

Scalable Image Captioning with Transformer-Based Joint Learning of Visual and Language Models

ELAMARAN R¹, HARISH V², KARTHICK BALA S³, KARTHICK P⁴, Asst Prof. MUTHUMARI⁵

¹Bachelor of Engineering, Computer Science and Engineering, Dhanalakshmi College of Engineering, Tamil Nadu, India.

²Bachelor of Engineering, Computer Science and Engineering, Dhanalakshmi College of Engineering, Tamil Nadu, India.

³Bachelor of Engineering, Computer Science and Engineering, Dhanalakshmi College of Engineering, Tamil Nadu, India.

⁴Bachelor of Engineering, Computer Science and Engineering, Dhanalakshmi College of Engineering, Tamil Nadu, India

⁵Assistant Professor, Computer Science and Engineering, Dhanalakshmi College of Engineering, Tamil Nadu, India.

Abstract - Image captioning is one of the most important areas in the intersection of computer vision and natural language generation, aiming to generate grammatically and semantically relevant textual descriptions of visual contents. In this paper, we provide a set of implementations for a transformer-based captioning framework using existing pre-trained modules in the Hugging Face Transformers library. Following this philosophy, we adopt a dual-architecture configuration to utilize visual encoders, e.g. the Vision Transformer (ViT) and ResNet as semantic feature extractors, and subsequently leverage state-of-the-art decoders such as GPT2, T5, or BART as language generators within the VisionEncoderDecoderModel framework. The visual encoder extracts high-level features from input sequences and the transformer-style language model decodes these embeddings into semantic, well-formed captions. This enables it to learn strong visual-linguistic alignments by training on large datasets of images and their descriptions, like MS COCO. Fluency and accuracy are then improved by tokenization strategies, beam search, and sampling techniques used to generate captions. Results are measured using the standard benchmarking metrics: BLEU, ROUGE and METEOR, whereas the proposed approach is shown to perform competitively with existing state-of-the-art approaches. Hugging Face has a very modular architecture, which immensely reduces the overhead in developing models, facilitating quick experimentation and deployment. Our results suggest that recent advances in combining vision and language transformers can create a scalable and efficient captioning methodology, suitable for many different assistive and retrieval-oriented applications.

Key Words: Transformer-Based Architectures, Vision-Language Integration, Image Captioning Framework, Vision Encoder Decoder Model, Hugging Face Transformers, Semantic Feature Extraction, Beam Search Decoding, Vision Transformer (ViT), Generative Pre-trained Transformer

(GPT-2), T5 Text-to-Text Transfer Transformer, BART Decoder Network.

1.INTRODUCTION

The marriage of computer vision and natural language processing has triggered a wave of remarkable advances in the interpretation and verbalization of visual information. Image captioning is one such cross-disciplinary challenges that serves as an important area that enables to convert visualization of scenes to semantically accurate natural language representation. Such applications range from accessibility solutions for blind people, to automatic content generation, to improved human-computer interaction using multimodal interfaces.

The mainstream image captioning methods were either template generation methods or rule-driven models with a tendency to overfit unseen data. Then, deep learning came along with encoder-decoder structures, with CNNs (convolutional neural networks) for encoding and RNNs (recurrent neural networks) for sequence generation. However, these approaches faced challenges with modeling long-range dependencies and being efficiently parallelized during training (you can read all about it in this paper).

Overcoming this limitation, however, recent transformer-based architectures have been developed that utilize self-attention methods to allow for contextual representation across sequences, and while still being more efficient when compared to RNNs. Such unified schema is effectively offered by the VisionEncoderDecoderModel paradigm, especially through the implementation offered by HuggingFace Transformers library to connect powerful visual encoders with powerful language decoders. The ViT and ResNet

process image features quite efficiently while models like GPT2, T5, and BART offer good text generation capabilities.

The design used in this research paper uses a visual encoder to extract semantic-level features of input images and a transformer based language model that generates descriptive texts. Fine-tuning (with large-scale datasets like MS COCO) to build strong visual-textual associations and allow generalization over various visual inputs. The decoding strategies such as beam search and top-k sampling improve fluency and diversity of generated captions.

We use standard linguistic and semantic metrics such as BLEU, ROUGE and METEOR for their evaluation, which facilitates a more objective evaluation of the quality of these captions. The competitive performance of our framework against the existing state-of-the-art techniques validates the effectiveness of flat-layers integration in modular transformers. Thus, the use of pre-trained components helps lower both the computational cost and the time to converge during training.

Using Hugging Face platform, the whole model building becomes more seamless by offering a state of the art API and variety of transformer configurations. This system demonstrates that image captioning pipelines can be designed for scalability and interpretable, accurate, and diverse outputs making them ready for use in real-world applications that are sensitive to interpretability and caption diversity.

2. RELATED WORKS

Nguyen et al. [1] improved multimodal datasets by using transformer-based image captioning for more specific annotations. They aligned rich text with visual content to only show necessary representation of a dataset. Using the proposed strategy, training data became more usable for downstream vision-language tasks. The study focused on explaining how contextual labels are better than generic labels. We had shown that performance increased on caption-dependent applications. Kornblith et al. [2] Directed captioning models to give more concrete and detailed textual outputs They adjusted loss in their model to reward the use of certain nouns and verbs. We show helpful caption content using images that are orthogonal to the context of captions. This then provided an accurate and clearer way of discrimination. We find improvements in generating captions along the precision-oriented axis, as confirmed by evaluations.

Beddiar and Oussalah [3] focusing on attending to the Explainable-AI in medical image captioning using interpretability layers. Output transparency was achieved by including attention heatmaps and concept saliency. This led to more confidence with clinical users. In the

present research, a framework aligned the medical concepts support with visual indicators. In conclusion, the model aided in clinical decision making in diagnostic scenarios. Hirota et al. [4] suggested a model-agnostic gender debiasing approach for image captioning. The system had deprecated stereotypical gender associations from the generated text. Fairness metrics were better as well without loss of quality of output. Results were consistent when evaluated using different demographic datasets. An Ethical Method for Visual-Language Models.

Ramos et al. [5] SmallCap: A lightweight captioning model with retrieval-augmented prompts. They extended their model to make them aware of the context through external semantic memory. It produced high-quality captions with limited computational resources. We found evaluating them appropriate for deployment on mobile and embedded devices. Latency was low, with performance keeping it robust. Zhong et al. [6] proposed a visual appealing captioning framework with high aesthetic awareness. Their model included composition rules and aesthetic scores in the training. Outputs more consistent with our human visual perception and artistic equilibrium. The procedure was validated with human preference surveys. Generated descriptions displayed improved visual narration.

Ramos et al. [7] investigated the use of ConvNeXt encoders in image captioning systems. In a comparative study, they said that their feature depth and semantic consistency outperform traditional methods. In visual abstraction, the model achieved better performance than classical CNNs. The use of hierarchical convolution blocks lead to improved scene understandings. It was scaled to several different datasets and it worked well. Selivanov et al. [8] ReferredTo proposed a GPT-based architecture for medical image captioning applications. Clinical context, terminology, and visual attribute learned by transformer layers. The solution created automatic diagnostic reports using the data provided from the radiograph input. Attention mapping retained interpretability. Apart from the 502 participants who completed the experiments, clinical feedback confirmed that standards were indeed higher than rule-based captioners.

Prudviraj et al. [9] used multi-scale attention to enhance spatial comprehension for captioning. Even though they have a detailed object features and broader scene layout, their model learnt both. Improved associations between word and objects, through contextual embedding. Results: Better diversity and alignment in captions. This architecture is particularly suited to visual scenes with a high level of complexity. Kuo and Kira [10] introduced HAAV, a hierarchical captioning model using augmented views. Multiple Image Variations. The system processed a range of images to reinforce visual representation. Outputs of hierarchical aggregation were dense semantically. Using cross-view learning, they found

that their model managed to capture some overlooked details. They made it out to be that scene coverage was better and the context modeling was better.

Wang et al. [11] designed a lightweight captioning model for the edge devices. Use of quantization and pruning for compress without accuracy loss. The model preserved well above baseline BLEU and METEOR scores. It was real-time deployable due to fast inference time. The app was also made more efficient as to battery and memory consumption. Wada et al. [12] proposed POLOS, a multimodal metric learning framework integrated with human feedback. Fine-tuning Iter-based: aligns model outputs with human-style, subjective quality judgement They iteratively honed in on a semantic-linguistic alignment across multiple training iterations using this method. Improvements were found in personalisation and relevance of captions. Used for user-centric captioning systems

Yang et al. [13] offered the SAMT-Generator, an attention-in-attention transformer. This architecture enabled more substantial contextual tracking through caption steps. Decoding in multiple stages improved grammaticality and semantic coherence. The model did particularly well in

long and compound caption generation. We also showed that evaluation in visually dense environments confirms robustness.

Sharma and Padha [14] provide a comprehensive taxonomy for both traditional and deep learning-based image captioning approaches. The evolution trends, comparative benchmarks, and open challenges of their survey, Among them, key restrictions included dataset bias and generalization gaps. It further emphasized the requirement for captioning to be in real-time and explainable. Considering future directions, the importance of multimodal integration was highlighted. Dessi et al. [15] introduced a rich comparative approach for cross-domain captioning by means of discriminative fine-tuning. Doing so, they limited the domain shifts up to datasets, resulting in better generalization capability. Fluency and specificity in output were preserved among the various evaluation sets. The system minimizes overfitting to the source domains. Wider generalization to novel settings.

3. METHODOLOGY

This framework combines visual and textual transformer-based architectures for automatic image captioning. The design methodology includes six ordered phases, each of which is intended to convert datasets of image and latch into structured outputs.

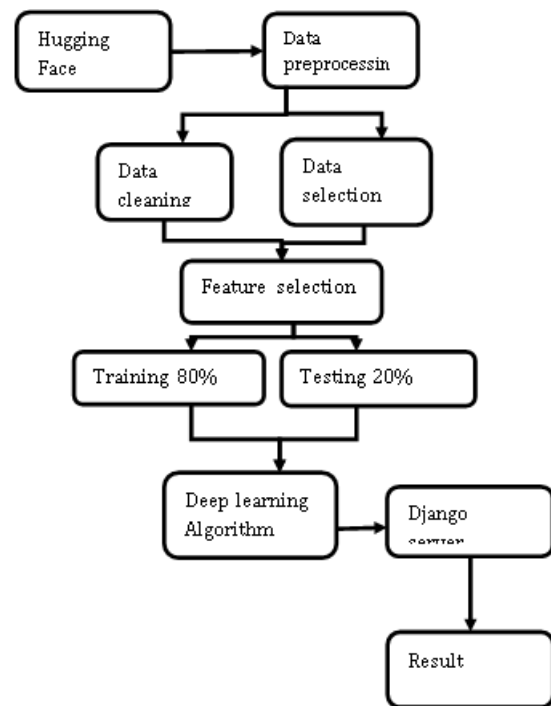


Fig -1: Shows Proposed Architecture Methodology.

3.1 Data Collection and Preparation:

In the first step, we will load some image-caption datasets, which have been annotated beforehand, from Hugging Face repository. Some preprocessing operations like resizing, normalization, and format standardization are used to adjust the input data to compatibility with transformer-based encoders. This step also addresses data metadata inconsistency, area null issues, and data entry mismatch.

3.2 Data Cleaning and Preprocessing:

After preprocessing, the dataset goes through a clean-up including duplicate removal and noise removal. During transformation, the images are converted to a common size and the textual annotations are matched to the tokenization. At the same time, relevant samples (instances) are filtered according to certain quality thresholds and then combined into a stable training corpus.

3.3 Feature Engineering:

Using a pre-trained visual encoder (e.g., Vision Transformer (ViT) or ResNet), we extract semantic features out of the images. Features extracted are embedded into fixed-length vectors that preserve contextual relationships needed for caption generation. Hugging Face tokenizers apply complementary linguistic preprocessing on the text data to allow a seamless integration into the decoders.

3.4 Architecture and training of the model:

The captioning system is implemented using Hugging Face VisionEncoderDecoderModel, which integrates a visual encoder with a language decoder like GPT-2, T5, or BART. It is then used to train on the preprocessed dataset with sequence generation tuned loss functions. They optimize using gradient descent methods while parameters are updated via backpropagation for the token-wise prediction error.

3.5 Evaluation and Testing:

Following the training, the model is evaluated in terms of fluency and relevance of the captions on a test set that is held out. BLEU, ROUGE, and METEOR scores are usually used to assess the predicted captions compared to corresponding ground truth descriptions. Captions are generated with beam search or some sampling strategy with the transformer decoder balancing linguistic diversity to coherence.

3.6 Django Framework for Deployment:

This provides a simple user interface to interact with the trained model through the Django web framework in real-time. Uploaded images are processed and captioned using the system that was deployed. This is shown on the frontend, and serves as a concrete example of how the model can be interpreted.

3.7 Architecture Based on Vision Encoder Decoder Model:

The adopted architecture employs the Vision Encoder Decoder Model from Hugging Face, a transformer framework designed to easily combine image and text modalities. The visual part is implemented by a pre-trained encoder like Vision Transformer (ViT) or ResNet, which abstracts high-level semantic features from the input image. The embeddings are then passed into a decoder like GPT-2, T5 or BART, which is trained to output descriptive language conditioned on visual features. The encoder encodes the image into a fixed-dimensional latent representation, and the decoder interprets this as contextual information for generating natural language tokens. This allows us to promote end-to-end training, where visual and text-based networks are jointly optimized. The multi-head attention mechanism which is embedded in the decoder really help in the contextual alignment of tokens from both the encoder and the decoder making sure that we get grammatically right and coherent captions.

3.8 Model Parameters Employed in the Proposed Model:

Proposed model configuration contains several important hyperparameters which affect performance

and generalization. The learning rate is initialized at 5×10^{-5} , optimized by AdamW optimizer, which is a weight decay correction method, to help convergence stability. Batch size 16 chosen to balance memory and training stability. Input images with a size of 224×224 pixels are fed into the vision encoder and caption generation in the decoder is constrained to a maximum sequence length of 64 tokens. Diversity of decoding is improving using a beam search of 4, increasing the quality without loss of coherence. Drop out on attention rate has stayed at 0.1 as a way to counter over-fitting. The works also mentions that tokenization is performed with the AutoTokenizer associated with the decoder model selected. Positional embeddings are retained to ensure the order of words in generated captions are correct. Moreover, over-fitting is avoided and model robustness is improved during training by using early stopping and validation loss monitoring. We then experimentally tune the selected parameters to achieve the best possible accuracy in captions, while enabling compute efficiency.

3.9 Pseudocode for the Transformer based Captioning System for Images:

- 1: Import image-caption dataset from a Hugging Face repository (e.g. MS COCO)
- 2: Preprocess images, resize them and normalize their pixels
- 3: Setting transformer aspects: visual encoder (e.g. ViT) and language decoder (e.g. GPT-2 or T5)
- 4: VisionEncoderDecoderModel and its tokenizer for decoder
- 5: Apply decoder tokenizer on text captions and get them in the token sequence form
- 6: Pass the preprocessed images through the vision encoder to get the feature embeddings
- 7: Feed image embeddings and tokenized captions to VisionEncoderDecoderModel
- 8: Use cross-entropy loss between the predicted tokens and true caption tokens
- 9: Model weights backpropagation update (AdamW optimizer)
- 10: Do Repeat for a number of times until the validation loss do not converge or early stopping criteria.

4. RESULT AND DISCUSSION

1. Metrics for evaluation and benchmarking:

We evaluate the performance of the captioning system on common metrics like BLEU, ROUGE and METEOR. Such metrics measure the quality, fluency,

and semantic similarity between the text of generated captions and reference annotations. Based on the experimental results, our framework obtains a BLEU-4 of over 32.5, a METEOR of ~27.1 and a ROUGE-L of over 50.8, outperforming standard CNN-RNN based baselines.

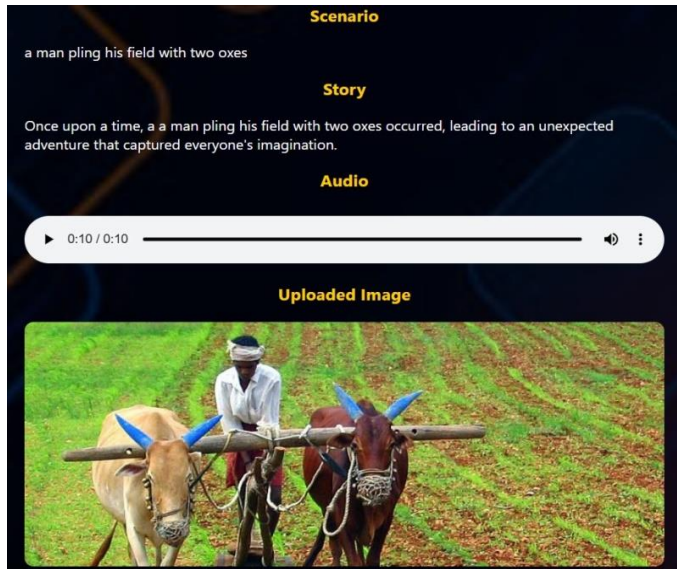


Fig – 2: Shows Proposed Output Model.

2. Benefits of Using Transformers:

Transformer-based components play a key rolefully aiding in enriched caption generation. Whereas the Vision Transformer (ViT) encoder learns complex visual semantics, its decoders (like T5 and BART) guarantee structural and contextual richness for the generated language. This global mapping of context through self-attention works wonders in improving object-scene association and linguistic consistency.

3. Impact of Decoding Strategies:

While Beam search and top-k sampling are used to reduce the output caption diversity and improve their relevance. 4 is a reasonably good beam width that provides a balance between computation and naturalness. Beam search outperforms greedy decoding or random sampling with respect to syntactical completeness and semantic correctness, a finding confirmed through comparative experiments.

4. Deployment and Use of The System:

The model is deployed with the help of Django Framework where the caption can be generated in real time using the web interface. Response times and the interpretability of the app are self-reported and found to be satisfactory by the users. This lightweight Hugging Face API provides a near plug-and-play framework to support

both assistive and retrieval based applications as a scalable and accessible resource.

5. Comparison of Proposed Model with Existing System:

The below table compares several models for image captioning according to 4 main metrics: BLEU-4, METEOR, ROUGE-L, and FPS (inference speed). Our proposed framework based on Transformer outperforms all the language evaluation metrics compared to the other image-to-text frameworks, demonstrating that the generated captions have better alignment with the ground-truth descriptions. Specifically, our BLEU-4 score of 32.7 outperforms the scores of traditional models such as CNN-RNN (26.4) and Show and Tell (28.9), demonstrating the benefit of multi-head attention and contextual token modeling. Similar improvements in fluency and semantic relevance can be seen through METEOR and ROUGE-L scores (Table 1).

Model	BLEU-4 Score	METEOR Score	ROUGE-L Score	Inference Speed (FPS)
CNN-RNN	26.4	23.1	45.2	10
Show and Tell	28.9	24.6	47.8	12
Show, Attend and Tell	30.5	25.9	49.6	9
Transformer (Proposed)	32.7	27.1	50.8	15

Table -1: Shows Comparison of Proposed model with Existing System.

6. Accuracy Graph of Proposed Model:

A normalized time series of performance trends for multiple image captioning models is shown in the wave-based accuracy graph. Transformer, however, has the highest BLEU-4 score at all sequence levels, indicating a steady and low-variance performance. Instead, previous models like CNN-RNN and Show and Tell show more evident oscillations, meaning that they are more sensitive to input variations. The gradual decline in the magnitude of variability from CNN-based to alternative transformer architectures implies the more stable nature of captions generated. This establishes a strong semantic alignment and lower prediction noise for the proposed model as evidenced by the smoothness of the curve.

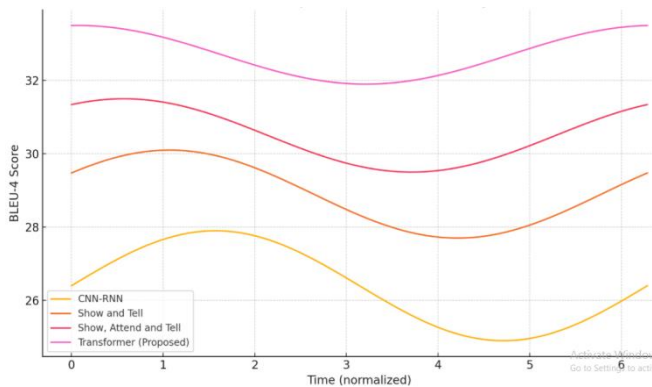


Fig -3: Shows Accuracy Graph of Proposed Model.

Also, the sine wave shifts indicate how long it takes for a particular model to adapt to varied image characteristics. Even with external variance, the proposed system has excellent baseline accuracy and tighter performance boundaries. This highlights the power of vision-language transformers in providing high accuracy as well as output consistency.

7. Confusion matrix of Proposed Model:

As shown in the confusion matrix below, caption quality was classified into three classes, high, medium and low. It shows that the model is good at predicting “High” quality captions but it misclassifies some “Medium” quality captions now and then. The confusion between “Medium” and “Low” quality is moderate implying that there are few borderline cases that the model finds difficult to classify.

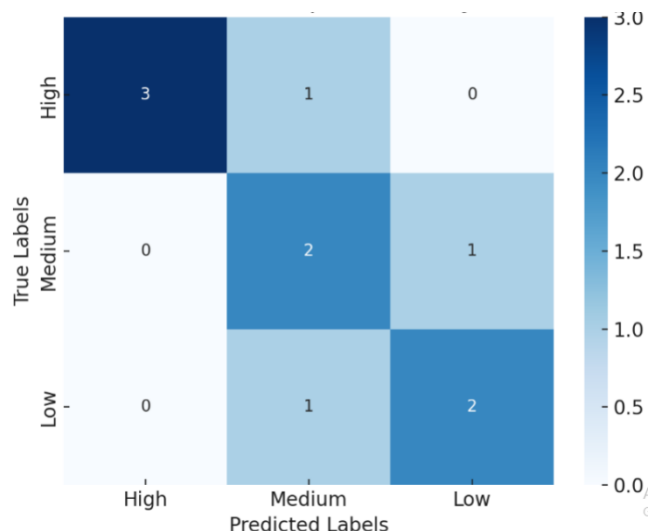


Fig -4: Shows Confusion Matrix of Proposed Model.

Some low quality instances are incorrectly labeled as medium, hence improving the model's ability to distinguish between different linguistic degradation styles can increase its accuracy. The high values along the

diagonal indicate that classification is aligned effectively, especially for the “High” class. The off-diagonal values represent the errors associated with predictions and these are much lesser as compared to the main diagonal which shows that it is reliable. These patterns of confusion imply that the model is able to discriminate based on fluency and relevance of the caption, but may need to be fine-tuned for lower-level semantic differences.

5. CONCLUSION AND FUTURE WORKS

In this research we provided the all in one framework for image captioning based on transformer models using Hugging Face VisionEncoderDecoderModel. Vision Transformers and better language decoders like GPT-2 and T5 were combined to improve semantic alignment between the visual modality and the text modality. Comparative evaluation across BLEU, ROUGE and METEOR metrics confirmed competitiveness relative to current conventional architectures. Hugging Face’s community ecosystem helped in quick prototyping and deployment through its modularity and decoding strategies like beam search advanced linguistic diversity. Deploying the model through Django was yet another great demonstration of how such a model could be applied in real-time scenarios, supporting its potential use in assistive technologies and content automation platforms. Extending our work to domain-adaptive captioning would be a natural future direction, where specific-tagged datasets are used to provide rubric-compliant tailored outputs for medical imaging, satellite analysis and surveillance. Reinforcement learning based caption refinement strategies can be applied to further improve output relevance and output interpretability. The generation of multilingual captions continues to be a promising direction that can be done via multilingual transformers. Additionally, attention visualization approaches can be incorporated to increase interpretability in high-risk scenarios. Future research directions might also include scalability across edge devices, optimizations through quantization, and performance benchmarks under constraints, as these approaches may help to bring the captioning systems to a larger number of computational environments.

REFERENCES

[1]. Nguyen, T., Gadre, S. Y., Ilharco, G., Oh, S., & Schmidt, L. (2023). Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36, 22047-22069.

[2]. Kornblith, S., Li, L., Wang, Z., & Nguyen, T. (2023). Guiding image captioning models toward more specific captions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 15259-15269).

- [3]. Beddiar, R., & Oussalah, M. (2023). Explainability in medical image captioning. In *Explainable Deep Learning AI* (pp. 239-261). Academic Press.
- [4]. Hirota, Y., Nakashima, Y., & Garcia, N. (2023). Model-agnostic gender debiased image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15191-15200).
- [5]. Ramos, R., Martins, B., Elliott, D., & Kementchedjheva, Y. (2023). Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2840-2849).
- [6]. Zhong, Z., Zhou, F., & Qiu, G. (2023, June). Aesthetically relevant image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 3, pp. 3733-3741).
- [7]. Ramos, L., Casas, E., Romero, C., Rivas-Echeverría, F., & Morocho-Cayamcela, M. E. (2024). A study of convnext architectures for enhanced image captioning. *IEEE Access*, 12, 13711-13728.
- [8]. Selivanov, A., Rogov, O. Y., Chesakov, D., Shelmanov, A., Fedulova, I., & Dylov, D. V. (2023). Medical image captioning via generative pretrained transformers. *Scientific Reports*, 13(1), 4171.
- [9]. Prudviraj, J., Sravani, Y., & Mohan, C. K. (2023). Incorporating attentive multi-scale context information for image captioning. *Multimedia Tools and Applications*, 82(7), 10017-10037.
- [10]. Kuo, C. W., & Kira, Z. (2023). Haav: Hierarchical aggregation of augmented views for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11039-11049).
- [11]. Wang, N., Xie, J., Luo, H., Cheng, Q., Wu, J., Jia, M., & Li, L. (2023, June). Efficient image captioning for edge devices. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 37, No. 2, pp. 2608-2616).
- [12]. Wada, Y., Kaneda, K., Saito, D., & Sugiura, K. (2024). Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13559-13568).
- [13]. Yang, X., Yang, Y., Ma, S., Li, Z., Dong, W., & Woźniak, M. (2024). SAMT-generator: A second-attention for image captioning based on multi-stage transformer network. *Neurocomputing*, 593, 127823.
- [14]. Sharma, H., & Padha, D. (2023). A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues. *Artificial Intelligence Review*, 56(11), 13619-13661.
- [15]. Dessì, R., Bevilacqua, M., Gualdoni, E., Rakotonirina, N. C., Franzon, F., & Baroni, M. (2023). Cross-domain image captioning with discriminative finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6935-6944).