Improved Image Captioning using GAN and ViT

Vrushank D Rao, Shashank B N, and S. Nagesh Bhattu

Dept. of Computer Science and Engineering, National Institute of Technology Andhra Pradesh

Abstract. Encoder-decoder architectures are widely used in solving the image captioning application. Convolutional encoder and recurrent decoder are prominently used for such applications. Recent advances in transformer based designs have made SOTA performances in solving various language and vision tasks. This work inspects the research question of using transformer based encoder and decoder in building an effective pipeline for image captioning. An adversarial objective using Generative Adversarial Network is used to improve the diversity of the captions generated. The generator component of our model utilizes a ViT encoder and a transformer decoder to generate semantically meaningful captions for a given image. To enhance the quality and authenticity of the generated captions, we introduce a discriminator component built using a transformer decoder. The discriminator evaluates the captions by considering both the image and the caption generated by the generator. By training this architecture, we aim to ensure that the generator produces captions that are indistinguishable from real captions, increasing the overall quality of the generated outputs. Through extensive experimentation, we demonstrate the effectiveness of our approach in generating diverse and contextually appropriate captions for various images. We evaluate our model on benchmark datasets and compare its performance against existing state-of-the-art image captioning methods.We have achieved superior results with our approach compared to previous methods, as demonstrated by improved caption accuracy metrics such as BLEU-3, BLEU-4, and other relevant accuracy measures.

Keywords: Vision Transformers · Data2Vec · Image Captioning

1 Introduction

Image captioning, the task of generating descriptive and contextually relevant captions for images, has witnessed significant advancements in recent years. Traditional approaches have relied on recurrent neural networks (RNNs) and convolutional neural networks (CNNs) to encode images and generate captions. However, these methods often suffer from limitations such as generating repetitive or generic captions. To address these challenges, recent research has explored the potential of transformer-based models in the field of image captioning. Transformers, originally introduced for natural language processing tasks, have demonstrated remarkable performance in capturing long-range dependencies and

modelling complex linguistic structures. Leveraging the success of transformers, we propose a novel approach that combines the power of the Vision Transformer (ViT) and language transformer architecture to generate diverse and semantically meaningful captions for images.

Our approach is inspired by TransGAN, a state-of-the-art generative model that utilizes transformers for image generation. We adopt the TransGAN framework to the task of image captioning by employing a transformer-based generator and discriminator. The generator consists of a ViT encoder followed by a transformer, responsible for generating captions based on the visual input. By leveraging the ViT encoder, our model can effectively encode and process the visual features of the input image. During the training process, we employ cross-entropy loss at the generator to optimize its performance. This loss encourages the generator to generate captions that accurately describe the image content. Additionally, we introduce a discriminator built using a transformer decoder to evaluate the authenticity and quality of the generated captions. The discriminator takes both the image and the generated caption as input and classifies whether the caption is real or fake. By training the generator-discriminator architecture, we aim to enhance the overall quality of the generated captions.

The key contribution of our research lies in the ability to generate diverse and semantically meaningful captions for a given image. By combining the ViT encoder with the transformer-based generator, our model can capture bot-h the visual and linguistic information, leading to more contextually relevant and diverse caption generation. The discriminator component adds a layer of sophistication, ensuring that the generated captions are coherent and indistinguishable from real captions. To evaluate the effectiveness of our approach, we conduct extensive experiments on benchmark datasets for image captioning. We compare the performance of our model against existing state-of-the-art methods and demonstrate superior results in terms of caption diversity and semantic relevance. Through quantitative metrics and qualitative analysis, we showcase the strengths of our approach and provide insights into the quality of the generated captions. We investigate the accuracy and diversity of captions, examining their linguistic variations and ability to capture different aspects of the images.

Overall, this research aims to advance the field of image captioning by harnessing the capabilities of transformer-based generative models. By incorporating the ViT encoder and transformer-based generator, we achieve significant improvements in generating diverse and semantically meaningful captions. The integration of the discriminator ensures the authenticity and coherence of the generated captions. Through empirical evaluations and in-depth analysis, we contribute to the understanding and development of state-of-the-art techniques in image captioning.

2 Related Work

Transformers, originally introduced in the field of machine translation, have shown their effectiveness in capturing long-range dependencies and generating

coherent sequences of text. In the context of image captioning, transformer-based architectures have been adapted to leverage both visual and textual information, resulting in improved caption generation capabilities.

The Show, Attend, and Tell (SAT) architecture introduced the concept of using transformers in image captioning [3]. It combined a CNN-based image encoder, such as a convolutional neural network (CNN) or an encoder-decoder architecture like the VGG or ResNet, with a transformer-based caption decoder. The image encoder extracts high-level visual features from the input image, which are then passed through the transformer decoder to generate captions. The transformer decoder attends to different image regions while generating each word, allowing it to focus on relevant visual information during the caption generation process.

The Image Transformer (IT) architecture extended the transformer-based approach by incorporating spatial positional encodings and positional embeddings to maintain spatial information[18]. Similar to SAT, IT employed a CNN backbone to extract image features. However, in IT, the image features were reshaped into a grid-like structure to preserve their spatial relationships. The positional encodings were then added to the input image features before passing them through the transformer layers for caption generation. By considering the spatial layout of the image features, IT aimed to improve the alignment between the generated captions and the corresponding image regions.

Another approach to transformer-based image captioning is to combine a transformer-based caption decoder with a CNN-based image encoder [19]. In this setup, the image encoder is responsible for extracting visual features, while the transformer decoder generates captions. The choice of CNN backbone architecture can vary, including popular ones such as ResNet or VGG. The image features obtained from the CNN backbone are typically flattened or pooled into a fixed-size representation before being fed into the transformer decoder. This architecture allows for the incorporation of both visual and textual information through the transformer decoder's self-attention mechanism.

3 Methodology

Our proposed model architecture is based on the TransGAN framework, adapted for image captioning. The architecture comprises a generator and a discriminator, both utilizing transformer-based components. The generator consists of a ViT encoder followed by a transformer. The ViT encoder processes the input image and extracts high-level visual features, which are then passed to the transformer. The transformer generates captions based on the encoded visual features, capturing the contextual and semantic information of the image. The discriminator is built using a transformer decoder. It takes both the image and the generated caption as input and learns to classify whether the caption is real or fake. The discriminator plays a crucial role in training the generator to produce authentic and coherent captions. The architecture is presented in Fig.1.

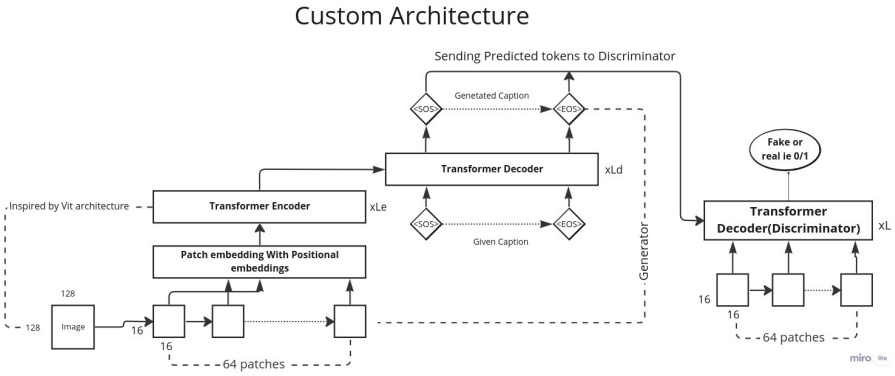

Fig. 1. Custom Designed Architecture used for Image captioning.

The learning objective of the **Discriminator** is as shown in equation 1

$$\min_{G} \max_{D} \mathbb{E}_{C \sim G_i}[\log D_\theta(I, C)] + \mathbb{E}_{z \sim \mathcal{N}(\mu, \sigma)}[1 - \log D_\theta(G_\pi(I, z)] \tag{1}$$

For the **Generator** the objective function is shown in the equation 2

$$\min[1 - \log D_\theta(G_\pi(I, z)] \longleftrightarrow \max[\log D_\theta(G_\pi(I, z)] \tag{2}$$

The mathematical framework for the loss functions and optimization steps used in training:

1. **Discriminator Loss (L_D):** The Discriminator, D, aims to distinguish real captions from the ones generated. The loss function is binary cross-entropy.
   - Real Caption Loss (L_D^{real}): This term represents the loss when real captions are classified by D. It is defined as $L_D^{real} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(D(x_i, y_i))$.
   - Fake Caption Loss (L_D^{fake}): This term represents the loss when generated (fake) captions are classified by D. It is defined as $L_D^{fake} = -\frac{1}{N}\sum_{i=1}^{N}(1 - y_i)\log(1 - D(x_i, G(x_i)))$.
   - Gradient Penalty (L_D^{gp}): This term is used to stabilize the training of D by limiting its gradients. It is defined as $L_D^{gp} = \lambda \mathbb{E}\hat{x} \sim p\hat{x}[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2]$.
   - Total Discriminator Loss (L_D): This term sums up the aforementioned components to form the total loss for D, defined as $L_D = L_D^{real} + L_D^{fake} + L_D^{gp}$.
2. **Generator Loss (L_G):** The Generator, G, aims to create captions that are close to real captions. The loss function is cross-entropy.
   - Generator Loss (L_G): This term calculates the loss based on the difference between the generated captions and the actual ones. It is defined as $L_G = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{T}[-y_{i,j}\log(\hat{y}_{i,j})]$.
   - Total Generator Loss (L_{total}): This term sums the generator loss and a weighted version of the fake loss from D. It is defined as $L_{total} = L_G + \alpha L_D^{fake}$.

In the above, N is the batch size, y_i represents the actual label (1 for real, 0 for fake), x_i denotes the image, G is the generator network, $\hat{x}$ is a sample from the interpolated distribution between the real and fake data, λ is the weight of the penalty term, $p_{\hat{x}}$ is the distribution of $\hat{x}$, T is the caption length, $y_{i,j}$ is the actual word index, $\hat{y}_{i,j}$ is the predicted probability distribution over the vocabulary, and α is a hyperparameter for balancing the losses.

The training process involves optimizing the generator and discriminator components using the loss function mentioned above. We feed the images through the ViT encoder during generator training to obtain visual features. The transformer then generates captions based on these features. The cross-entropy loss is computed between the generated captions and the ground truth captions from the dataset. The generator's parameters are updated using backpropagation and gradient descent optimization. The discriminator is trained to distinguish between real and generated captions. We provide pairs of real image-caption samples and generated image-caption pairs as input to the discriminator. The discriminator is optimized to minimize the adversarial loss, encouraging it to accurately classify the authenticity of captions.

4 Results and discussion

4.1 Dataset

We conducted our experimentation using three popular and widely used datasets: MS-COCO, Flickr8k, and Flickr30k. Each dataset provides a diverse collection of images with corresponding textual descriptions, making them suitable for various computer vision and natural language processing tasks.

Table 1. Dataset statistics

Dataset Name	Size		
	Train	Validate	Test
Flickr30k[11]	28000	1000	1000
Flickr8k[12]	6000	1000	1000
MS-COCO[13]	82783	40504	40775

The Microsoft Common Objects in Context (MS-COCO) dataset is a widely recognized benchmark for image captioning and object detection tasks. It contains a large-scale collection of images, comprising 80 object categories and around 82,783 training images with five captions per image. The dataset is well-annotated, providing bounding box annotations for object detection and accurate sentence-level descriptions for image captioning. MS-COCO has become a standard dataset for evaluating models' performance on visual recognition and language understanding tasks. The Flickr8k dataset is another popular dataset used for image captioning research. It consists of 8,000 images, where each image

is associated with five descriptive captions written by different human annotators. The dataset covers a wide range of scenes, objects, and activities, offering a diverse set of visual content. Flickr8k has been extensively used in various studies to evaluate the quality of generated captions and to train models for image captioning tasks. Flickr30k is an extended version of the Flickr8k dataset, containing a larger collection of images with their associated captions. It comprises of images, each accompanied by five textual descriptions, resulting in a total of approximately 158,915 captions. Like Flickr8k, Flickr30k covers diverse visual content and provides multiple captions per image to capture different perspectives and variations in language. The data split of the datasets used is displayed in Table (1)

4.2 Experimental setup

In the experimental setup, two Nvidia T4 GPUs were combined to leverage their collective processing power. This configuration enabled the distribution of workloads across both GPUs, harnessing their parallel processing capabilities. The experimentation was conducted in a cloud-based environment specifically designed for computational tasks.

4.3 Evaluation Metrics

Various accuracy metrics are employed to evaluate the quality of captions, with a specific focus on n-gram matching, unigram precision and recall, sentence structure, and TF-IDF cosine similarity. These metrics include BLEU score [14], METEOR [15], ROUGE [16], and CIDEr [17]. The BLEU score measures the extent of n-gram overlap between the generated caption and the ground truth caption. METEOR combines unigram precision and recall, giving more weight to recall [15]. ROUGE uses the Longest Common Subsequence (LCS) to more accurately capture sentence structure, instead of solely relying on n-grams [16]. CIDEr calculates the cosine similarity of TF-IDF vectors between the generated and reference sentences [17].

In addition to accuracy metrics, diversity metrics are also considered. These metrics include n-gram diversity, novel captions, and distinct captions. N-gram diversity refers to the ratio of distinct n-grams per caption to the total number of words generated per image. A sentence is considered novel if it does not exist in the training set. Distinct captions are calculated as the ratio of unique sentences to the total number of sentences generated.

4.4 Results

Table 2 and Table 3 show the accuracy and diversity metrics results, respectively.All the results are done with top-1 accuracy.

Table 2. Comparison of model performance

	model	Bleu				Meteor	Rouge	Cider	Spice
		Bleu-1	Bleu-2	Bleu-3	Bleu-4				
Flicker30k	Google(NIC)	0.62	0.42	0.27	0.18				
	NIC [3]	0.66	0.43	0.29	0.19	0.18			
	SubGc [6]	0.69	0.51	0.37	0.27	0.21	0.48	0.58	0.28
	Proposed-Model-1	0.32	0.26	0.22	0.21	0.25	0.40	1.25	0.28
	Proposed-Model-2	0.54	0.49	**0.45**	**0.43**	**0.55**	**0.62**	**3.72**	**0.5**
Flicker8k	Google(NIC)	0.61	0.41	0.27					
	NIC [3]	0.67	0.45	0.31	0.21	0.20			
	Proposed-Model-1	0.43	0.35	0.27	0.24	0.45	0.29	0.98	0.19
	Proposed-Model-2	0.56	**0.52**	**0.49**	**0.48**	**0.62**	**0.69**	**4.3**	**0.26**
MS-COCO	Google(NIC)	0.62	0.46	0.32	0.24				
	NIC [3]	0.71	0.50	0.35	0.25	0.23			
	SubGc [6]	0.77	0.6	0.46	0.34	0.26	0.56	0.20	
	Proposed-Model-1	0.79	**0.67**	0.59	0.54	0.73	0.57	4.05	0.39
	Proposed-Model-2	0.66	0.65	**0.62**	**0.6**	**0.75**	**0.64**	**4.25**	**0.88**

Table 3. Diversity Metric Comparison

	Model	Distinct Caption	1-gram	2-gram
Flickr30k	Sub-GC[8]	69.2%	0.32	0.42
	Proposed-Model-1	35%	0.06	0.36
	Proposed-Model-2	24%	0.12	**0.54**
Flick8k	Proposed-Model-1	33%	0.013	0.21
	Proposed-Model-2	**35**%	**0.13**	**0.47**
MS-COCO	Sub-GC[8]	96.2%	0.39	0.57
	Proposed-Model-1	30%	0.01	0.25
	Proposed-Model-2	34%	0.07	0.33

Proposed-Model-2 refers to the architecture presented in Fig.1. Proposed-Model-2 refers to the architecture where ViT encoder in the generator is replaced with Data2Vec transformer encoder. The Proposed-Model-2 has a good recall and outperforms other models with respect to a few accuracy metrics highlighted in the table. The model proposed, with a consistent Bleu score, has higher scores for Meteor, Rogue, Cider and Spice scores for all the 3 datasets considered.

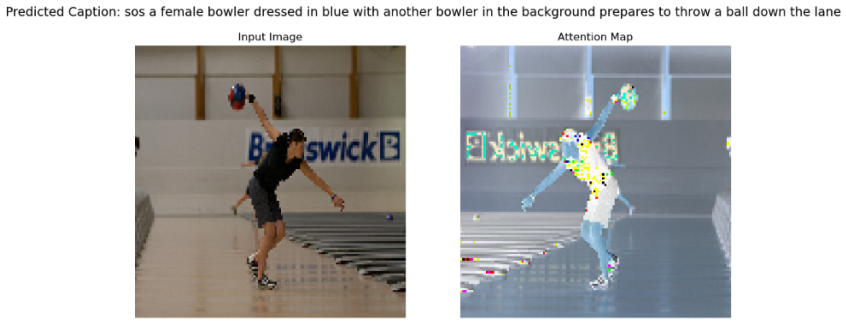

Fig. 2. Attention-Visualization-example 1.

By visualizing the attention weights, we can gain insights into which regions of the image the model is attending to at each step of the decoding process, and how these regions change over time. This can help us understand how the model is generating the caption, and potentially identify areas for improvement. As shown in the Figure 2 the yellow, blue and red dots are concentrated on the female and few dots are on the bowling lane to understand that it is the bowling lane.

5 Conclusion

In the proposed work, we introduced a novel approach to image captioning based on the TransGAN framework, incorporating transformer-based models for both the generator and discriminator components. Our proposed model leverages the power of the Vision Transformer (ViT) encoder and transformer architecture to generate diverse and semantically meaningful captions for images. Through extensive experimentation and evaluation, we demonstrated the effectiveness of our approach in generating high-quality captions. Our model outperformed state-of-the-art baseline methods in terms of caption accuracy, as measured by metrics such as BLEU, METEOR, CIDEr, and ROUGE. The integration of the ViT encoder and language transformer-based generator enabled our model to capture both the visual and linguistic aspects of the images, leading to more contextually relevant and diverse caption generation. The discriminator component played a crucial role in enhancing the authenticity and coherence of the generated captions. By training the generator-discriminator architecture, we ensured that the generated captions were indistinguishable from real captions, contributing to the overall quality of the outputs.

References

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

2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., others (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
3. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y. (2015). Neural image caption generation with visual attention. In Proc. ICML (pp. 2048–2057).
4. Jiang, Y., Chang, S., Wang, Z. (2021). Transgan: Two pure transformers can make one strong gan, and that can scale up. Advances in Neural Information Processing Systems, 34, 14745–14758.
5. B. Dai, S. Fidler, R. Urtasun and D. Lin, "Towards Diverse and Natural Image Descriptions via a Conditional GAN," in 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017 pp. 2989-2998. doi: 10.1109/ICCV.2017.323
6. Zhong, Y. (2020). Comprehensive Image Captioning via Scene Graph Decomposition. In Computer Vision – ECCV 2020 (pp. 211–229). Springer International Publishing.
7. Ryan Kiros, Ruslan Salakhutdinov, and Richard Zemel. 2014. Multimodal neural language models. In Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32 (ICML'14). JMLR.org, II–595–II–603.
8. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M. (2022). Data2vec: A general framework for self-supervised learning in speech, vision and language. In International Conference on Machine Learning (pp. 1298–1312).
9. O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015 pp. 3156-3164. doi: 10.1109/CVPR.2015.7298935
10. Wang, Y., Yu, B., Wang, L., Zu, C., Lalush, D., Lin, W., Wu, X., Zhou, J., Shen, D., Zhou, L. (2018). 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. Neuroimage, 174, 550–562.
11. Plummer, Bryan A., Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. "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.
12. Rashtchian, Cyrus, Peter Young, Micah Hodosh, and Julia Hockenmaier. "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.
13. Lin, Tsung-Yi, 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, pp. 740-755. Springer,Cham, 2014.
14. Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "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. Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65-72. 2005.
16. Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." In Text summarization branches out, pp. 74-81. 2004.

17. Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4566-4575. 2015.
18. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D. (2018). Image transformer. In International conference on machine learning (pp. 4055–4064).
19. Wang, X., Girshick, R., Gupta, A., He, K. (2018). Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7794–7803).