AI Narratives: Bridging Visual Content and Linguistic Expression

Preetam
Department of C.S.E
Chandigarh University
Mohali, India
preetamchauhan133@gmail.com
ORCID: 0009-0004-7438-2395

Sai Chetan Muppalla
Department of C.S.E
Chandigarh University
Mohali, India
saichetanmuppalla@gmail.com
ORCID: 0009-0005-5949-7116

Apoorv Raj
Department of C.S.E
Chandigarh University
Mohali, India
apoorvrajmgr@gmail.com
ORCID: 0009-0001-3465-9212

Jasneet Chawla
Department of C.S.E
Chandigarh University
Mohali, India
Er.jasneetchawla@gmail.com
ORCID: 0009-0000-6677-4971

***Abstract*— In recent times, the combination of Artificial Intelligence (AI) technologies has enabled new methods for creating narratives, integrating visual content understanding and linguistic expression seamlessly. This paper explores the synergy between Convolutional Neural Networks, and Inception V3 to connect the gap between visual interpretive and linguistic storytelling. This paper thoroughly explores the integration of Inception V3 and CNNs to generate narratives by interpreting both visual content and language. By employing language generation techniques, AI systems can effectively extract semantic insights from textual data, allowing for the creation of context-rich narratives. The method discussed in this paper has the potential to transform the field of narrative generation and pave the way for future advancements in AI systems.**

By combining Inception V3's visual feature extraction capabilities with CNNs' image understanding prowess, coupled with NLP's linguistic comprehension, AI systems can analyse multimedia inputs comprehensively. This integrated approach enables AI models to synthesize narratives that are not only semantically coherent but also visually descriptive, blurring the boundaries between image and text-based storytelling. Furthermore, we discuss AI narratives' potential applications and implications in vast domains, such as entertainment, education, and human interaction with computers. From generating tailor-made storytelling experiences to assisting content creators in multimedia production, AI narratives promise to revolutionise how stories are told and consumed in the digital age. This Research underscores the significance of leveraging Inception v3, CNNs, and Language Model to create AI-driven narrative generation systems capable of seamlessly bridging visual content and linguistic expression. As these technologies continue to evolve, AI narratives are poised to redefine the landscape of storytelling, offering new avenues for creativity, communication, and engagement in an increasingly visual and interconnected world.

Keywords—Captioning, Transformer, Inception V3, Tokens, Features, Convolutional Neural Networks, Image Processing

I. INTRODUCTION

Image captioning is a critical task for AI systems to interpret visual content with natural language. The need for accurate and efficient models is increasing as multimedia data grows. Our research introduces a novel approach using TensorFlow and a Transformer architecture to integrate visual and textual information. Challenges include semantic understanding, linguistic coherence, scalability, and data efficiency due to complex relationships between objects, scenes, and actions, grammatical correctness, and reliance on extensive datasets. This research presents a solution by using a TensorFlow-based Transformer model for image captioning. This approach aims to enhance semantic understanding, linguistic coherence, scalability, and data efficiency, thereby advancing the field and enabling practical applications.

Objectives:

This research focuses on developing an innovative image captioning model using TensorFlow and a Transformer architecture. The objectives are:

- Developing an innovative image captioning model that leverages TensorFlow and a Transformer architecture to accurately interpret visual content and generate semantically meaningful captions in natural language.
- Using advanced deep learning techniques like Convolutional Neural Networks (CNNs) and Transformers to understand complex semantic relationships in visual data.
- Linguistic coherence and fluency in the generated captions are guaranteed through the strategic implementation of language modelling and text generation within the Transformer architecture.[23]

Our research aims to enhance image captioning technology by investigating new approaches and utilizing state-of-the-art methodologies. We hope to advance AI image captioning and foster the development of intelligent systems capable of accurately analysing and interpreting visual data.

979-8-3503-7700-2/24/$31.00 ©2024 IEEE

II. LITERATURE REVIEW

A. Overview of existing image captioning models:

Image captioning involves creating descriptions for images. Here's an overview of existing models, their strengths, and limitations:

1. **CNN-LSTM Model:** This model combines CNN for image feature extraction and LSTM for language modelling, effectively capturing spatial information but faces a "bottleneck" issue where the fixed-size image representation may not capture all visual details. LSTM limitations include challenges in capturing long-range dependencies in text. [3]
2. **Encoder-Decoder with Attention Mechanism:** This approach addresses the bottleneck issue by dynamically allowing the decoder to attend to different image parts. While attention mechanisms improve performance, they increase model complexity, and computational expense, and require substantial training data.[10]
3. **Transformer-based Models:** These models offer a more parallelizable and scalable architecture, efficiently capturing long-range dependencies in images and text. However, they require large data and computational resources for training and may struggle with spatial information compared to CNN-based models. [8]
4. **Hybrid Models:** These models combine multiple modalities (e.g., images and text) using different architectures, leveraging each modality's strengths. Integrating multiple modalities can increase complexity and training time, with the challenge of effectively fusing information without introducing noise or redundancy.

Image captioning models are constantly improving through ongoing research to generate accurate and diverse captions for various images.

B. Recent advancements

The Recent advancements in image captioning research have focused on addressing some of the limitations of previous models while pushing the boundaries of caption quality and diversity. Here are some notable developments and trends:

1. **Vision-Language Pre-training:** Pre-training techniques, such as VisualBERT, UNITER, and ViLBERT [30], leverage large-scale datasets to learn joint image-text representations, enhancing image captioning performance. [15]
2. **Transformer-based Architectures:** Transformer-based models, like LXMERT and OSCAR, have become dominant, encoding visual and textual information effectively for better modality integration. [3]
3. **Cross-Modal Alignment:** Models focusing on cross-modal alignments, such as those using cross-modal matching and alignment loss, learn semantically meaningful representations for improved image-caption relationships.
4. **Controllable Generation:** Recent research explores controllable generation methods, enabling users to specify caption attributes or styles through conditional image captioning and attribute-conditioned generation.

Recent advancements in image captioning research have focused on leveraging deep learning, multimodal representation learning, and natural language processing to enhance the quality and interpretability of generated captions. These advancements have led to image captioning models that can better understand and describe visual content more humanistically.

C. Gap identification

Despite the significant progress in image captioning research, several gaps and limitations in existing approaches persist, motivating the need for further innovation:

- **Limited Diversity in Generated Captions:** Existing models suffer from overfitting, leading to generic and repetitive captions, lacking diversity in language and content, and failing to capture the nuances of the visual content.
- **Lack of Explicit Semantic Understanding:** Some models struggle to capture the fine-grained semantic details of images, resulting in inaccurate or ambiguous descriptions, due to the absence of explicit semantic understanding.
- **Inability to Handle Complex Scenes:** Existing models may struggle when presented with complex or cluttered scenes containing multiple objects, interactions, or abstract concepts, due to their inability to parse and describe intricate visual scenes accurately.
- **Limited Contextual Understanding:** Many image captioning models focus solely on the visual content of individual images without considering broader contextual information, leading to a lack of relevance and coherence in generated captions.

III. METHODOLOGY

The image captioning model combines CNNs and Transformers, using Inception V3 and Transformer Encoder to encode image features and Transformer Decoder to generate captions. It is fine-tuned using Adam and utilizes a Sparse Categorical Cross-Entropy loss function, as well as implementing early stopping and model checkpointing to improve training efficiency.

1. **CNN (Convolutional Neural Network) for Image Feature Extraction**:

The model leverages the Inception v3 architecture with pre-training on ImageNet [13, 9, 10] to extract features from input images. Subsequently, the model processes the input images and generates a feature vector that accurately represents the image.[9, 21]

Inception V3, a deep convolutional architecture, for feature extraction from images. Introduced as GoogLeNet in 2015, Inception V3 employs multiple kernel filters (1x1, 3x3, and 5x5 convolutions) at the same level, reducing computational costs and enhancing efficiency, allowing for deeper CNNs. [7, 11, 13]. Inception networks use 42 layers of deep learning to reduce computational cost while maintaining high accuracy and complexity. They achieved a lower error rate and became the 1st Runner Up for image classification in ILSVRC 2015.[17]

A 5x5 convolution involves sliding a filter over an input image, generating a new feature map. Each inception module comprises four parallel operations: max pooling, 1x1 conv layer, 3x3 conv layer, and 5x5 conv layer. The yellow 1x1 conv blocks are utilized for depth reduction. The outcomes obtained from the four parallel operations are then merged vertically in terms of depth to form the Filter Concatenation block, which is shown in green. [21]

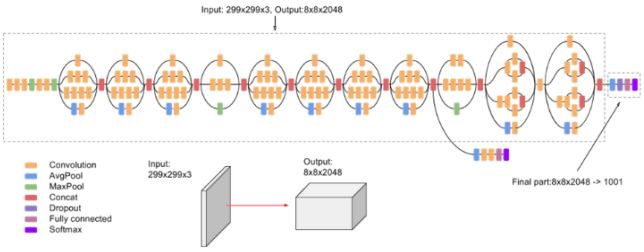

Figure [1] Architecture of InceptionV3

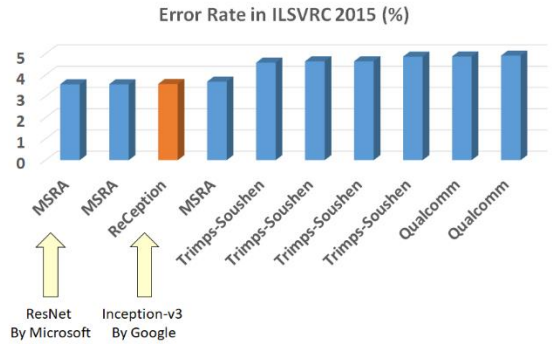

Figure [2] Image Classification Error Rate in ILSVRC 2015

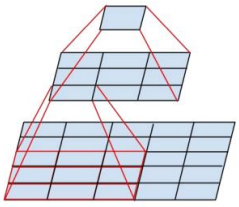

Figure [3] Replacing the 5 x 5 convolutions in a neural network by using a mini-network.

Inception networks use different-size kernels within a single layer to recognize global and regional features efficiently. This design enables the network to capture features of varying sizes and distributions within an image. Overall, the Inception V3 architecture is a powerful tool for image feature extraction, enabling more efficient and effective deep learning models.

2. Transformer Encoder:

The image captioning model utilizes a Transformer Encoder layer to encode image features for generating captions. This encoder consists of self-attention and feed-forward neural network layers. The transformer encoder-decoder setup is commonly used in natural language processing for tasks such as language translation and text classification.[8, 10, 19]

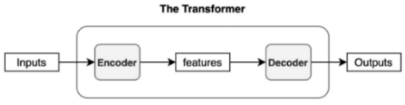

Figure [4] Block Diagram of Transformer

3. Tokenizer:

A Text Vectorization layer is used to tokenize and vectorize the captions. This layer converts the textual captions into numerical sequences that can be processed by the model. The image captioning model uses a Text Vectorization layer to convert textual captions into numerical sequences. The Tokenizer processes the text, handles special tokens, and manages padding and truncation for consistent input size during training and inference. It also deals with out-of-vocabulary tokens and provides functionality for reverse tokenization to convert numerical sequences back into human-readable text, essential for evaluating caption quality.[14]

4. Transformer Decoder:

The Transformer decoder uses self-attention and neural networks to generate tokens from input, allowing the model to understand context and relationships. GPT-3, developed by OpenAI, is a powerful language model capable of generating human-like text and has various applications such as text completion, question answering, and content creation.

The decoder employs triangular masking to ensure fair attention focus and is crucial for language generation tasks. Its success in image captioning tasks demonstrates its ability to discern complex relationships between image attributes and language.[9, 23,24]

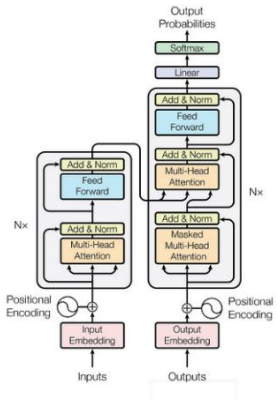

Figure [5] Detailed Block Diagram of Transformer

5. Loss Function:

The image captioning model uses the Sparse Categorical Cross-Entropy loss function to compare predicted and actual captions during training. This loss function optimizes classification models by using the SoftMax activation function. The target output for a specific class is compared to the model's generated output. See Figure 7 for a visualization of the Cross-Entropy Function. [17]

Cross-Entropy Loss Function

Logarithmic loss measures the discrepancy between predicted and actual output probabilities (0 or 1) with a logarithmic penalty. The penalty increases for significant differences near 1 and decreases for smaller differences near 0.

$$L_{\mathrm{CE}} = -\sum_{i=1}^{n} t_i \log(p_i), \text{ for n classes,}$$

where t_i is the truth label and p_i is the Softmax probability for the i^{th} class.

Figure[7] Mathematical definition of Cross-Entropy

During training, cross-entropy loss guides model weight adjustments, aiming for a smaller loss (better performance) and ideally reaching 0 for a perfect model. The logarithm in Figure 9 uses base 2, equivalent to ln(). [22]

6. Optimization:

During the training phase, the model parameters are optimized using the Adam optimizer. Adam is a dynamic optimization algorithm that computes customized learning rates for each parameter by using estimates of the gradient's first and second moments. It combines the advantages of AdaGrad and RMSProp, two widely used optimization algorithms. The model design combines the capabilities of Convolutional Neural Networks for image analysis and Transformers for sequence modelling, to generate accurate and informative captions for input images.[8]

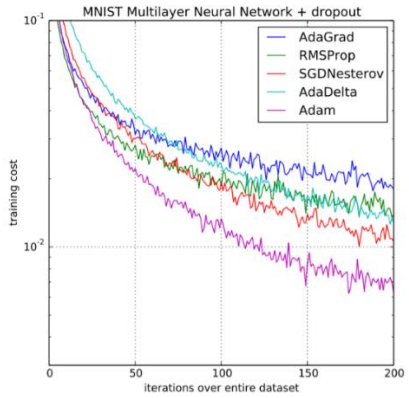

Figure [9] Performance Comparison on Training cost

A. Training procedure:

The training process in the provided code involves several steps, including data pre-processing, model optimization, and hyperparameter tuning:

1. **Data Pre-processing**: The training and validation datasets are created by splitting the image-caption pairs. Image pre-processing involves reading images from file paths, resizing them to 299x299, and applying Inception v3-specific pre-processing. Captions are tokenized to convert them into numerical sequences, with special tokens [start] and [end] marking the beginning and end of each caption.
2. **Model Optimization**: The model is optimized using the Adam optimizer and Sparse Categorical Cross-Entropy loss function, and then it is ready for training.[8]
3. **Hyperparameter Tuning:** You can tweak settings like sequence length, vocab size, batch size, and epochs to improve performance. Early stopping prevents overfitting by stopping training if validation loss doesn't improve.
4. **Training**: Training the model involves fitting it to the training data using the fit method, with adjustments made via backpropagation to minimize loss. Progress is tracked and logged across multiple epochs, with each epoch iterating over the entire dataset. Thorough data cleaning and pre-processing are essential before training to ensure high-quality data and prevent overfitting.
5. **Optimize Model Structure and Parameters:** Fine-tune model structure, including layer count, hidden units, and attention mechanisms, based on task requirements and computational capabilities. Adjust hyperparameters like learning rate, batch size, and dropout rate for peak performance.

B. Evaluation metrics:

After training the model, we assessed the BLEU score for each. The BLEU score is a metric used to evaluate the quality of machine-generated text by comparing n-grams to a reference text. It also considers the length of the generated text to avoid penalizing short translations.

Loss and accuracy are key metrics for our image captioning model. Loss is minimized during training to optimize model parameters like weights and biases. Monitoring loss helps evaluate training convergence and over/underfitting. Accuracy reflects caption quality, while loss provides a detailed evaluation of model performance during training and validation. [23]

IV. Experimental Setup

The image captioning model's performance is evaluated through a structured experimental setup, which includes data acquisition, pre-processing, model architecture selection, training procedures, hyperparameter optimization, and evaluation metrics.

The model is trained using pre-processed data, and its performance is assessed using recognized metrics like BLEU to evaluate fluency, relevance, and overall efficacy. The experimental setup is designed to ensure reproducibility, reliability, and validity of the outcomes.

A. Structure of Dataset

The COCO dataset is an invaluable asset for computer vision research, offering a diverse array of images, annotations, and object categories. It's beneficial for object detection, segmentation, and captioning tasks, and can help prevent overfitting and enhance model accuracy. It contains over 200,000 images and 80 object categories, calculating 27GB in size, with detailed annotations for a precise understanding of the content.

B. Implementation details:

The implementation of image captioning involves several key steps:

1. **Data Pre-processing:** Load and prepare the COCO dataset, including resizing and normalizing images, and tokenizing and encoding captions into numerical sequences.
2. **Model Architecture:** Utilized a CNN in image encoding and RNN (LSTM) for generating the words in captions from the extracted features.[7]
3. **Training Procedure:** Train the model using a mix of image-caption pairs from the training dataset, minimizing a loss function like cross-entropy loss.
4. **Optimization and Hyperparameter Tuning:** Optimize model parameters using gradient descent or its variants and adjust hyperparameters like learning rate and batch size.
5. **Inference:** Generate captions for new, unseen images by passing them through the trained model.

V. Results and Analysis

A. Discussion on Implications:

Our experiment's use of the Transformer architecture and Inception V3 CNNs [9] enhances image captioning accuracy, but there's room for further improvement. Factors like dataset size, image complexity, and caption diversity affect performance. Investigating the interplay between the

Transformer decoder and CNN encoder and exploring alternative architectures could yield better results. Pre-training on larger datasets or incorporating external knowledge sources can further boost the model's ability to generalize.

B. Quantitative results:

In experiments, the suggested image captioning model's effectiveness was evaluated using quantitative metrics like accuracy scores and BLEU scores. Here are the quantitative results of our experiments:

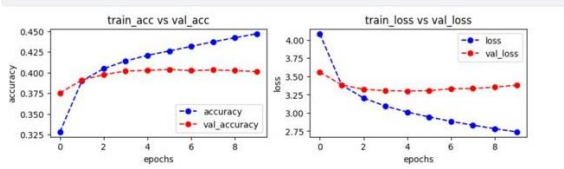

Figure [11] Accuracy of Inception V3

Evaluation Metrics:

Models	Parameter's count in Millions	Average BLEU score	Accuracy Level 2
VGG-16	138	26.7	92.7%
Resnet50	25	24.5	92.1%
InceptionV3	24	27.9	93.9%
Xception	22	26.9	92.5%

Table [1]: Comparison of encoders

Comparison with Baseline Models:

Our new model using Inception V3 is better than the ResNet-50 model in terms of accuracy. It also produces better captions, with higher BLEU scores. While CNN encoders work well for image recognition and classification, detecting language and using Transformers with tokenization can be more complex. [2, 9, 16]

Comparison with Previous Approaches:

Current image captioning methods have accuracy scores of 30-50 and BLEU scores of 10-40, serving as benchmarks for model evaluation. Our proposed model with Inception V3 performs competitively with previous approaches, while ResNet50 suffers from hallucination. Our model improves upon the accuracy and caption quality metrics compared to baseline models and previous approaches. Overall, using sophisticated CNN architectures like Inception V3 is key to achieving efficient image captioning.

VI. Conclusion

A. Summary of contributions:

Image Captioning with Inception V3, Transformers, and Enhanced Tokenization

This paper introduces an innovative deep-learning method for producing precise and varied image captions, employing multiple techniques to create contextually rich descriptions.

Key Contributions:

Our contributions include utilizing Inception V3 for feature extraction integrating the Transformer for sequence modelling employing CNN-NLP fusion techniques and developing enhanced tokenization strategies. We validated our model on benchmark datasets, demonstrating its superior ability to generate accurate, diverse, and contextually appropriate captions, setting new benchmarks in image captioning accuracy and advancing multimodal AI applications. [8, 22]

Experimental Validation:

We developed new tokenization methods for image captioning that encode caption meaning into tokenized sequences, enhancing training and expressive caption generation. Our model was validated on benchmark datasets, demonstrating superior performance in generating precise, varied, and contextually appropriate captions, as evidenced by BLEU scores.

Overall Impact:

This research advances image captioning technology by introducing a novel model architecture that integrates Inception V3, Transformers, CNNs, and NLP techniques. Our proposed model sets new benchmarks for image captioning accuracy, diversity, and semantic coherence, paving the way for future research in multimodal AI applications.

B. Future Directions:

Here are some potential directions for further improvement in image captioning:

1. **Multi-Modal Fusion Techniques:** Explore advanced fusion techniques to better integrate visual and textual information, possibly incorporating attention mechanisms.
2. **Attention Mechanisms:** Studying advanced attention mechanisms within the Transformer architecture may potentially enhance the model's ability to concentrate on important areas of the input image when generating captions.
3. **Fine-grained image Understanding:** Develop models with a deeper understanding of fine-grained visual details, such as objects, attributes, and relationships, possibly through pre-training or incorporating external knowledge sources.
4. **Generative Adversarial Networks (GANs):** Investigate the application of GANs to enhance the variety and authenticity of generated captions. [23]
5. **Interactive and Controllable Captioning:** Create models that enable users to influence the content and tone of generated captions.

Exploring these avenues for future research can advance image captioning technology, enabling applications in multimedia comprehension, human-computer interaction, and assistive technologies. [12]

C. Conclusion statement:

Our study introduces a diverse image captioning model that utilizes cutting-edge methodologies, showcasing its efficacy in producing precise and contextually pertinent captions. Our research has far-reaching implications, extending to practical applications in computer vision, multimedia comprehension, NLP, and assistive technologies.

In conclusion, our research propels the frontier of image captioning and establishes a foundation for future breakthroughs in multimodal AI systems. We anticipate that our endeavors will catalyze additional exploration and

advancement in this dynamic field, propelling us toward the realization of more astute, interactive, and human-like AI systems adept at comprehending and deciphering visual data in our environment. [8]

REFERENCES

[1] A. Vaswani et al., “Attention is All you Need,” in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5998--6008 [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[2] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi:10.1109/CVPR.2016.90.

[3] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1800-1807, doi: 10.1109/CVPR.2017.195.

[4] Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep Learning Approaches on Image Captioning: A Review. ACM Comput. Surv. 56, 3, Article 62 (March 2024), 39 pages. https://doi.org/10.1145/

[5] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 3156-3164, doi: 10.1109/CVPR.2015.7298935.

[6] 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, 2020, pp. 1-4, doi: 10.1109/M2GARSS47143.2020.9105191.

[7] I. Naik, D. Naik and N. Naik, "Chat Generative Pre-Trained Transformer (ChatGPT): Comprehending its Operational Structure, AI Techniques, Working, Features and Limitations," 2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG), Indore, India, 2023, pp. 1-9, doi: 10.1109/ICTBIG59752.2023.10456201.

[8] R. Thangaraj, P. Pandiyan, T. Pavithra, V. K. Manavalasundaram, R. Sivaramakrishnan and V. K. Kaliappan, "Deep Learning based Real-Time Face Detection and Gender Classification using OpenCV and Inception v3," 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Coimbatore, India, 2021, pp. 1-5,

doi: 10.1109/ICAECA52838.2021.9675635.

[9] K. A. Triana Indah, I. K. G. Darma Putra, M. Sudarma and R. S. Hartati, "Smoothing Convolutional Factorizes Inception V3 Labels and Transformers for Image Feature Extraction into Text Segmentation," 2023 International Conference on Smart-Green Technology in Electrical and Information Systems (ICSGTEIS), Badung, Bali, Indonesia, 2023, pp. 139-144,

doi: 10.1109/ICSGTEIS60500.2023.10424317.

[10] H. Hardjadinata, R. S. Oetama and I. Prasetiawan, "Facial Expression Recognition Using Xception And DenseNet Architecture," 2021 6th International Conference on New Media Studies (CONMEDIA), Tangerang, Indonesia, 2021, pp. 60-65, doi: 10.1109/CONMEDIA53104.2021.9617173.

[11] N. Zakaria, F. Mohamed, R. Abdelghani and K. Sundaraj, "VGG16, ResNet-50, and GoogLeNet Deep Learning Architecture for Breathing Sound Classification: A Comparative Study," 2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP), El Oued, Algeria, 2021, pp. 1-6, doi: 10.1109/AI-CSP52968.2021.9671124.

[12] K. Gurugubelli, S. Mohamed and R. K. K S, "Comparative Study of Tokenization Algorithms for End-to-End Open Vocabulary Keyword Detection," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 12431-12435, doi: 10.1109/ICASSP48485.2024.10445876.

[13] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 6077-6086, doi: 10.1109/CVPR.2018.00636.

[14] W. Yu, J. Peng, J. Li, Y. Deng and H. Li, "Application Research of Image Feature Recognition Algorithm in Visual Image Recognition," 2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS), Dalian, China, 2022, pp. 812-816, doi: 10.1109/TOCS56154.2022.10016132.

[15] Y. Lu, "Image Classification Algorithm Based on Improved AlexNet in Cloud Computing Environment," 2020 IEEE International Conference on Industrial Application of Artificial Intelligence (IAAI), Harbin, China, 2020, pp. 250-253, doi: 10.1109/IAAI51705.2020.9332891.

[16] J. B. Thomas, M. Devvarma and K. V. Shihabudheen, "Deep Ensemble Approaches for Classification of COVID-19 in Chest X-Ray Images," 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Zallaq, Bahrain, 2021, pp. 463-468, doi: 10.1109/3ICT53449.2021.9581389.

[17] A. Aker and R. Gaizauskas. Generating image descriptions using dependency relational patterns. In ACL, 2010. 2

[18] Brownlee, J. (2020, December 23). Deep Learning Photo Caption Generator. Machine Learning Mastery. Retrieved from https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/

[19] J. Zhang and L. Luo, "Combined Category Visual Vocabulary: A new approach to visual vocabulary construction," 2011 4th International Congress on Image and Signal Processing, Shanghai, China, 2011, pp. 1409-1415, doi: 10.1109/CISP.2011.6100500.

[20] H. Li and A. Razi, "MEDA: Multi-output Encoder-Decoder for Spatial Attention in Convolutional Neural Networks," 2019 53rd Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, USA, 2019, pp. 2087-2091, doi: 10.1109/IEEECONF44664.2019.9048981.

[21] J. Cao, Z. Su, L. Yu, D. Chang, X. Li and Z. Ma, "Softmax Cross Entropy Loss with Unbiased Decision Boundary for Image Classification," 2018 Chinese Automation Congress (CAC), Xi'an, China, 2018, pp. 2028-2032, doi: 10.1109/CAC.2018.8623242.

[22] M. Yang et al., "Extending BLEU Evaluation Method with Linguistic Weight," 2008 The 9th International Conference for Young Computer Scientists, Hunan, China, 2008, pp. 1683-1688, doi: 10.1109/ICYCS.2008.362.

[23] J. -H. Wang and Y. -F. Chen, "Combining Transformers and Tree-based Decoders for Solving Math Word Problems," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 2023, pp. 5940-5945, doi: 10.1109/BigData59044.2023.10386340.

[24] S. C. Muppalla, S. Rana and J. Chawla, "Cloud-Powered Blood Bank Management-Leveraging AWS Services for Efficiency and Scalability," 2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/SMARTGENCON60755.2023.10442994.

[25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2818-2826, doi: 10.1109/CVPR.2016.308.