"Image Captioning using VGG-16 Deep Learning Model"

[1]VANSHIKA SINGH, [2]SATYANKAR , [3]YUVIKA SINGH, [4]VAIBHAV TIWARI
MENTOR: [5]MR. VISHAL JAYSWAL
[2]2000270100180, [2]2000270100137, [2]2000270100197, [2]2000270100178
[12345]DEPTARTMENT OF COMPUTER SCIENCE AND ENGINEERING
[12345]AJAY KUMAR GARG ENGINEERING COLLEGE, GHAZIABAD

ABSTRACT

- Image captioning integrates computer vision, natural language processing, and deep learning to generate descriptive captions for images, utilizing datasets like Flickr8k and Flickr30k for evaluation.
- The study employs the VGG-16 CNN for feature extraction and a RNN to create coherent and contextually relevant captions.
- Generated captions are evaluated against human-provided references using standard metrics such as BLEU to assess quality.
- The findings have practical applications in content indexing, assistive technology for the visually impaired, and improving user interfaces on image-centric systems, with the comparison of datasets providing insights for future research.

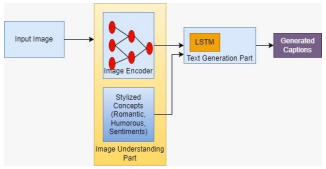

Fig1 .Block Diagram of Image Captioning process

OBJECTIVES

- Implement a caption generation model using a CNN to condition a LSTM language model.
- Focuses on advancing image captioning using deep learning model.
- Accuracy improvement by training methods to create precise and contextually relevant descriptions

SCOPE

- Image captioning, driven by deep learning, benefits healthcare, education, entertainment, social media, autonomous systems, e-commerce, and cultural heritage by enhancing diagnostics, accessibility, search functionalities, and historical preservation.
- Advances in deep learning models and large datasets promise increasingly sophisticated and context-aware image captioning systems, transforming how we interact with visual information.

PROBLEM DEFINITION

- In the realm of image caption generation using deep learning, researchers have grappled with the challenge of achieving high Bleu score accuracy. Numerous research papers have highlighted the under performance of models combining Recurrent Neural Networks (RNN) with Convolutional Neural Networks (CNN).
- It has been observed that the traditional RNN-CNN pairing lags behind LSTM-CNN architectures in terms of accuracy, as evidenced by various studies.

LSTM Model

We are defining LSTM model as decoder , in this an image is given as input to CNN, output of CNN is sent to LSTM which generates output sequences which are further used to generate captions.

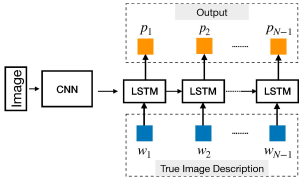

Fig 2. LSTM as a decoder

IMPLEMENTATION

This process for generating image captions integrates deep learning and image processing techniques. It begins by inputting a dataset, dividing it into manageable batches, and preprocessing the images for quality enhancement. CNNs extract detailed features from the preprocessed images. These features serve as input for LSTM networks, which generate textual descriptions or captions for each image. Simultaneously, an image repository and caption database are maintained for easy retrieval, indexing captions with their corresponding images. A text similarity engine compares newly generated captions with existing ones to enhance relevance and accuracy. The final output consists of images paired with contextually appropriate captions. This method optimally utilizes CNNs for feature extraction and LSTMs for sequential data processing, ensuring the production of robust and meaningful image descriptions. Through this process, the integration of deep learning and image processing techniques facilitates the generation of accurate and relevant captions for images.

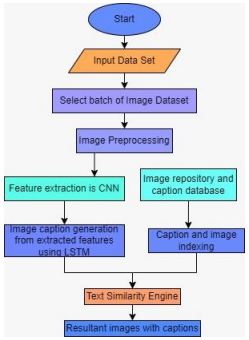

Fig3. Flow diagram of image captioning

RESULTS AND FINDINGS

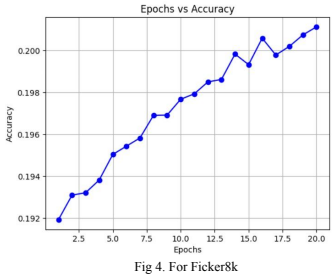

Fig 4. For Ficker8k

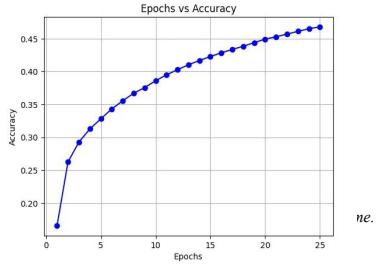

Fig 5. For Ficker8K on Ficker30k

TABLE I
BLEU SCORES FOR DATASETS

Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Flickr8k	0.53	0.30	0.11	0.17
Flickr30k	0.53	0.27	0.16	0.08
Flickr8k trained on Flickr30k	0.56	0.34	0.22	0.13

TABLE II
CAPTIONS MAPPING

Image Id	Image	Caption
1002674143_1b742ab4b8		• **Reference Caption:** A little girl covered in paint sits in front of a painted rainbow with her hands inside bowl • **Predicted Caption:** two children enjoy the grass with fingerpaints in the background
101669240_b2d3e7f17b		• **Reference Caption:** A man in hat is displaying pictures next to a skier in a blue hat • **Predicted Caption:** Two people are the snow white in the snow

OBSERVATIONS

- The transition from traditional machine learning to deep learning, especially CNN-based models, has significantly improved image captioning by leveraging hierarchical feature learning, though it requires substantial computational resources.
- Techniques like data augmentation, GANs, encoder-decoder architectures with attention mechanisms, and reinforcement learning have further refined image captioning models, resulting in more accurate, detailed, and contextually appropriate captions.
- With continuous improvements in computational resources and algorithm sophistication, the performance and applicability of image captioning systems are expected to advance, offering even greater accuracy and relevance in generated captions.

CONCLUSIONS

- This review highlights the effectiveness of a CNN-LSTM hybrid model in image captioning, detailing modern advancements in feature extraction and object localization, and their smooth integration into language models.
- The study demonstrates the progress in combining computer vision and natural language processing, showing how the CNN-LSTM model enhances captioning performance and provides insights for future research in this evolving field.

REFERENCES

- Ariyo Oluwasammi, Muhammad Umar Aftab, Zhiguang Qin, Son Tung Ngo, Thang Van Doan, Son Ba Nguyen, Son Hoang Nguyen, and Giang Hoang Nguyen. Features to text: a comprehensive survey of deep learning on semantic segmentation and image captioning. Complexity, 021:1–19, 2021.
- Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. IEEE transactions on pattern analysis and machine intelligence, 45(1):539–559, 2022.
- Licheng Jiao and Jin Zhao. A survey on the new generation of deep learning in image processing. Ieee Access, 7:172231–172263, 2019

ACKNOWLEDGEMENT

We extend our heartfelt gratitude to Ajay Kumar Garg Engineering College for their exceptional support and fostering environment. Special thanks to Mr. Vishal Jayaswal(Assistant Professor, CSE Dept.) for his insightful guidance, Prof. Anu Chaudhary(HOD, CSE Dept.) for her leadership, and Mr. Shashank Sahu(Professor In-Charge) for his invaluable advice. We also thank our families and friends for their unwavering support. This research's success is a testament to the collective efforts and contributions of all involved. Thank you.