IMAGE CAPTIONING:AUTOMATING THE ART OF WORDS

Alan munigety
School of Computer Science
and Engineering
KLE Technological University
Hubli, Karnataka
Email: alanmunegety501@gmail.com

Darshan B Garagatti
School of Computer Science
and Engineering
KLE Technological University
Hubli, Karnataka
Email: darshanbg1967@gmail.com

Ritwik Bhagat
School of Computer Science
and Engineering
KLE Technological University
Hubli, Karnataka
Email: contactbhagatritwik@gmail.com

Subhangee Rai
School of Computer Science
and Engineering
KLE Technological University
Hubli, Karnataka
Email: aradhya81702002@gmail.com

Priyadarshini Patil
School of Computer Science
and Engineering
KLE Technological University
Hubli, Karnataka
Email: priyadarshini.patil@kletech.ac.in

Meena S M
School of Computer Science
and Engineering
KLE Technological University
Hubli, Karnataka
Email: msm@kletech.ac.in

***Abstract*—Image captioning, the process of generating textual descriptions for images, has become a prominent and evolving research challenge in recent times. Despite the existence of numerous solutions, there remains a significant need for continuous attention to achieve more accurate and precise results. Consequently, our initiative involves developing an image captioning model by exploring various combinations of Convolutional Neural Network (CNN) architectures in conjunction with Long Short-Term Memory (LSTM) to enhance performance.In this pursuit, we have implemented three different combinations of CNN and LSTM for model development. The proposed model is trained using three distinct Convolutional Neural Network architectures—namely, Inception-v3, Xception, and ResNet50—for extracting features from images.we have also used cnn-rn architecture The caption generation aspect is handled by Long Short-Term Memory. The selection of the optimal combination among the three pairs of CNN and LSTM is determined based on the model's accuracy. This systematic approach aims to improve the overall efficacy of image captioning models and contribute to the ongoing advancements in this field.**

I. INTRODUCTION

The process of creating text descriptions for photographs that adhere to human language cognition is known as image captioning.This has important applications in the industrial and medical fields. It can help visually impaired persons and foster visual intelligence in human-robot interaction in the industrial setting. Accurate thinking in clinical settings necessitates years of professional development and training. It is crucial to implement automatic medical picture captioning in order to lessen the strain of clinicians and increase diagnostic efficiency and accuracy An encoder-decoder framework can be used to accomplish earlier picture captioning, which is inspired by Recurrent Neural Network (RNN)in machine translation. However, in contrast to human descriptions, the results were largely general and lacked diversity. Given the growing progress picture captions of deep models have shown encouraging gains in naturalness and variety. Convolutional Neural Networks (CNNs) are used to conduct picture captioning by extracting high-level and effective feature representation from source data using layers that range from several to thousands. With its strong generating and fitting capabilities, generating Adversarial Network (GAN) has made significant advancements in picture captioning. It can approximate maximum likelihood estimation and remove Markov restrictions or unrolled approximation inference networks. In addition to the previously stated networks, the attention mechanism. is crucial in determining which features will be input into a later language model or in allocating scarce resources to key sections. At the moment, captioning cannot follow the cognitive patterns of natural language in humans, but can comply with clinical language style of medical diagnosis report.

II. LITERATURE SURVEY

While previous surveys have explored statistical language-based and deep learning-based image and video captioning, our work distinguishes itself by incorporating the substantial advancements in Generative Adversarial Networks (GANs) and the rapid progress in the medical field.In particular, we found a survey closely related to our scope, providing an overview encompassing visual encoding, text generation, training strategies, and evaluation metrics in deep neural network (DNN)-based captioning. However, these earlier works primarily centered on DNN-based approaches, overlooking the noteworthy contributions of GANs and the advancements in the medical domain. To the best of our knowledge, our paper marks the initial endeavor to offer a comprehensive survey of deep captioning, covering the entire spectrum from simulating workflow and caption algorithms to application perspectives. This includes the latest advancements in both natural and medical fields.

The importance of generating natural language descriptions from visual data, particularly in the context of image captioning using deep learning techniques. It highlights the historical background of the problem, the use of complex systems combining visual primitive recognizers with structured formal languages, and the recent surge of interest in still image description with natural text. The survey also covers related work, including experimental results using the MSCOCO dataset and the incorporation of a guiding network in the encoder/decoder framework. Additionally, it delves into the methodology, detailing the use of CNN and LSTM for image caption generation, and the challenges faced in choosing the dataset for training the model. The survey also provides references to various publications and resources related to deep learning, CNN, LSTM, and image captioning.In summary, the literature survey in the document provides a comprehensive overview of the historical context, related work, methodology, and references related to image captioning using CNN and LSTM, offering valuable insights for researchers and practitioners in the field of deep learning and computer vision

Early models for image captioning, such as those referenced in, often encoded visual information using a single feature vector representing the entire image. However, this global representation approach neglected valuable information about objects and their spatial relationships. Karpathy and Fei-Fei [11], in an exception to this trend, utilized an R-CNN object detector to extract features from multiple image regions, generating separate captions for each. Despite this, the spatial relationships between detected objects were not explicitly modeled, a limitation also present in their subsequent dense captioning work .Attention-based approaches in image captioning sought to ground words in the predicted caption to specific image regions. However, visual attention, often derived from higher convolutional layers, had limitations in spatial localization and semantic meaning. Anderson et al. addressed this by combining a ”bottom-up” attention model with a ”top-down” LSTM, demonstrating superior performance in visual question answering and image captioning. Yet, spatial information was not explicitly incorporated. Geometric attention, introduced by Hu et al. for object detection, leveraged bounding box coordinates and sizes to infer the importance of relationships between pairs of objects based on their proximity and size. Subsequent successful works followed the paradigm of obtaining image features through an object detector and generating captions via an attention LSTM.Yao et al. introduced Graph Convolutional Networks to add global context, incorporating semantic and spatial relationship graphs. Our proposed approach, in contrast, directly utilizes bounding box size ratios and coordinate differences, implicitly capturing these relationships. used graph structures, including a semantic scene graph, but lacked the ability to capture visual geometry specific to each image.Transformers, a breakthrough in NLP , have shown performance improvements in image captioning . Previous studies explored global image features or uniform sampling of image partitions. Our proposed Object Relation Transformer improves upon this by adopting a bottom-up approach, leveraging the Transformer’s ability to model sequential data without a predefined order. The Object Relation Transformer aims to encode spatial relationships between objects for enhanced image captioning.

Used graph structures, including a semantic scene graph, but lacked the ability to capture visual geometry specific to each image.Transformers, a breakthrough in NLP , have shown performance improvements in image captioning [22]. Previous studies explored global image features or uniform sampling of image partitions. Our proposed Object Relation Transformer improves upon this by adopting a bottom-up approach, leveraging the Transformer’s ability to model sequential data without a predefined order. The Object Relation Transformer aims to encode spatial relationships between objects for enhanced image captioning.

III. Gaps and Challenges

A. Difficulty in precisely articulating the intricate semantics present in images.

Existing image captioning models exhibit limitations in conveying the accurate count of objects and often lack sensitivity to terms like ”two” or ”group.” Additionally, these models struggle with pinpointing focal points in complex scenes, leading to a diminished understanding of crucial content within the image.

B. Inconsistency between objects during the training and testing phases

This inconsistency arises from the dependence on specific datasets during training, introducing errors when describing novel objects during testing. This disparity may result in inaccuracies in image descriptions, thereby impacting the overall efficacy of the image captioning process.

C. Cross-language text description for images.

Current image captioning methods, reliant on deep learning or machine learning, face a scarcity of annotated training samples in languages beyond English and Chinese. This scarcity poses a hurdle in generating text descriptions in multiple languages for images, demanding substantial manual effort and time.

IV. Proposed Work

1) *CNN-CNN based framework:*

- Despite LSTM networks having memory cells that excel at retaining long-term information during sequence generation, their continuous updating at each time step poses challenges for achieving truly extensive long-term memory. Recent research has drawn inspiration from machine learning, showcasing the advantages of incorporating Convolutional Neural Networks (CNNs) into image captioning tasks. The application of CNNs in Natural Language Processing (NLP) for text generation has proven to be highly effective.In the realm of neural machine translation, CNNs have demonstrated superiority over traditional RNNs, exhibiting both higher accuracy and significantly increased training speed. Many image captioning approaches draw parallels with machine translation, considering an image as a sentence in a source language within a sequence-to-sequence architecture.
- A notable advancement in text generation within image captioning is the CNN-CNN framework, pioneered by Aneja et al. This framework consists of three main components, resembling the structure of an RNN. The first and last components involve word embeddings, while the central component, instead of employing LSTM or GRU units, utilizes masked convolutions. Unlike the recurrent function in RNNs, this component in the CNN-based approach is feed-forward.Aneja et al. demonstrated that the CNN-CNN framework boasts faster training times per parameter but incurs a higher loss compared to RNN. The CNN model's accuracy is attributed to its penalization for producing less-peaked word probability distributions. However, less-peaked distributions are considered acceptable, offering the flexibility to predict diverse captions.illustrates the generated descriptions of CNN-RNN and CNN-CNN models, showcasing the diversity achievable with CNN-based approaches. Despite differing means, both convolutional models and recurrent models aim to abstract layered information, emphasizing important content while ignoring minor details. In terms of accuracy, there isn't a substantial difference between convolutional and recurrent models. However, the notable speed advantage of CNNs over RNNs in training can be attributed to two factors: the parallel processing capability of convolutions versus the sequential processing of recurrent models, and the availability of GPU acceleration for convolution models with no comparable hardware for speeding up RNN training.The CNN-CNN framework represents a synergistic blend of CNNs and RNNs in the domains of machine translation and image captioning. Given the widespread effectiveness of CNNs in computer vision, their successful application in machine translation, and subsequently in image captioning, underscores their versatility. Future research should delve deeper into the CNN-CNN-based attention mechanism and explore the potential synergies between CNNs and RNNs in the decoder phase.

2) *Reinforcement based framework:*

- Reinforcement learning has found widespread application in areas such as gaming and control theory, where concrete optimization targets are inherent. However, applying reinforcement learning to image captioning poses a unique challenge due to the nontrivial task of defining an appropriate optimization goal. In the context of image captioning with reinforcement learning, the generative model (typically an RNN) is conceptualized as an agent interacting with the external environment, receiving words and a context vector as input at each time step. The agent's parameters define a policy, and the execution of this policy involves the agent selecting an action, which, in sequence generation, corresponds to predicting the next word at each time step. Following an action, the agent updates its internal state (the hidden units of the RNN). Upon reaching the end of a sequence, the agent observes a reward, and the RNN decoder functions as a stochastic policy during this process.In training, the policy gradient (PG) method is employed, choosing actions based on the current policy and only receiving a reward at the end of the sequence (or after reaching the maximum sequence length). Training aims to find agent parameters that maximize the expected reward.equation used J X t=1 log p(at—st) (r v(st))

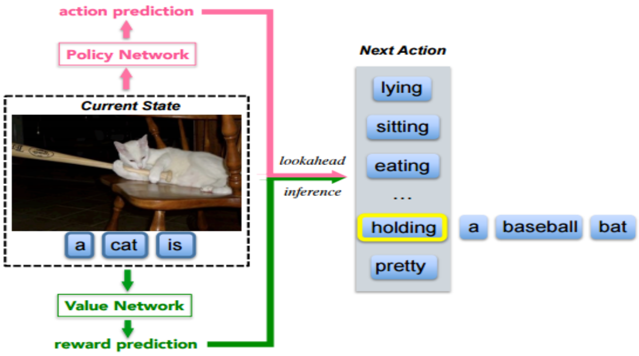


Fig. 1. Reinforcement learning ex

V. Results

.

The study directly compared the results with other models using the comparable Xception and VGG16 features by the BLUE Score metric. The evaluation focused on the quality of the generated captions and the differences between the Xception and VGG16 models. .

The deep learning architecture utilizing the Xception model demonstrated a commendable BLEU score of 55.01The study conducted a comparative analysis between the Xception and

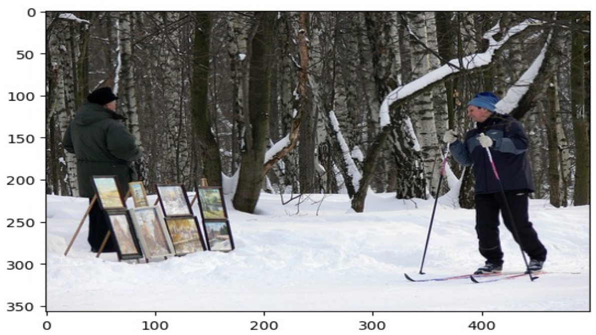


Fig. 2. CNN-CNN MODEL

——————Actual——————
startseq man in hat is displaying pictures next to skier in blue hat endseq startseq man skis past another man displaying paintings in the snow endseq startseq person wearing skis looking at framed pictures set up in the snow endseq startseq skier looks at framed pictures in the snow next to trees endseq startseq man on skis looking at artwork for sale in the snow endseq
——————Predicted——————
startseq two people are hiking up snowy mountain endseq

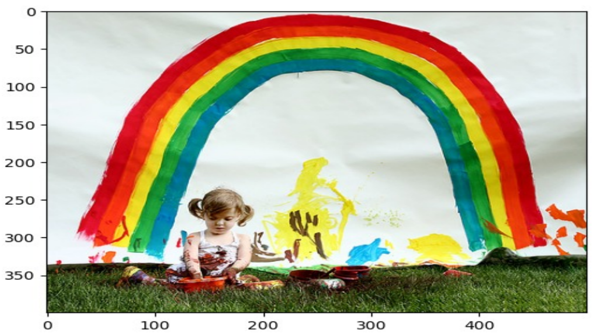


Fig. 3. CNN-NN MODEL

——————Actual——————
startseq little girl covered in paint sits in front of painted rainbow with her hands in bowl endseq startseq little girl is sitting in front of large painted rainbow endseq startseq small girl in the grass plays with fingerpaints in front of white canvas with rainbow on it endseq startseq there is girl with pigtails sitting in front of rainbow painting endseq startseq young girl with pigtails painting outside in the grass endseq -
——————-Predicted——————
startseq little girl in pink dress is lying on the side of the grass endseq

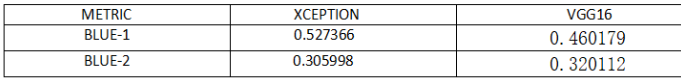
VGG16 models, evaluating their performance based on the BLEU Score metric, which assesses the quality of generated captions. The research emphasized potential enhancements through diverse modifications, including the utilization of a larger dataset, adjustments to the model architecture (e.g., incorporating an attention module), extensive hyperparameter tuning, employing cross-validation sets to address overfitting concerns, and the adoption of Beam Search as an alternative to Greedy Search during the inference phase. In terms of experimental comparison, the study directly pitted results against other models, focusing on the comparable features extracted from Xception and VGG16, assessed by the BLEU Score metric. The evaluation centered on discerning differences in the quality of generated captions between the Xception and VGG16 models. For implementation, Python was employed, with Keras 2.0 serving as the framework for deep learning model implementation. The TensorFlow library acted as the backend for the Keras framework. The neural network training took place on Google Colab, leveraging various Keras APIs, including the Keras Model API Keras.

VI. Methodology

We have used Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) to achieve the goal of image caption generation. The study utilized the Flickr8k dataset for training the model. The image feature extraction was performed using the Xception model, which provided a 2048 vector element representation of the photos. The sequence processor, acting as a word embedding layer, handled the text input and was connected to an LSTM for the final phase of image captioning. Additionally, the study applied preprocessing steps for both images and captions, including converting all descriptions to lowercase, removing special tokens, and tokenization with a fixed vocabulary size of 8,464. The implementation of the model was carried out using Python, with Keras 2.0 used for deep learning model implementation, and TensorFlow as the backend for the Keras framework. The study also evaluated the model's performance using the BLEU score and outlined potential improvements, such as using a larger dataset, changing the model architecture, conducting more hyperparameter tuning, and using Beam Search instead of Greedy Search during inference.

VII. Conclusion

We employed CNN and LSTM models in our study, showcasing the implementation of a deep learning approach for image captioning. We utilized the sequential API of Keras, with TensorFlow serving as the backend for model implementation. The resulting deep learning architecture achieved a notable BLEU score of 55.01particularly with the Xception model. Our investigation identified potential avenues for improvement, including the exploration of larger datasets, adjustments to the model architecture (e.g., incorporating an attention module), extensive hyperparameter tuning, the use of cross-validation sets to address overfitting, and the adoption of Beam Search in lieu of Greedy Search during inference. Furthermore, we highlighted the distinctions between the VGG16 and Xception models, shedding light on challenges encountered, such as the selection of datasets and convolutional feature extractors. The study delved into the advantages



METRIC	XCEPTION	VGG16
BLUE-1	0.527366	0. 460179
BLUE-2	0.305998	0. 320112

TABLE I
RESULTS

of image description, especially in assisting visually impaired individuals in comprehending online images. In summary, our conclusion underscores the successful implementation of the deep learning approach for image captioning, while also identifying areas for potential refinement and emphasizing the significance of image description across various applications.

.

References

[1] Kiros, R.; Salakhutdinov, R.; Zemel, R. Multimodal neural language models. International Conference on Machine Learning, Beijing, China, 21–26 June 2014; pp. 595–603.

Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.

Yan, S.; Xie, Y.; Wu, F.; Smith, J.S.; Lu, W.; Zhang, B. Image captioning based on a hierarchical attention mechanism and policy gradient optimization. J. Latex Cl. Files 2015, 14, 8.

Karpathy, A.; Li, F.-F. Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; Volume 39, pp. 3128–3137.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning (ICML'15), Lille, France, 6–11 July 2015.