding for vision transformer." In Procsof the IEEE/CVF Int Conf on Computer Vision, pp. 10033-10041. 2021.