International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 10 | Oct 2022 www.irjet.net p-ISSN: 2395-0072

IMAGE CAPTIONING USING TRANSFORMER: VISIONAID

Ishaan Shivhare[1], Joy Purohit[2], Vinay Jogani[3], Prof. Pramila M. Chawan[4]

[1,2,3] B. Tech Student, Dept. of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India
[5]Associate Professor, Dept. of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India

---

Abstract - *Image captioning is the fascinating task of generating concise and accurate descriptions of a given image. Common approaches to tackling this challenge contain various shortcomings, such as a lack of diversity in training data, lengthy training time, lack of fine-grained information, and internal covariate shifts. These are just some of the issues being faced by researchers in the field today. In addition, many recent implementations do not consider the geometric positioning of the objects in the image [6], thus missing out on critical geometrical relationships between the concerned objects. To understand the domain and the aforementioned issues more deeply, a literature review has been carried out in this paper, followed by the proposal of an image captioning model "VisionAid" that integrates solutions to many of these problems.*

Key Words: **Image captioning, transformers, sequence to sequence, attention, enoder, decoder.**

1. INTRODUCTION:

Image captioning is a critical task in the field of artificial intelligence which deals with the automatic generation of textual descriptions of an image. It leverages techniques from the field of Computer Vision and Natural Language Processing to extract features from an image to generate corresponding sequentially descriptive textual explanations. These generated captions give an insightful explanation of the image by describing the objects, their properties, pairwise relationships between these distinct objects and the actions taking place in the image. The task of image captioning has several uses such as aiding visually impaired people, image indexing, social media analysis, and product recommendations. The applications are endless. There are a variety of techniques that researchers have used to create image captioning models. Inspired by the success of the sequence-to-sequence models in the domain of machine translation [21,29], encoder-decoder architecture remains a popular approach to tackling the task of image captioning. Initially, research work such as [23], [24] adopted the use of a pre-trained Convolutional Neural Network (CNN) as an encoder and a Recurrent Neural Network (RNN) as a decoder. In this model, an image is given as an input to the encoder, where a Convolutional Neural Network is used to extract visual information from the image. This information is fed to the decoder, which is a Recurrent Neural Network. The decoder is responsible for language modeling up to the word level. The RNN takes the encoded word and previously vectorized word embeddings as inputs to predict the next word of the sentence, thus generating the caption.

Further improving upon the previously described architecture, [25] demonstrated that the adoption of the Faster R-CNN to extract region-level features is overwhelmingly advantageous. Most researchers followed suit and grid-level features extracted by CNNs were no longer in use. However, this technique of using a Faster-RCNN as the encoder of the object detector is not faultless, and some of the problems it has are as follows. Firstly, since region-level features may not necessarily cover the entire image, there will likely be a lack of fine-grained information [17]. Secondly, this type of object detector needs an extra dataset for pre-training (Visual Genome Dataset [26]) due to the time-consuming nature of region-level feature extraction. This leads to an overall increase in difficulty in training the image captioning model, thereby limiting potential applications.

For the decoder, the use of a long short-term memory (LSTM) RNN has become a common approach. However, LSTM decoders have shortcomings in training efficiency and expression ability. The complex addressing and overwriting mechanism along with inherent sequential processing problems lead to challenges during training. This can be easily understood by considering the following example. Consider an LSTM using the hidden state h_t to memorize historical information. Therefore, to generate the current hidden state, it needs the previously hidden state h_{t-1} as an input. By design, this mechanism functions to make a good relationship across time. Unfortunately, it leads to the sequence training problem. The training time will increase with the increment of the sequence length, which inherently influences the training in parallel. The attention mechanism has become a commonly used approach to deal with the sequence-to-sequence problem by memorizing the relation between the input and output.

Used in an RNN network, this mechanism has yielded excellent performance, helping the model attend to salient objects in the image, as demonstrated in the Soft-Attention model[24].

Recently, self-attention networks (SAN) have been introduced as a replacement for conventional RNNs in image captioning. For instance, the Transformer[7] model contains a stacked attention mechanism instead of recurrence, which can draw global dependencies between the input and the output. The mechanism in the standard transformer model consists of self-attention modules and multi-head attention modules. The former can correlate different positions to compute a representation for the whole sequence, while the latter correlates different multi-modal representations, establishing contact with the image and text. This form of self-attention and its variants that followed have shown state-of-the-art performance in image captioning, however, there are two problems with vanilla self-attention (henceforth denoted SA). Firstly, SA is susceptible to the "Internal Covariate Shift" [28] problem, i.e the tendency that the distribution of activations drifts during training in a feed-forward network. Secondly, SA is unable to model the geometric relationships between input elements. Vanilla SA treats its inputs as a "bag of features", neglecting their inherent geometric structure, which can play a critical role in understanding image content.

Another issue that has arisen as models for image captioning become more and more complex, is the difficulty in obtaining sufficiently large quantities of labeled data to achieve high-quality training of these models. Even though commonly used image captioning datasets have a relatively large number of training examples, they still lack different descriptions of each image. Many of these captions are very similar, leading to a lack of diversity in the captions generated by the models trained on these datasets.

Hence, to remedy the issues mentioned above in the usage of a pre-trained CNN as the encoder and an RNN as the decoder, usage of LSTM as the decoder, employment of vanilla self-attention networks and the lack of diversity in captions of training data, we conduct a literature review of different transformer models to gain a more thorough understanding of the different mechanisms and modules that can be incorporated into our proposed solution. Following that, we propose a transformer model "VisionAid" which integrates solutions to the issues mentioned previously. In our model, firstly, using BERT [22] word embeddings we augment the diversity of captions. Secondly, we use Swin-Transformers [13] to extract grid-level features from the input image, which is the initial input vector to the encoder. Thirdly, we replace the vanilla self-attention module with a combination of Normalized self-attention (NSA) and Geometric-Aware self-attention (GSA) module to eliminate the "Internal Covariate Shift problem" [28] and to regard pairwise geometric relationships between the elements in the image.

The main contributions of our paper can be stated as follows:

1. We conducted research in the domain of image captioning, and reviewed literature across the field, analyzing their approaches and contributions.
2. We proposed a model for image captioning "VisionAid" keeping in mind the problems we set out to solve and the insights gained from the research conducted.

2. LITERATURE REVIEW:

Performed below is a literature review of the image captioning domain, to understand the different mechanisms and modules that can be incorporated into our proposed model.

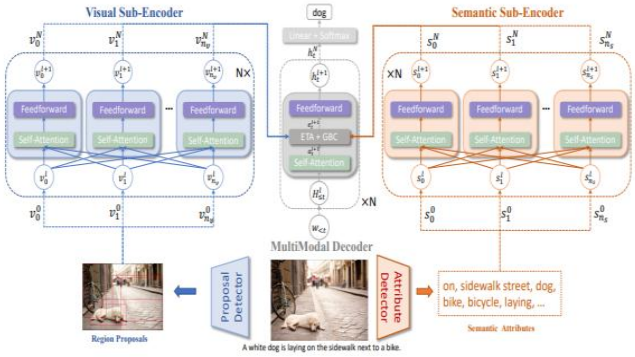

Fig -1: ETA Transformer [1]

[1] proposes an encoder-decoder architecture. [1] consists of attention [7] and feed-forward layers. The encoder is divided into two sub-encoders, each responsible for extracting visual and semantic features respectively through region proposals and semantic attributes. [1] adopts a multimodal decoder to regard both visual and semantic information simultaneously and generate captions for an image. A novel entangled attention (ETA) module is integrated which enables attention on both the visual and semantic features simultaneously guiding each

other. Additionally, a Gated Bilateral Controller (GBC) guides the decoder to integrate the semantic and visual features. These novel modules can be integrated into the proposed system to combine semantic and visual information better.

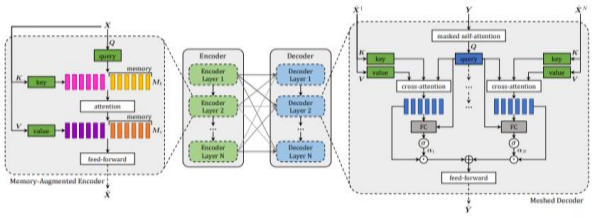

Fig -2: Meshed-Memory Transformer [2]

Similar to [1], [2] is an encoder-decoder architecture with attention [7] and feed-forward layers. However, the encoder in [2] uses additional memory-augmented attention through memory vectors leveraging the prior knowledge learnt. Each encoder layer is connected to every decoder layer forming a mesh structure enabling the decoder to leverage multi-scale information, and each decoder layer uses cross-attention with all the encoding layers. The decoder first receives a sequence of vectorized words and contextual embeddings from the encoder as input, then each decoder layer uses a masked self-attention operator to consider only previous words for the prediction of the next word. Multiple decoders are stacked together to refine the prediction of the captions for the image.

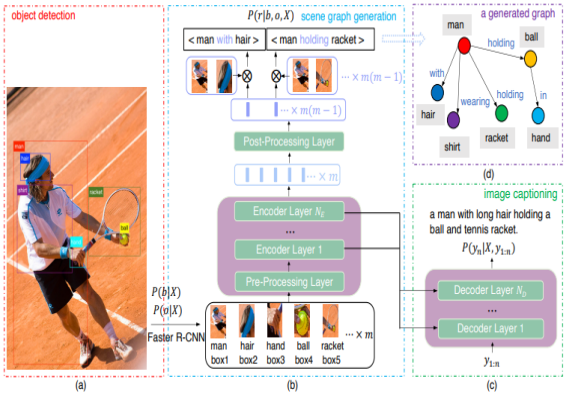

Fig -3: Relational Transformer [4]

[4] proposes a transformer model which generates captions with the objective of scene generation to elaborate pairwise relationships between words, this allows the model to extract relational features between the objects helping it to generate finer image captions. The encoder consists of a Pre-Processing layer which takes bounding boxes, labels and image features as the input generated by [6], encoding both object and spatial information. Further, a stack of encoder layers [7] using multi-head self-attention generates vectors representing the contextual meaning of the object by exploiting pairwise relationships between the objects. Lastly, a Post-Processing Layer pairs "n" object features to all possible "n(n-1)" object features for the graph generation. The decoder layers consider a weighted sum of all the outputs of each encoder layer and word embeddings of previously computed words to determine the image captions.

[5] aims to improve the self-attention (SA) module [7] by integrating a Normalization Self-Attention (NSA) and Geometry-aware Self-Attention module (GSA). The NSA helps the training performance by additionally performing normalization on the hidden units of SA. Further, the geometry of an object helps in representing visual characteristics and information. Hence, the GSA regards pairwise geometric relationships between things to improve attention performance.

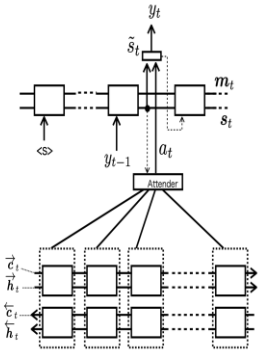

Fig -4: The encoder-decoder architecture used in [9]

[9] proposes methods for keeping a check on the output sequence length for neural encoder-decoder models. Four methods have been introduced. The first method, "Beam Search without EOS Tags" is an encoder similar to NSS methods. The second method, "Discarding Out-of-range Sequences" works on the principle of discarding out-of-range sequences and, unlike the EOS tag, it allows one to decide when to stop generation. The third method, "Length Embedding as Additional Input for the LSTM" is an encoder mainly specified to keep a check on the length of the output sequence and inputs the remaining length l_t to the decoder at each step of the decoding process. Lastly, the method "Length-based Memory Cell Initialization" inputs the desired length once at the initial state of the decoder.

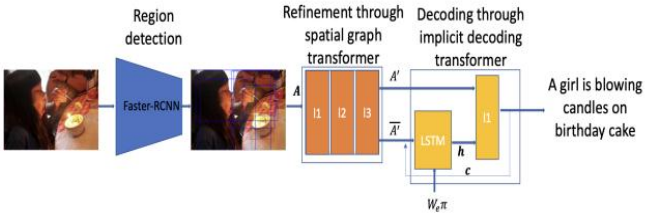

Fig -5: Image Transformer[10]

Image Transformer[10] proposes a novel internal architecture for the transformer layer, better adapted to the task of image captioning than the standard transformer layer model [7]. The original transformer layer[7] consists of a stack of multi-head dot product-based transformer refining layers and a decoding part that is also a stack of refining layers, taking the output of the encoding layers as well as the embedded features of the previous predicted word. In alternative to [7], [10] proposes to use a spatial graph encoder transformer in the encoding part which considers three common categories of spatial relationships parent, neighbor and child for each query region in a graph structure. The transformer layer uses a dot product attention mechanism to infer the most relevant areas of the image. The decoding part, [10] consists of an LSTM[11] layer and an implicit transformer decoding layer. The novelty introduced by the spatial graph encoder transformer is something that can be used to improve the performance of the proposed system.

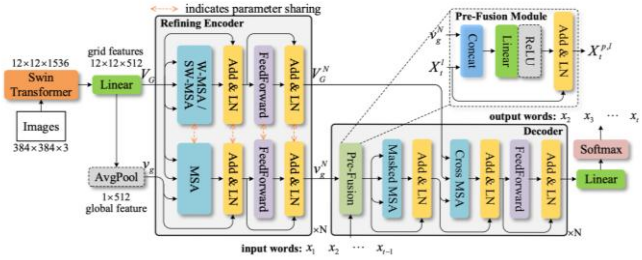

Fig -6: PureT [12]

PureT[12] proposes a pure Transformer based image captioning model. This model does not use a Faster R-CNN[6] as the backbone encoder. Instead, a SwinTransformer[13] is used for the feature extraction of grid-level features from the images, and the average pooling is computed as the initial image global feature. A refining encoder is constructed similar to [14] using Shifted Window MSA by SwinTransformer[13] for the refinement of initial grid features and the global feature. The refining decoder directly adopts the Transformer Decoder[7] for the generation of captions. In addition, the word embedding vector is pre-fused with the global image feature from the Encoder before the MSA to increase the inter-model feature interactions.

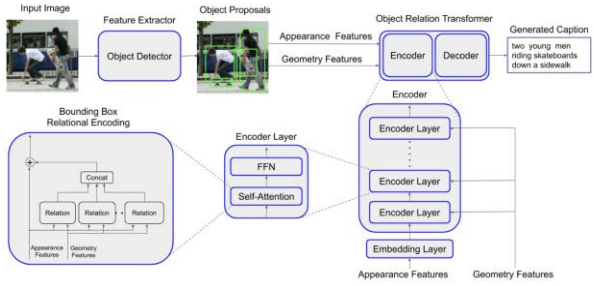

Fig -7: Object Relation Transformer[15]

Object Relation Transformer[15] proposes the use of object spatial relationship modeling to achieve the task of image captioning, by incorporating the object relationship module seen in [16] within the standard Transformer encoder[7] of the Transformer encoder-decoder architecture. [15] incorporates relative geometry by modifying the attention weight matrix, multiplying the appearance-based attention weights by a learned position of their relative position and size. Thus [15] encodes 2D position and size relationships between objects detected in images.

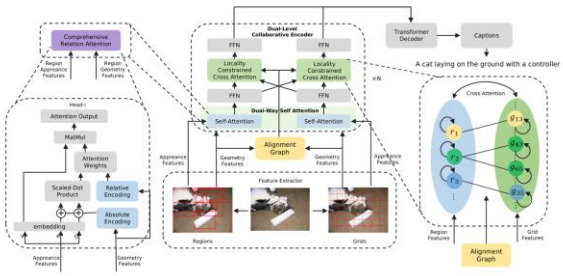

Fig -8: Dual-Level Collaborative Transformer[17]

[17] proposes a novel Dual-Level Collaborative Transformer network that uses both region and grid features of an image, to leverage the complementarity of these features to improve overall performance in the image captioning task. These two features are first processed using a Dual-Way Self Attention (DWSA) module, where a Comprehensive Relation Attention (CRA) scheme embeds absolute and relative geometry information of these input features. In addition, a Locality-Constrained Cross Attention (LCCA) module is used, where a constructed geometric alignment graph guides the semantic alignment between the sources of the two

features. This addresses the degradation issue caused by semantic noises produced during the direct use of the two sources of features during the attention process.

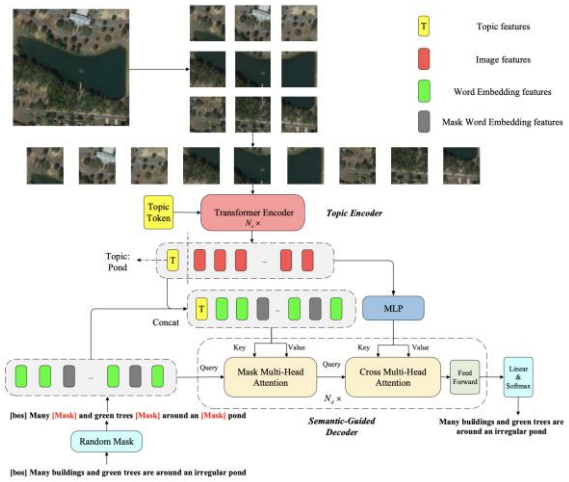

Fig -9: Mask guided Transformer network [18]

[18] proposes a mask-guided Transformer network with a topic token. The model consists of three main parts. A topic encoder is used for the extraction of features and taking note of the relationship between objects multi-head attention is used. A semantic-guided component consisting of multiple attention modules is used along with semantic-guided attention and cross-attention modules for the generation of good captions. Lastly, a Mask-Cross-Entropy strategy is designed for the improvement of a variety of generated captions. This helps in increasing the model's learning ability.

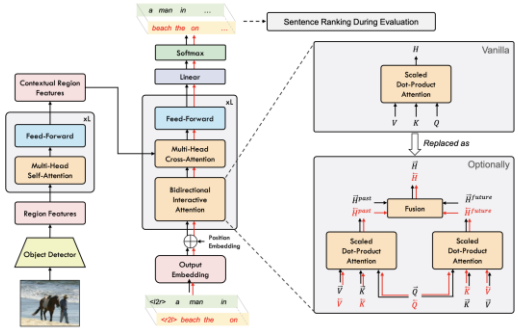

Fig -10: CBT model [19]

[19] introduces a Compact Bidirectional Transformer model for image captioning. The model leverages bidirectional context implicitly and explicitly with the decoder executing parallely. The architecture consists of the model CBTIC, which is built on a famous transformer [7]. The CBTIC model consists of an image feature encoder and captioning decoder. The architecture also features 2 training stages. In the first training stage given a triple, both are padded to equal length without loss of meaning. In the second stage, fine-tuning of the model is conducted using self-critical training (SC) [20] for both left-to-right and right-to-left directions and the gradient.

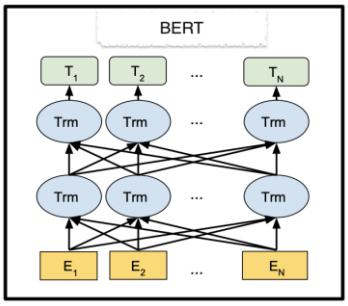

Fig -11: BERT model [8]

[8] proposes Bidirectional Encoder Representations from Transformers (BERT). The architecture is proposed as the datasets available currently are very small in terms of the different captions per image. The BERT model using text augmentation methods expands the training dataset. BERT uses a transformer, an attention mechanism that learns relations between words in a text. The transformer consists of two mechanisms, an encoder that reads the text input and a decoder that produces a prediction.

3. PROPOSED SYSTEM:

3.1 Flowchart

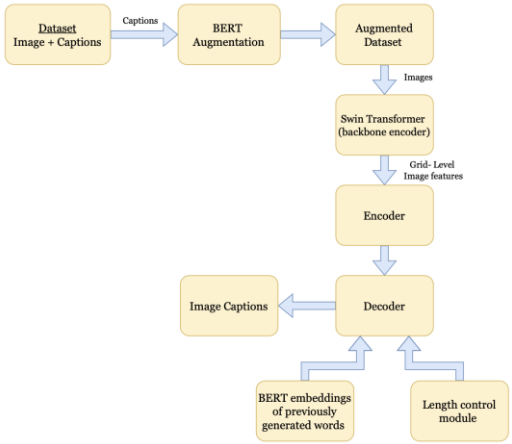

Fig -12: Flowchart of the proposed system VisionAid

3.2 Problem Statement

Based on the research conducted in our literature review, we propose the model "VisionAid". The relevant flowchart diagram can be observed in figure 12 The modules contained in VisionAid are as follows:

1. SwinTransformer as the backbone Encoder.
2. Refining encoder consisting of novel Window Normalized Geometric Self Attention (W-NGSA)/Shifted Window NGSA (SW-NGSA) attention modules.
3. Refining decoder consisting of the Geometric-Aware Self Attention (GSA) module.

In addition to these modules we have also explored augmentation of the captions of the training images using BERT [22], as a training technique to further enhance the performance of VisionAid, and the usage of length embeddings, to encourage the model to generate captions of the desired length, depending on the application.

3.3 Architecture

The architecture of VisionAid follows a typical encoder-decoder structure, with a SwinTransformer as the backbone encoder to extract the grid-level image features. The refining encoder consists of N stacked refining encoder blocks. Each block contains a combination of W-NGSA and SW-NGSA attention modules. Similarly, the decoder also is made up of N-stacked decoder blocks. Each block consists of Cross-GSA and GSA attention modules. The decoder then takes the extracted image features and a concatenation of word embeddings, length embeddings and positional embeddings as the input to generate sequential captions for the image. This can be observed in Figure 13 and will be further elaborated upon below.

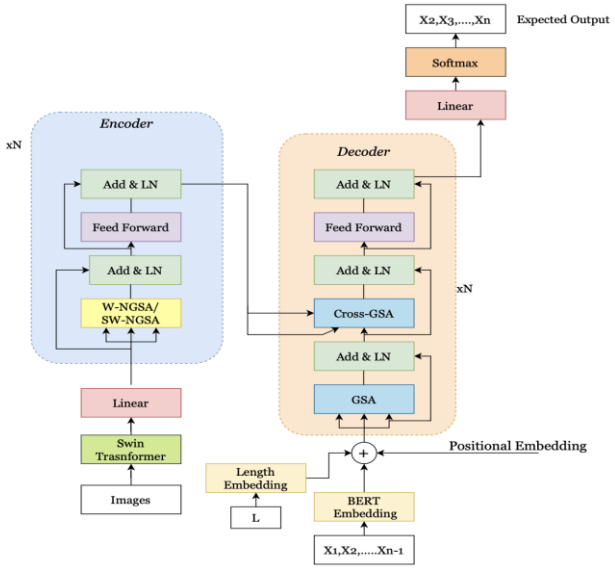

Fig -13: Architecture Diagram of our proposed model Vision Aid.

3.3.1 Encoder:

The image is first passed through the backbone Encoder of VisionAid which is a SwinTransformer [13]. This enables the extraction of grid-level features from the given input image, which becomes the initial vectorized feature input. These vectorized inputs are then passed through a linear projection layer to be fed into the refining encoder.

Each refining encoder block is constructed using Window Normalized-Geometric Self-Attention (W-NGSA) followed by Shifted Window NGSA attention modules to refine the extracted image grid features. VisionAid takes inspiration from the Window and Shifted Window partitioning schemes described in [13] and the novel attention mechanisms Normalized Self-Attention (NSA) and Geometric-Self Attention (GSA) for the attention modules used in the proposed architecture. The details of these methods are further explained as follows.

In [13], both W-MSA and SW-MSA are used in the encoder blocks, in which inputs of Query (Q), Key (K) and Values (V) are given from the image grid features, hence the length and the dimension are the same for these values. These inputs are first partitioned into multiple windows by the W-MSA and SW-MSA mechanisms. Then the application of MSA is performed on each window separately.

In SW-MSA, new windows are generated as a result of shifting the window partitions. By spanning the previous window's borders in figure 14. (a), the self-attention computation in the new windows creates links between them. These links across the windows are significant to further enhance the modeling capability. Hence, the addition of SW-MSA after W-MSA is a good strategy to augment the modeling power.

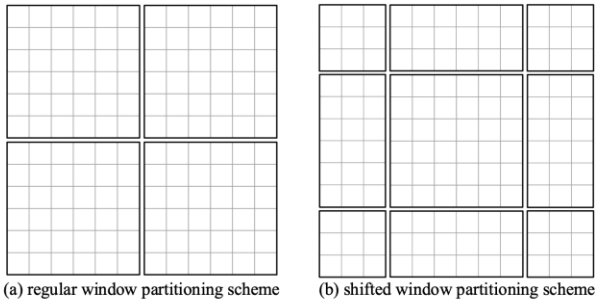

Fig -14: The regular window partitioning scheme and shifted window partitioning scheme of W-MSA and SW-MSA respectively. [12]

[5] makes the case that internal covariate shift[28] issues might arise with standard Shifted-Attention (SA). Typically, SA is viewed as a mapping of a collection of query and key-value pairs. In SA, the generation of the attention weights may be compared to feeding the queries into a fully connected layer, whose parameters are dynamically constructed based on the inputs. Network parameters are thus altered during training. The distribution of queries could shift as a result of this. SA may not be learnt properly as a result of the subsequent layers' ongoing need to adjust to the altered input distribution. [5] presented Normalized Self-Attention to address the internal covariate shift issue inside SA (NSA). To rectify their distributions, NSA applies a cutting-edge normalization technique to the SA's concealed activations. This cutting-edge technology can mean the effective decoupling of the fully-connected layer's parameters from the parameters of the other layers This normalization method differs from Layer Normalization (LN) in that LN normalizes along all channels of each element, while the method proposed in [5] normalizes along each channel of all input elements in an instance (Instance Normalization [3]).

Input element geometric connections cannot be modeled by Vanilla SA. Understanding the image's content depends heavily on these inherently geometric interactions between things, which are both critical and complex. As is frequently done in the case of 1D sentences, adding representations of absolute positions to each element of the inputs is one way to include position information in SA. However, as it is more difficult to deduce the 2D geometry connections between objects from their absolute locations, this technique is not very practical for image captioning. As a result, [5] introduces GSA, which enhances the SA module by accounting for pairwise geometry relationships and image content data.

To fully leverage the advantage realized by the methods proposed in [13] and [5], VisionAid's attention mechanism makes use of W-NGSA followed by SW-NGSA in the encoder. Hence, the inputted feature vectors are output as refined image features that provide more information about the relationships between the objects in the image than a standard implementation of SA.

3.3.2 Decoder:

The decoder predicts the captions for the image word-by-word, taking into account the refined grid-level image features computed by the encoder and the several embeddings fed to the decoder. The description of the embeddings is as follows:

1. Contextualized-word embeddings of the previously generated words are calculated using the BERT transformer [22], enabling the model to predict the next word considering the context of the previous words.
2. Length embeddings guide the model to produce captions of the desired length.
3. Positional encoding information to maintain the order of the words.

Each decoder block makes use of an attention module consisting of Cross Geometric-Aware Self-Attention (Cross-GSA) modules and GSA modules. The decoder does not use the NSA attention module due to the autoregressive nature of the decoder and has variable-length inputs. This is not desirable for the normalization technique used (Instance Normalization[3]) since the mean and variance would become statistically insignificant when the length of the sequence is 1.

The output vectors are then passed through a linear projection layer, before finally being fed into the softmax activation function to generate the image captions.

4. FUTURE SCOPE:

As part of further exploration of the domain of image captioning, we can expand our approach in the following ways. Firstly, an implementation of the proposed model VisionAid can be carried out. Following that, a comparative analysis of VisionAid with other standard transformer-based image captioning models can be conducted to increase our understanding of the model. Finally, different approaches to length control can be explored to achieve different types of captions for the image such as concise captions for image indexing or more descriptive captions where they may be required.

5. CONCLUSION:

To sum up, in this paper, we first conducted a literature review of distinct Transformer-based models in the image-captioning domain space to gain a deeper understanding of the Transformer implementations used and the problems faced by these models. To effectively deal with the issues faced we then proposed a model VisionAid. VisionAid uses a SwinTransformer as the backbone encoder, ensuring grid-level feature extraction from the image. Furthermore, a refining encoder is integrated which uses regular and shifted window partitioning schemes to further enhance the modeling capability, along with normalized self-attention modules and geometric self-attention modules to account for internal covariate shift and to fully utilize geometric relationships between the objects in the image. The refining decoder also uses geometric self-attention and cross-geometric self-attention mechanisms along with BERT embeddings to further contextualize the image captions concerning the previous words of the caption. Length embeddings can be used to guide the model toward generating captions of the desired length. The refining encoder will thus output more accurate captions as compared to standard implementations, and VisionAid will be able to demonstrate state-of-the-art performance using these techniques.

ACKNOWLEDGEMENT:

This work is supported in part by Prof. Pramila M. Chawan. We thank the reviewers for their valuable discussions and feedback.

REFERENCES:

[1] Li, G., Zhu, L., Liu, P., & Yang, Y. (2019). Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp.8928-8937).

[2] Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp.10578-10587).

[3] Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalisation: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

[4] Yang, X., Liu, Y., & Wang, X. (2021). Reformer: The relational transformer for image captioning. *arXiv preprint arXiv:2107.14178*.

[5] Guo, L., Liu, J., Zhu, X., Yao, P., Lu, S., & Lu, H. (2020). Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10327-10336).

[6] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

[7] Vaswani, A., Shazeer, N., Parmar, N.,Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I.(2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[8] Atliha, V., & Šešok, D. (2020). Text augmentation using BERT for image captioning. *Applied Sciences*, *10*(17), 5978.

[9] Kikuchi, Y., Neubig, G., Sasano, R., Takamura, H., & Okumura, M. (2016). Controlling output length in neural encoder-decoders. *arXiv preprint arXiv:1609.09552*.

[10] He, S., Liao, W., Tavakoli, H. R., Yang, M., Rosenhahn, B., & Pugeault, N. (2020). Image captioning through image transformer. In *Proceedings of the Asian Conference on Computer Vision*.

[11] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

[12] Wang, Y., Xu, J., & Sun, Y. (2022). End-to-End Transformer Based Model for Image Captioning. *arXiv preprint arXiv:2203.15350*.

[13] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision (pp. 10012-10022).

[14] Ji, J., Luo, Y., Sun, X., Chen, F., Luo, G., Wu, Y., ... & Ji, R. (2021, May). Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI conference on artificial intelligence*(Vol. 35, No. 2, pp. 1655-1663).

[15] Herdade, S., Kappeler, A., Boakye, K., & Soares, J. (2019). Image captioning: Transforming objects into words. *Advances in Neural Information Processing Systems*, *32*.

[16] Hu, H., Gu, J., Zhang, Z., Dai, J., & Wei, Y. (2018). Relation networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3588-3597).

[17] Luo, Y., Ji, J., Sun, X., Cao, L., Wu, Y., Huang, F., ... & Ji, R. (2021, May). Dual-level collaborative transformer for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 3, pp. 2286-2293).

[18] Ren, Z., Gou, S., Guo, Z., Mao, S., & Li, R. (2022). A mask-guided transformer network with topic token for remote sensing image captioning. *Remote Sensing*, *14*(12), 2939.

[19] Zhou, Y., Hu, Z., Liu, D., Ben, H., & Wang, M. (2022). Compact Bidirectional Transformer for Image Captioning. *arXiv preprint arXiv:2201.01984*.

[20] Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7008–7024.

[21] Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural Machinetranslation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

[22] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[23] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3156-3164).

[24] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048-2057). PMLR.

[25] Anderson, P., He, X., Buehler, C., Teney, D.,Johnson, M., Gould, S., & Zhang, L. (2018). 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).

[26] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Fei-Fei, L. (2017). Visual genome:Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, *123*(1), 32-73.

[27] Zhu, X., Li, L., Liu, J., Peng, H., & Niu, X. (2018). Captioning transformer with stacked attention modules. *Applied Sciences*, *8*(5), 739.

[28] Ioffe, S., & Szegedy, C. (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). PMLR.

[29] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.