Bangla Image Caption Generation through CNN-Transformer based Encoder-Decoder Network

Md Aminul Haque Palash | MD Abdullah Al Nasim | Sourav Saha | Faria Afrin | Raisa Mallik | Sathishkumar samiappan

Introduction

What do you see in the picture?

Well some of you might say “**A girl is waving a flag in the middle of a field** ”, Some may say “**A girl is walking and holding a flag**”

Definitely all of those captions are relevant for this image and there may be some other also. But the point i want to make is ; it’s so easy for us, as human beings, to just have a glance at a picture and describe it in an appropriate language. Even a 5 year old could do this with utmost ease.

But, can you write a computer program that takes an image as input and produces a relevant caption as output?

Motivation

We must first understand how important this problem is, to real world scenarios. Let’s see few applications where a solution to this problem can be very useful:

Automatic Caption can help, make **Google Image Search** as good as Google search, as then every image could be first converted into a caption and then search can be performed based on caption.

Self driving cars: Automatic driving is one of the biggest challenges and if we can properly caption the scene around the car, if can give a boost to the self driving system.

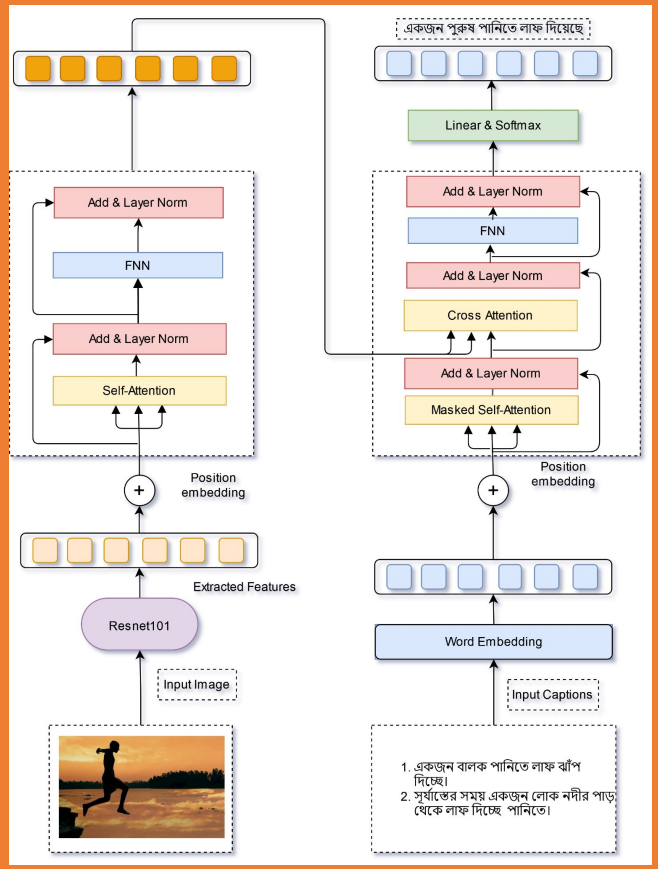
PREVIOUS WORK

- Previously many methods have been proposed for Bangla image captioning, most of them are based on Convolutional neural Network and Recurrent neural Network.

- Long Short-Term Memory (LSTM) or RNN models are sequential and need to be processed in order. That is not very efficient in handling long sequences.

- The model tends to forget the contents of the distant position or, in some cases, mixes the contents of adjacent positions: the more the steps, the more challenging for the recurrent network to make decisions.

our architecture



Why Transformer ?

Transformer were introduced in the context of machine translation with the purpose to avoid recursion in order to allow parallel computation (to reduce training time) and also to reduce drop in performances due to long dependencies. The main characteristics are:

- **Non sequential**: sentences are processed as a whole rather than word by word.
- **Self Attention**: this is the newly introduced 'unit' used to compute similarity scores between words in a sentence.
- **Positional embeddings**: another innovation introduced to replace recurrence. The idea is to use fixed or learned weights which encode information related to a specific position of a token in a sentence.

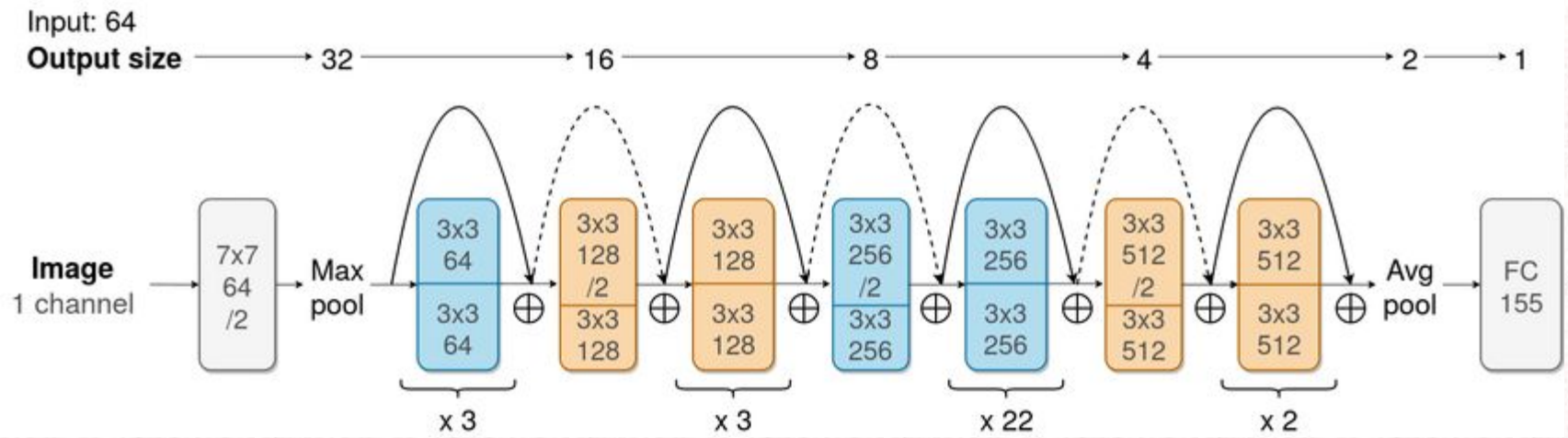
OUR PROPOSED MODEL

- A captioning model reels on three main component a CNN layer, Encoder layer and a Decoder layer.

- Our captioning model is all about merging the three to combine their most powerful attributes i.e.

  - **Convolutional Neural Networks** extract features from image
  - **Transformer Encoder layer** encodes spatial information from features extracted from CNN layer
  - **Transformer Decoder layer** work well with any kind of sequential data, such as generating a sequence of words

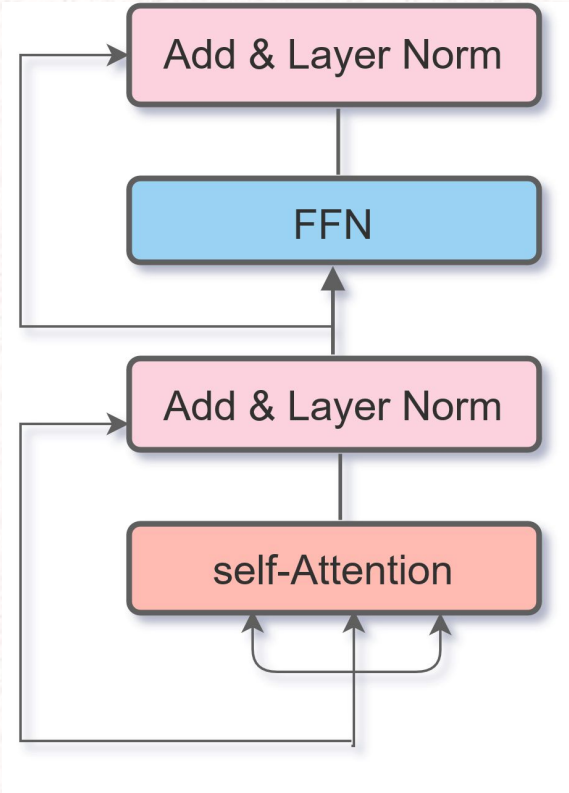
CNN (ResNet-101)



Encoder layer

Our encoder layer consists of three layer

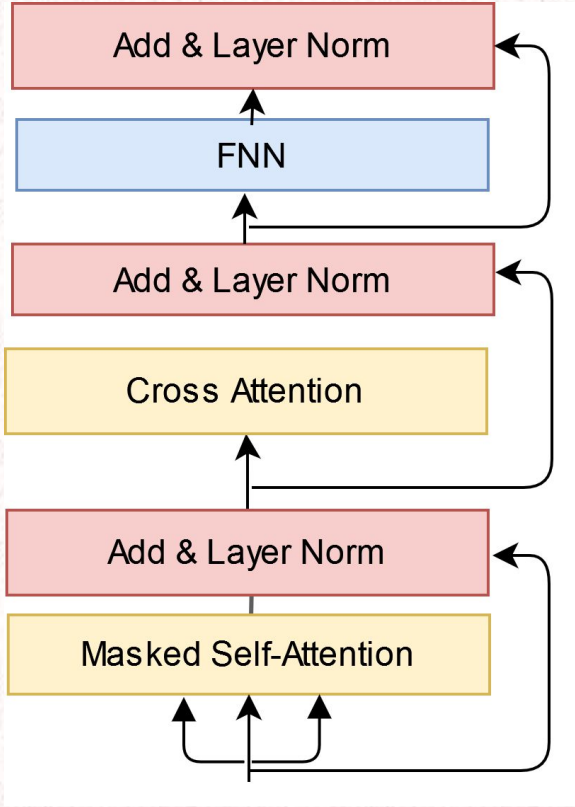
- Self-Attention layer
- Add & layer Norm
- FFN(feed-forward network)



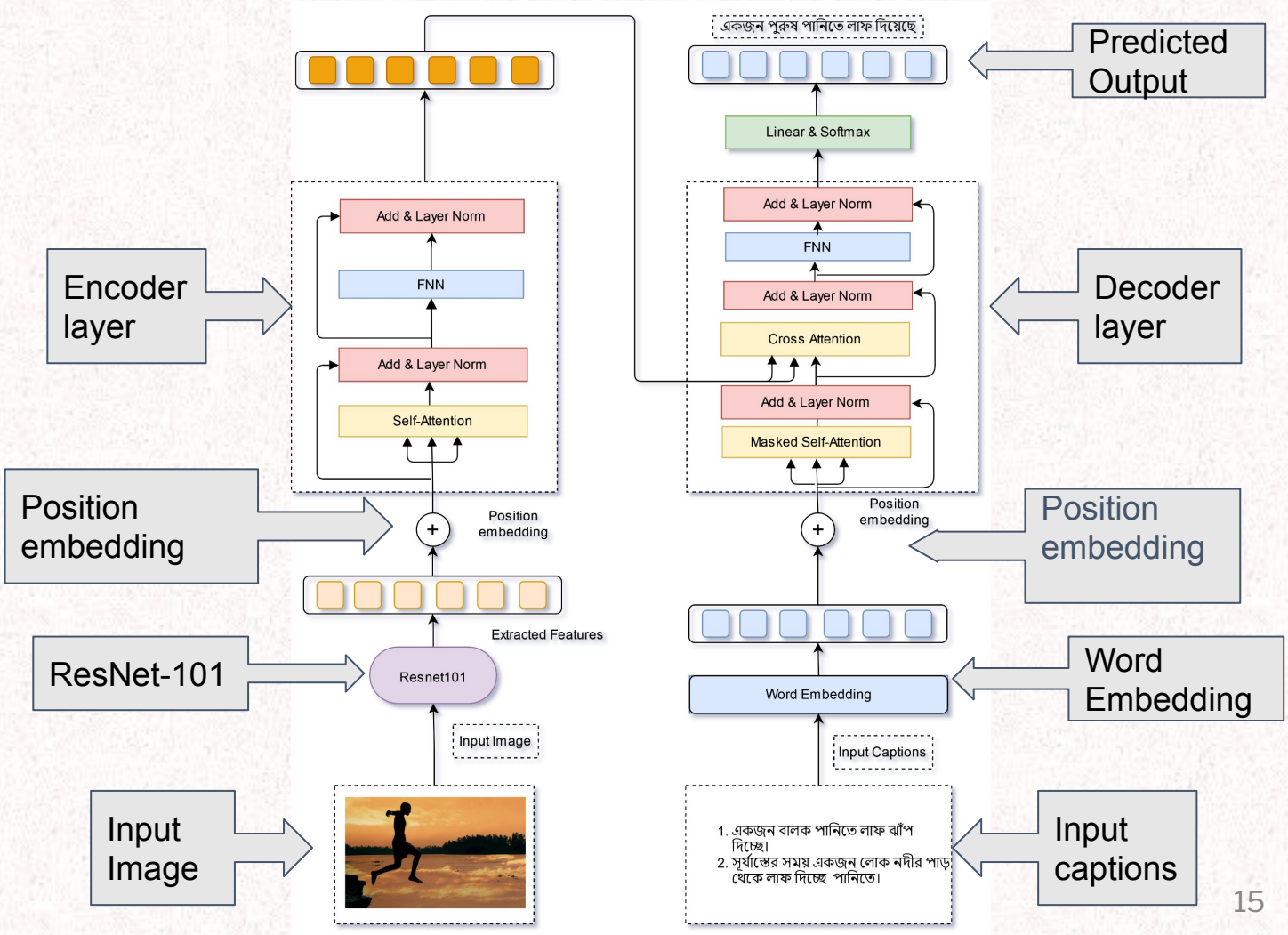
Decoder layer

Our Decoder layer consists of four layer

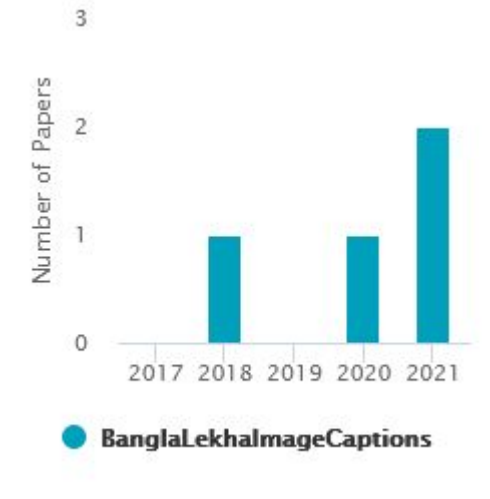
- Masked self-attention
- Add & Layer Norm
- Cross Attention
- FNN (feedforward neural network)



Full architecture



DATA COLLECTION



There are many open-source datasets are available. Though our motive is to generate bangla caption from image. Therefore,

- We have tested and trained our model on the **BanglaLekhaImageCaptions** dataset.
- There are 9,154 images in the dataset presented are collected from the public domain.
- It is relevant to Bengali culture

SOME EXAMPLES

Caption 1: সমুদ্রের দিকে একজন মানুষ হেঁটে যাচ্ছে ।

Caption 2: একজন মানুষ সমুদ্র সৈকত দিয়ে হাঁটছে আর সূর্য অস্ত যাচ্ছে।

Caption 1: চারটি মূর্তি আছে। উপরে একজন মানুষ দাড়িয়ে আছে।

Caption 2: একটি ছেলে ৪ টি মূর্তির চোখে কাল কাপড় বাধে দিচ্ছে।

Caption 1: সামনে কয়েকজন নারী ও পিছনে অনেকগুলো পুরুষ দাড়িয়ে আছে।

Caption 2: 'লাল সাদা শারি পরে ৬ টি মেয়ে দারিয়ে আছে আর তাদের পেছনে আরও মানুষ হেঁটে আসছে।

Caption 1: দুইজন বালক পানি থেকে ফুল তুলছে।

Caption 2: একটি বাচ্চা থালি গায়ে বিল থেকে শাপলা তুলছে।

PERFORMANCE

Quantitative Analysis

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Our proposed model	0.694	0.580	0.505	2.22e-308	0.337
CNN + ResNet-50 [merged]	0.651	0.426	0.278	0.175	0.297
CNN + LSTM [mixture	0.632	0.414	0.269	0.168	0.291
CNN + Bi-LSTM [inject]	0.619	0.403	0.261	0.163	0.296

- You can see our proposed model achieved better results than the traditional architecture.

Original caption 1(Bangla): দুইজন মানুষ সাইকেল চালাচ্ছে
(Two people are cycling.)
Original caption 2(Bangla): দুইজন ছেলে দুইটা সাইকেল চালিয়ে রাস্তা দিয়ে যাচ্ছে।
(Two boys are riding two bicycles on the road.)
Predicted caption(Bangla): দুইজন পুরুষ সাইকেল চালিয়ে যাচ্ছে।
(Two men are riding bicycles.)

BLEU 1: 0.882
BLEU 2: 0.833
BLEU 3: 0.795
BLEU 4: 1.508e-231

Original caption 1(Bangla): বৃষ্টিতে একজন পুরুষ রিক্সা টানছে।
(A man pulls a rickshaw in the rain.)
Original caption 2(Bangla): বৃষ্টির মাঝে পুরুষ যাত্রী নিয়ে রিক্সা চালিয়ে যাচ্ছেন একজন পুরুষ।
(A man is driving a rickshaw with male a passenger in the rain.)
Predicted caption(Bangla): একজন পুরুষ রিক্সা চালিয়ে যাচ্ছেন।
(A man is driving a rickshaw.)

BLEU 1: 0.969
BLEU 2: 0.906
BLEU 3: 0.870
BLEU 4: 2.225e-308

Original caption 1(Bangla): পানিতে একটি শিশু আছে।
(There is a baby in the water.)
Original caption 2(Bangla): লাল-সবুজ শাড়ি পরে হাতে একটি ফুলের ঝুড়ি নিয়ে একটি বাচ্চা মেয়ে পানিতে বসে হাত ভেজাচ্ছে।
(A baby girl in a red-green sari with a flower basket in her hand is sitting in the water and getting her hands wet.)
Predicted caption(Bangla): একজন ছোট মেয়ে পানিতে দাড়িয়ে পানি ছিটা দিচ্ছেন।
(A little girl is standing in the water and sprinkling water.)

BLEU 1: 0.755
BLEU 2: 0.395
BLEU 3: 0.215
BLEU 4: 2.225e-308

Original caption 1(Bangla): একজন রিক্সাচালক একটা দেয়াল লিখন পড়তে পড়তে যাচ্ছে।
(A rickshaw puller is going and reading graffiti.)
Original caption 2(Bangla): একজন পুরুষ আছে। দেয়ালে ছবি আঁকা আছে।
(There is a man. There are pictures on the wall.)
Predicted caption(Bangla): একজন পুরুষ দাড়িয়ে দেয়ালে চিত্রকর্ম দেখছে।
(A man is standing and looking at the paintings on the wall.)

BLEU 1: 0.664
BLEU 2: 0.382
BLEU 3: 0.282
BLEU 4: 2.225e-308

Future Plan

Forward plan.....

- Develop a fully transformer architechture
- Increase our data by combining BanglaLikhaImageCaption dataset with MScoco dataset

Conclusion

Summary

- A brand new CNN-Transformer base architechture
- striking performance

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)

2. Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B.: Automatic description generation from images: A survey of models, datasets, and evaluation measures. Journal of Artificial Intelligence Research 55, 409–442(2016)

3. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12299–12310 (2021)

4. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)

References

5. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10578–10587 (2020)

6. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation. pp. 376–380 (2014)

7. Ghosal, P., Nandanwar, L., Kanchan, S., Bhadra, A., Chakraborty, J., Nandi, D.: Brain tumor classification using resnet-101 based squeeze and excitation deep neural network. In: 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP). pp. 1–6. IEEE (2019)

8. Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image captioning: Transforming objects into words. arXiv preprint arXiv:1906.05963 (2019)

References

9. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research 47, 853–899 (2013)

10. Khan, M.F., Sadiq-Ur-Rahman, S., Islam, M.S.: Improved bengali image captioning via deep convolutional neural network based encoder-decoder model. In: Proceedings of International Joint Conference on Advances in Computational Intelligence. pp. 217–229. Springer (2021)

11. Liu, W., Chen, S., Guo, L., Zhu, X., Liu, J.: Cptr: Full transformer network for image captioning. arXiv preprint arXiv:2101.10804 (2021)

12. Mansoor, N., Kamal, A.H., Mohammed, N., Momen, S., Rahman, M.M.: Banglalekhaimage-captions (2019), Mendeley Data, V2, doi: 10.17632/rxxch9vw59.2

13. Naim, F.A.: Bangla handwritten digit recognition based on different pixel matrices. In: International Conference on Innovative Computing and Communications. pp. 325–341. Springer (2022)

References

14. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
15. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
16. Rahman, M., Mohammed, N., Mansoor, N., Momen, S.: Chittron: An automatic bangla image captioning system. Procedia Computer Science 154, 636–642 (2019)
17. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using amazon's mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. pp. 139–147 (2010)
18. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7008–7024 (2017)
19. Saifullah, K.: neuropark/sahajbert (2021), https://github.com/khalidsaifullaah/BERTify
20. Tanti, M., Gatt, A., Camilleri, K.P.: What is the role of recurrent neural networks (rnns) in an image caption generator? arXiv preprint arXiv:1708.02043 (2017)

References

21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)

22. Wang, H., Zhang, Y., Yu, X.: An overview of image caption generation methods. Computational intelligence and neuroscience 2020 (2020)

23. Wu, X., Jackson, D.J., Chen, H.C.: Novel fractal image-encoding algorithm based on a full-binary-tree searchless iterated function system. Optical Engineering 44(10), 107002 (2005)

Thank You!

stay safe and healthy