Content

Figure List

Table List

1. Introduction

Image captioning merges deep learning and Natural Language Processing (NLP) to convert images to text. It enables computers to recognize and describe images in a human-like manner, enhancing accessibility, optimizing digital marketing, generating visual translations, and providing insights through automated analysis (Wang *et al.*, 2022). We compare the performances of convolutional neural network (CNN) plus long short-term memory (LSTM) model (Figure 1) versus Transformer architecture (Figure 2). We will explore their functionalities and further business applications in image captioning.

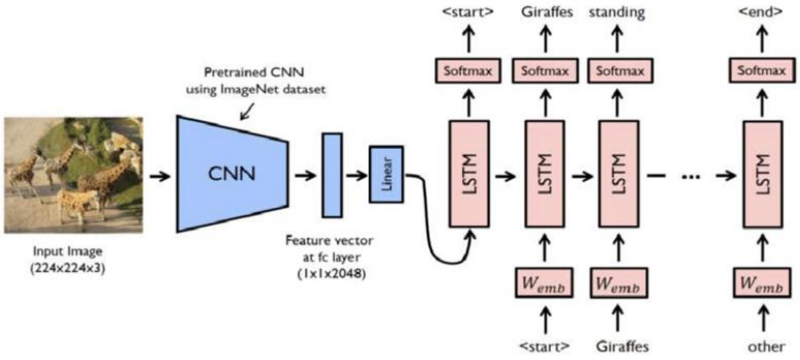


Figure 1: CNN + LSTM Model Structure (Analytics Vidhya, 2018)

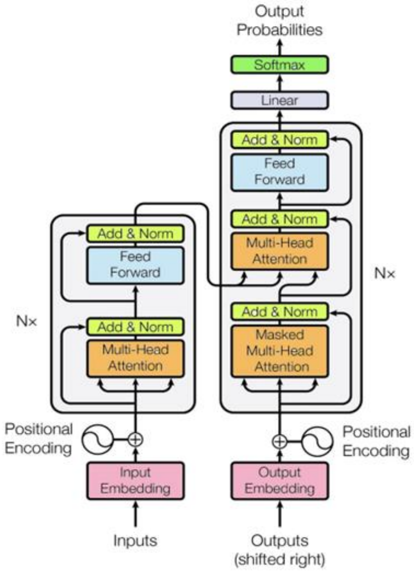


Figure 2: Transformer Structure (Dandwate et al., 2023)

2. Methodology

2.1 Pre-processing of Data

We used the public Flickr8k dataset from Kaggle (8,000 generic images), where one image matched with five differently worded captions. We conducted the following text-cleaning steps:

- Created a random subset of the dataset for train and test
- Words were tokenised
- Punctuations were removed
- Converted to lowercase
- To generate grammatically accurate English, stop words were not removed, and stemming was not conducted.

2.2 Generative Image Transformer (GIT)

GIT is a transformer decoder for image captioning, conditioned on image and text tokens to predict the next text token. The model is trained using "teacher forcing" on many image and text pairs (Wang *et al.*, 2022).

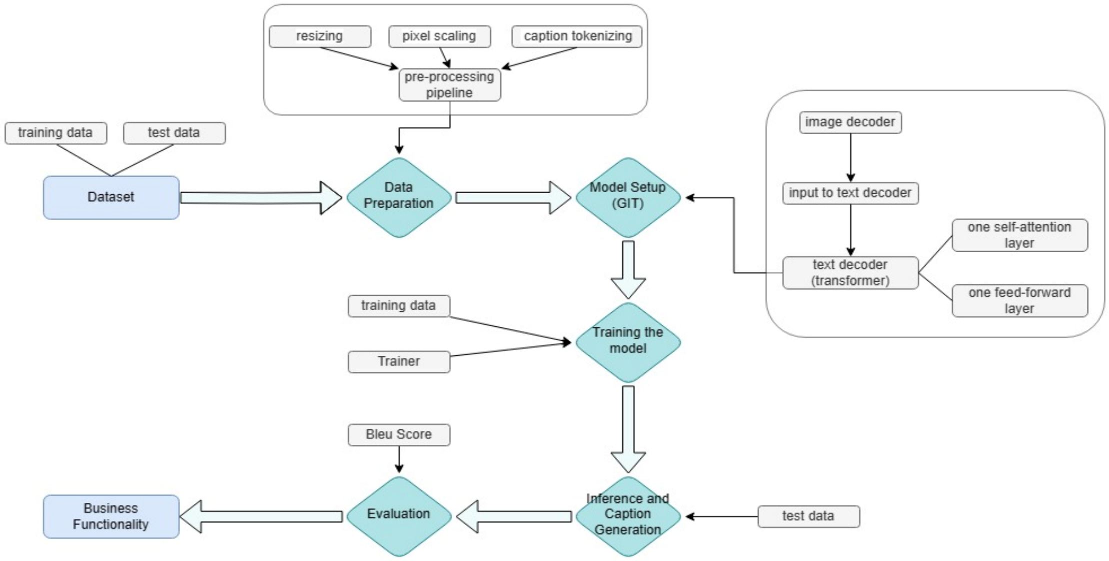


Figure 3: Workflow of GIT approach

From Figure 3, the image captioning model leverages a pre-trained language model GIT from HuggingFace Hub, which offers significant advantages (Wolf *et al.*, 2020). Using transfer learning, less training data reduces training time while a processor converts images into feature vectors.

Training and testing datasets, organized as JSON Lines files with image-caption pairs, ensure consistent mapping between images and their metadata.

During training, the model refines its parameters over multiple epochs using techniques like AdamW, learning to map image features to text by minimizing loss function comparing predicted captions to ground-truth references. In the inference phase, the model generates captions for new test images by encoding them into feature vectors and passing them through the trained model. These captions are decoded and saved for further evaluation.

2.3 CNN-LSTM Multi-Model Approach

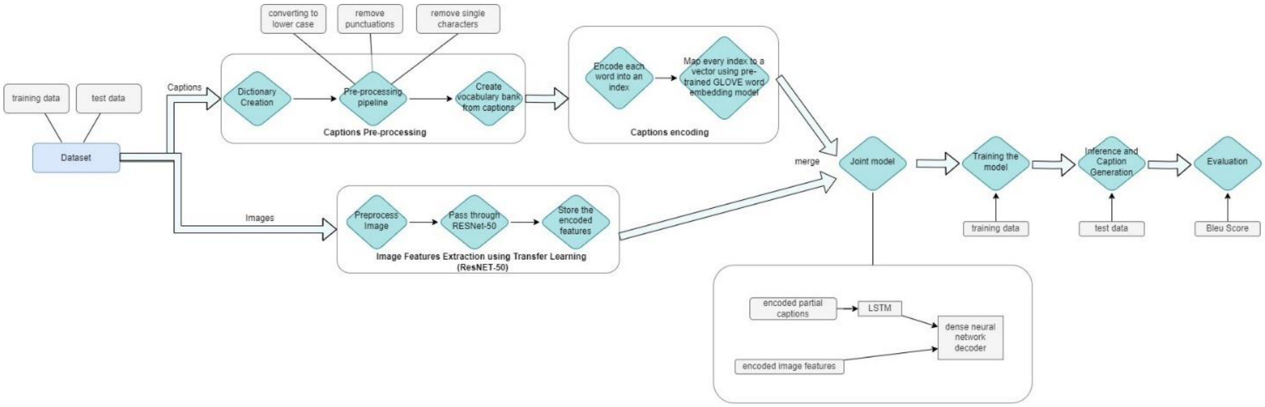


Figure 4: Workflow of CNN-LSTM multi-model approach

From Figure 4, the captions and images are processed from the dataset separately. After running the captions through pre-processing pipeline, we created a vocabulary bank of unique words. We index the words in a sorted dictionary. The text is inputted into the "Sequence Processor," where it uses pre-trained GLOVE word embeddings which are then encoded by an LSTM layer, resulting in vector representations of the text (Gautam, 2021).

We choose a pre-trained CNN called ResNET-50 for image processing, which extracts and encodes features into vector format from image inputs. Alam *et al.* (2021) found that this is a robust model with high accuracy and low value-loss.

Finally, the outputs, vectorized image features, and word embeddings from the earlier two input models are concatenated together into a joint model. Here, a feed-forward neural network makes up the decoder section of this ‘encoder-decoder model’ to predict the next word in the caption based on the extracted features of the image and the preceding words. The final output dense layer has several nodes equal to the vocabulary size and makes a probability prediction using the Softmax activation function for the next word in the sequence.

3. Experimenting Evaluation

3.1 Metrics Interpretation

BLEU is a metric for evaluating machine-generated translations by comparing with reference translations (Papineni *et al*., 2002), and it is used to check the accuracy of captions generated by both models (see Appendix A). GIT performs better than CNN-LSTM with a BLEU score of approximately 0.3 instead of 0.16, since CNN-LSTM may struggle to capture fine-grained details or handle long-range dependencies effectively, leading to lower similarity with reference captions (Kelvin *et al.*, 2015; see Table 1 and Appendix C).

The BLEU score ranging from 0.3 - 0.5 is the expected performance level of machine translation systems (Vaswani et al., 2017; Devlin et al., 2018). The BLEU score for GIT falls short of the optimal range given that our model is trained on a small dataset (Bahdanau *et al*.,2014).

Although CNN+LSTM effectively learns hierarchical features and captures spatial-temporal information, GIT aligns better with our specific needs with its superior capabilities, because it excels at understanding complex inter-item relationships, capturing detailed visual features, and producing coherent, contextually relevant descriptions (see Appendix B and D). Hence, considering our focus on accurate and context-sensitive descriptions, GIT is the more appropriate choice for our applications.

Table 1: Comparison of Models

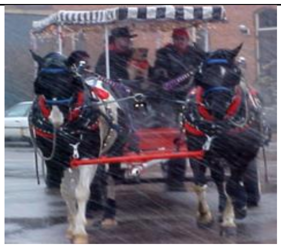
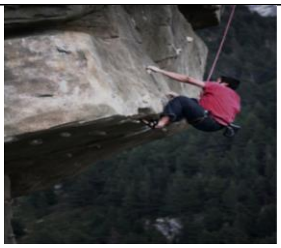
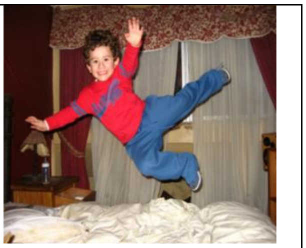
Sr.No	Model	Test dataset BLEU Scores (AVG)	Suitable range
1	GIT	0.28	0.3-0.5
2	CNN + LSTM	0.16	

3.2 Qualitative Analysis

We compared the standard captions and those generated by GIT and CNN+LSTM (see Table 2). Although the captions generated by GIT are not flawless, they display more human-readable content compared to those produced by CNN+LSTM.

Due to computing capability constraints, both models were limited to running on only 30 images in the training dataset. However, CNN+LSTM generated captions that were too short to be valid, leading us to retrain it on a more extensive training dataset of 300 images. Despite efforts to expand the vocabulary of CNN+LSTM, it failed to outperform GIT. Techniques such as data augmentation, ensemble approaches, hyperparameter tuning, and regularization strategies are suggested to improve the CNN+LSTM and achieve better captioning performance.

Table 2: Caption Comparison CNN + LSTM vs GIT

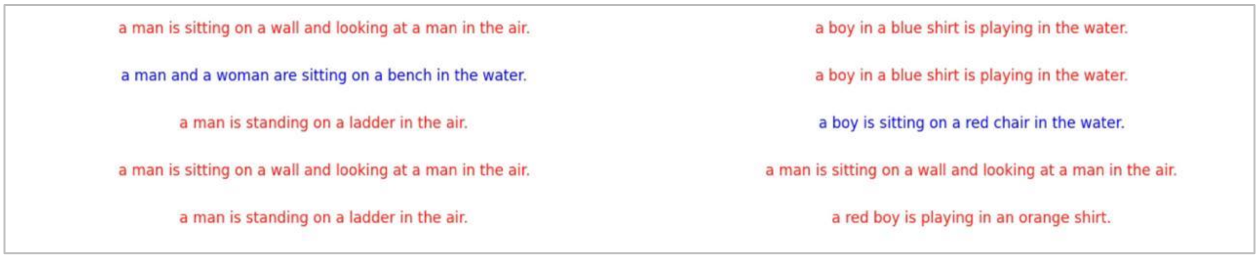
Images			
Standard	The horses pull the carriage	A rock climber practices on a rock-climbing wall	The young boy in red and blue outfit is poised in midair
CNN + LSTM	Several men in	Man in red shirt and shirt up the	Young girl wearing white stands in
GIT	A boy is sitting on a red chair in the water	A man is sitting on a rock in the water	A boy is playing with a red shirt and a blue shirt in the air

4. Innovative Business Applications

The base GIT architecture can be further utilized for business applications by improving efficiency and achieving Environmental, Social, and Governance (ESG) targets.

4.1 Digital Marketing (Topic Modelling/Data Catalogues):

Image captioning enables marketing firms to manage their visual data more efficiently and reduce data management costs. Since images take up more storage space than text, image captioning can enable marketing firms to store vast amounts of visual data in centralized, easily searchable, and scalable textual data catalogues (Harzig *et al.*, 2018). It has been adopted by large-scale E-commerce firms like Amazon, where they have automated caption creation and enterprise-scale search for images using generative AI and Amazon Kendra (Srinivasan, 2023). Using GIT workflow, small and medium enterprises can enhance the streamlining search results for users and increase product discoverability while reducing search friction, considering the limited budget.



a man is sitting on a wall and looking at a man in the air.

a man and a woman are sitting on a bench in the water.

a man is standing on a ladder in the air.

a man is sitting on a wall and looking at a man in the air.

a man is standing on a ladder in the air.

a boy in a blue shirt is playing in the water.

a boy in a blue shirt is playing in the water.

a boy is sitting on a red chair in the water.

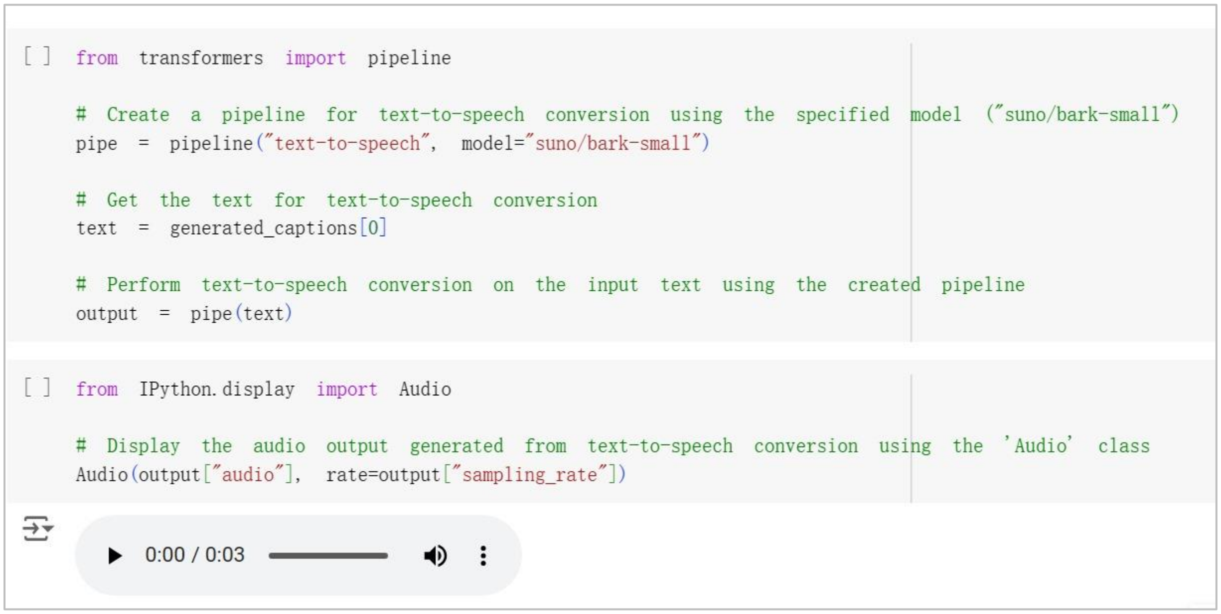
a man is sitting on a wall and looking at a man in the air.

a red boy is playing in an orange shirt.

Figure 5: Topic modelling output

4.2 Accessibility of digital content for blind and low vision (Text-to-Audio):

Captioning aids visually impaired individuals by translating visual data into audio descriptions through text intermediaries (see Figure 6). Image captioning is currently being used in non-profit sectors, such as Microsoft's partnership with the University of Texas at Austin under Microsoft Ability Initiative, which aims to develop an image captioning dataset to foster new accessible technologies for people who are blind or visually impaired (Blog M.A., 2019). Incorporating ESG principles is crucial for modern organizations (Cheema & Langa, 2022; Fenwick *et al.*, 2023). This report argues that captioning for the visually impaired aligns with ESG by fostering inclusivity on the Internet where everyone can access online resources regardless of their abilities, which would increase mandates for companies to ensure equitable digital access. It would enable the company to build trust with consumers and stakeholders, further increasing brand reputation ethically and socially.



```
[ ] from transformers import pipeline

    # Create a pipeline for text-to-speech conversion using the specified model ("suno/bark-small")
    pipe = pipeline("text-to-speech", model="suno/bark-small")

    # Get the text for text-to-speech conversion
    text = generated_captions[0]

    # Perform text-to-speech conversion on the input text using the created pipeline
    output = pipe(text)

[ ] from IPython.display import Audio

    # Display the audio output generated from text-to-speech conversion using the 'Audio' class
    Audio(output["audio"], rate=output["sampling_rate"])
```

Figure 6: Output for Text to Audio

4.3 Other Applications

Beyond marketing and accessibility, image captioning has applications across many sectors. For example, it can be used in healthcare sector to analyze medical images to detect conditions like tumours, education sector to automatically generate visual summaries that aid in teaching and research, and security sector to assist in monitoring and identifying objects in surveillance footage (Srinivasan, 2023).

5. Conclusion

In conclusion, this report compares two architectures, with GIT performing better. However, limitations exist. Firstly, our models struggle with feature recognition in highly complex images and tend to perform better with more straightforward images. Secondly, acquiring accurate quality labels for images, especially in specialized domains like the medical field, is costly, time-consuming, requires expert input, and involves privacy concerns.

References

Abdal Hafeth, D. and Kollias, S. (2024) ‘Insights into object semantics: Leveraging transformer networks for Advanced Image captioning’, *Sensors*, 24(6), p. 1796. doi:10.3390/s24061796.

Alam, M.S. *et al*. (2021) ‘Comparison of different CNN model used as encoders for image captioning’, *2021 International Conference on Data Analytics for Business and Industry (ICDABI)* [Preprint]. doi:10.1109/icdabi53623.2021.9655846.

Bahdanau, D., Cho, K. and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. [online] arXiv.org. Available at: https://arxiv.org/abs/1409.0473.

Blog, M.A. (2019). *Working together to improve image captioning for people who are blind*. [online] Microsoft Accessibility Blog. Available at: https://blogs.microsoft.com/accessibility/working-together-to-improve-image-captioning-for-people-who-are-blind/.

Cheema, S. and Langa, M., 2022. Environment, social, and governance (ESG) and sustainability. In *A Director's Guide to Governance in the Boardroom* (pp. 135-171). Routledge.

Dandwate, P., *et al.* (2023). Comparative study of Transformer and LSTM Network with attention mechanism on Image Captioning. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2303.02648.

Devlin, J., *et al.* (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [online] arXiv.org. Available at: https://arxiv.org/abs/1810.04805.

Dosovitskiy, A., *et al.* (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv:2010.11929 [cs].

Fenwick, M., *et al.* (2023). ESG as a business model for small and medium-sized enterprises. In *The Elgar Companion to UNCITRAL* (pp. 392-409). Edward Elgar Publishing

Gad, G. et al. (2022). ‘Deep learning-based context-aware video content analysis on IOT devices’, Electronics, 11(11), p. 1785. doi:10.3390/electronics11111785.

Gautam (2021). Image-Captioning. [online] GitHub. Available at: https://github.com/gautamgc17/Image-Captioning .

Harzig, P., *et al.* (2018). Multimodal Image Captioning for Marketing Analysis. [online] IEEE Xplore. doi:https://doi.org/10.1109/MIPR.2018.00035.

Herdade, S., Kappeler, A., Boakye, K. and Soares, J. (2020). Image Captioning: Transforming Objects into Words. arXiv:1906.05963 [cs]. [online] Available at: https://arxiv.org/abs/1906.05963.

Lester, B., Al-Rfou, R., Constant, N. and Research, G. (n.d.). The Power of Scale for Parameter-Efficient Prompt Tuning. [online] Available at: https://arxiv.org/pdf/2104.08691.

Ng, J.Y.-H., *et al.* (2015). *Exploiting Local Features from Deep Networks for Image Retrieval*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.1504.05133.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2001). BLEU: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* - ACL '02. [online] doi:https://doi.org/10.3115/1073083.1073135.

Shaikh, J. (2021). Automatic Image Captioning Using Deep Learning. [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2018/04/solving-an-image-captioning-task-using-deep-learning/.

Sennrich, R., *et al.* (2015). Neural Machine Translation of Rare Words with Subword Units. [online] arXiv.org. Available at: https://arxiv.org/abs/1508.07909.

Srinivasan, B. (2023). Automate caption creation and search for images at enterprise scale using generative AI and Amazon Kendra | AWS Machine Learning Blog. [online] aws.amazon.com. Available at: https://aws.amazon.com/tw/blogs/machine-learning/automate-caption-creation-and-search-for-images-at-enterprise-scale-using-generative-ai-and-amazon-kendra/ .

Vaswani, A., *et al.* (2017). Attention Is All You Need. [online] arXiv.org. Available at: https://arxiv.org/abs/1706.03762.

Wang, J., et al. (2022). GIT: A Generative Image-to-text Transformer for Vision and Language. arXiv:2205.14100 [cs]. [online] Available at: https://arxiv.org/abs/2205.14100.

Wolf, T., *et al.* (2020). Transformers: State-of-the-Art Natural Language Processing. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics.

Xu, K., *et al.* (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. [online] arXiv.org. Available at: https://arxiv.org/abs/1502.03044.

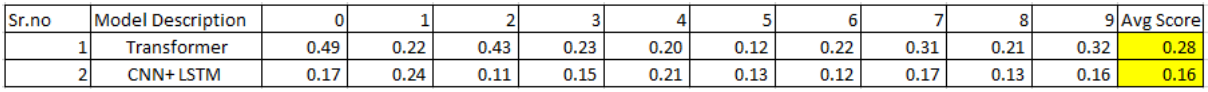
Appendix A: BLEU Technique

Definition	Advantages	Disadvantages	Formulas	References
Evaluation metrics assess the likeness between machine-generated text and reference text, gauging n-gram overlap, where elevated scores suggest greater resemblance.	It achieves strong correlations with human assessments by averaging individual sentence judgements throughout a test corpus, giving general consistency precedence over exact judgement for each sentence: More improves quality	May not always accurately convey the meaning of translations, especially when dealing with grammatical irregularities or fine-grained semantic differences.	• BLEU-1: BLEU-1=∑i=1Ncount i∑i=1NtotaliBLEU-1=∑i=1N totali ∑i=1N counti • BLEU-n: BLEU-n=∑i=1Ncount i∑i=1NtotaliBLEU-n=∑i=1N totali ∑i=1N counti • Geometric Mean: BLEU=BP×exp(1n∑i=1nlog(BLEU-n))BLEU=BP×exp(n1 ∑i=1n log(BLEU-n))	(Doe & Smith, 2023; Papineni, Roukos, Ward & Zhu, 2002; Liu 2016)

Appendix B: CNN+ LSTM and GIT

Model	Definition	Advantage	Disadvantage	Reference
CNN+LSTM model	A deep learning architecture that combines CNN and LSTM networks	1. hierarchical feature learning capability 2. capture both spatial and temporal information	1. challenges posed by long-term dependencies in sequential modeling through the integration of attention mechanisms. 2. Training requires significant volumes of data.	(Xu et al., 2015; Ng et al., 2015)
GIT	A neural network framework crafted to convert images into detailed textual descriptions.	Ability to understand complex item relationships, grasp complex visual characteristics, and produce coherent, contextually relevant captions.	The actual implementation of GIT models in resource-constrained situations may be impeded by their computational complexity and resource needs.	(Dosovitskiy et al., 2020; Vaswani et al., 2017)

Appendix C Comparison of BLEU score for 10 images



Sr.no	Model Description	0	1	2	3	4	5	6	7	8	9	Avg Score
1	Transformer	0.49	0.22	0.43	0.23	0.20	0.12	0.22	0.31	0.21	0.32	0.28
2	CNN+ LSTM	0.17	0.24	0.11	0.15	0.21	0.13	0.12	0.17	0.13	0.16	0.16

Appendix D: Comparison of CNN-LSTM model vs Generative Image Transformer (GIT)

Image Captioning Models	
CNN-LSTM	**Generative Image Transformer (GIT)**
Older	Newer, better performance
Multi-modal model (CNN + LSTM). Output from CNN is fed into LSTM, after going through a word embedding layer.	Integrated single model, simplified architecture of one image encoder and one text decoder
For the small training data size, CNN-LSTM is faster (Gad et al., 2022).	For the small training data size, GIT is slower.
Depends on the vocabulary bank of training data. Limited training data will hinder the model's ability to generalize, leading to suboptimal performance.	Image encoder is a Swin-like vision transformer pre-trained on 0.8B image-caption pairs, and has a wide vocabulary bank.
Requires large amounts of data for effective training and good model performance	Require less training data for good model performance
Vanishing gradient problem is reduced in LSTM as compared to a typical RNN	Does not suffer from vanishing gradient problem
Effective for shorter captions, but may struggle with capturing long-range dependencies and understanding complex contexts.	Better at handling longer captions, more adaptable to different datasets as it utilizes self-attention mechanisms to capture long-range dependencies, allowing for better understanding of semantic context
There is a sequential processing of information.	Transformer architecture enables parallel processing of information. This parallel approach is more efficient for sequence-to-sequence tasks compared to LSTM models. (Abdal Hafeth & Kollias, 2024)