e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 11, November 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

© 2024 IJMRSET | Volume 7, Issue 11, November 2024| DOI: 10.15680/IJMRSET.2024.0711086

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Automatic Image Captioning using NLP

K.Siddartha[1], B.Sowmya[2], Y.Sravya Reddy[3], P.Shravya[4], P. Sowmya[5], Prof. Thanish Kumar[6]

Department of AIML, School of Engineering, Malla Reddy University, India[1,2,3,4,5,6]

ABSTRACT: An image caption is something that describes an image in the form of text. It is widely used in programs where one needs information from any image in automatic text format. We analyze three components of the process: convolutional neural networks (CNN), recurrent neural networks (RNN) and sentence production. It develops a model that decomposes both images and sentences into their elements, regions of intelligent languages in photography with the help of LSTM model and NLP methods. It also introduces the implementation of the LSTM Method with additional efficiency features. The Gated Recurrent Unit (GRU) and LSTM Method are tested in this paper. According to tests using BLEU Metrics LSTM is identified as the best with 80% efficiency. This method enhances the best results in the Visual Genome role-caption database and flicker8kdatabase.

I. INTRODUCTION

Image captioning research is looking beyond single-sentence descriptions and exploring paragraph-length captions for richer image understanding. This "image for narrative" approach tackles the limitations of single sentences by using datasets like Visual Genome to train models that generate coherent, multi-sentence narratives about the image content. Researchers are employing various deep learning techniques like Long Short-Term Memory(LSTM)networks and Gated Recurrent Units(GRUs) to achieve this, addressing repetitive sentence generation issues through beam search and attention models. This shift towards paragraph captions holds promise for a more comprehensive description of image content.

II. LITERATUREREVIEW

The project's literature survey delves into the current state of paragraph caption generation, uncovering both its potential and limitations. Existing research, as detailed in works by Krause et al. (2016),highlights the inadequacy of single-sentence captions infully comprehending an image. To address this, the survey explores approaches like "Image for Narrative" (Chen et al., 2017), which aims to create coherentmulti-sentence descriptions. How ever, a key limitation identified, particularly by Liuetal.(2019),is the tendency of current models to generate repetitive sentences.

The survey acknowledges the established role of deep learning techniques like LSTMs and GRUs (Mehta et al., 2023) in handling sequential data for sentence generation. Additionally, it sheds light on the promising potential of Visual Language Transformers (VLITs), which leverage pre-trained language models for richer captions(Mehtaetal.,2023).Looking ahead, the survey emphasizes two main challenges: ensuring factual accuracy and maintaining paragraph coherence (Mehta et al., 2023).. It identifies limitations, explores promising deep learning techniques, and suggests future directions for overcoming current challenges and achieving more accurate and informative multi-sentence image description.

III. PROBLEM STATEMENT

The challenge of image captioning is to design a deep learning model that can analyze an image and automatically create an natural language description, bridging the gap between computer vision and natural language processing. This requires the model to extract key features from the image and translate them into a grammatically correct and informative sentence, with success measured by the accuracy, fluency, and richness of the generated captions.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. METHODOLOGY

Data preprocessing is a critical step in automatic image captioning using natural language processing(NLP). Effective preprocessing ensures that the model can learn from the data in a meaningful way, leading to improved performance. Here are some common data preprocessing techniques used in this context:

- **Resizing**: Images are typically resized to a consistent dimension to ensure uniformity in input size for the model.
- **Normalization**: Pixel values are normalized (e.g., scaled to a range of 0 to 1) to improve convergence during training.
- **Augmentation**: Techniques such as rotation, flipping, cropping, and color adjustments
- **Tokenization**:Captions are split into individual words or tokens, making it easier for the model to process language data.
- **Lowercasing**: Text is often converted to lowercase to reduce vocabulary size and maintain consistency.
- **Removing Punctuation and Stop Words**: Unnecessary characters and common stop words may be removed to focus on meaningful words that contribute to the understanding of the image context.
- **Building a Vocabulary**:A vocabulary is created from the processed captions ,often including only the most frequent words to limit complexity and improve training efficiency.
- **Encoding**: Words in the vocabulary are converted into numerical representations(e.g.,using one-hot encoding or embeddings like Word2Vec or GloVe) for input into the model.

V. EXPERIMENTALRESULTS

Our goal is to make sure the models work efficiently and effectively. We use advanced techniques like Long Short-Term Memory (LSTM) models and Natural Language

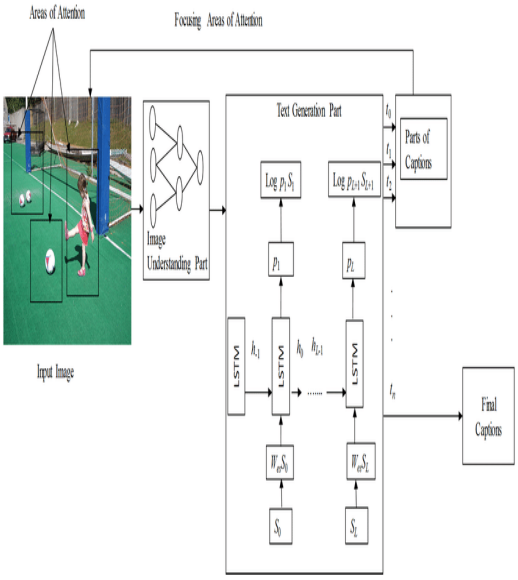

Fig 6.1 Comparison of Model evaluation Metrics

© 2024 IJMRSET | Volume 7, Issue 11, November 2024| DOI: 10.15680/IJMRSET.2024.0711086

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

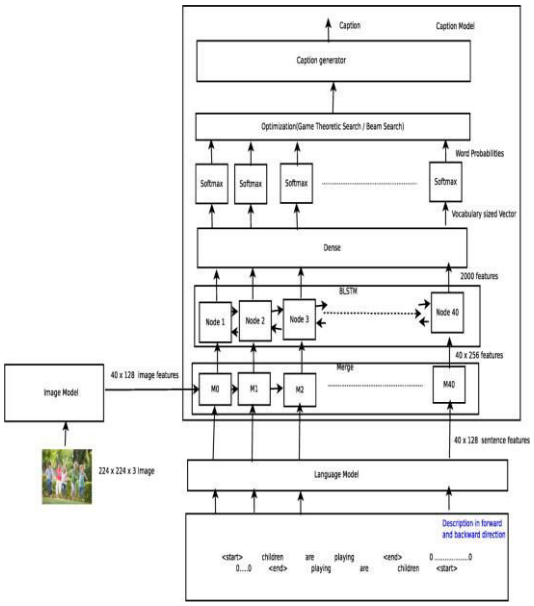

Fig6.2Architecture

Fig6.3Output screen

VI. CONCLUSION

In conclusion, image captioning plays a crucial role in converting visual information into textual form, facilitating its accessibility and utilization in various automated programs. This abstract highlights the analysis of three key components involved in the image captioning process:convolutional neural networks(CNN), recurrent neural networks (RNN), and sentence production. By leveraging advanced models such as the Long Short-Term Memory (LSTM) model and employing natural language processing (NLP) techniques, the study proposes a novel approach that decomposes both images and sentences into their constituent elements. Moreover, the implementation of the LSTM Method with additional efficiency features demonstrates promising results, particularly in conjunction with the Gated Recurrent Unit (GRU). Through rigorous testing using BLEU Metrics, the LSTM Method emerges as the most efficient, achieving an impressive 80% efficiency rate. These findings signify a significant advancement in image captioning technology, with the proposed method outperforming existing approaches, particularly in datasets such as the Visual Genome role-caption database.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VII. FUTUREENHANCEMENT

The future of image captioning with deep learning holds exciting possibilities that promise to push the boundaries of current capabilities. Here are some key areas of potential enhancement:

- Embracing Multimodality: Integrating additional modalities beyond vision, such as audio descriptions or haptic feedback, can create richer and more comprehensive captions, particularly for complex scenes.
- Fine-Grained Attention Mechanisms: Advanced attention mechanisms could enable the model to not only focus on relevant image regions but also understand the relationships between these regions, leading to more detailed and nuanced captions.
- Explainable AI: Developing interpretable models that explain there as oning behind caption generation would foster trust and transparency in applications like medical image analysis.
- Lifelong Learning: Continuously learning models that can adapt to new data and situations would enhance robustness and real-world applicability.
- Generative Adversarial Networks (GANs): Utilizing GANs for caption generation could lead to more diverse and creative captions, potentially surpassing human-written descriptions in some cases.

REFERENCES

Machine Learning

[1] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson. nocaps: novel object captioning at scale. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 8948–8957, Manhattan, New York, U.S., 2019.

[2] H. Ahsan, D. Bhatt, K. Shah, and N. Bhalla. Multi- modal image captioning for the visually impaired. In Proceedingsofthe2021ConferenceoftheNorthAmerican Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 53–60, Stroudsburg, PA, USA, jun 2021. Association for Computational Linguistics.

[3] R. Al Sobbahi and J. Tekli. Comparing deep learning models for low-light natural scene image enhancement and their impact on object detection and classification: Overview, empirical evaluation, and challenges. Signal Processing: Image Communication, page 116848, 2022

[4] R. Al Sobbahi and J. Tekli. Low-light image enhancement using image-to-frequency filter learning. In Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II, pages 693–705. Springer, 2022.

[5] J.-B.Alayrac, J.Donahue, P.Luc, A.Miech, I.Barr,Y. Hasson, K.Lenc, A.Mensch, K.Millican, M.Reynolds,et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35:23716–23736, 2022.

[6] P.Anderson, B. Fernando, M. Johnson, and S. Gould. Spice:Semantic propositional image captione valuation. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, Computer Vision – ECCV 2016, pages 382–398, Manhattan, NewYork, USA, 2016. Springer International Publishing.ISBN978-3-319-46454-1.doi:10.1007/978-3- 319- 46454-1_24.

[7] P.Anderson, X.He, C.Buehler, D.Teney, M.Johnson, S.Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6077–6086, Los Alamitos, CA, USA, jun 2018. IEEE Computer Society. doi:10.1109/CVPR.2018.00636. URL https://doi.ieeecomputersociety.org/10.1109/CVPR. 2018.00636.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com