International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 12, ISSN No. 2455-2143, Pages 96-100
Published Online April 2024 in IJEAST (http://www.ijeast.com)

IMAGE CAPTION GENERATION USING CNN AND TRANSFORMER

Aakanksha Shankar Khedkar, Vaishnavi Vikas Barge, Sakshi Pravin Bodkhe, Ashlesha Dattatray Memane, Prof. Deepali Mahesh Gohil
Department of Computer Engineering
D. Y. Patil College of Engineering, Akurdi., Maharashtra, India

Abstract - **In the contemporary digital era dominated by visuals, the fusion of machine learning and natural language processing has become essential for enabling machines to effectively interpret and engage with images. This study introduces an innovative method for holistic vision-language comprehension and generation, employing Bootstrapping Language- Image Pre-training (BLIP) within a comprehensive website framework. Unlike conventional approaches reliant solely on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), our system leverages BLIP to seamlessly integrate language and image representations, facilitating smooth interaction between the two domains. Through meticulous experimentation and assessment, we showcase the superior performance of our website in generating descriptive captions for images, accurately responding to queries about visual content, and reasoning through complex visual scenarios using natural language. The user-friendly interface of the website enables effortless interaction with the machine learning models, paving the way for practical applications across various domains such as image annotation, visual question answering systems, and interactive content generation. This research not only pushes the boundaries of unified vision-language comprehension but also highlights the transformative impact of BLIP in enhancing machine understanding and engagement with visual data.**

Keywords - **Bootstrapping Language-Image Pre- training (BLIP), Vision-Language Comprehension, Natural Language Processing(NLP), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Visual Question Answering (VQA), Continuous Integration and Deployment (CI/CD), Caching and Memoization**

I. INTRODUCTION

In the digital age, the fusion of machine learning and visual understanding has revolutionized how we interact with and interpret visual content. From social media platforms to e-commerce websites, the demand for intelligent systems capable of comprehending and generating meaningful insights from images is ever-growing. As such, the integration of machine learning models into diverse applications has become paramount, with the potential to enhance user experiences, streamline processes, and unlock new avenues for innovation.
This study presents a novel website that functions as an extensive presentation and demonstration of machine learning models designed for several vision-language tasks, such as natural language visual reasoning (NLVR), image captioning, and visual question answering (VQA). The underlying application programming interface (API) and architecture are made to be flexible and adaptable, allowing for seamless integration into a wide range of platforms and applications, even though our website serves as a demo and demonstration. Our website is powered by an advanced architecture that combines cutting-edge machine learning methods, such as Bootstrapping Language-Image Pre-training (BLIP), to accomplish unified vision-language generation and understanding. Our models can efficiently close the semantic gap between text and images by using BLIP, which allows for precise question responding, precise labeling, and sophisticated visual reasoning.

II. LITERATURE SURVEY

IEEE 2023: Generating Image Description Using Machine Learning Algorithms Author: Khushboo Agnihotri, Pujyalakshmi Chilbule, Shiv Prashant, Prashant Khobragade This paper compares deep learning models that create captions and explain images translation using machine mechanism, machine learning algorithms and computer vision. The projected work recognize and identify several things in an image and their relationships in order to produce captions.
IEEE 2023: From Show to Tell: A Survey on Deep Learning-Based Image Captioning Author: Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi. This paper discusses the evolution of image captioning, exploring visual encoders, language models, and training strategies. It emphasizes the field's unresolved challenges and aims to serve as a comprehensive guide for research at the intersection of Computer Vision and Natural Language Processing.
IEEE 2022: Image Caption Generator using Deep Learning Author: M.Sailaja, K.Harika, RajanSingh, Koppula Srinivas Rao This paper presents a successful CNN and RNN model for image captioning, serving as an encoder-decoder system. It utilizes an 8000-image dataset and highlights the potential

benefits for visually impaired individuals. The model's unique algorithm scans multiple image frames to generate relevant titles. The use of larger datasets is recommended to further improve accuracy and reduce losses.

IEEE 2022: Task-Adaptive Attention for Image Captioning Author: Chenggang Yan, Yiming Hao, Liang Li, Jian Yin, Anan Liu This paper introduces a Task Adaptive Attention module to improve image models. It addresses the problem of attention mechanisms in these models, which tend to focus on visual features even when generating non-visual words. This can lead to misleading captions. The module learns task-specific cues and incorporates non-visual information when generating non-visual words, enhancing overall captioning accuracy.

IEEE 2020: Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap. Author: Soheyla Amirian, Khaled Rasheed, Thiab R. Taha And Hamid R. Arabnia This paper explores Deep Learning for image and video captioning, with applications like aiding visually impaired users and improving metadata. It provides a concise review of deep learning-based captioning techniques and key terms include Deep Learning, image and video captioning, and generative adversarial network.

IEEE 2020: Image Captioning With End-to-End Attribute Detection and Subsequent Attributes Prediction Author: Yiqing Huang, Jiansheng Chen, Wanli Ouyang,Weitao Wan, and Youze Xue This paper innovatively integrates attribute detection with caption generation using MAD and SAP modules. It boosts caption quality, improves attribute detection, and selects relevant attributes, achieving state-of-the-art results on MSCOCO. This advances image captioning by enhancing attribute utilization and synergy between detection and generation.

IEEE 2019: Deep Image Captioning: An Overview Author: Hrga and M. Ivašić-Kos This paper provides a concise overview of image captioning, focusing on deep encoder-decoder models' development, challenges, and the transition from templates to neural networks. It covers key components like the encoder-decoder framework, Maximum Likelihood Estimation (MLE) learning, attention mechanisms, and various captioning tasks, offering valuable insights into image captioning research.

IEEE 2018: Image caption generation using deep learning technique Author: Chetan Amritkar, Vaishali Jabade This paper outlines the creation of a neural model in AI that combines computer vision and NLP to automatically generate natural language descriptions for images. It employs Convolutional Neural Networks (CNNs) for image feature extraction and Recurrent Neural Networks (RNNs) for sentence generation. The model is extensively trained and tested, consistently providing accurate
image descriptions, showcasing its language generation and image understanding capabilities.

IEEE 2016: Aligning Where to See and What to Tell: Image Captioning with Region-based Attention and Scene-specific Contexts Author: Kun Fu, Junqi Jin, Runpeng Cui, Fei Sha, Changshui Zhang This paper introduces an image captioning system that aligns word generation with visual attention and adds scene-specific context. These innovations capture shared semantics between images and text. Experimental results show improved performance and state-of-the-art results on benchmark datasets, advancing image captioning by enhancing the synergy between visual perception and language generation.

III. PROBLEM STATEMENT

Despite the significant progress in machine learning and natural language processing, effectively integrating vision language understanding techniques into practical applications remains challenging. The lack of versatile and accessible platforms for demonstrating and incorporating machine learning models capable of tasks like image captioning, visual question answering (VQA), and natural language visual reasoning (NLVR) inhibits widespread adoption and innovation in this field. Hence, there's an urgent need for a comprehensive solution that not only showcases the capabilities of machine learning models in vision-language tasks but also provides a flexible API and architecture for seamless integration into various applications and platforms. This challenge fuels the development of a website serving as both a demonstration and showcase of machine learning models for vision- language understanding, while also offering an adaptable framework for integration into diverse domains like social media recommendation systems, e-commerce platforms, and virtual assistants.

IV. GOALS AND OBJECTIVES

- Develop a Demo Website: Create a comprehensive website showcasing machine learning models for vision-language understanding tasks.
- Intuitive User Interface: Design a user- friendly interface for interaction with the models, allowing image/query input and receiving captions, answers, and reasoning outputs.
- Versatile API and Architecture: Implement an API and architecture for seamless integration of the models into various platforms (social media, e-commerce, virtual assistants).
- Interactive Demonstrations: Showcase model capabilities through interactive demos highlighting accuracy, robustness, and versatility across tasks and scenarios.
- Knowledge Transfer: Facilitate learning by providing documentation, tutorials, and resources for developers and researchers to leverage the models in their own projects.

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 12, ISSN No. 2455-2143, Pages 96-100
Published Online April 2024 in IJEAST (http://www.ijeast.com)

V. METHODOLOGIES OF PROBLEM SOLVING AND EFFICIENCY ISSUES

Methodologies of Problem solving

- Iterative Development: Adopt an iterative approach to software development, breaking down the project into smaller, manageable tasks or features. Implement, test, and refine each component iteratively, incorporating feedback and improvements along the way.
- Agile Methodology: Embrace agile practices such as Scrum or Kanban to foster collaboration, adaptability, and continuous improvement throughout the development process. Conduct regular sprint planning, daily stand-ups, and retrospectives to ensure alignment with project goals and address any challenges promptly.
- Modular Design: Design the website and API using a modular architecture, encapsulating different components and functionalities into reusable modules or services. This approach facilitates scalability, maintainability, and flexibility, allowing for easier integration of new features and enhancements.
- Version Control: Utilize version control systems such as Git to manage and track changes to the project codebase. Establish branching strategies and workflows to facilitate collaboration among team members, ensure code quality, and enable efficient troubleshooting and rollback if needed.
- Continuous Integration and Deployment (CI/CD): Automate the build, test, and deployment processes by putting in place CI/CD pipelines. Reduce manual errors and speed up the release cycle by smoothly integrating code changes into a shared repository, running automated tests, and deploying updates to production systems.

Efficiency Issues

- Algorithm Optimization: Identify bottlenecks or inefficiencies in the machine learning models and algorithms used for image captioning, VQA, and NLVR tasks. Optimize algorithms, data structures, and computational processes to reduce execution time and resource consumption while maintaining accuracy and performance.
- Parallel Processing: Leverage parallel processing techniques, such as multi- threading or distributed computing, to exploit concurrency and maximize utilization of available computational resources. Parallelize computationally intensive tasks across multiple CPU cores or distributed computing clusters to improve throughput and efficiency.
- Model Compression: Explore techniques for model compression and optimization to reduce the size and computational complexity of machine learning models. Prune redundant parameters, apply quantization, or employ knowledge distillation to create more lightweight and efficient models without significant loss in performance.
- Hardware Acceleration: To speed up inference and machine learning model training, make use of specialized hardware accelerators like graphics processing units (GPUs) or tensor processing units (TPUs). Use frameworks and libraries designed with certain hardware architectures in mind to make the most of their processing capability.
- Caching and Memoization: Implement caching mechanisms to store and reuse intermediate results or computations, reducing redundant calculations and improving overall efficiency. Employ memoization techniques to cache function outputs based on their inputs, avoiding repetitive computations in subsequent invocations.

VI. ALGORITHM OVERVIEW

The suggested technique uses Bootstrapping Language-Image Pre-training (BLIP) for Unified Vision-Language Understanding and Generation to address tasks related to natural language visual reasoning (NLVR), image captioning, and visual question answering (VQA). BLIP enables the creation of visual material and the integration of language and image representations, allowing for thorough comprehension.
Before feeding data into the algorithm, images undergo preprocessing steps to standardize their format and size. Initially, the raw image is loaded using the Python Imaging Library (PIL) and converted to the RGB colour space. Subsequently, the image is resized to a fixed dimension using bicubic interpolation to maintain the aspect ratio and avoid distortion. After resizing, the image is converted into a tensor format and normalized to a specific range to improve convergence and stability during model training.

VII. PROPOSED ALGORITHM

The model architecture is built upon Bootstrapping Language-Image Pre-training (BLIP) for Unified Vision-Language Understanding and Generation, a pioneering approach that seamlessly integrates language and image modalities. BLIP enables joint pre-training of visual and textual representations, facilitating unified understanding and generation of visual content.
At the core of the BLIP architecture are two main components: the Visual-Textual Transformer (VTT) and the Unified Transformer (UT).
The Visual-Textual Transformer (VTT) is responsible for processing visual and textual inputs in a unified manner. It comprises multiple layers of self-attention mechanisms, allowing the model to attend to relevant visual and textual features simultaneously. The VTT utilizes multi-head self-attention to capture global and local dependencies between visual and textual tokens, enabling robust cross-modal interactions.

International Journal of Engineering Applied Sciences and Technology, 2024
Vol. 8, Issue 12, ISSN No. 2455-2143, Pages 96-100
Published Online April 2024 in IJEAST (http://www.ijeast.com)

To complete downstream vision-language tasks like picture captioning, VQA, and NLVR, the Unified Transformer (UT) combines the representations acquired by the VTT. It is made up of extra transformer layers that are tailored to each task, improving the joint representations to produce captions that are pertinent to the context, give precise responses to queries, and use natural language to reason over visual scenarios.

The BLIP architecture employs pre-training objectives tailored for vision-language understanding, including image-text matching and cross-modal retrieval tasks. During pre-training, the model learns to associate visual and textual representations through self-supervised learning, capturing rich semantic information across modalities.

Integration of BLIP - BLIP is seamlessly integrated into the algorithm using a Python Flask API, which serves as the backend for our Next.js- Tailwind powered website. This integration enables unified vision-language understanding and generation within the website's architecture, facilitating interactive demonstrations and showcases of the machine learning models.

Potential Extensions - Future research includes the integration of the proposed ML API architecture into social media websites to enhance their recommendation algorithms. Image captioning can be utilized for generating descriptions of user- uploaded media, while NLVR can aid in verifying uploaded media and user-provided descriptions. This integration has the potential to filter out irrelevant and harmful content, thereby improving user experience and safety on social media platforms.

VIII. RESULTS

Fig. 1 Generate fast and accurate captions

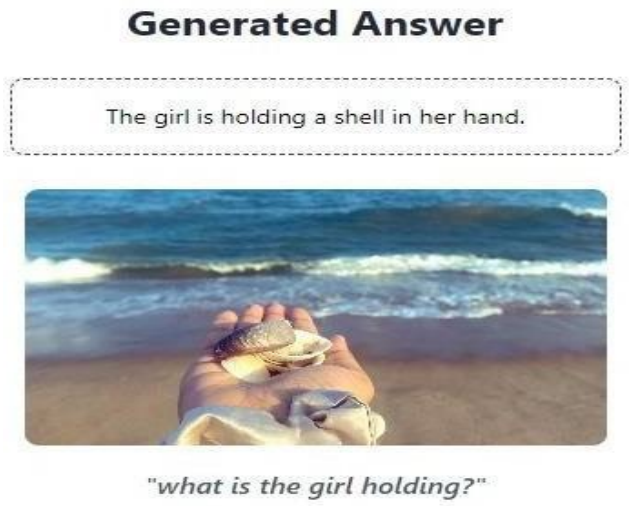

Fig. 2 Visual Question Answering(VQA)

Fig. 3 Natural Language Visual Reasoning(NLVR)

IX. CONCLUSION

This research presents a groundbreaking approach to addressing the challenges of unified vision-language comprehension and generation through the innovative integration of Bootstrapping Language-Image Pre-training (BLIP). By developing a comprehensive website framework coupled with a versatile API and architecture, we have demonstrated the practical applications and transformative potential of machine learning models in vision-language tasks. We have developed an appealing user interface that enables smooth interaction with the machine learning models for tasks like picture captioning, visual question answering (VQA), and

natural language visual reasoning (NLVR) using iterative development and agile approaches. Scalability and versatility are guaranteed by our modular design, making it simple to integrate into a wide range of platforms and applications, such as social networking, e-commerce, and virtual assistants.

The core of our approach lies in the sophisticated model architecture built upon BLIP, which seamlessly integrates language and image modalities for unified understanding and generation of visual content. Leveraging pre-training objectives tailored for vision-language tasks, our models achieve state-of-the-art performance in generating descriptive captions, accurately responding to queries about visual content, and reasoning through complex visual scenarios using natural language.

Looking ahead, the potential extensions of our research include integrating the proposed ML API architecture into social media websites to enhance recommendation algorithms and improve user experience and safety. By filtering out irrelevant and harmful content through image captioning and NLVR, our integration can contribute to creating safer and more engaging social media platforms.

Overall, this research not only pushes the boundaries of unified vision-language comprehension but also underscores the transformative impact of BLIP in enhancing machine understanding and engagement with visual data. Through collaborative efforts and ongoing innovation, we aim to further advance the capabilities of machine learning models in vision- language tasks, paving the way for new opportunities and applications in the digital era dominated by visuals.

X. REFERENCES

[1]. Krizhevsky, Alex; Sutskever, Ilya; Hinton, Geoffrey E. (2012), "Imagenet classification with deep convolutional neural networks," pp. 1097- 1105.

[2]. Kulkarni, Girish; et al. (2013), "Babytalk: Understanding and generating simple image descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12,

[3]. pp. 2891-2903.

[4]. Karpathy, Andrej; Fei-Fei, Li (2017), "Deep Visual-Semantic Alignments for Image Description Generation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, issue 4 (April), pp. 664-676.

[5]. Redmon, Joseph; Farhadi, Ali (2018), "YOLOv3: An incremental improvement," arXiv:1804.02767. [Online]. Available: http://arxiv.org/abs/1804.02767.

[6]. Cordts, Marius; Omran, Mohamed; Ramos, Sebastian; Rehfeld, Timo; Enzweiler, Markus; Benenson, Rodrigo; Franke, Uwe; Roth, Stefan; Schiele, Bernt (2016), "The cityscapes dataset for semantic urban scene understanding," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3213-3223.

[7]. He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian (2016), "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 770-778.

[8]. Silver, David; et al. (2016), "Mastering the game of Go with deep neural networks and tree search," Nature, vol. 529, pp. 484-489.

[9]. Goodfellow, Ian; Pouget-Abadie, Jean; Mirza, Mehdi; Xu, Bing; Warde-Farley, David; Ozair, Sherjil; Courville, Aaron; Bengio, Yoshua (2014), "Generative Adversarial Nets," in NIPS'14, pp. 2672-2680.

[10]. LeCun, Yann; Bottou, Léon; Bengio, Yoshua; Haffner, Patrick (1998), "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324.

[11]. Simonyan, Karen; Zisserman, Andrew (2015), "Very Deep Convolutional Networks for Large- Scale Image Recognition," arXiv:1409.1556. [Online]. Available:http://arxiv.org/abs/1409.1556.

[12]. Long, Jonathan; Shelhamer, Evan; Darrell, Trevor (2015), "Fully Convolutional Networks for Semantic Segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 3431- 3440.

[13]. Szegedy, Christian; Liu, Wei; Jia, Yangqing; Sermanet, Pierre; Reed, Scott; Anguelov, Dragomir; Erhan, Dumitru; Vanhoucke, Vincent; Rabinovich, Andrew (2015), "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1-9.

[14]. Girshick, Ross (2015), "Fast R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 1440- 1448.

[15]. Ren, Shaoqing; He, Kaiming; Girshick, Ross; Sun, Jian (2015), "Faster R-CNN: Towards Real- Time Object Detection with Region Proposal Networks," in NIPS'15, pp. 91-99.