Subject: DEEP Learning MDM

Guided by:
Dr. Sunita Barve

MIT | Academy of Engineering
(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

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

IMAGE CAPTIONING

Image captioning uses computer vision and language processing to describe images in text. It typically has an encoder to extract features and a decoder to generate captions. In areas like remote sensing, it's harder due to complex visuals and specialized terms.

REFERENCE PAPER

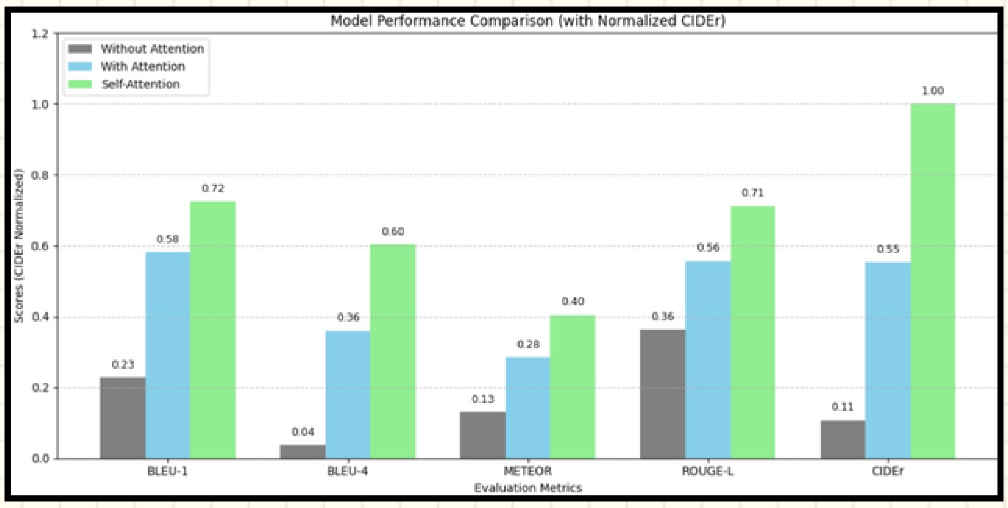
The paper "**A TextGCN-Based Decoding Approach for Improving Remote Sensing Image Captioning**" introduces a TextGCN-based encoder-decoder model with multi-layer LSTMs and comparison-based beam search to improve remote sensing image captioning, achieving top results on the RSICD dataset.

DATASET DESCRIPTION

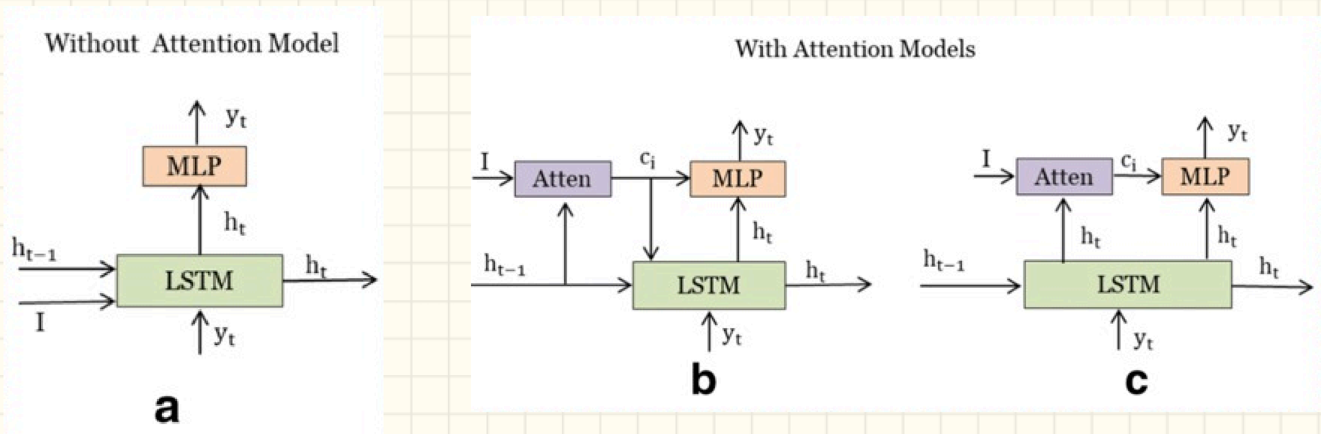
- RSICD dataset (Remote Sensing Image Captioning Dataset) Contains 10921 images
- Categories: Includes diverse land-use scenes like airports, residential areas, farmlands, forests, etc.

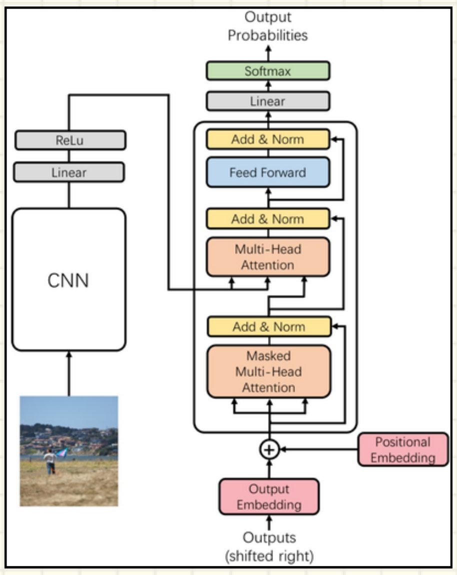
MODELS

- Without attention: LSTM
- With Attention: BAHADANU
- Self Attention



ARCHITECTURE





Transformer Architecture

EVALUATION METRICS

- BLEU 1
- BLEU 4
- ROGUE-L
- METEOR
- CIDEr

- **Without Attention**
  ```
  BLEU-1: 0.2268513466036563
  BLEU-4: 0.03741031958502291
  METEOR: 0.12974948176486473
  ROUGE-L: 0.36337815014538283
  CIDEr: 0.6167209552633228
  ```
- **With Attention (Bahadanu)**
  ```
  BLEU-1: 0.58176645556193
  BLEU-4: 0.3592453719294569
  METEOR: 0.2847761155768596
  ROUGE-L: 0.5569654760906465
  CIDEr: 3.1942525941217625
  ```
- **Self Attention (Transformer)**
  ```
  BLEU-1: 0.7245303513310414
  BLEU-4: 0.6033369125995696
  METEOR: 0.40481117861914684
  ROUGE-L: 0.7116814029470848
  CIDEr: 5.772107762349921
  ```

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

CONCLUSION

- Self-attention-based models significantly improved image captioning quality.
- Attention mechanisms enhance the contextual relevance and accuracy of generated captions.
- Transformer model achieved higher metric scores than the model in the base paper.