IMAGE CAPTIONING

MIT Academy of Engineering

A deep learning task where a model generates descriptive captions for images

PROBLEM STATEMENT

To compare image captioning models and assess how attention mechanisms enhance the accuracy and relevance of generated captions.

ABSTRACT

This project explores image captioning using deep learning architectures: CNN+LSTM (no attention), CNN+RNN with Bahdanau attention, and Transformer with self-attention. The models were trained and evaluated on the MS-COCO dataset using BLEU, METEOR, ROUGE-L, CIDEr, and SPICE metrics. Results show that attention-based models generate more fluent and accurate captions, with Transformer outperforming the rest in all key metrics.

DATASET

MS-COCO 2017

- Contains 118,000+ images with 5 human-written captions each.
- https://www.kaggle.com/datasets/awsaf49/coco-2017-dataset

Flickr8k

- Contains 8,000 images, each with 5 captions describing people and actions in the scene.
- https://www.kaggle.com/datasets/adityajn105/flickr8k?select=Images.

INTRODUCTION

This project investigates the effectiveness of attention-based encoder-decoder architectures in generating image captions. We implement and compare two models: a traditional CNN + LSTM (without attention) and a Transformer-based model (with self-attention).

METHODOLOGY

1. Baseline (CNN + LSTM):
   - This model simply encodes the image and feeds it to an LSTM decoder without focusing on specific parts of the image at each word-generation step. Hence, no attention is applied.
2. Attention models (Bahdanau or Luong):
   - These introduce a mechanism to focus on different image regions dynamically during caption generation.
3. Transformer (Self-Attention):
   - Uses self-attention throughout the encoder and decoder, allowing it to model long-range dependencies and fine-grained spatial understanding.

RESULT

- **BLEU** (Bilingual Evaluation Understudy)
- **METEOR, ROUGE-L, CIDEr, SPICE**

Metric	LSTM/GRU	Attention	Transformer
BLEU-4	0.1101	0.3360	0.5795
METEOR	0.2154	0.2510	0.3969
CIDEr	0.2895	1.068	1.7444
SPICE	0.1541	0.1810	0.3279

APPLICATIONS

- E-commerce

Describe product images for cataloging, accessibility, and SEO enhancement.

- Human-Robot Interaction

Enable robots to narrate what they "see" for better contextual understanding.

- Automatic Image Description for the Visually Impaired

Generate natural-language descriptions to help visually impaired users understand images.

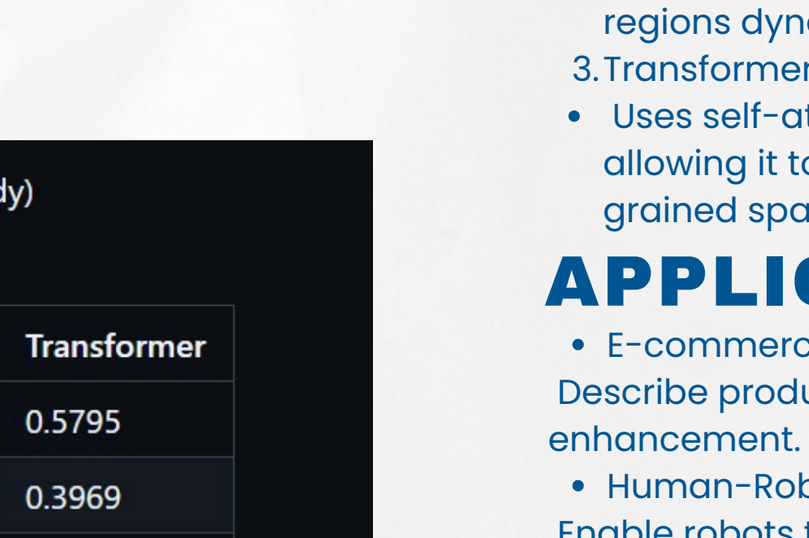
ANALYSIS

Performance

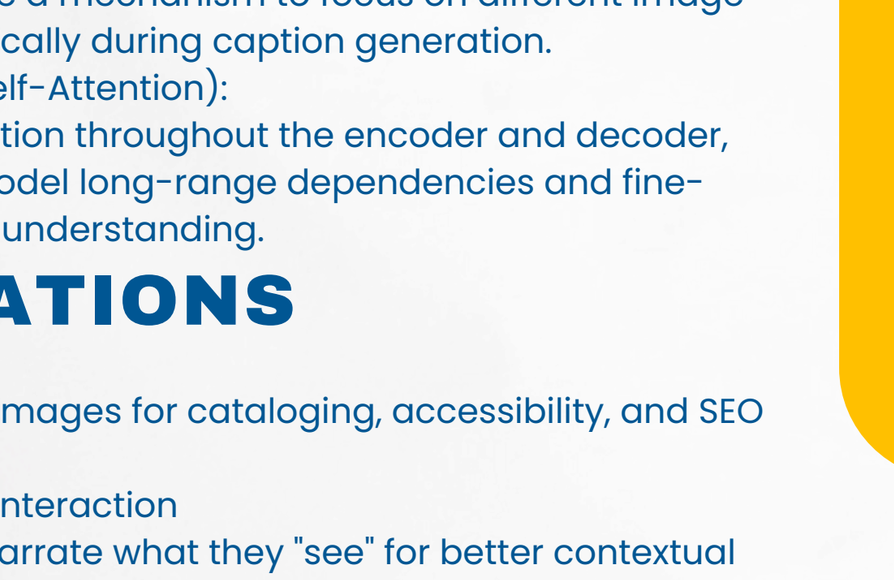
- Transformer achieves the highest BLEU, CIDEr, and SPICE scores. Attention LSTM performs better than baseline with improved context understanding.
- Baseline LSTM has the lowest scores, suitable only for basic tasks.

Efficiency

- Baseline LSTM is fastest and lightest.
- Attention LSTM offers a balance of speed and accuracy. Transformer is slowest but most accurate.



WITHOUT ATTENTION



SELF ATTENTION

CONCLUSION

- Use Transformer for best results when accuracy is key.
- Use Attention LSTM for balance between performance and interpretability.
- Use Baseline LSTM for resource-limited environments.

Team:

1. **Priyanka Kadam-202201060018**
2. **Aditi Kulkarni-202201070046**
3. **Yathang Tupe-202201070076**