MIT | Academy of Engineering

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

# Image Captioning using Encoder Decoder

Project Guide:
Dr. Sunita Barve

Team Members:
Suryansh Ambekar (202201090042)
Kaustubh Mahajan (202201070128)
Ayush Fating (202201070127)

# Introduction

Image captioning is the task of automatically generating descriptive text for images by combining computer vision and natural language processing. It typically uses an encoder-decoder framework, where the encoder extracts visual features and the decoder generates a caption. This technology is useful in applications like image retrieval, accessibility tools, and automated content generation.

# Objectives

1. Implement and compare three architectures:
- Without Attention
- With Bahdanau Attention
- Self Attention Transformer

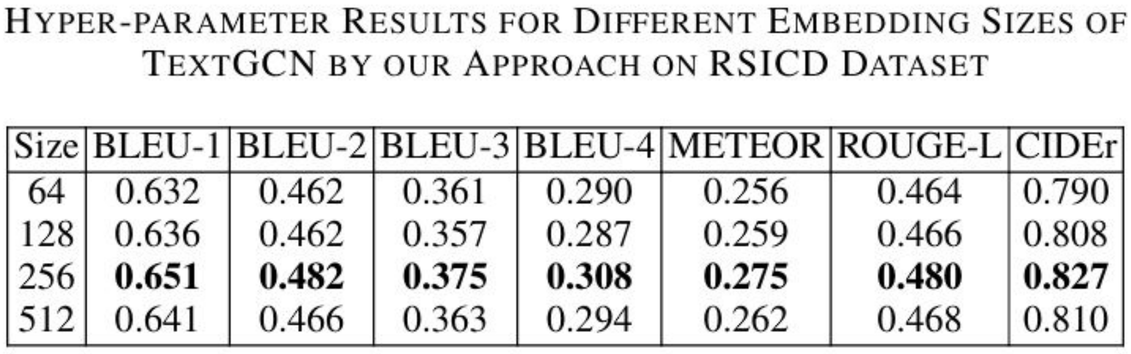
2. Train and evaluate models on RSICD(Remote Sensing Image Captioning Dataset) dataset.

3. Analyze performance using standard performance metrics like BLEU 1, BLEU 4, ROGUE-L, METEOR, CIDEr.

# Baseline Research Paper

**A TextGCN-Based Decoding Approach for Improving Remote Sensing Image Captioning**

Year: 2024 (arXiv version posted October 2024)
Objective: Improve the quality of captions for remote sensing images
Method: Combines TextGCN (for word embeddings) with a multi-layer LSTM decoder



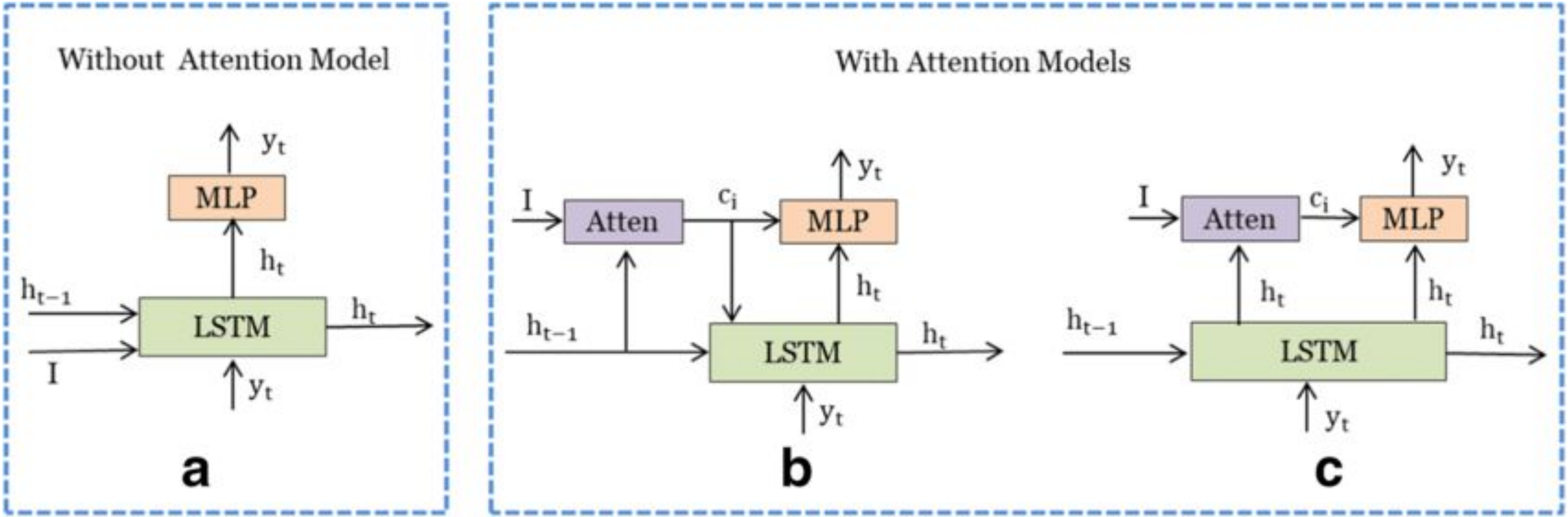
HYPER-PARAMETER RESULTS FOR DIFFERENT EMBEDDING SIZES OF TEXTGCN BY OUR APPROACH ON RSICD DATASET

| Size | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|---|---|
| 64 | 0.632 | 0.462 | 0.361 | 0.290 | 0.256 | 0.464 | 0.790 |
| 128 | 0.636 | 0.462 | 0.357 | 0.287 | 0.259 | 0.466 | 0.808 |
| 256 | **0.651** | **0.482** | **0.375** | **0.308** | **0.275** | **0.480** | **0.827** |
| 512 | 0.641 | 0.466 | 0.363 | 0.294 | 0.262 | 0.468 | 0.810 |

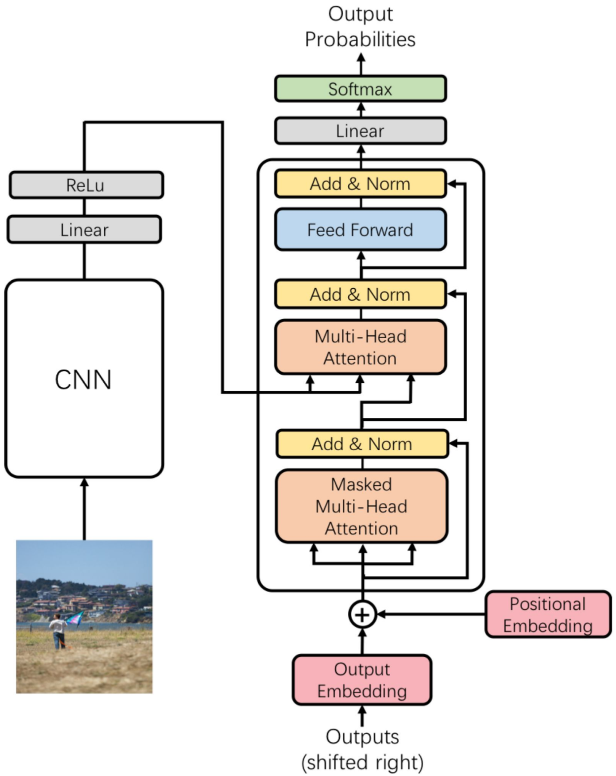
# Dataset Description

- RSCID(Remote Sensing Image Captioning Dataset)
- Link: https://paperswithcode.com/dataset/rsicd
- Total records: 1,0921 images
- Content: High-resolution remote sensing (satellite) images
- Categories: Includes diverse land-use scenes like airports, residential areas, farmlands, forests, etc.

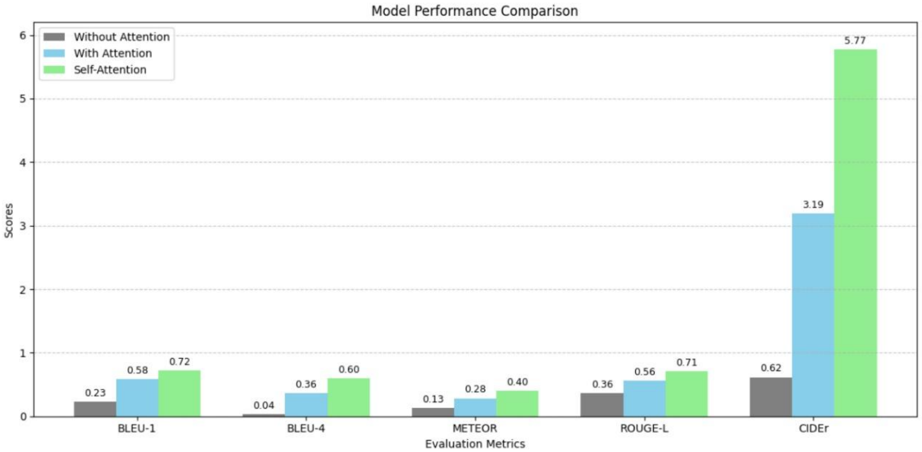
# Architecture Model

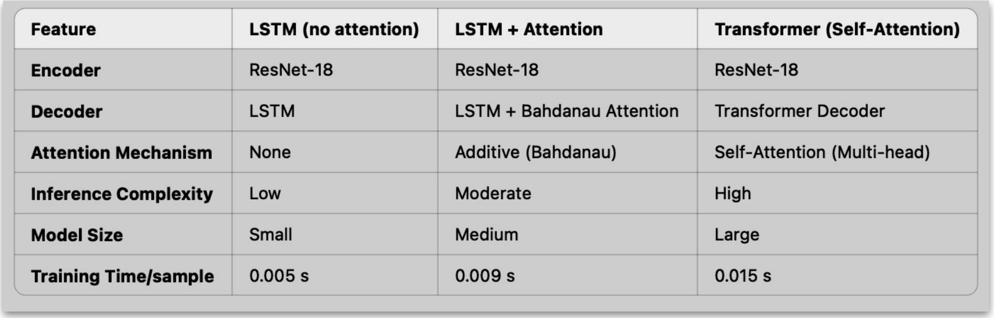


# Architecture Model



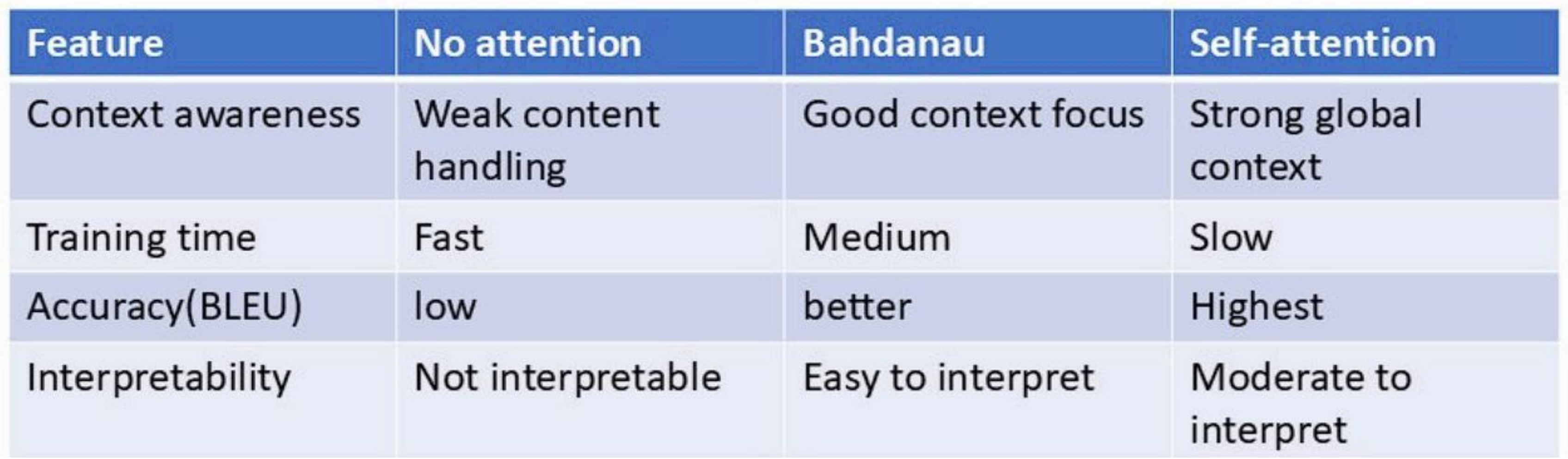
# Results & Graphs





| Feature | LSTM (no attention) | LSTM + Attention | Transformer (Self-Attention) |
|---|---|---|---|
| **Encoder** | ResNet-18 | ResNet-18 | ResNet-18 |
| **Decoder** | LSTM | LSTM + Bahdanau Attention | Transformer Decoder |
| **Attention Mechanism** | None | Additive (Bahdanau) | Self-Attention (Multi-head) |
| **Inference Complexity** | Low | Moderate | High |
| **Model Size** | Small | Medium | Large |
| **Training Time/sample** | 0.005 s | 0.009 s | 0.015 s |

# Analysis table and discussion



| Feature | No attention | Bahdanau | Self-attention |
|---|---|---|---|
| Context awareness | Weak content handling | Good context focus | Strong global context |
| Training time | Fast | Medium | Slow |
| Accuracy(BLEU) | low | better | Highest |
| Interpretability | Not interpretable | Easy to interpret | Moderate to interpret |

# Conclusion

The integration of self-attention mechanisms, particularly through Transformer models, has led to significant improvements in image captioning performance. These models enhance the contextual relevance and accuracy of generated captions by effectively capturing long-range dependencies. In comparison to the baseline model presented in the paper, the Transformer model achieved higher BLEU-1 and BLEU-4 scores, demonstrating its superior capability in generating more accurate and coherent image descriptions.