AUTHORS

*Kartikey Sapkal 202201040077*
*Yash Gunjal 202201040106*
*Arhant Nitanware 202201040062*

*Guided by-Diptee Chikmurge*

# Image Captioning

*"Image Captioning", comparing Without Attention(LSTM),With Attention(Bahdanau )andWith Self-Attention Transformer architectures*

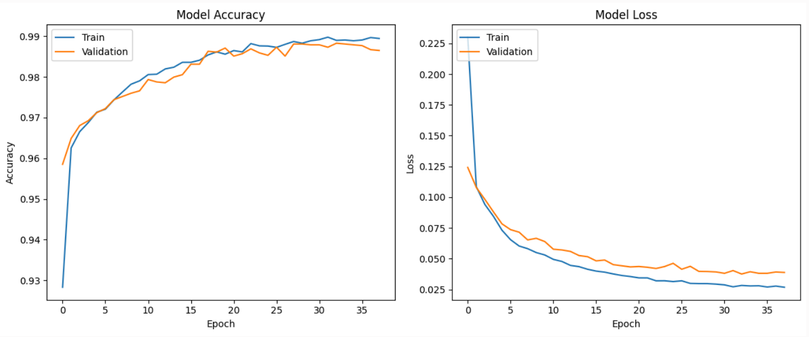
DEEP LEARNING

*Dataset Link :*
*https://www.kaggle.com/datasets/adityajn105/flickr8k*

*Reserch paper -*
*https://ijrti.org/papers/IJRTI2208183.pdf*

## *01. Abstract*

Image captioning, an essential vision-language task, aims to generate natural language descriptions for images. This project evaluates and compares three neural architectures—Vanilla LSTM, Bahdanau Attention, and Transformer-based models—on the Flickr8k dataset. Preprocessing includes tokenization, sequence padding, and normalization. Performance is assessed using standard metrics like BLEU and loss accuracy curves.




## *02. Objective*

- To implement and compare three deep learning-based image captioning models:
1. Vanilla Encoder-Decoder
2. Attention-based Encoder-Decoder
3. Self-Attention (Transformer-based) Decoder
- To evaluate their performance using standard captioning metrics.

## *Related literature*

Early image captioning used rule-based or retrieval methods. Encoder-decoder architectures brought major advances, and attention mechanisms improved focus on key image regions. Transformers further enhanced performance through self-attention and parallelism.

## *03. Methodology*

Dataset: Flickr8k

- Contains 8000 images, each with 5 human-annotated captions
- Labels created for normal vs. anomalous captions (optional for anomaly study)

.

### *Preprocessing:*

- Text cleaning, lowercasing
- Vocabulary building and tokenization
- Sequence padding and embedding matrix generation

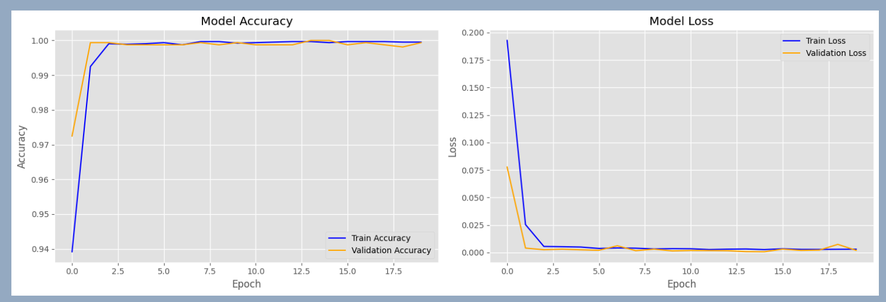
- **Vanilla LSTM: Basic encoder-decoder structure with fixed context vector**
- **Bahdanau Attention: Enables dynamic context vector with weighted importance**
- **Transformer: Self-attention layers capture global dependencies efficiently**

## *04. Results/Findings*

- Vanilla Model: Baseline with limited contextual accuracy
- Attention Model: Better localization and relevant captions
- Self-Attention Model: Best performance, especially with longer captions

## *05. Analysis*

The LSTM model performed well but showed slight overfitting and struggled with long-range dependencies. Bahdanau Attention improved focus and convergence, achieving the highest accuracy. The Transformer generalized well but needed more training time. Overall, attention mechanisms significantly enhanced anomaly detection.




## *06. Conclusion*

Transformer-based models deliver superior results in terms of both contextual relevance and fluency. Attention mechanisms significantly improve performance by allowing the model to focus on key regions in images. This research emphasizes architectural advancements in caption generation. Future work may explore larger datasets and multimodal fusion to further enhance captioning accuracy. Additionally, integrating feedback loops or reinforcement learning could make caption generation more adaptive and user-aligned.

## *07. Future Scope*

- Train on larger datasets like MS-COCO
- Integrate audio or location data (multimodal)
- Use Reinforcement Learning for adaptive feedback
- Optimize for mobile/edge deployment

*"A picture is worth a thousand words, but deep learning teaches machines to say them."*