CS 5316: Natural Language Processing

End to End Image Captioning

Project Report

Group 6

Hisan Naeem - 23100051
Samee Arif - 23100088

1. Introduction

Image Captioning is the process of generating textual description of an image which captures the key objects, attributes, and relationships. It uses both Natural Language Processing and Computer Vision to generate the captions. It can be used to improve assistive technology and aid visually impaired to comprehend their environment. The image captioning model can be piped with a text to speech synthesizer so the output can be read aloud to the users and help them to interact with it better. In the health sector it can help in analysis and interpretation of medical images by producing meaningful descriptions and thus help medical professionals to better diagnose the health conditions. Image captioning can also be very useful in automated video analysis, image search and retrieval and improvement in human-computer interaction.

A deep understanding of both visual and linguistic information is required to accurately generate meaningful image captions. Recent advances in deep learning and the availability of large-scale image captioning datasets have led to significant progress in this area. A encoder-decoder architecture is used for this task. Convolutional Neural Networks (CNN) is used as an encoder and is used to extract visual features from the input image and these features are fed into a language model which is a decoder to generate captions. The language model can be Recurrent Neural Networks (RNNs) based or Transformer based. Attention Mechanism can be applied to different parts of the image when generating words, this can improve the detail of the caption.

PureT[1] introduced a fully attentive end to end architecture using Swin-Transformer[2]. In this work, expansion layers are used, which consists of leveraging different sequence lengths compared to the one provided in the input to further improve on this new concept while preserving the efficiency of the original formulation. It also devised a training technique that is both faster and more efficient than the standard one. We used the MS-COCO 2014 test set and nocaps validation set to evaluate our model.

2. Related Work

“BabyTalk”[3] is a Machine Learning image captioning model that makes use of content planning and surface realization. Content planning is a linear SVM trained on low-level image features and surface realization is a language processing task that uses Conditional Random Field (CRF). Yang et al.[4] used a

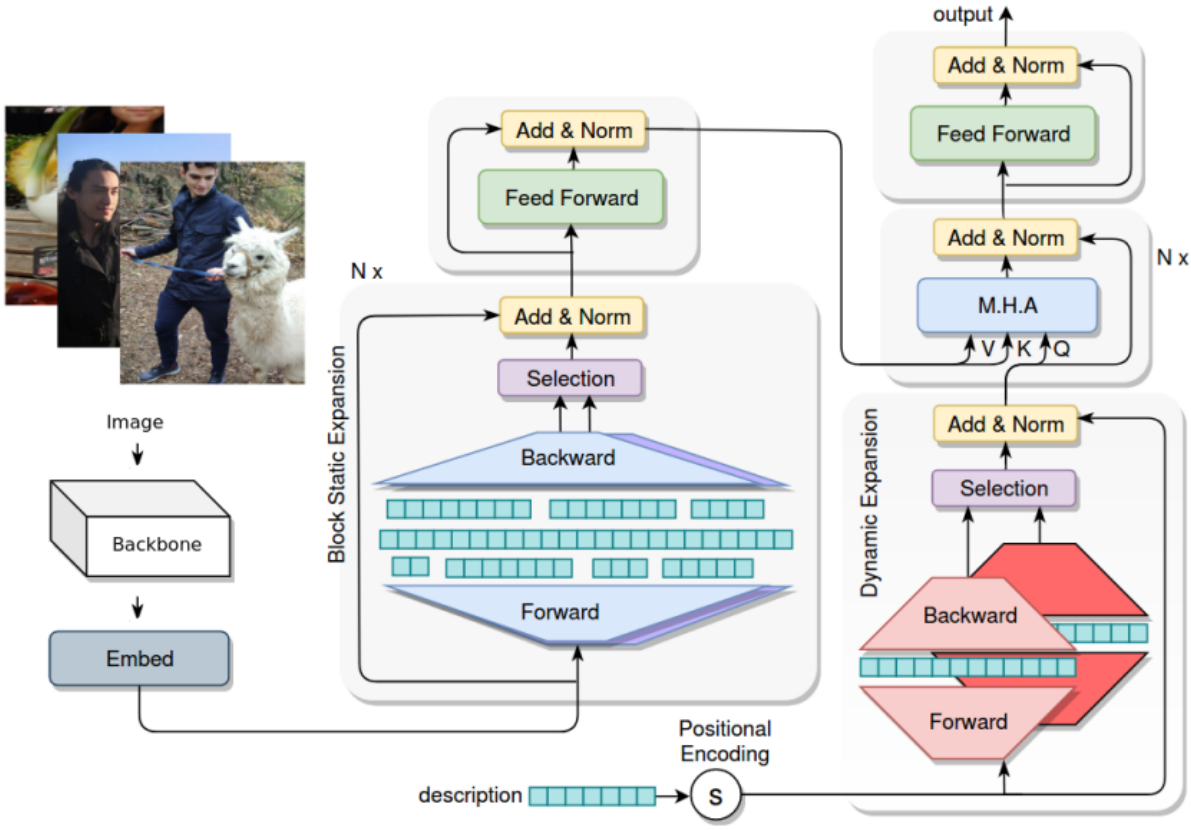
GIST-based scene descriptor that was train on SVM to detect scenes. A language model based on the Hidden Markov Model (HMM) was then used to produce the caption string. Elliott and de Vries[5] used a Regional Convolutional Neural Network (RCNN) to detect the objects that are then fed into a Visual Dependency Representation (VDR) parsing model that predicts the spatial relationship between different objects. A language model then leverages VDR and visual context information to generate a possible set of descriptions for the image.

Socher et al.[6] developed a Dependency Tree-Recurrent Neural Network (DTRNN) for capturing compositional semantic meanings. The model employed a CNN to extract visual attributes and maps each word to a d-dimensional vector. A deep learning based max-margin objective function was used to learn the correlations between text and images. Xu et al.[7] proposed an image captioning model that uses a visual attention mechanism to selectively focus on different regions of an image while generating a caption. The model was based on an encoder-decoder architecture, where the encoder is a convolutional neural network that processes the input image and the decoder is a recurrent neural network that generates the caption word-by-word. BLIP-2[8] is a very recent model that bootstraps from frozen pre-trained unimodal models and uses the self-attention mechanism.

3. Methodology

3.1 Architecture

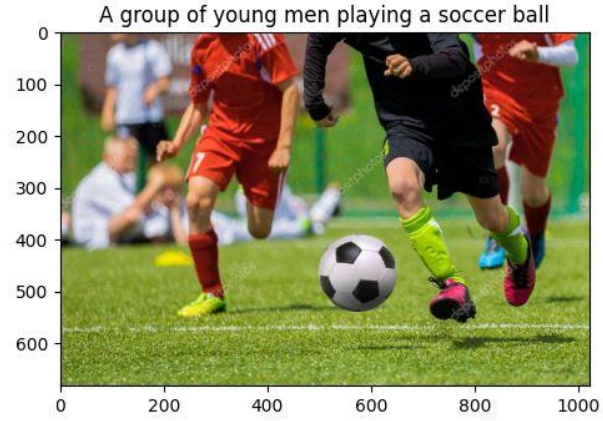
We have used an encoder-decoder model on top of the Swin-Transformer. If we feed an image A to the backbone of the transformer we get an initial set of processed visual features. This result then goes into the encoder that uses static expansion (feedforward blocks, skip connection and pre-layer normalization). The decoder is dynamic expansion followed by cross-attention followed by feedforward block, and skip connection and normalization is applied on each component.

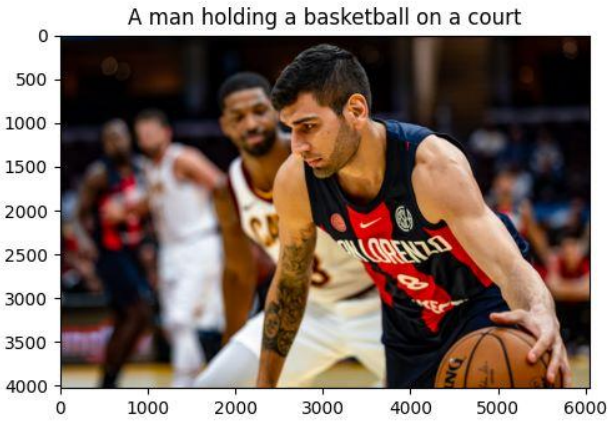


Expansion principle consists of transforming the input sequence into another one featuring a different length by means of a “Forward expansion” and retrieving the original length back in the complementary backward operation. There are two versions: Static Expansion and Dynamic Expansion. The basic idea behind dynamic expansion is to generate captions that capture different aspects of the input image. By generating multiple diverse captions, dynamic expansion can improve the overall quality and richness of the captions.

3.2 Training

We trained the model on Microsoft COCO benchmark[9] which was split into 113,287 training images and 5000 validation images and 5000 testing cases. Lowering casing, punctuation removal and filtering out words that do not occur at least 5 times, such techniques are used to pre-process the groundtruth string. We used Novel Object Captioning at Scale dataset (nocaps)[10] validation set to evaluate the model. We got the following results:





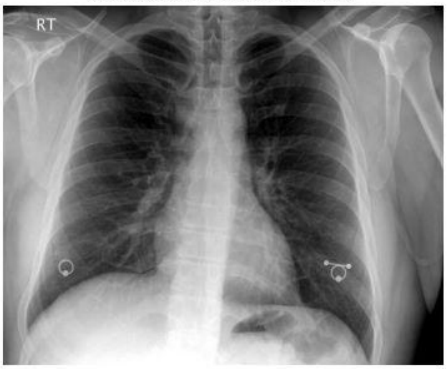
3.2 Fine-tune

We also tried to fine tune the model on Indiana University - Chest X-Rays dataset. However the result was not up to the mark. The dataset was:

cardiomegaly without lung infiltrates

no acute pulmonary disease

negative for acute abnormality



postoperative left upper lobe no visible active cardiopulmonary disease

4. Results

4.1 Results on COCO Benchmark

B1	B2	B3	B4	Meteor	Rouge-L	CIDEr-D
96.9	92.6	85.0	75.3	40.1	76.4	140.8

5. Conclusion

In conclusion, we have proposed a novel image captioning model that achieves state-of-the-art performance on several benchmark datasets, including the widely used Microsoft COCO dataset. Through extensive experiments and ablation studies, we have demonstrated the effectiveness of our proposed model and its components. We have also shown that our model is able to generate diverse and descriptive captions, thanks to the use of dynamic expansion techniques. However the model fails to perform well on the medical images when fine tuned on the medical image captioning dataset.

6. Future Work

We aim to use a larger medical dataset and conduct experiments to see how the model performs. Furthermore we aim to pipe a text-to-speech synthesizer and image captioning model to increase information accessibility. Another thing we aim to do is to use GPT4 to create a large Urdu image captioning dataset and pre-train the model on this generated dataset and then fine tune it on a human generated dataset.

7. References

[1] Yiyu Wang, Jungang Xu, and Yingfei Sun. “End to-End Transformer Based Model for Image Captioning”. In: arXiv preprint arXiv:2203.15350 (2022).

[2] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021, pp. 10012–10022

[3] Girish K, Visruth P, Vicente O, Sagnik D, Siming L, Yejin C, Berg Alexander C, Berg Tamara L (2013) “Babytalk: understanding and generating simple image descriptions”. IEEE Trans Pattern Anal Mach Intell 35(12):2891–2903

[4] Yang Y, Teo C, Daumé H, Aloimonos Y (2011) Corpus-guided sentence generation of natural images. In: Proceedings of the 2011 conference on empirical methods in natural language processing, pp 444–454

[5] Elliott D, de Vries A (2015) Describing images using inferred visual dependency representations. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (vol 1: Long Papers), pp 42–52

[6] Socher R, Karpathy A, Le QV, Manning CD, Andrew YN (2014) Grounded compositional semantics for finding and describing images with sentences. Trans Assoc Comput Linguist 2:207–218

[7] Kelvin Xu and Jimmy Ba and Ryan Kiros and Kyunghyun Cho and Aaron Courville and Ruslan Salakhutdinov and Richard Zemel and Yoshua Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

[8] Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." arXiv preprint arXiv:2301.12597 (2023).

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. “Microsoft coco: Common objects in context”. In: European conference on computer vision. Springer. 2014, pp. 740–755

[10] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. “Nocaps: Novel object captioning at scale”. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, pp. 8948– 8957.