JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A Multilingual Model for Text Generation of Visuals

[1]Ramya B N, [2]BanuShree T, [3]Greeshma P M, [4]Pallavi K G, [5] Pranitha T Manur

[1]Asst.Prof, [2,3,4,5]UG Student
[1,2,3,4,5]Artificial Intelligence & Machine Learning
[1,2,3,4,5]Jyothy Institute of Technology, Bengaluru, India

Abstract: Image captioning refers to a machine generating human-like captions describing the image. With the recent surge of interest in the field, deep learning models have been proved to give state-of-the-art. This research presents a model integrating convolutional neural networks (CNNs) for image feature extraction and transformer-based architectures for sequence generation, trained on the Flickr8k dataset. Results indicate varying accuracy across languages, with Tamil and Telugu achieving high levels (~90%), followed by English (~80%), and French, Spanish, and Kannada (~60%). These findings underscore the potential of multilingual captioning in enhancing accessibility and cross-cultural understanding, while highlighting the need for continued research to address challenges in linguistic diversity and accuracy, ultimately advancing inclusivity in visual perception.

***Index Terms* - Multilingual image captioning, Natural language processing, Deep Learning, Convolution Neural Network**

I. INTRODUCTION

In today's interconnected world, the fusion of computer vision and natural language processing has sparked innovative avenues, notably in the realm of multilingual image captioning. This paper delves into the pioneering domain of multilingual image captioning, presenting a novel approach that transcends linguistic boundaries to foster global communication. At the heart of this research lies the integration of convolutional neural networks (CNNs) for image feature extraction and transformer-based architectures for sequence generation, a unique synthesis aimed at revolutionizing how we perceive and interpret visual content across diverse languages. The novelty of this research lies in its holistic approach, which not only endeavours to generate descriptive captions for images but also strives to do so seamlessly across multiple languages, encompassing English, Spanish, French, German, and South Indian languages like Kannada, Telugu, and Tamil. By breaking down linguistic barriers, this approach opens doors to enhanced accessibility and cross-cultural understanding, paving the way for a more inclusive visual landscape.

Text generation holds profound significance in numerous domains, transcending mere captioning to become a cornerstone of modern communication and accessibility tools. The ability to automatically generate descriptive text for images facilitates comprehension for individuals with visual impairments, enabling them to access and interpret visual content independently. Moreover, in an era inundated with multimedia data, text generation plays a pivotal role in content-based image retrieval systems, enhancing searchability and discoverability across diverse linguistic contexts. Furthermore, text generation fosters inclusivity by enabling content creators to reach global audiences effortlessly, transcending language barriers and facilitating cross-cultural engagement. Thus, the importance of text generation extends far beyond mere captioning, encompassing accessibility, searchability, and global communication.

Several notable contributions in the literature underscore the significance and potential of multilingual image captioning and text generation. Xu et al. (2015) introduced a pioneering architecture employing visual attention mechanisms to enhance caption quality, setting new standards in image caption generation [1]. Li et al. (2018) proposed an entangled transformer model for image captioning, pushing the boundaries of accuracy and performance in this domain [3]. Mishra et al. (2020) explored transformer-based models for generating descriptive captions in Hindi, demonstrating promising results in multilingual captioning tasks [8]. These seminal works, among others, serve as foundational pillars in the research landscape, inspiring and guiding advancements in multilingual image captioning and text generation.

II. LITERATURE SURVEY

In the realm of image captioning, numerous significant contributions have emerged, each pushing the boundaries of research and innovation. Xu et al.'s pioneering work introduced the concept of visual attention mechanisms, revolutionizing caption quality by enabling selective focus on various parts of input images [1]. Following this, Faraby et al. tackled the challenge of expanding captioning to diverse languages with their model for Bengali caption generation, achieving remarkable BLEU scores that surpassed existing methods. This paper obtained BLEU1-42.6, BLEU2-27.95, BLEU3-23.66, BLEU4-16.41 and Meteor-28.7 respectively [2]. Li et al. further elevated captioning performance by introducing the Entangled Transformer, a model that intricately integrates spatial and channel-wise attention mechanisms, setting new benchmarks in caption quality metrics such as METEOR and CIDEr [3].

Saha's comprehensive guide to convolutional neural networks (CNNs) provided invaluable insights into the fundamental architectures underpinning many image captioning models, laying the groundwork for understanding CNNs' role in this domain [4]. Xiao et al.'s hierarchical encoder-decoder network introduced a novel approach capturing intricate visual features through multiple encoding and decoding layers, resulting in significant improvements in caption quality metrics such as BLEU and ROUGE-L [5]. Adhikari and Ghimire's exploration of Nepali image captioning demonstrated the feasibility of generating captions in less-represented languages, opening avenues for broader multilingual captioning efforts [6].

The introduction of the transformer architecture by Vaswani et al. marked a paradigm shift in NLP tasks, including image captioning, achieving state-of-the-art performance through its reliance on self-attention mechanisms. BLEU score of 41.8 after training for 3.5 days on eight GPUs [7]. Mishra et al.'s investigation into transformer-based models for Hindi captioning showed promising results, indicating the potential of such architectures in multilingual captioning tasks. The proposed model has attained the BLEU-1 score of 62.9, BLEU-2 score of 43.3, BLEU-3 score of 29.1, and BLEU4 score of 19.0. [8]. Emami et al.'s work on Arabic image captioning leveraged pre-training of deep bidirectional transformers to achieve competitive performance, further highlighting the effectiveness of pre-training in multilingual settings [9]. Finally, Liu et al.'s CPTR introduced a full transformer network for image captioning, demonstrating superior performance across various datasets and solidifying its position as a leading architecture in the field's achieves 129.4 [10].

Amidst this extensive array of research endeavours, our study distinguishes itself through its comprehensive approach to multilingual image captioning. By combining the strengths of CNNs for image feature extraction and transformer-based architectures for sequence generation, our model not only achieves accurate and high-quality captions across multiple languages but also addresses key challenges such as linguistic diversity and accuracy. Through rigorous experimentation and evaluation, our research advances the state-of-the-art in image captioning, offering a robust and versatile solution for bridging linguistic barriers and promoting inclusive communication in the visual realm.

III. Methodology

In this paper, we propose CNN, Transformer based and Image Captioning Model

CNN-EfficientNetB0

EfficientNetB0 stands out as a leading convolutional neural network (CNN) architecture, excelling in both performance and computational efficiency. Its core principle, "compound scaling," finely adjusts the network's depth, width, and resolution in unison, resulting in superior performance compared to models with fewer parameters.

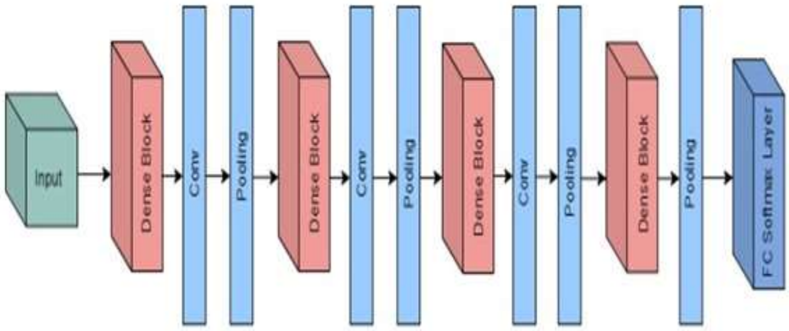

fig.1. EfficientNetB0

(1) Convolutional Layers

At the heart of EfficientNetB0 are convolutional layers that progressively extract features of varying abstraction levels from input data, starting with basic features like edges and progressing to more complex patterns.

$$\mathrm{Output(i,j,k)} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} \mathrm{Input}(i+m, j+n, l) \text{ x Filter } (m,n,l,k) + \mathrm{b(k)} \qquad (1)$$

This equation represents the calculation performed at each neuron in the convolutional layer, where **Input** is the input feature map, **Filter** is the filter (or kernel), b is the bias term, and M, N, and L are the dimensions of the filter.

(2) Depth-wise Separable Convolution
By employing depth-wise separable convolution, EfficientNetB0 splits the standard convolution into two distinct phases: depth-wise and point-wise convolution. This strategy significantly reduces computational overhead while preserving network efficiency.

$$\text{Depth-wise(i,j,k)}= \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} Input(i+m, j+n, l) \text{ x Filter}(m, n, l) \tag{2}$$

(3) Scaling
The cornerstone of EfficientNetB0's architecture, compound scaling, ensures a balanced adjustment of depth, width, and resolution as the network scales. This holistic approach enables the network to adapt across various tasks without compromising computational efficiency.
Depth = α^{ϕ}
α is a scaling factor controlling the number of layers.
Width = β^{ϕ}
β is a scaling factor controlling the number of channels in each layer.
Resolution = γ^{ϕ}
γ is a scaling factor controlling the input resolution.

(4) Global Average Pooling
Towards the end of the network, global average pooling condenses spatial dimensions into a single value per channel, mitigating overfitting and providing a succinct representation of extracted features.

$$\text{Output}(i,j,k) = \frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} \text{Input}(i,j,k) \tag{3}$$

(5) Fully Connected Layers
In certain configurations, EfficientNetB0 may incorporate fully connected layers for tasks like classification, mapping extracted features to desired output classes. EfficientNetB0 undergoes rigorous training using techniques such as backpropagation and stochastic gradient descent (SGD) with adaptive learning rate methods like Adam. Through iterative refinement, the network learns to extract relevant features and optimize parameters to minimize a specified loss function tailored to the task at hand.
In essence, EfficientNetB0 epitomizes a synergy of architectural innovation and mathematical precision, crafting a formidable framework for efficient and potent feature extraction in the domain of computer vision.

Transformers

The Transformer architecture represents a groundbreaking departure from traditional recurrent and convolutional architectures, relying solely on self-attention mechanisms and feed-forward neural networks. This design offers unprecedented capabilities in processing sequential data, garnering widespread adoption across various domains, particularly in natural language processing tasks

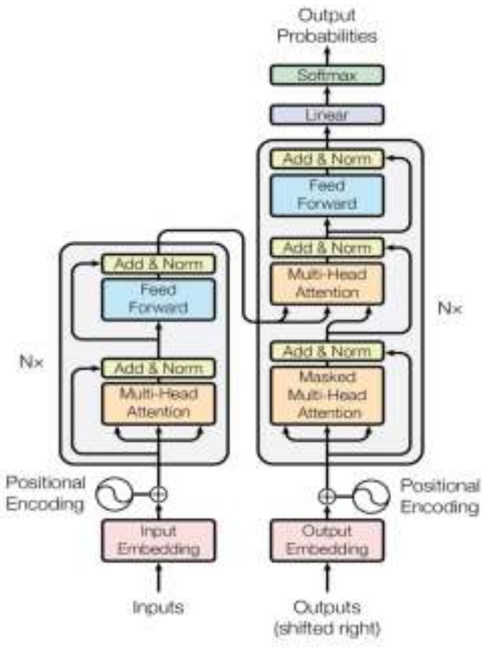

fig.2: Transformer architecture

Transformer Encoder

The Transformer encoder forms the basis for processing input features, transforming them into contextualized representations suitable for downstream processing. It consists of a stack of identical layers, utilizing self-attention mechanisms to analyze relationships within the input sequence and feed-forward neural networks for further processing. Layer normalization stabilizes training by ensuring consistent activations across layers, contributing to the model's robustness and efficiency.

(1) Self-Attention Mechanism: The encoder begins by leveraging self-attention mechanisms to analyze the relationships between different elements within the input sequence. This mechanism allows the model to weigh the importance of each element based on its relevance to others, enabling effective information aggregation and contextualization.

(2) Feed-Forward Neural Networks: Following self-attention, the encoder utilizes feed-forward neural networks to process the attended representations further. These networks consist of multiple layers of fully connected units, typically augmented with activation functions like Rectified Linear Units (ReLU). This hierarchical processing facilitates the extraction of complex features and patterns from the input data. The second sublayer is a fully connected feed-forward network consisting of two linear transformations with Rectified Linear Unit (ReLU) activation in between:

$$\mathrm{FFN}(x) = \mathrm{ReLU}(W_{1x} + b_1)W_2 + b_2 \tag{4}$$

The six layers of the Transformer encoder apply the same linear transformations to all the words in the input sequence, but each layer employs different weight (W1,W2) and bias (b1,b2) parameters to do so.

(3) Layer Normalization: At each step of processing, layer normalization is applied to stabilize the training process by ensuring consistent activations across layers. This normalization step contributes to the overall robustness and efficiency of the model by mitigating issues related to vanishing or exploding gradients. layernorm(.), which normalizes the sum computed between the sublayer input, x, and the output generated by the sublayer itself, sublayer(x):

$$\mathrm{layernorm}(\, x + \mathrm{sublayer}(x)) \tag{5}$$

The output of the encoder module comprises a set of contextualized representations, capturing both local and global dependencies within the input data. These representations serve as a rich source of information for subsequent stages of processing within the Transformer architecture. Our unified multimodal architecture strikes a balance between simplicity and performance, offering promising prospects for further advancements in multimodal systems.

Transformer Decoder

In contrast to the encoder, the Transformer decoder focuses on generating coherent outputs based on the encoded representations provided by the encoder. Its role is particularly prominent in tasks such as language translation and text generation, where the model must produce meaningful sequences of tokens or symbols.
The decoding process unfolds through a series of steps mirroring those of the encoder, albeit with some key differences:

(1) Self-Attention and Cross-Attention: Similar to the encoder, the decoder employs self-attention mechanisms to capture intra-sequence dependencies. Additionally, it incorporates cross-attention mechanisms to attend to the encoded representations generated by the encoder, enabling the model to leverage context from the input features effectively.

(2) Feed-Forward Neural Networks: Following attention mechanisms, the decoder utilizes feed-forward neural networks to process the attended representations and generate output sequences. These networks operate in a manner akin to those in the encoder, facilitating feature extraction and transformation.

(3) Positional Encoding: To account for the sequential nature of the output generation task, the decoder incorporates positional encoding to provide information about the relative positions of tokens within the output sequence. This enables the model to understand and preserve the sequential order of generated tokens, crucial for producing coherent outputs.

(4) Layer Normalization: Similar to the encoder, layer normalization is applied after each sub-layer within the decoder to stabilize training and ensure consistent activations.

Overall, the Transformer decoder plays a pivotal role in generating contextually relevant outputs based on the encoded representations provided by the encoder. Through a combination of attention mechanisms, feed-forward networks, and positional encoding, it achieves remarkable performance in tasks requiring sequence generation and synthesis.

VI. WORKING OF IMAGE CAPTIONING MODEL

Our research paper introduces an innovative caption generator model aimed at advancing image captioning technology. At its core, the model adopts an encoder-decoder architecture, combining the strengths of a pre-trained Convolutional Neural Network (CNN), EfficientNetB0, for extracting detailed visual features from input images, and a Transformer-based decoder. This encoder-decoder setup enables the model to generate descriptive captions by comprehensively understanding both the visual content of the images and the linguistic structure of the captions.

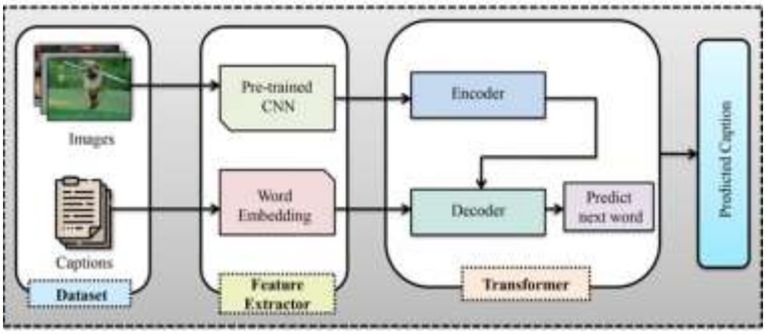

fig.3: Captioning model

During the training phase, our model learns to associate input images with their corresponding captions. This process involves optimizing the model's parameters using gradient descent techniques, particularly employing the Adam optimizer. By minimizing the Sparse Categorical Cross-Entropy loss, which measures the disparity between predicted and actual captions, the model refines its caption generation capabilities. Additionally, training on a meticulously curated dataset split into training and validation sets ensures the model's robust generalization to unseen data scenarios.

The introduction of Transformers in the decoder represents a significant breakthrough in our research. Transformers excel at capturing intricate dependencies in sequences, making them exceptionally well-suited for tasks like language generation. Leveraging Transformers empowers our model to adeptly capture the nuanced relationships between visual features and textual descriptions, surpassing the capabilities of conventional architectures. The impact of our research extends beyond the technical advancements in image captioning. By generating descriptive captions for images, our model significantly enhances accessibility and comprehension of visual content across various applications. This includes aiding image search, facilitating content recommendation, and providing accessibility tools for the visually impaired. Furthermore, the model's demonstrated ability to generate captions in multiple languages underscores its versatility, making it applicable across diverse linguistic landscapes and fostering cross-cultural communication and inclusivity in multimedia understanding.

In summary, our research paper presents a pioneering approach to image captioning, integrating cutting-edge technology to produce descriptive captions for images. This innovation holds significant promise for advancing image captioning research and has broad applications in enhancing accessibility, facilitating cross-cultural communication, and fostering inclusivity in multimedia understanding.

V.RESULTS

Our paper presents an image captioning model that integrates convolutional neural networks (cnns) for image feature extraction and transformer-based architectures for sequence generation. Utilizing the Flickr8k dataset, we preprocess the data to establish connections between image filenames and their corresponding captions. The model architecture comprises an Efficient Net CNN as the image encoder and a transformer-based encoder-decoder architecture for caption generation. During training, we minimize the cross-entropy loss between predicted and ground-truth captions, employing sparse categorical cross-entropy loss and optimization through the Adam optimizer with a customized learning rate schedule. To enhance generalization and prevent overfitting, we implement early stopping techniques and data augmentation methods such as random flipping, rotation, and contrast adjustment.

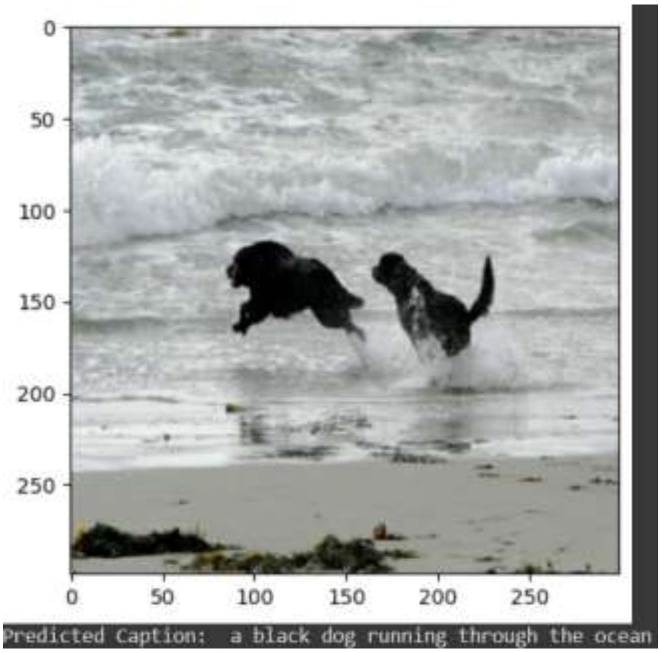

fig.4: Text Generated in Single Language

fig.5: Text Generated in Multiple Language

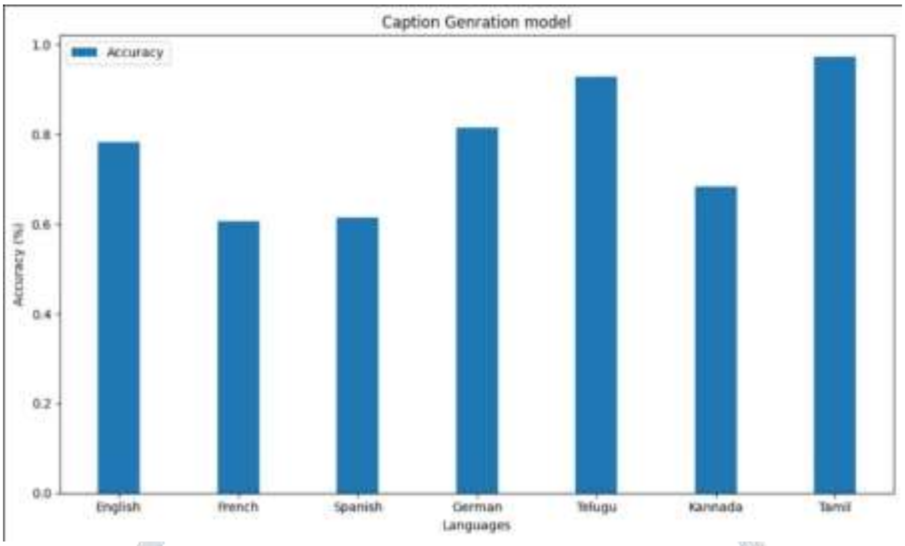

fig.6: Accuracy across Languages graph

Upon evaluation, our model demonstrates varying levels of accuracy across different languages. Notably, Tamil and Telugu achieve high accuracies around 90%, indicating strong comprehension and accuracy in caption generation. English follows closely behind with an accuracy of 80%, showing proficient performance. However, French, Spanish, and Kannada exhibit slightly lower accuracies around 60%, suggesting moderate success in caption generation. Overall, our model's performance varies across languages, highlighting potential disparities in language comprehension and caption generation capabilities. Through the integration of CNNs and transformer-based architectures, along with careful preprocessing and training techniques, our paper presents a robust image captioning model capable of generating accurate captions across diverse languages.

VI. CONCULSION

Our research underscores the profound impact of multilingual image captioning in transcending language barriers and promoting global communication. Leveraging the capabilities of AI, our study not only demonstrates the generation of descriptive captions across multiple languages but also addresses the imperative need for accuracy and inclusivity. By meticulously overcoming challenges such as linguistic diversity and precision, our methods pave the way for a more accessible and inclusive digital landscape. Moreover, our paper significantly contributes to the advancement of AI and machine learning by presenting innovative approaches to multilingual captioning and establishing stringent evaluation criteria. Through rigorous experimentation and meticulous analysis, we have demonstrated the efficacy of our techniques in producing accurate and culturally sensitive captions, thus bridging the gap between diverse linguistic communities.

As we continue to navigate the complexities of language and culture in the digital age, our research serves as a beacon of progress, advocating for inclusivity and cross-cultural understanding. By championing the transformative potential of multilingual captioning, we envision a future where technology not only facilitates communication but also fosters empathy and mutual respect across borders. In essence, our work signifies a crucial step towards a more connected and harmonious global society.

VII. References

[1] Liu, Haixia & Brailsford, Tim. (2023). Reproducing “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”. Journal of Physics: Conference Series. 2589. 012012. 10.1088/1742-6596/2589/1/012012
[2] Faruk, Al & Faraby, Hasan & Azad, Md & Fedous, Md & Morol, Md. Kishor. (2020). Image to Bengali Caption Generation Using Deep CNN and Bidirectional Gated Recurrent Unit.
[3] G. Li, L. Zhu, P. Liu and Y. Yang, "Entangled Transformer for Image Captioning," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 8927-8936, doi: 10.1109/ICCV.2019.00902. keywords: {Semantics;Visualization;Decoding;Encoding;Logic gates;Snow;Proposals},
[4] X. Chen, T.-Y. L. Hao Fang, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325, 2015. 1
[5] X. Xiao, L. Wang, K. Ding, S. Xiang and C. Pan, "Deep Hierarchical Encoder–Decoder Network for Image Captioning," in IEEE Transactions on Multimedia, vol. 21, no. 11, pp. 2942-2956, Nov. 2019, doi: 10.1109/TMM.2019.2915033.
keywords: {Visualization;Semantics;Hidden Markov models;Decoding;Logic gates;Training;Computer architecture;Deep hierarchical structure;encoder–decoder;LSTM;image captioning;retrieval;vision-sentence},
[6] A. Adhikari and S. Ghimire, "Nepali Image Captioning," 2019 Artificial Intelligence for Transforming Business and Society (AITB), Kathmandu, Nepal, 2019, pp. 1-6, doi: 10.1109/AITB48515.2019.8947436. keywords: {Training;Visualization;Task analysis;Recurrent neural networks;Decoding;Data models;Grammar},

[7] Vaswani, Ashish, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. “Attention is All you Need.” Neural Information Processing Systems (2017).
[8] Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha, Pushpak Bhattacharyya, Amit Kumar Singh, Image captioning in Hindi language using transformer networks, Computers & Electrical Engineering, Volume 92, 2021, 107114, ISSN 0045-7906, https://doi.org/10.1016/j.compeleceng.2021.107114.
[9] Emami, Jonathan & Nugues, Pierre & Elnagar, Ashraf & Afyouni, Imad. (2022). Arabic Image Captioning using Pre-training of Deep Bidirectional Transformers. 40-51. 10.18653/v1/2022.inlg-main.4.
[10] Liu, Wei & Sihan, Chen & Guo, Longteng & Zhu, Xinxin & Liu, Jing. (2021). CPTR: Full Transformer Network for Image Captioning.
[11] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In EMNLP, 2016. 1, 2, 8
[12] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In CVPR, 2017. 1, 6, 8
[13] O’Shea, Keiron and Ryan Nash. “An Introduction to Convolutional Neural Networks.” ArXiv abs/1511.08458 (2015): n. pag.
[14] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. arXiv preprint arXiv:1603.05027, 2016. 7
[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 1997. 3
[16] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. arXiv preprint arXiv:1606.08390, 2016. 2, 5
[17] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In NIPS, 2015. 2
[18] J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang. Aligning where to see and what to tell: image caption with region based attention and scene factorization. arXiv preprint arXiv:1506.06272, 2015. 2
[19] A. Karpathy and F.-F. Li. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015. 6
[20] V. Kazemi and A. Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. arXiv preprint arXiv:1704.03162, 2017. 2, 5
[21] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. arXiv preprint arXiv:1602.07332, 2016. 3, 5, 10
[22] C. Lin. Rouge: a package for automatic evaluation of summaries. In ACL Workshop, 2004. 6
[23] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. In ECCV, 2014. 5, 6
[24] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of spider. In ICCV, 2017. 7
[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.- Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. arXiv preprint arXiv:1512.02325, 2015. 2
[26] J. Lu, X. Lin, D. Batra, and D. Parikh. Deeper lstm and normalized cnn visual question answering model. https://github.com/VT-vision-lab/ VQA_LSTM_CNN, 2015. 8
[27] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In CVPR, 2017. 1, 2, 3, 7
[28] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In NIPS, 2016. 1, 2
[29] K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In ACL, 2002. 6
[30] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek. Areas of attention for image captioning. In ICCV, 2017. 2
[31] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global Vectors for Word Representation. In Proceedings of the Conference on Empirical Methods for Natural Language Processing (EMNLP), 2014. 10
[32] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In CVPR, 2016. 2