Image Captioning

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF

BACHELOR OF TECHNOLOGY

IN

MODERN MACHINE LEARNING

FOR THE ACADEMIC YEAR 2024-2025

Introduction

- Image captioning is the process of automatically generating textual descriptions for images. This technology is crucial for assisting visually impaired individuals, enhancing image search capabilities, and supporting content creation for multimedia platforms.
- Image captioning combines natural language processing and computer vision to produce textual descriptions for images.
- This technology has key applications, such as aiding the visually impaired, improving image search, and enhancing multimedia content creation.
- Techniques like CNNs for feature extraction and RNNs (LSTM networks) for sequence generation have shown significant progress.

Introduction

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

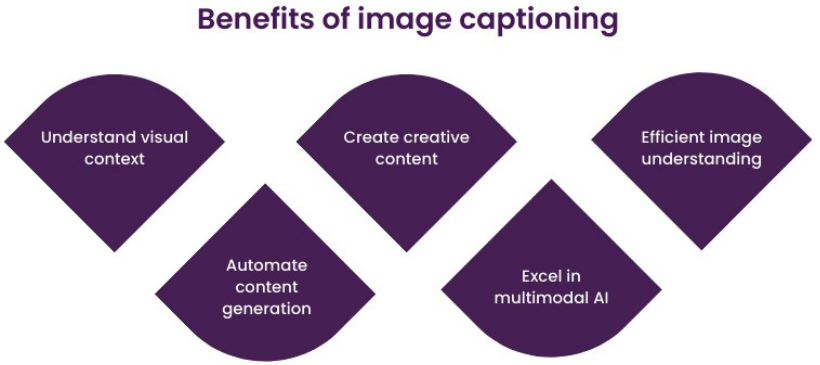
A giraffe standing in a forest with trees in the background.

Objectives of Image Captioning

- Automatic Description Generation: Create meaningful captions without human intervention.
- Contextual Understanding: Recognize relationships and contexts of objects within images.
- Application Domains: Accessibility tools, improved user experience, and enhanced image retrieval systems.

Advancements in Image Captioning

- Advancements in image captioning have been driven by deep learning models. Attention mechanisms in models like 'Show, Attend, and Tell' allow models to focus on important image areas, improving caption quality. BLEU-1 scores on Fickr8k benchmark datasets show significant improvements, moving from 60% to over 75% accuracy in recent models.



Challenges in Image Captioning

- Challenges include captioning images with complex scenes, understanding abstract concepts, and addressing dataset biases. Ongoing research aims to improve accuracy, diversity, and contextual relevance of generated captions.

Literature Survey

1. Zhenghang Yuan et al. (2020): Proposed a multi-level attention module for spatial and scale features extraction, enhancing remote sensing image captioning.
2. Jingqiang Chen et al. (2019): Developed a multi-modal attentional mechanism for news image captioning using an RNN-based decoder.
3. Soheyla Amirian et al. (2020): Focused on encoder for object and feature extraction in CNN, using G-MLE and G-GAN strategies for image captioning.
4. Edy Mulyanto et al. (2019): Utilized CNN and LSTM models for image caption generation across multiple languages, achieving competitive BLEU scores.

Literature Survey

5. Shiru Qu et al. (2017): Implemented LSTM with attention for image captioning, enhancing interpretability and alignment with human intuition.
6. Feng Chen et al. (2019): Addressed challenges in neural network-based models using template-based methods with attributes, scenes, and objects.
7. Minsi Wang et al. (2016): Proposed a parallel-fusion RNN-LSTM architecture for image caption generation, utilizing multiple evaluation metrics.
8. Xinru Wei et al. (2017): Focused on image retrieval by dense caption reasoning, comparing with baseline methods on a large-scale dataset.

Literature Survey

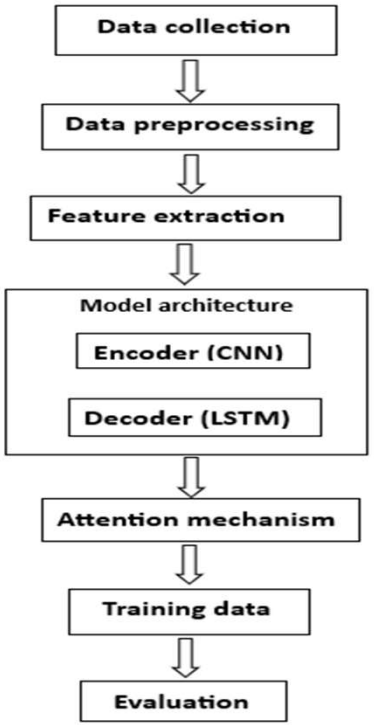
10. Abu Musa Sakib et al. (2024): Proposed CNN and LSTM layers for image caption generation and feature extraction using pre-trained CNN.
11. Genc Hoxha et al. (2020): Developed a CNN-RNN framework for remote sensing image captioning, reviewing existing methods in the field.
12. Binqiang Wang et al. (2020): Focused on image captioning using memory cells and topics, employing template-based and retrieval-based methods.
13. Varsha Kesavan et al. (2019): Proposed a transfer learning approach with a pre-trained VGG16 model encoder and attention mechanisms.

Comparison of Different Models

- **Model Comparison**:

1. CNN + RNN: Combines visual features with language generation - Effective for simple images.
2. LSTM with Attention: Focuses on relevant parts of images - Better interpretability.
3. Template-based: Uses predefined templates - Fast generation but limited diversity.
4. GCN-based: Graph convolution for attributes - Effective for remote sensing.

Flowchart


Data collection
Data preprocessing
Feature extraction
Model architecture
Encoder (CNN)
Decoder (LSTM)
Attention mechanism
Training data
Evaluation

Dataset (flickr8k)

A man and child kayak through gentle waters .

A dog chases a dog toy on the grass .

Two swans glide on river .

Two girls crouch in a small stall .

A group of people are sitting on the steps outside .

Two guys walking , one carrying a skateboard

A white and black dog chases after a decoy-animal on a string .

A little girl on a swing .

A tan dog and a black dog fight .

A chicken and a white dog in the mulch .

Two guys riding skateboards with one of them performing a jump trick .

Two girls dressed like waitresses dance along with a man dressed as a chef .

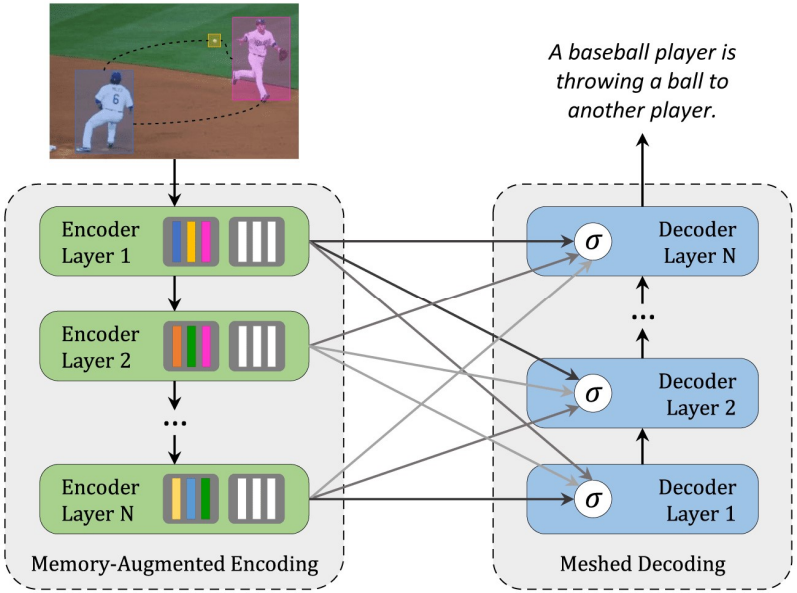
A boy wearing a blue parka , green pants , and white shoes is leaping in a ramp .

A brown dog and a black dog are together in tall grass .

Two women talking on cellphones , with posters behind them .

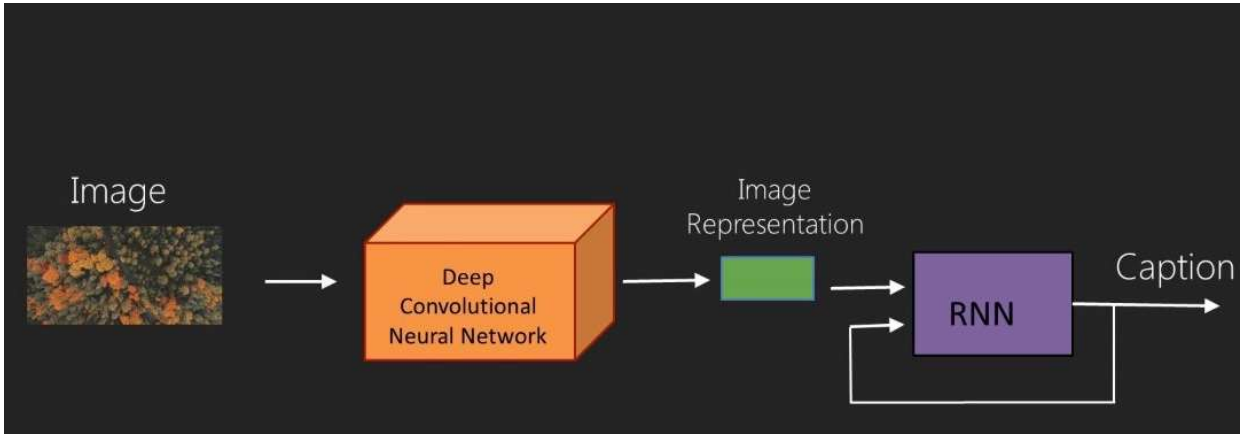
Methodology

- The methodology involves using deep learning frameworks that combine CNNs for visual feature extraction with RNNs/LSTMs for language generation, incorporating attention mechanisms to enhance caption relevance.



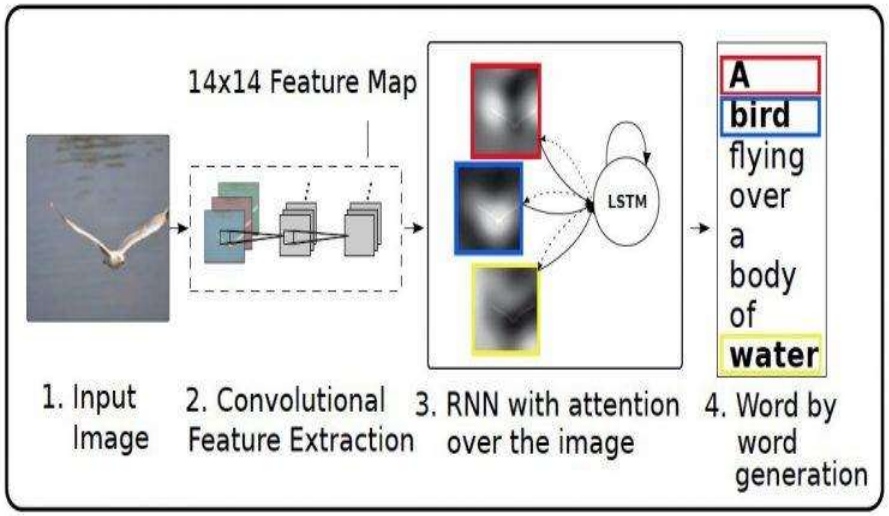
Methodology

- The approach involves CNNs for image feature extraction and RNNs or transformers for generating captions, with attention mechanisms to enhance focus on relevant regions within images. The model will be trained on diverse datasets with captions.



Methodology

- Attention mechanisms allow the model to focus on specific parts of an image, generating contextually relevant captions.
- This approach improves accuracy, especially for complex images.



Data Collection

- Data for image captioning is typically collected from large datasets such as Flickr8k-Images-Captions, which contains images paired with human-generated captions.
- This data is essential for training models to understand the context and relationships within images.

Training Process

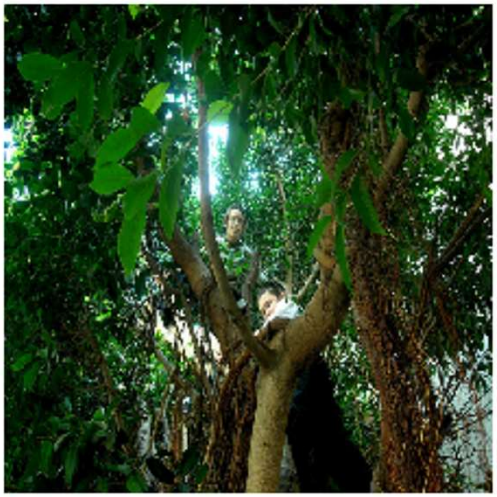
- The training process involves feeding the model pairs of images and their corresponding captions.
- The model learns to generate captions by minimizing the difference between its predictions and the actual captions using loss functions like cross-entropy.

Evaluation Metrics

BLEU (Bilingual Evaluation Understudy) Score:

- Measures the overlap between generated and reference captions. Bleu 1 and Bleu 2 score is calculated for our Image captioning model.
- Evaluates the quality of generated caption by comparing them to reference caption.
- The BLEU-1 and BLEU-2 score specifically measure the precision of single words and word pair respectively.

Results



```
----------Actual----------
startseq children climbing huge tree endseq
startseq two children are sitting in some tree branches endseq
startseq two children look down from green tree they have climbed endseq
startseq two kids in green tree endseq
startseq two people standing in tree on sunny day endseq
----------Predicted----------
startseq two children are standing on tree endseq
```

```
----------Actual----------
startseq black dog is running in the water endseq
startseq black dog running into the water endseq
startseq black dog running through water endseq
startseq black dog runs through the water endseq
startseq black dog splashes through water endseq
----------Predicted----------
startseq black dog is running in the water endseq
```

Challenges in Image Captioning

Despite advancements, image captioning faces challenges such as:

1. **Complex Scenes**: Difficulty in generating accurate captions for images with multiple objects.
2. **Abstract Concepts**: Struggles with understanding and describing abstract ideas.
3. **Bias in Datasets**: Training data may contain biases that affect model performance.

Future Scope:

Future research in image captioning may focus on:

- Improving model interpretability and alignment with human intuition.
- Developing models that can generate diverse and contextually relevant captions.
- Addressing ethical concerns related to bias and fairness in AI.

Applications of Image Captioning

- **Content Creation**: Automating the generation of captions for social media and blogs.
- **Image Search**: Enhancing search engines by providing textual descriptions of images.
- **Education:** Provide educational materials with textual descriptions of visual content, helping to bridge the gap for students with disabilities or those in remote areas with limited access to visual media.

References

1. Z. Yuan, X. Li, and Q. Wang, 'Exploring Multi-Level Attention and Semantic Relationship for Remote Sensing Image Captioning,' IEEE Access, vol. 8, pp. 2608-2620, 2020, doi: 10.1109/ACCESS.2019.2962195.
2. J. Chen and H. Zhuge, 'News Image Captioning Based on Text Summarization Using Image as Query,' 2019 15th International Conference on Semantics, Knowledge and Grids (SKG), Guangzhou, China, pp. 123-126, doi: 10.1109/SKG49510.2019.00029.
3. S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, 'Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap,' IEEE Access, vol. 8, pp. 218386-218400, 2020, doi: 10.1109/ACCESS.2020.3042484.
4. E. Mulyanto, E. I. Setiawan, E. M. Yuniarno, and M. H. Purnomo, 'Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset,' 2019 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Tianjin, China, pp. 1-5, doi: 10.1109/CIVEMSA45640.2019.9071632.

References

5. S. Qu, Y. Xi, and S. Ding, 'Visual attention based on long-short term memory model for image caption generation,' 2017 29th Chinese Control And Decision Conference (CCDC), Chongqing, China, pp. 4789-4794, doi: 10.1109/CCDC.2017.7979342.
6. F. Chen, S. Xie, X. Li, S. Li, J. Tang, and T. Wang, 'What Topics Do Images Say: A Neural Image Captioning Model with Topic Representation,' 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shanghai, China, pp. 447-452, doi: 10.1109/ICMEW.2019.00083.
7. M. Wang, L. Song, X. Yang, and C. Luo, 'A parallel-fusion RNN-LSTM architecture for image caption generation,' 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, pp. 4448-4452, doi: 10.1109/ICIP.2016.7533201.
8. X. Wei, Y. Qi, J. Liu, and F. Liu, 'Image retrieval by dense caption reasoning,' 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, pp. 1-4, doi: 10.1109/VCIP.2017.8305157.

References

9. A. M. Sakib, S. A. Mukta, and M. Y. Hossain, 'Automated Image Captioning System,' April 2024, License CC0, DOI: 10.13140/RG.2.2.24966.79689.
10. G. Hoxha, F. Melgani, and J. Slaghenauffi, 'A New CNN-RNN Framework For Remote Sensing Image Captioning,' 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Tunis, Tunisia, pp. 1-4, doi: 10.1109/M2GARSS47143.2020.9105191.
11. B. Wang, X. Zheng, B. Qu, and X. Lu, 'Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning,' IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 256-270, 2020, doi: 10.1109/JSTARS.2019.2959208.
12. V. Kesavan, V. Muley, and M. Kolhekar, 'Deep Learning based Automatic Image Caption Generation,' 2019 Global Conference for Advancement in Technology (GCAT), Bangalore, India, pp. 1-6, doi: 10.1109/GCAT47503.2019.8978293.

Submitted by

Name	Roll No
Sakshi Warkari	20230015
Himanshi Hatwar	20230008
Prachi Dhore	20230004
Himanshu Katrojwar	20230002
Sakshi Jiwtode	20230006

Under the guidance of

Amit Pandey Sir