

















## Image Captioning with Convolutional Neural Networks and Autoencoder-Transformer Model

Selvani Deepthi Kavila<sup>1</sup>, Moni Sushma Deep Kavila<sup>1</sup>, Kanaka Raghu Sreerama<sup>2</sup>, Sai Harsha Vardhan Pittada<sup>1</sup>,  
Krishna Rupendra Singh<sup>3</sup>, Badugu Samatha<sup>4</sup> and Mahanty Rashmita<sup>5\*</sup>



<sup>1</sup>Department of CSE (Artificial Intelligence & Machine Learning and Data Science) Anil Neerukonda Institute of Technology and Sciences(A), Visakhapatnam, Andhra Pradesh, India; <sup>2</sup>Department of AI & ADS, GST, GITAM University, Visakhapatnam, Andhra Pradesh, India; <sup>3</sup>Department of Computer Science and Engineering, Vignan's Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India; <sup>4</sup>Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh, India; <sup>5</sup>Department of Basic Sciences and Humanities, Vignan's Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India

E-mail/Orcid Id:

SDK,  selvanideepthi14@gmail.com,  <https://orcid.org/0000-0001-5307-3113>; MSDK,  monisushmakavila@gmail.com,  <https://orcid.org/0009-0000-6457-3500>; KRS,  ksreeram@gitam.edu,  <https://orcid.org/0000-0003-1168-237X>; SHVP,  saiharsha9897@gmail.com,  <https://orcid.org/0009-0004-5871-1427>; KRS,  rupukrishna@gmail.com,  <https://orcid.org/0009-0007-6402-9194>; BS,  bsamatha@kluniversity.in,  <https://orcid.org/0000-0003-1353-2797>; MR,  rashmitamoon@gmail.com,  <https://orcid.org/0000-0001-9247-8295>

### Article History:

Received: 11<sup>th</sup> Jul., 2024

Accepted: 15<sup>th</sup> Oct., 2024

Published: 30<sup>th</sup> Dec., 2024

### Keywords:

Image Captioning, Deep Learning, Transformers, Autoencoders, Convolutional Neural Networks, Machine Learning

### How to cite this Article:

Selvani Deepthi Kavila, Moni Sushma Deep Kavila, Kanaka Raghu Sreerama, Sai Harsha Vardhan Pittada, Krishna Rupendra Singh, Badugu Samatha and Mahanty Rashmita (2024). Image Captioning with Convolutional Neural Networks and Autoencoder-Transformer Model. *International Journal of Experimental Research and Review*, 46, 297-304.

### DOI:

<https://doi.org/10.52756/ijerr.2024.v46.023>

**Abstract:** This study deals with emerging machine learning technologies, deep learning, and Transformers with autoencode-decode mechanisms for image captioning. This study is important to provide in-depth and detailed information about methodologies, algorithms and procedures involved in the task of captioning images. In this study, exploration and implementation of the most efficient technologies to produce relevant captions is done. This research aims to achieve a detailed understanding of image captioning using Transformers and convolutional neural networks, which can be achieved using various available algorithms. Methods and utilities used in this study are some of the predefined CNN models, COCO dataset, Transformers (enc-BERT, dec-GPT) and machine learning algorithms which are used for visualization and analysis in the area of model's performance which would help to contribute to advancements in accuracy and effectiveness of image captioning models and technologies. The evaluation and comparison of metrics that are applied to the generated captions state the model's performance.

### Introduction

Image captioning in the sector of image processing is a difficult task from the first, Natural Language Processing(NLP) with deep-learning (DL) (Bahdanau et al., 2016), which involves the generation of textual captions for images. With the growth of image data on the internet, the need for machines to process and understand images and make descriptions has become increasingly important. This task needs a combination of

processing the given images to understand image content and use transformers(NLP) to generate effective and contextually relevant descriptions (Dekvin et al., 2019).

In recent years, there has been a growing interest in using improvements in Deep Learning and neural network models for image captioning tasks (Vedantam et al., 2015). According to the mentioned reference, these models have improvised the traditional implementation and created a foundation that increased the ability to



generate accurate and meaningful descriptions for images, which can significantly improve the accessibility and usability of visual information in a number of domains (Rao et al., 2021).

This study aims to use the essence of machine-based learning and CNN models (Johnson et al., 2016) for image captioning, which focuses on enhancing understanding of the outcomes of these models through metrics like BLEU, ROUGE (Lin et al., 2004; Keerthana et al., 2024), METEOR, CIDER, SPICE (Lavie et al., 2007). Additionally, this study uses these metrics to provide insights into the workings of these models, explaining how and why certain captions are generated. Another key objective is to implement Transformers (Vaswani, 2017) to improve the efficiency and accuracy of captions. Finally, the study aims to evaluate the performance of the proposed system using a range of metrics, comparing results with models to establish its effectiveness and potential advantages. Our Major Contributions outlines:

- ✧ Predicting image captions using machine learning and neural network models.
- ✧ Execution of Transformers while dealing with captioning input image.

### Related work

Image captioning (Zhou et al., 2020), a subset of artificial intelligence, aims to generate a descriptive caption that describes the image provided as input. Over recent years, this task has attracted substantial attention, driving the emergence of diverse research approaches and methodologies. Early image captioning methods relied on template-based techniques and statistical approaches. Template-based approaches utilized predefined structures filled with detected objects, actions, and scene descriptions, providing a simplistic yet limited method of generating captions (Bengio et al., 1994). On the other hand, statistical machine translation treated image description as a language translation task, where images were considered source languages and captions as target languages (Vinyals et al., 2015). Also, the approach uses computer vision (Rennie et al., 2017) to get the captions. These methods built a foundation but faced challenges in processing the complexity and variability of natural images.

The advancements in deep learning made a significant improvement in image captioning. Convolutional Neural Networks (CNNs) (Johnson et al., 2016; Bahdanau et al., 2016) enabled and LSTM (Gers et al., 2019; Hochreiter et al., 1997) extraction of important visual features from images, capturing related representations that help in

meaningful image understanding (Szegedy et al., 2017). Concurrently, Recurrent Neural Networks (RNNs), particularly sequential models, are mostly used for this task as they produce valid outputs that are up to the mark (Hochreiter et al., 1997). Integration of CNNs and RNNs in an encoder-decoder architecture became an efficient approach, enabling systems to learn to generate related captions based on visual inputs effectively.

To enhance relevance and compatibility of generated captions, relation-based mechanisms were introduced. Visual attention mechanisms selectively focus on relevant regions of an image during the captioning process, using parallel integration of textual generation with visual content (Lin et al., 2004). Relational attention models expanded on this concept by capturing both local details and global context within images.

Recent advancements in image captioning have seen a rise of transformer-based models, such as Bidirectional Encoder Representation from Transformer (BERT) visual attention (Kelvin et al., 2015) and ViLBERT (visual and language BERT), which have significantly improved performance. These models use transformer architectures to capture long-range dependencies and contextual relations in visual and textual data of inputs (Vaswani et al., 2017). Techniques integrating visual and textual features parallel achieve more accurate and context-relatable captioning.

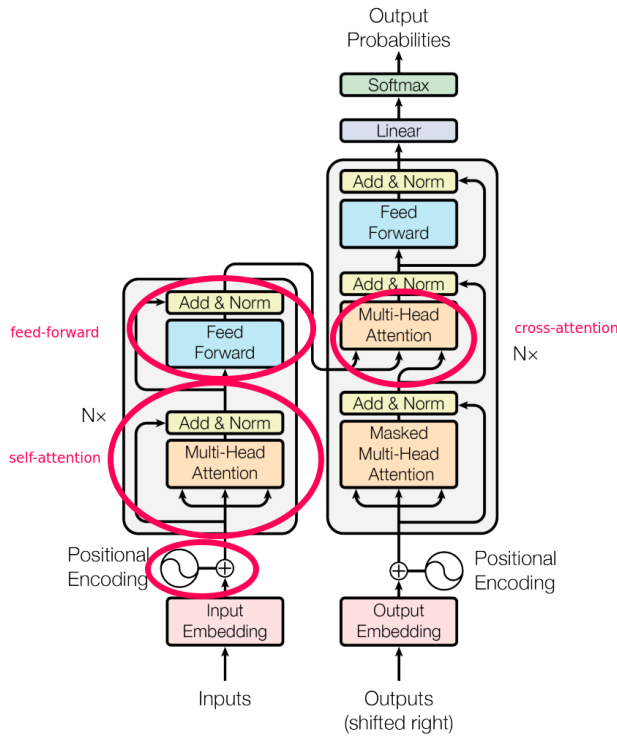
Looking forward, future research in image captioning will be used to address important challenges and explore new implementations. One critical area of focus is on developing more explainable and interpretable models. Reliability is crucial for understanding and confirming decision-making processes of image captioning systems, improving trust and usability in practical tasks, which was helped by metrics (Mikolov et al., 2013). The growing importance of creating deployable solutions for real-world applications, domains such as assistive technologies, content retrieval, and human-computer interaction is growing.

### Existing System

#### Transformer-based Models

Transformer models with a more relative-based approach performed well in various NLP tasks and are increasingly applied in image Captioning.

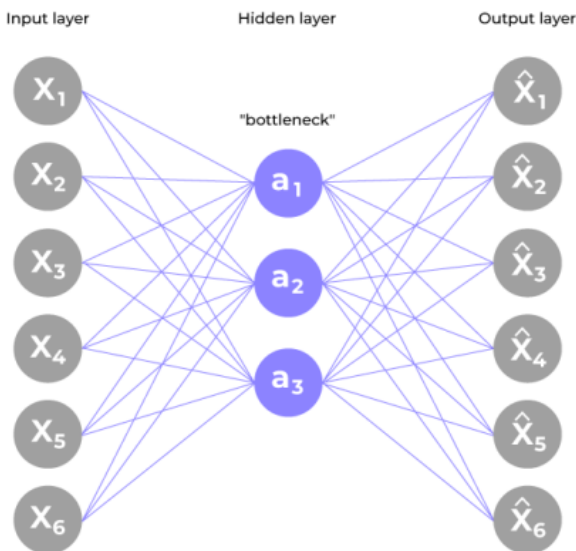
A transformer model is a deep learning model with the implementation of encode and decode processes, which have quickly become fundamental models for many tasks in natural language processing (NLP), and have been applied to a wide range of tasks in other relevant sectors. The whole encoder-decoder combination of the transformer is shown in Figure 1.



**Figure 1. Transformer Architecture (Vaswani et al., 2017).**

**Auto Encoders:**

Autoencoders are used to fine-tune the input and representation vectors in the generation of more coherent image captions. Give the current output as input and iterate until a certain threshold of satisfaction metrics is met. Simple implementation of Autoencoders, as shown in Figure 2 (Kingman D et al., 2014).



**Figure 2. Layers in Autoencoders implementation.**

**Materials and Methods**

The proposed system employs advanced techniques to generate descriptive captions for images using the COCO dataset. Initially, captions are converted into word embeddings, while images are processed through the

InceptionV3 CNN to extract feature vectors. To handle the complexity of these inputs, auto-encoders are used for dimensionality reduction. During training, the model aligns image features with captions, refining the encoder-decoder architecture within a transformer model to minimize error. Finally, the trained model generates captions for new images, starting with a predefined tag and continuing until the end tag occurs.

**Data Accession**

Data Acquisition will be crucial in building a strong image captioning model. This study used the COCO (Common Objects in Context) dataset, which includes everyday scenes, objects, and actions.

The COCO dataset is a base dataset for this, with over 200,000 image inputs, each with at least five different captions. The dataset is divided into three splits for training (80% of images) and validation (20% of images).

The data set is obtained from Kaggle, and the contents are placed in the project structure, from which the data is accessed and worked at run time.

**Feature Extraction**

As we are using a transformer-based model, the features that are to be considered and given more acknowledgment will be dealt with by the attention mechanism that is being applied at the time of the training.

The images that were fed will get manipulated at the time of training by the model of CNN in which inception\_v3\_MODEL, a pre-trained model used for classification, is used except for the fact that we are using up to the last layer, which is not included and we will use the output feature vector created for the part which will be relating to the images.

At the time the images are fed to the CNN\_Encoder model, the features will be given the weights according to their importance and then the processed vector form is generated at the pre-final layer of the inception\_v3 model used.

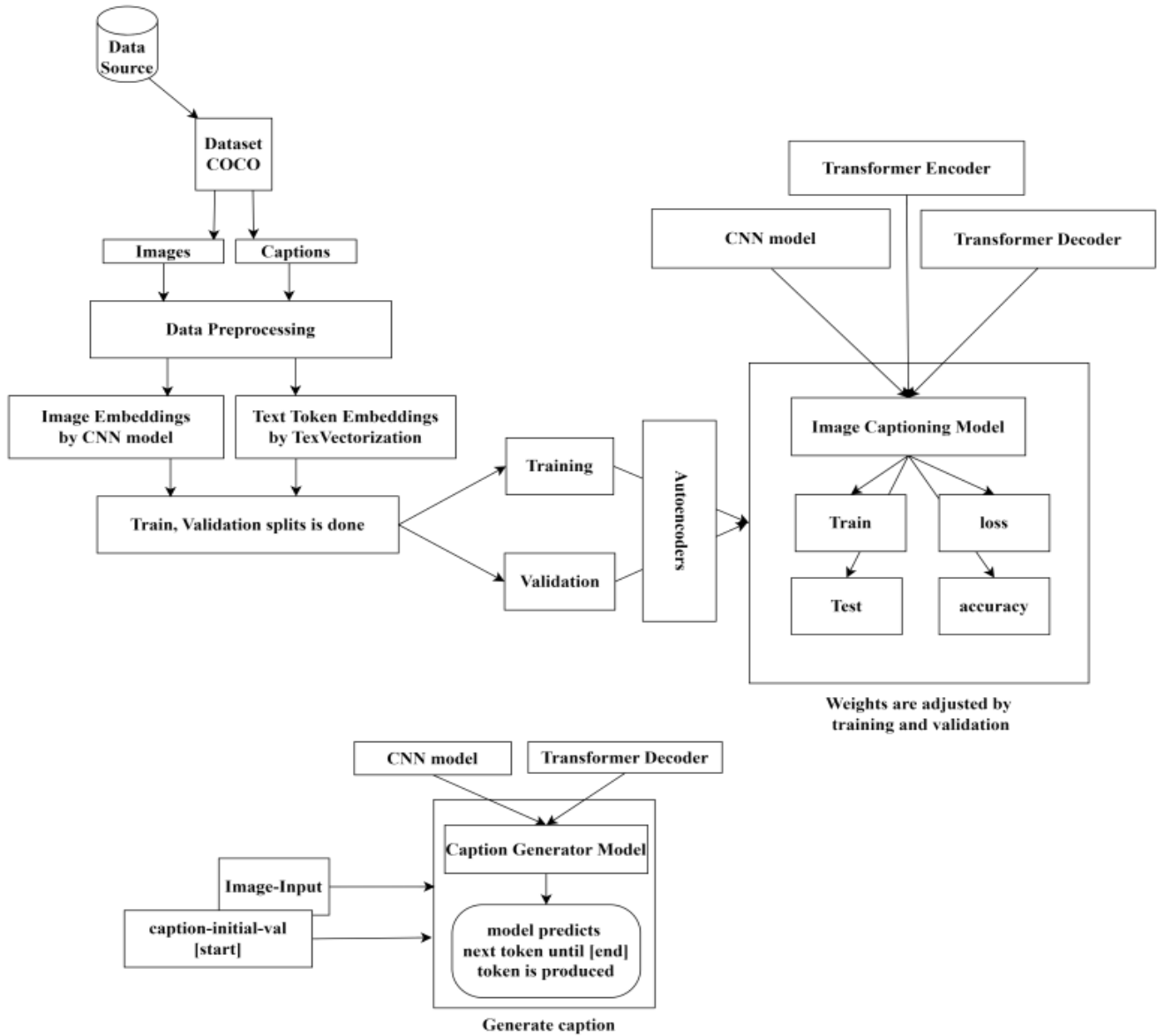
**Training Image Captioning Models**

**CNN Encoder**

Fully connected layer - Flattening - The feature map is flattened into 1-dimensional vector - with length  $H*W*C$ . Using an autoencoder, a fully connected neural network layer is applied to reduce dimensionality (Krizhevsky et al., 2017).

$$f' = FC(Flatten(F)) \tag{1}$$

Here, in equation (1) FC is the fully connected layer, F is the feature map, Flatten() - function to convert the 3\*d output to a single flattened input,  $f'$  encoded output after applying fully connected layer.



**Figure 3. System Model to approach Image Captioning.**

The feature map is denoted with dimensions  $H \times W \times C$  with h-height, w-width, and c-number of channels. The feature map is 3 dimensional representation that helps capture all the important features of the images that are trained (Johnson et al.,2016).

**Transformer Architecture**

**Transformer Encoder**

The Transformer Encoder processes the input sequence, usually using tokenized text or embeddings, and updates the weights in the model encoder. It consists of 2 normalization layers and an Attention network layer with a dense layer with fully connected neurons to provide the attention output (Cho et al., 2014; Kelvin et al., 2015).

**Input Embeddings**

Initially, input tokens or words are embedded into high-dimensional vectors. These embeddings capture semantic meaning and syntactic information of the input sequence.

**Multi-Head-Self-Attention Methodology:**

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \tag{2}$$

Here, in equation (2),  $z_1, z_2, \dots, z_k$ : vector output by a neural networks layer,  $\sigma(z_i)$  is the probability corresponding to the input  $z_i$ , The sum of exponentials of all K logits acts as the normalization factor.

This is the most important part of the Transformer model. Self-attention allows each word in the input sentence to address all other words at a time, captures dependency levels and relationships between words and makes weights representing the relations.

Attention Scores: Determine how much focus each word should place on other words. The softmax (equation 2) function normalizes these scores to obtain the attention weights.

**Layer Normalization**

Normalizes the output of the attention mechanism across the feature dimension, which reduces computation complexity while training, ensuring stable training and faster convergence.



### Feed-Forward Neural Network

A fully connected network will be applied to each position separately and identically, and the outputs will be forwarded as inputs for the next layer. Helps capture complex patterns in the input sequence.

### Residual Connection and Normalization In Network Layers

Add residual connections around each sub-layer to make the system more diverse, followed by layer normalization. This helps in mitigating the vanishing gradient problem and speeds up training.

### Transformer Decoder

The Transformer Decoder processes with embedding starting the process followed by 2 attention layers, 3 normalization layers, 2 fully layers with fully connected nature and 2 dropout layers. Finally, a normalize layer which produces the output through a fully connected layer (Vaswani et al., 2017).

### Layer Normalization, Feed-Forward Neural Network, and Residual Connections:

Similar to the encoder, these components are used to process the decoder inputs and generate the final output sequence.

### Output Layer

The output of the decoder is passed through a linear transformation and a softmax (equation 2) activation function to generate probabilities over the target vocabulary. This predicts the next token in the output sequence. This is performed until the [end] token is predicted.

### Performance Analysis and Results

We use a range of metrics to assess the quality of the generated captions, including BLEU, METEOR, CIDEr and ROUGE.

#### Precision (P):

Ratio of correctly generated relevant captions to total generated captions.

#### Recall (R):

The ratio of correctly generated relevant captions to total relevant captions in dataset.

#### F1-Score:

Harmonic means of precision and recall.

### BLEU (Bilingual Evaluation Understudy)

This measures the similarity by taking into consideration of the generated caption and the reference caption using n-gram precision. We report BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores, which correspond to 1, 2, 3, and 4-gram precision, respectively (Papineni, 2002).

$$BLEU_{-n} = BP \cdot \exp(\sum_{n=1}^N w_n \log p_n) \quad (3)$$

Here, in equation 3,

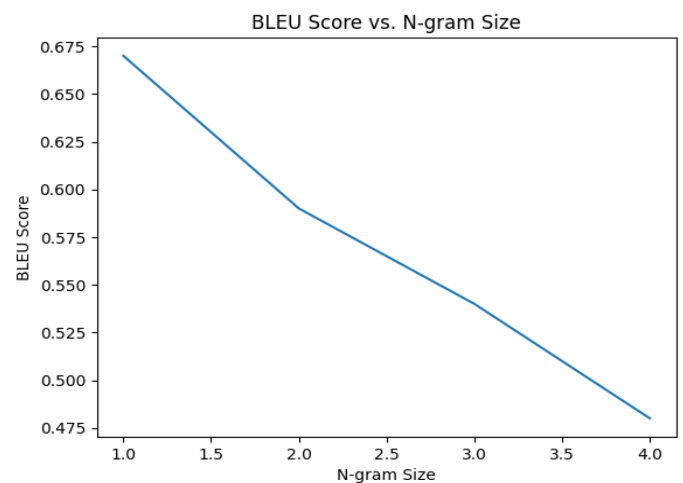
$w_n$ - the weight assigned to each n-gram

$p_n$ - the precision of n-gram.

$BP$  - Brevity Penalty.

**Table 1. Scores obtained by BLEU Metric.**

N-gram Size	Precision (%)	Recall (%)	F1-Score (%)	BLEU Score
1-gram	85.7	90.1	87.9	0.857
2-gram	78.3	82.5	80.3	0.782
3-gram	70.2	74.1	72.1	0.702
4-gram	61.8	65.4	63.5	0.618



**Figure 4. Line plot of how BLEU scores vary w.r.t N-Gram size.**

In BLEU metric, the scores are calculated according to n-grams. Figure 4 shows that the score is high if we consider the n value to be one. This means we will be calculating 1-grams precision, recall and f1 score and this process will be comparing the correctly predicted word present in the real caption.

### METEOR (Metric for Evaluation of Translation with Explicit Ordering)

Calculates the similarity between the generated caption and the reference caption based on the harmonic mean(precision) and harmonic mean(recall) of uni-grammatches (Lavie et al., 2007).

$$METEOR = F_{mean}(1 - P_{frag}) + \gamma \cdot P_{frag} \cdot (1 - \frac{F_{mean}}{P_{frag}}) \quad (4)$$

Here, in equation 4,

$F_{mean}$ - harmonic mean of precision and recall.

$P_{frag}$ - the fragmentation penalty.

Gama ( $\gamma$ ) - weight parameter.

$1 - \frac{F_{mean}}{P_{frag}}$  - captures the relation between the f-score and fragmentation penalty.

Table 2. Scores obtained by METEOR Metric.

Caption Length (Tokens)	Precision (%)	Recall (%)	Harmonic Mean (F-score)	Fragmentation Penalty	METEOR Score
5-10	85.1	80.4	82.7	0.12	0.726
11-15	78.2	75.6	76.9	0.15	0.693
16-20	74.3	70.1	72.2	0.19	0.664
21-25	70.9	66.5	68.6	0.21	0.639

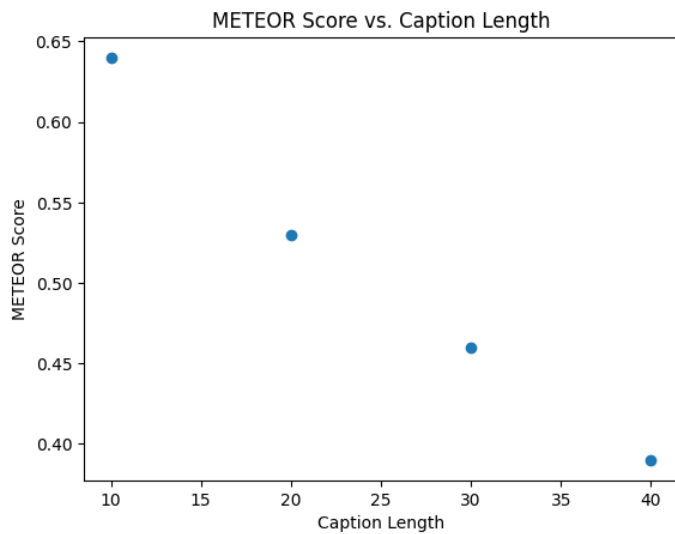


Figure 5. Scatter plot of how METEOR scores vary with caption length.

**CIDEr (Consensus-based Image Description Evaluation)**

It is the measure of similarity between generated caption and reference caption using TF-IDF weighted n-gram similarity(Vedantam R et.,2015).

$$CIDE_r = \frac{1}{m} \sum_{i=1}^m \frac{C_i \cdot C_i^{ref}}{||C_i|| ||C_i^{ref}||} \quad (5)$$

Here, in equation 5,

$C_i$ - TF-IDF vector of the predicted caption.

$C_i^{ref}$  - TF-IDF vector of the reference or actual caption.

m represents the size of these n-grams.

Table 3. Scores obtained by CIDEr Metric.

Caption Length (Tokens)	TF-IDF (Predicted)	TF-IDF (Reference)	CIDEr Score
5-10	0.85	0.90	0.86
11-15	0.78	0.82	0.80
16-20	0.71	0.74	0.72
21-25	0.65	0.69	0.67

CIDEr Score vs. N-grams

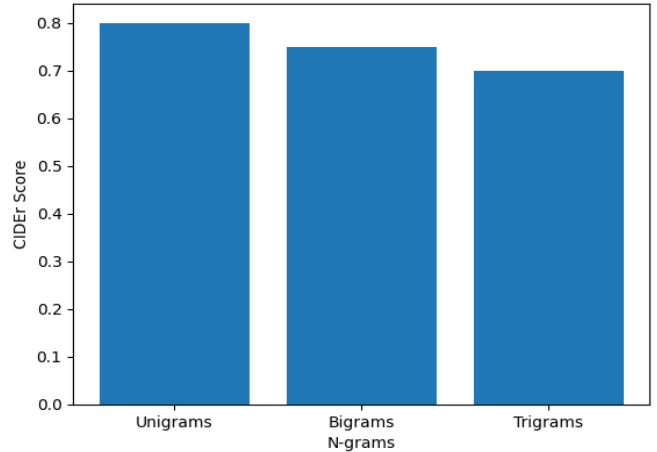


Figure 6. Bar Plot of how CIDEr scores vary the value of m- 1,2,3.

**SPICE F-score Evaluation**

The SPICE F-score is a metric used to evaluate the semantic propositional content of generated captions. It measures the similarity between the generated caption and the reference caption regarding semantic propositional content (Anderson et al., 2016).

$$SPICE = \frac{1}{|G|} \sum_{G_i \in G} F1(G_i^{pred}, G_i^{ref}) \quad (6)$$

In equation 6,  $G_{pred}$  and  $G_{ref}$  are the semantic proportional graphs in the generated and actual captions.

Table 4. Scores obtained by SPICE Metric.

Caption Length (Tokens)	Precision (%)	Recall (%)	SPICE F-score
5-10	79.2	83.5	81.3
11-15	73.8	78.2	75.9
16-20	68.9	73.3	71.0
21-25	62.7	66.9	64.7

SPICE F-score Confusion Matrix

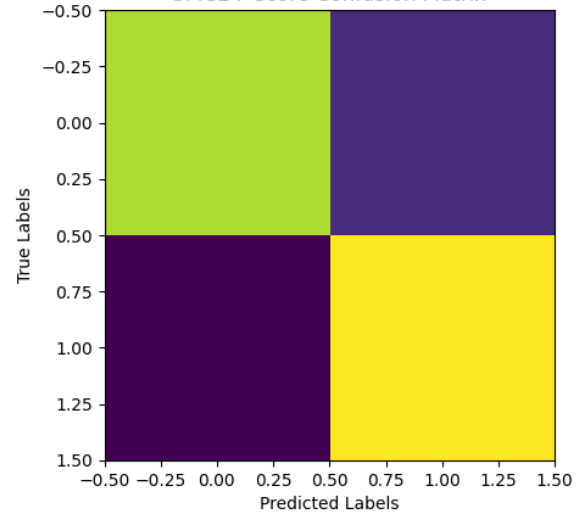
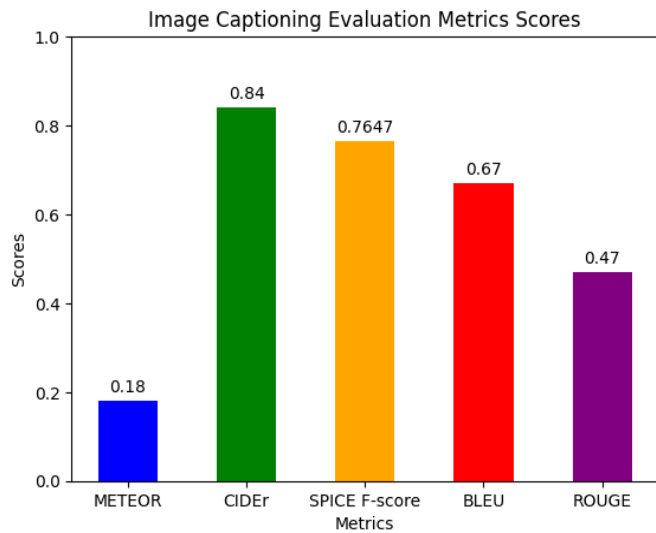


Figure 7. Heat Map representing SPICE Predicted vs True Values.



**Figure 8. Bar graph visualizing the models' scores on different metrics.**

### Conclusion

In this study, we have dealt with improving the world of image captioning by utilizing Transformers, Auto-encoders, and Convolutional Neural Networks (CNNs). Through our research, we aimed to get our hands on methodologies, algorithms, and advancements in this field, paving the way for improved understanding and application of cutting-edge techniques. Our investigation commenced with an overview of related research, tracing the evolution of image caption generation from early template-based methods to the latest transformer-powered models. We used fundamental components of these models, including CNNs, RNNs, attention mechanisms, and transformer architectures, with their roles in generating accurate and relevant captions for images.

In our quest for model evaluation, we used the metrics related to NLG assessment. These methods provided invaluable insights into the contributions of individual features, enhancing our understanding of model predictions and the trustfulness of the model. As we conclude this study, we reflect on strides made in the field of image captioning, recognizing the potential for later advancements and innovations. Looking ahead, the future scope of this research deals with several possibilities for exploration and enhancement.

**Enhanced Model Architectures:** Continual refinement and optimization of model architectures, leveraging advancements in deep learning and neural network techniques to improve captioning accuracy and efficiency. **Integration of Inputs:** Exploring the integration of textual and visual models to produce more coherent, relevant and contextually rich captions gives a deeper understanding of image content.

### Conflict of interest

None

### References

- Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). SPICE: Semantic propositional image caption evaluation. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 382-398. <https://doi.org/10.48550/arXiv.1607.08822>
- Bahdanau, D., Cho, K., & Bengio, Y. (2016). Neural machine translation by jointly learning to align and translate. *ICLR 2015*. <https://doi.org/10.48550/arXiv.1409.0473>
- Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 157-166. <https://doi.org/10.1109/72.279181>
- 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. Doha, Qatar. Association for Computational Linguistics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724-1734. <https://doi.org/10.48550/arXiv.1406.1078>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1*, 4171-4186. <https://doi.org/10.48550/arXiv.1810.04805>
- Gers, F.A., Schmidhuber, J., & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12, 2451-2471. <https://doi.org/10.1162/089976600300015015>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Johnson, J., Karpathy, A., & Fei-Fei, L. (2016). DenseCap: Fully convolutional localization networks for dense captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4565-4574. <https://doi.org/10.1109/CVPR.2016.494>.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, Yoshua Bengio (2015). Show, attend and tell: Neural image caption generation with visual

- attention. *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 37, 2048-2057. <https://dl.acm.org/doi/10.5555/3045118.3045336>
- Keerthana, B., Vamsinath, J., Kumari, C. S., Appaji, S. V. S., Rani, P. P., & Chilukuri, S. (2024). Machine Learning Techniques for Medicinal Leaf Prediction and Disease Identification. *International Journal of Experimental Research and Review*, 42, 320–327. <https://doi.org/10.52756/ijerr.2024.v42.028>
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ICLR 2015* <https://doi.org/10.48550/arXiv.1412.6980>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84 – 90. <https://doi.org/10.1145/3065386>
- Lavie, A., & Agarwal, A. (2007). METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. *Proceedings of the Second Workshop on Statistical Machine Translation, Association for Computational Linguistics*, pp. 228-231. <https://dl.acm.org/doi/10.5555/1626355.1626389>
- Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain*, pp. 74–81, 25–26. <https://aclanthology.org/W04-1013.pdf>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 3111-3119. <https://doi.org/10.48550/arXiv.1310.4546>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311-318. <https://doi.org/10.3115/1073083.1073135>
- Rao, M. S., Sekhar, C., & Bhattacharyya, D. (2021). Comparative analysis of machine learning models on loan risk analysis. *In Advances in intelligent systems and computing*, pp. 81–90. [https://doi.org/10.1007/978-981-15-9516-5\\_7](https://doi.org/10.1007/978-981-15-9516-5_7)
- Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1179-1195. <https://doi.org/10.1109/CVPR.2017.131>
- Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 4278-4284. <https://dl.acm.org/doi/10.5555/3298023.3298188>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 1-11. <https://doi.org/10.48550/arXiv.1706.03762>
- Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDER: Consensus-based image description evaluation, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566-4575. <https://doi.org/10.1109/CVPR.2015.7299087>
- 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. <https://doi.org/10.48550/arXiv.1411.4555>
- Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J. J., & Gao, J. (2020). Unified vision-language pre-training for image captioning and VQA. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07), 13041-13049. <https://doi.org/10.1609/aaai.v34i07.7005>

### How to cite this Article:

Selvani Deepthi Kavila, Moni Sushma Deep Kavila, Kanaka Raghu Sreerama, Sai Harsha Vardhan Pittada, Krishna Rupendra Singh, Badugu Samatha and Mahanty Rashmita (2024). Image Captioning with Convolutional Neural Networks and Autoencoder-Transformer Model. *International Journal of Experimental Research and Review*, 46, 297-304.

DOI : <https://doi.org/10.52756/ijerr.2024.v46.023>



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.