Advances in Deep Learning Image Captioning and Their Application for Caption-Guided Sentiment-Aware Hybrid Transformer

Rauki Yadav[1], Dr. Vineet Kumar Goel[2], Dr. Sohit Agarwal[3]
[1]*Research Scholar, Computer Science and Engineering, Bhagwan Mahavir Centre for Advanced Research, Vesu, Surat, 395007, Gujarat, India*
[2]*Dean, Faculty of Engineering, Bhagwan Mahavir University, Vesu, Surat, 395007, Gujarat, India*
[3]*Professor, School of Computer Science, Suresh Gyan Vihar University, Jaipur 302017, Rajasthan*

***Abstract*—The generation of image captions has emerged as a promising domain of research that merges the field of visual understanding with natural language processing. The challenge of generating an image caption requires the model in question to identify the most salient objects in an image, evaluate the relationship among the items, and translate that evaluation into coherent language. The proposed work will assess a subset of the deep learning approaches that have contributed to advances in image caption generation, including early convolutional and recurrent networks, cutting-edge attention-based decoder methods, and transformer-based networks. It expands to assess recent efforts that are focused on emotional types of caption generation, fusion of multiple modalities, optimizing caption generation with reinforced learning, hybrid encoders, and transformer-based methods. Most current models have shown good performance across benchmarks, including MS COCO. However, issues concerning accurate representations of abstract scenes, the hallucination of content, and alternate application contexts persist. This paper suggests an increase in significant improvement through the use of grounding, providing valid evaluation metrics, and architectures that depict semantic structure more accurately. This paper proposes a novel multimodal framework called the Caption-Guided Sentiment-Aware Hybrid Transformer-CNN model.The proposed architecture integrates semantic scene understanding, sentiment-aware reasoning, and deep visual feature learning within a unified framework. The model is composed of three complementary modules.1.Vision-Language Model (VLM) backbone, inspired by architectures such as BLIP-2 and LLaVA, to generate scene-descriptive natural language captions from sampled classroom video frames.2.RoBERTa-based sentiment analysis network fine-tuned on educational discourse corpora, including classroom conversations, student feedback, and instructor-student interactions.3.EfficientNet-B4 convolutional neural network used as the primary visual feature extractor. The proposed framework establishes that combining semantic caption understanding, sentiment-aware contextual reasoning, and transformer-based multimodal fusion can significantly enhance automatic classroom engagement recognition, particularly in real-world and visually ambiguous learning environments.**

***Index Terms*—Image Caption Generation, Deep Learning, Attention Mechanisms, Vision Transformers, Multimodal Learning**

I. INTRODUCTION

Generating natural language descriptions from images involves two domains - visual recognition and linguistic generation, which is developed in a single computational process. In a sophisticated language generation system, the images must be capable of recognizing objects, deducing spatial and semantic relations, awareness of situational context, and transforming the findings into coherent sentences. The intersection of perception and language reasoning motivates the research surrounding image captioning as a problem in computer vision and natural language processing. Early systems used convolutional networks for visual encoding and recurrent networks for sentence generation; however, the emergent need for increasingly complex visual scenes in generating coherent and contextually relevant captions led to more complex deep learning models. More recent studies indicate advances in attention mechanisms,

multimodal fusion models, and transformers that have further improved semantic alignment between images and text caption descriptions [1].

The importance of image captioning has increased, given the various possibilities for accessible and interpretable visual comprehension in a variety of contexts. Technologies to assist people visual impairment leverage a captioning system to turn visual media into a structured description of content. These technologies rely on a very specific type of model that can interpret content with both factual relevance and context. The tonal or emotional representation of image captioning is evident in the work of affective captioning [1]. Affective captioning proposes that models can go beyond objective attribution by adding interpretability to the emotional properties or style when describing artistic or sentiment driven images. Computer vision researchers have also written on reinforcement driven captioning systems that include training or learning mechanics for the best description by refining the quality of descriptions according to language and meaning evaluation [2]. Development of reinforcement technologies illustrates that captioning is not only about perception of images, but also involves learning processes that develop because of evaluation enactments and performance requirements of users with these models.

Deep learning (DL) has supplanted conventional methods owing to its mechanism of allowing models to navigate the learning of hierarchical representations of features directly from the data. Convolutional encoders can map global and regional visual features to the model, and recurrent and transformer-based decoders design fluent linguistic sequences. Baseline architectures established with convolutional encoders and recurrent decoders offer a framework to consider the critical visual workflow of caption generation. An example of this is provided by study [3] which demonstrates a classical pipeline of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) trained on several benchmark datasets and articulates how feature extraction, vocabulary, and decoding methods affect the intelligibility and grammaticality of captions. Emotional captioning research extends this thinking, as CSPDenseNet based encoders and BiLSTM decoders are integrated to capture fine grained affective and structural cues to affect descriptive richness [4]. This is an important evolution as captioning is becoming more concerned with concatenating multiple visual and linguistic signals and building these outputs in a unified framework.

Attention mechanisms have transformed the facets of captioning by allowing models to selectively attend to the salient portion of the image when generating a single word. Attention helps align visual features to the linguistic output, while also maximizing the model's vision to interpret complicated scenes that involve interactions between multiple objects. Even the evidence of transducer-based captioning demonstrates that self-attention-based architectures improve performance as they learn to predict long range dependencies and represent the global relationship in the image's context [5]. This has shifted a sequence-based processing paradigm embodied by recurrent neural networks, to parallel attention-based models that also afford the model's ability to represent rich visual semantics. Context aware attention models can further extend this by allowing the model dynamic adjustment of its focus when depicted with more complex semantic needs [6].

This paper provides a systematic implementation of deep learning techniques used for image captioning. Specifically, the models from traditional CNN and Recurrent Neural Network (RNN) baselines to advanced attention based and transformer-based architectures. Further, we include the study of multimodal 88o deep learning approaches for image captioning. Subsequent sections review the components of architectures that support state-of-the-art captioning systems, including encoders, decoders, attention mechanisms, and transformer architecture. The contributions of individual studies are synthesized and placed in the broader trajectory of the field. The discussion section assesses some common challenges to dataset biases, hallucination, evaluation challenges, and domain sensitivity. The applications and future directions consider the local context of real-world usage, potential new research areas, and the requirements for safe deployment. The paper concludes by encapsulating the evolution of image captioning and examines the future trajectory of multimodal based and transformer-based systems.

II. BACKGROUND

The generation of image descriptions constitutes a multi-modal task implementing visual understanding

and language production. An effective captioning system must perceive objects, make spatial and semantic inferences, and generate a coherent description in natural language. This ultimately involves converting high dimensional visual stimulus to more structured and ordered linguistic representation. This process requires a translation between visual information that becomes increasingly abstracted in very high dimensional spaces to linguistic definitions and structured forms of representation. Examples from emotional captioning are provided that illustrate how understanding an image must go beyond detecting the objects within the image to affective interpretation. The examples suggest visual perception and linguistic abstraction must be working in conjunction within a unified schema [1]. Likewise, where reinforcement driven captioning approaches have been used, linguistic quality description relies on the relationship between visual grounding and language fluency [2]. Deep learning pipelines representing these multimodal systems based on convolutional encoders with recurrent decoders provide the basis for multimodal integration. The deep learning models allow for extraction of visual features using CNNs, whereby visual features can be mapped to sequences produced by the language models [3][4].

The generation of image captions is complicated, since there is an intrinsic discrepancy between visual and linguistic types of information. Images contain spatial layout, texture, interactions between different objects, and contextual information, while language conveys meaning through grammatical structure, syntax, and a relationship of semantic dependences. The transition of information from images to text means that models will need to learn adequate representations that relate these disparate modes of information. Classically, encoder–decoder frameworks that apply convolutional architecture for global or regional feature extraction can generate populations of captions using recurrent decoder architectures or transformers. Research in captioning with transformer-based architecture posits that self-attention assists with more precise representations of visual and linguistic relationships and supports a structured mapping between image tokens and textual outputs [5]. Context aware attention mechanisms reinforce the need to selectively attend and contextually weight regions of images, which helps ensure that embeddings are enlivened with relevance, attention, and descriptiveness for the to-be-generated captions [6].

Vision and language unite within the domain of end-to-end captioning methods. Initial encoders of visual content encode an image into feature representations. To accomplish this, convolutional networks are employed, where ResNet, DenseNet, and similar architectures uncover hierarchical mapping that is critical for identifying objects and scene-level contextualization. Hybrid encoder–decoder systems indicate how the choice of encoder influences the depth and accuracy of generated captions. For instance, ResNet50 is included in a hybrid model with LSTM and Gated Recurrent Unit (GRU) decoders, and the authors found that feature extraction quality and recurrent contextual mechanisms can enhance fluency and structural consistency in created captions [10]. The development of transformer architectures and other new ways to represent the visual information using Generative Adversarial Network (GAN) allows for improved global context representation and encourages the usage of different linguistically responsive constraints that improve visual representation and linguistic realism [7].

There are two principal features to the caption generation task. The second feature involves 'language decoders.' Early efforts at caption generation have contained recurrent neural networks simply because the organization of natural language is sequential. In addition to using recurrent models on a bidirectional level to enhance dependencies across the sentence structure and improve grammatical health and coherence, decoders utilizing LSTM and GRU have addressed the long-standing vanishing gradient dilemma that exists when using recurrent networks and can accommodate longer contexts [4]. Attention-enhanced decoders increased flexibility with utterances developed by means of explicit attention on groupings of visual regions alongside generated language. Research focusing on object-level attention have consistently found that associating specific regions of the image with language tokens has increased accuracy and have also produced human-like patterns of interpretation when exploring attention in caption models [9]. As these examples demonstrate, the models for decoders are as much a part of the language as the effectiveness of the output and how visual information is used to frame the language and captions.

Conventional image captioning methods were reliant on template based or retrieval-based approaches, which exhibited limited flexibility to generalize across varying scenes. Deep learning methods went far beyond these early approaches by allowing models to learn visual and language mode patterns directly from data. Studies involving military and aerial imagery have shown that domain specific captioning requires encoders capable of hosting low altitude perspectives, densely cluttered object arrangements, and domain specific visual categories [8]. These studies clearly demonstrate the importance of dataset features and domain adaptation in the design of generalized capable captioning systems. Transformer based captioning studies are a further evolution to this by demonstrating how the parallel processing multilevel head attention can resolve recurrent bottlenecks to allow models to reason about semantic relationships at a global scale [5].

Joint representation learning is emerging as yet another important development in image captioning. Using a multimodal approach, joint representation learning strives to align visual and linguistic representations within the same embedding space to improve semantic grounding and relevancy to the captioned image, and potentially improve caption utilization of descriptive language. Reinforcement based approaches optimize captioning policy against evaluation metrics and linguistic characteristics to enhance descriptive accuracy and fluency that surpass teacher forcing paradigms [2]. Context aware attention models illustrate that some form of dynamic feature weighting is a necessity for establishing meaning of complicated scenes. Hybrid CNN–RNN systems and transformer-based systems also illustrate the benefits of using local feature extraction and contextualizing features globally [6][7]. All in all, these advances historically represent the movement from traditional handcrafted approaches to current complex architectures that learn multimodal relationships directly from data.

The transition toward transformers indicates a significant departure from earlier captioning methods. Transformer based architectures process images as organized sets of features, using self-attention to model dependencies without sequential expectations. This offers the opportunity for transformers to model nuanced semantic relations, and ensure solid contextual harmonicity between images and natural language descriptions. Research combining vision transformers with additional generative elements indicates improvements in descriptive expressiveness and stylistic naturalness [7]. Encoder–decoders combining CNN encoders and transformers at their decoders further illustrate that transformers improve generalization and grammatical richness even in concert with traditional convolutional encoders [10]. These developments firmly situate transformer architectures as the touchstone for captioning research and overall, as the foundation for future multimodal systems.

III. DEEP LEARNING IN IMAGE CAPTION GENERATION

Deep learning has shifted the prospects of image captioning by allowing models to learn visual and linguistic representations within a single computational framework. The process involves transforming high dimensional visual inputs into text descriptions with semantic coherence, and identifying an architecture is required which can extract discriminative features, offer a sequential modelling capability, and align two modal representations of visual and linguistic elements. Initial work in emotional captioning recognizes the need of having a model with an architecture that accommodates both objective and affective properties of the image. These studies also provided evidence that cross attention model architectures were able to provide better semantic grounding in a work of art [1]. Reinforcement based captioning systems provided data to indicate that training models to optimize their behaviour according to some reward structure driven by the metric would influence linguistic fluency and descriptive accuracy of the captions generated [2]. Encoder–decoder pipelines architecture has represented traditional workflows in various deep learning studies, where Convolutional encoders extract the visual features and recurrent decoders generate the captions, for example. While these systems have provided examples of data preprocessing, vocabulary creation and evaluation considerations as part of the visual language modelling and generation process, they also guide subsequent progression of environmental architecture supporting continued deep learning advancements [3].The addition of more complex convolutional encoders such

as CSPDenseNet plus bidirectional recurrent decoders to interpret sentiment rich scenes extended this scheme in emotion captioning systems [4]. Later advances of transformer-based methods expanded these schemata with self-attention mechanisms capable of modelling long-ranged dependencies between image features and text sequences [5]. Attention models that were contextually aware contributed more enhancements by adapting attentional focus based upon changing semantic needs improving interpretability at the level of regions for image captions [6].

3.1 Deep Learning Architectures

3.1.1 Encoder–Decoder Frameworks

The fundamental framework for contemporary, deep learning-based captioning systems is the encoder-decoder architecture. In this architecture, the encoder transforms the input image into a compact feature representation, whereas the decoder utilizes the encoder's representation to output a natural language description. Hybrid frameworks also exemplify the adaptability of the encoder-decoder architecture. For instance, models that leverage classical convolutional backbones with GANs support the idea that adversarial loss can improve visual embedding accuracy while enhancing linguistic naturalness [7]. Recently, domain specific applications have emerged in the field of military surveillance backup encoder-decoder systems that tried to integrate specific detection modules into the architecture to allow the interpretation of low altitude images which contain small objects and complex spatial arrangements [8]. Attention based encoder-decoder models based on deep learning which enhance the decoder behaviour by learning correspondence between the image region and generated tokens are developed to further enhance their approach. This ability to enhance the alignment between visual and linguistic domains improves the capabilities of handling complex scenes composed of multiple and interacting entities [9]. Hybrid recurrent decoders that combined LSTM and GRU adapted from natural language robotics further suggest that encouraging complementary gating mechanisms in a hybrid decoder associates stability and expressiveness to the generative model var [10]. Collectively, these advances suggest that the encoder-decoder architecture can be flexibly adapted to become a learning framework capable of addressing differences in visual conditions, language objectives and performance aspects

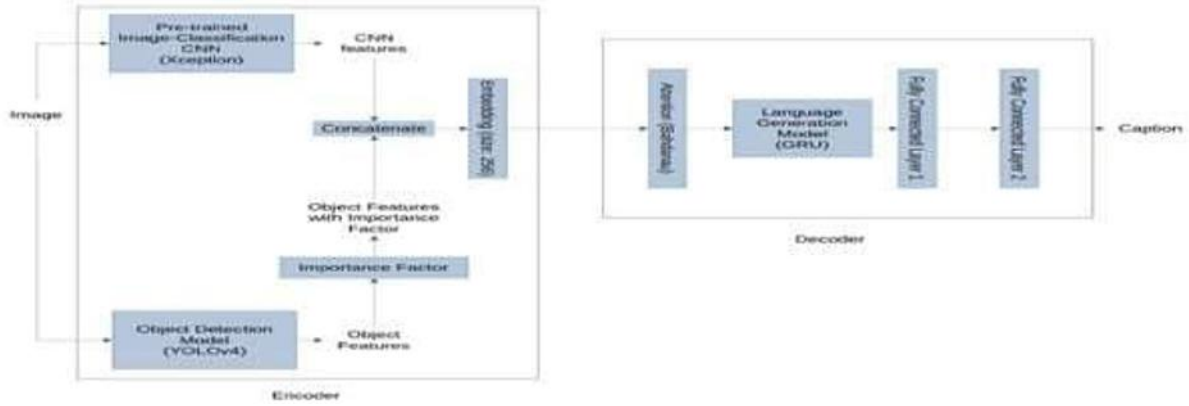

Fig.1. Block Diagram of Image Captioning model of Encoder-Decoder Framework [9]

3.1.2 CNN Based Image Feature Extractors

Convolutional neural networks are central to the extraction of visual features across most captioning systems. The hierarchical structure of convolutional networks facilitates the recognition of textures, edges, shapes, and essential text semantics to caption the item of interest. Bi-directional LSTM recurrent decoder systems depend on high-quality convolution features to obtain continuity in language modelling [11]. General deep learning pipelines with standard CNN encoders suggest that high-quality preprocessing steps, such as resizing, normalization, and data augmentation, will likely yield good recognition and caption quality overall [12]. Multimodal frameworks indicate strong support for disparate visual encoders by showing that learned image features interact with learned text embeddings directly within joint representation spaces [13]. Systems designed for application types (e.g., web-based captioning systems) reinforce the need for lightweight and efficient

convolutional encoders and utilize a CNN-style model to provide real-time inferences [14].Comparative investigations into machine learning and deep learning techniques lend credence to the advantage of convolutional feature extractors over more traditional hand engineered pipelines because the ability of convolutional networks to learn complex visual patterns is ubiquitous [15]. Hybrid approaches, including CNN based encoders with recurrent decoders, show the significant impact that encoder choice has with respect to the richness of descriptive language and diversity of vocabulary [16]. Convolutional encoders have also been used with emotion focused framework with affective processing modules [17]. Studies of pretrained encoders show that utilizing robust initialization a priori significantly improves resilience and cross domain generalization [18]. Integrated systems show that even when paired with transformers or reinforcement driven decoders, CNN based features remain a consistent and critical component [19]. Comparative syntheses suggest that while transformer encoders are emerging, CNNs still have strong applicability to captioning tasks [20]. Recent work has highlighted that accurate visual recognition remains challenging and often requires DL models capable of extracting complex and discriminative features from images related to agriculture and Healthcare domain [21] [22].

3.1.3 Sequence Modelling Using RNN, LSTM, and GRU

Comparative studies have identified a benefit of convolutional feature extractors over hand engineered pipelines in that convolutional networks are well known for learning a variety of complex visual patterns [15]. Hybrid systems, like CNN based encoders with recurrent decoders, demonstrate the influence encoder choice has on the richness of descriptive language and variety of vocabulary [16]. Convolutional based encoders have also been used in an emotion-based framework with affective processing modules [17]. Studies of pretrained encoders demonstrate benefits of using robust initialization a priori to improve resilience and cross-domain generalization [18]. Integrated systems have shown that CNN based features persist as a consistent and important component, while integrated with either transformers or reinforcement-based decoders [19]. Comparative syntheses establish that while transformer encoders are starting to emerge, CNNs have a considerable value in captioning tasks [20].

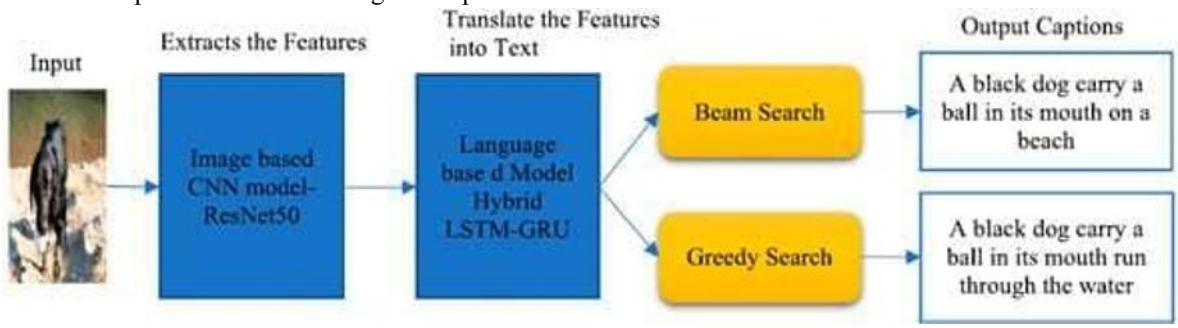

Fig.2. Image Captioning model with LSTM-GRU [10]

3.1.4 Attention Mechanisms and Transformer Models

Emotional cross attention mechanisms demonstrate attention's ability to align visual features with emotional descriptors to yield richer interpretative captions [1]. Reinforcement based models utilize attention in their policy optimization methods, which leads to better grounding and less incoherent phrasing [2]. Classical encoder–decoder frameworks include attention to better incorporate and align visual tokens with generated word tasks [3]. Emotion based systems reinforce feature extraction in their attention driven CSPDenseNet encoders [4]. Transformer based frameworks incorporate multi head self-attention to better model global dependencies [5]. Context aware attention provides adaptive weighting of regions based upon change in semantic demands [6]. Models leveraging transformers and GAN objectives contribute to improved stylistic fluency and semantic consistency with perceived visual coherence [7]. Domain specific transformer encoders seek to address aerial imagery challenges by employing scene level reasoning in cases of limited contextual regions [8]. Object level attention helps create human-like interpretive behaviour as attention aligns specific

detected regions to token labels [9]. Hybrid encoder–decoder systems show that transformer decoders provide better global coherence even with CNN encoder [10]. Recurrent transformers incorporate contextual recurrence into self-attention for improved sequential modelling [11]. A complete examination establishes that transformer architecture constitutes the leading edge of captioning investigations [20].

3.2 Datasets for Image Captioning

Datasets fundamentally impact model performance, by establishing linguistic variation, scene mixing, and object distributions. Affective analysis and subjective interpretation are supported through emotional captioning datasets [1]. Methods based in reinforcement rely on caption corpora with a rich distribution of linguistic variability to provide stable signals of reward, and avoid signalling rich variables [2]. Classical deep learning pipelines are very reliant on MS COCO and Flickr datasets, which provide multiple reference captions per image [3]. Emotion focused datasets can support increasingly granular learning of stylistic features [4]. Transformer based approaches support scale of datasets due to their increased parameter complexity [5]. Context aware approaches rely on datasets included diverse spatial and relational arrangements in the data [6]. Transformers–GAN hybrid approaches often utilize large corpora for stable convergence in adversarial learning [7]. Military datasets, and specialized visual patterns can support domain-specific modelling [8]. Object focused datasets provide detailed local region annotations for use in attention-based captioning [9]. The exploration of hybrid (Encoder-Decoder) systems is conducted using balanced datasets to assess the impact of encoder depth and decoder expressiveness [10]. Bidirectional recurrent systems are modelled with datasets that are linguistically rich to represent forward and reversal dependencies [11]. Multimodal datasets are beneficial to explore the evaluation of cross modal alignment [13]. In comparative studies associated with machine learning, the advantages of deep learning datasets mentioned above highlight the advantages of deep-learning datasets in comparison to traditional labeled corpora [15]. Integrated systems substantiate as fitting points of consideration when contemplating the inference of appropriateness and deployment of the datasets [19]. Survey evidence suggests that the datasets influence model generalization and bias patterns. [20]

Fig.3. Sample image from the Flickr30k dataset with captions [23]

3.3 Evaluation Metrics

Reliable evaluative metrics are important for evaluating both linguistic quality and visual grounding. The standard metrics Bilingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit ORdering (METEOR), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Consensus-based Image Description Evaluation (CIDEr), and Semantic Propositional Image Caption Evaluation (SPICE) remain the standards established with early systems and are still used with modern approaches. Emotion and stylistic captioning approaches use these metrics to achieve a balance between factual correctness and diversity and

expressivity, framing the tuition of these approaches [1]. Reinforcement learning motivated methods optimize policies directly against metric-based reward signals [2]. Baseline CNN- RNN models have used BLEU and METEOR as measures of sentence structural quality and semantic overlap [3]. Emotion - oriented architectures study how metrics behave under specific stylistic transformations while minimizing any effect of the source image and text combinations [4]. Transformer - based systems demonstrate improved BLEU scores compared to CNN - RNN models due to improved ability to capture stronger alignment and global semantics [5]. Context - aware models assess precision of alignment using region - sensitive or metric definitions [6]. Vision transformers apply adversarial evaluators to assess diversity of linguistic outcomes [7]. Domain - specific models use task - oriented metrics to evaluate robustness of language to notate specific imagery or domains [8]. Region - focused attention models use SPICE and other grounding metrics to measure semantic relationships to any or all of the images region captions [9]. Hybrid (encoder-decoder) architectures consider trajectories of metric performance over time during decoder optimization [10]. Studies of bidirectional recurrent modelling (short) evaluate measures of enhanced linguistic cohesiveness using METEOR and ROUGE scores [11]. Multimodal frameworks incorporate retrieval - based metrics for determining levels of correspondence across modalities [13]. Comparative analysis of machine learning generation notes incremental performance improvements throughout the expansion in capabilities and use of deep learning [15]. Surveys note that much consideration of metric limitations can be addressed through semantically reliable evaluation tools [20].

IV. RELATED WORK

The growth of image captioning applications began to move away from traditional convolutional-recurrent architectures towards attention-based, multimodal aligned frames, and affective reasoning transformer-based models. For instance, recent work in emotional captioning framed image captioning that competes with factual recognition, by taking capture of cross attention, integrating stylization and emotions [1]. This legitimized the idea that it can be appropriate to learn emotional cues from a visuals representation, not only demonstrating to add semantic richness to captions produced [1]. In other early work on reinforcement-based performance in captioning shifted the frame to captioning outside a supervised video model and to those who could approach appreciation-based reward functions. That is, it opened possibilities to model generation, not only through supervised learning, but shaped with feedback signals on compliments, based on perceived fluency, performance against metrics, and user aligned interpretability processes [2]. Earlier work focused on deep learning processing approaches, included convolutional encoding with recurrent decoding. These approaches tended to be commonly adopted as base architecture and evaluations supporting potential model architectures. The work in preprocessing procedures, vocabulary calibration and decoding procedures for generative captions was mentioned as ways to develop fluent and coherent captions [3]. This consideration included insights from earlier models that are limited to long range co-dependencies and area-specific interpretability, reflected in image compatibility.

The introduction of affective modelling into caption generation expanded what is possible in terms of captioning systems expressive capacity. The use of architecture for emotion-based tasks with CSPDenseNet encoders and bi-directional recurrent decoders have illustrated that using fine-grained visual cues can provide a more enriched sentiment interpretation than what is achievable using basic visual cues alone [4]. Those developments also indicate that captioning is not limited to a literal description of what an image contains, but can also include an inferred emotional context. The introduction of the transformer-based captioning model and self-attention mechanism into the architecture of the transformer model allowed for the modelling of the global relationships between visual features and linguistic tokens; hence improving the semantic coherence of the captions generated using the transformer model while at the same time, reducing the inherent bottlenecks of processing sequential inputs associated with recurrent models [5]. In addition, the implementation of dynamic region weighting (using the Attention mechanism) within a context-aware Attention Model is able to improve both detail resolution and interpretive alignment of images in complex scenes [6]. As a result of the

aforementioned work, it has set the stage for developing new architecture to combine Global Semantic Structure and Localised Interpretive Signals. Multiple models combining vision transformers with generative adversarial networks (GANs), or generative adversarial processes, have been able to utilise adversarial objectives as a way of enhancing stylistic qualities of generated captions and mitigating the generation of generic phrases while utilising the increased representational capacity of the caption generator provided through generative adversarial training and the transformer encoder architecture [7]. Domains specific studies on the military imagery have shown that when creating captioning models, it is necessary to use tailored visual encoding methods that can appropriately capture Low-Level View, Small Target, and Cluttered Object distributions, and that, when working with non-standard visual domains, it is critical to employ Data Base Specialization and Domain Adaptation methods [8].The objective of Integrating Object Features using Attention improved Interpretive Fidelity, aligning distinct visual areas of the object with the associated linguistic output of that object. The approach also served as further evidence of the alignment between the human perceptual model of an object with the computer-generated description [9]. The introduction of Hybrid Encoder-Decoder architectures with ResNet50 Backbones and recurrent decoders resulted in increased Fluency and Feature Stability. These findings reinforce the importance of Interrelated Encoder-Decoder Components of Hybrid Architectures in Captioning [10].

Sequence Modelling continues to dominate current research efforts, particularly with the development of Bidirectional Recurrent Networks. Bidirectional LSTMs have improved Grammatical Cohesion and Semantic Continuity by enriching the information incorporated within a sentence by including past and future context information [11]. The development of General Deep Learning Pipelines using classical CNN-LSTM configurations further advanced the evolution of Training Techniques, including Teacher Forcing and Beam Search, that affected the Reliability and Diversity of output Captioning [12]. Multimodal Frameworks have further developed the ability to integrate the image features and language embeddings into a unified Representation Space to enhance the Semantics of each Modality [13]. Application-Based Research also identified several Engineering Problems in deploying Captioning Applications, including Model Size Compression, Inference Time, and User-Centred Design Considerations [14].

Hybrid and multi-source captioning models expanded the potential solutions, using either multiple encoders or fused neural networks from complementary mechanisms. Research with hybrid recurrent models, such as mixtures of LSTM and GRU decoders, observed gains in stabilizing behaviour and adaptive gating when generating sentences [16]. Captioning systems that incorporated emotional stimuli to provide caption models sensitivity to sentiment created affective modelling; which helps improve description range, specifically for scenes which often include subjective interpretive mechanisms [17]. Resiliency studies on models offered useful insight in the role of pre-trained encoders, since they demonstrated models initialized with strong pretrained backbones have a greater degree of cross domain generalization and are less affected by dataset biases [18]. More practically available, Vision Speak has operationalized the integration of captioning pipelines into user facing applications similarly utilizing the combination of effective CNN encoders, recurrent or transformer-based decoders, and fine-tuning via reinforcement [19].

Important literature reviews and integrations have summarized advances across CNN–RNN, attentional, and transformer architectures. The literature noted methodological changes, empirical approaches, and new challenges, and discussed the need for better grounding, hallucination metrics, and more comprehensive evaluation metrics. The papers also discussed the next research trajectories in multilingual captioning, multimodal pretraining, and the need for efficient models for edge deployment [20]. In sum, all of literature reviewed shows a consistent trajectory in which image captioning has developed from basic CNN–RNN systems to multimodal architectures that can capture semantics, sentiment, context, and a global structure to the image.

V. DISCUSSION AND CHALLENGES

The research work indicates that deep learning has enhanced the effectiveness and expressive power of image captioning systems. However, these advancements have occurred simultaneously with challenges that continue to limit the systems'

reliability, domain-generalizability, explainability, and real-world use. One of the most notable challenges is caused by dataset bias and annotation imbalance. Large datasets developed for captioning reflect an overabundance of common object categories and scene structures that happen often, and ultimately lead models to overfitting valuing repetitions and default interpretations. We see this problem articulated in CNN–RNN pipelines that leverage the distributional properties of benchmark datasets to produce captions or summaries, and struggle to generalize to novel settings or object combinations [3][12]. We can see studies examining model robustness that show an overall bias when models just rely on encoders pretrained on large-scale datasets and domain-inductive biases to lessen aspects of bias, but as a longer-term conclusion still reinforce the original data-informed bias [18]. The domain-specific literature around military-based imagery shared across the studies indicates the continued use of datasets to inform models trained on images across different settings may inhibit model performance due to domain mismatch in datasets, particularly for low altitude images where small objects, odd angles or perspective and specific equipment are present and only trained or retrain to meet that context [8].

The phenomenon of hallucination is another central aspect of image captioning. Hallucination occurs when models predict objects, attributes, or actions that are not present in the image. Hallucination is not isolated to recurrent or transformer-based architectures. Affective captioning approaches that draw from emotional reasoning might invoke excessive hallucination when interpretive reasoning is valued ahead of factual grounding [1][4][17]. Reinforcement based captioning frameworks may increase fluency in sentences, but at the cost of increasing hallucination risk by utilizing reward functions that privilege stylistically appealing or metrically aligned sentences relative to visual fidelity [2]. Object attention and context aware models show improved hallucinations by improving alignment between image regions and generated tokens, but these models also experience hallucination [6][9]. Transformer models exhibit hallucination despite better ability to model global context because text priors more heavily influence predictions rather than visual evidence, such as influenced in ambiguous or cluttered scenes [5][20]. The presence of hallucination highlights the need for grounding aware learning strategies that impose a consistency between visual inputs and the textual outputs.

The evaluation methodology is a further area of difficulty. Metrics like BLEU, METEOR, ROUGE, and CIDEr which are commonly used to measure success in language generation tasks use token overlap rather than semantic or grounded quality. This means captioning systems can be optimized for metrics but not provide factually accurate or descriptively relevant captions. Foundational CNN–LSTM systems rely heavily on these metrics for validation [3][12] and reinforcement models explicitly optimize towards these metrics [2], thus reinforcing their impact, despite noted constraints. SPICE tries to include semantic evaluation based on scene graphs, but its quality depends on the reliability of external detectors, which adds in more sources of error. Though transformers and multimodal systems report improved BLEU or CIDEr scores due to improved linguistic coherence, these improvements don't necessarily correspond with less hallucination and human preference [5][13][20]. Numerous studies find the need for composite evaluation frameworks that incorporate grounding checks, semantic similarity measures, and human based assessments in order to get a better understanding of true model performance.

The computational requirements of contemporary captioning architectures pose an additional obstacle to adoption. Vision transformers, GAN assisted encoders, and multimodal fusion architectures are demanding in terms of the issue of compute power for their pre training and fine tuning [5][7][13]. Applied research indicates that limited resources will ultimately restrict the opportunities to deploy captioning systems in practice. For example, web-based captioning systems demonstrate the need for less resource intensive encoders and efficient decoders to enable real time inference [14]. Hybrid recurrent models that combine GRU and LSTM units perform reliably at decoding, and provide improved efficiency while producing reasonable performance [16]. Nevertheless, creating a high-quality captioning system that works well at low resource and energy availability situations e.g. mobile phones, embedded platforms or autonomous systems, remains an open issue.

The application of abstract, relational, or culturally responsive content remains limited. Although models

are effective in object naming styles, performance is lacking with high level reasoning, spatial relationships, and narrative interpretation. Approaches that introduce emotional captioning attempt to work through these interpretation challenges, although their development takes much curation of labelled affect data which is subjective and expensive to build [1][4][17]. Multimodal approaches have improved alignment of relationships, but still account for the vision encoders and linguistic embeddings established representational limits which do not address subtle contextual cues [13]. Resiliency studies in domains emphasize how pretrained encoders help generalize across domains, but also demonstrate vast performance gaps remain when transferring learning across heterogeneous visual contexts [18]. The above problem also illustrates the need to develop architectures that include a mechanism to reason about structured inputs, utilize external sources of information about the world, and ground information with relational.

Researchers utilizing bidirectional LSTMs or transformer-based encoders acknowledge that captioning performance varies depending on the complexity and structure of a language's morphology which limited generalization beyond sets based on English [11][20]. Low resource languages will require specialized tokenization strategies, cross linguistic transfer learning solutions, and consideration of the context of annotation. Future systems will be expected to include multi-lingual pretraining, or provide adaptive embedding schemes to encourage captioning capabilities across diverse linguistic communities around the globe.

Lastly, safety, ethics, and user-cantered design are more important than ever in the approaches to deploying captioning systems in the wild. Assistive applications require captions to be highly reliable and unambiguous, especially for visually impaired users who get access to their awareness of the world through the textual description for them. Studies conducted about field studies emphasize conservatively captioning behaviours, modelling uncertainty in the captions, and asking for user feedback to address risk [14][19]. Military and surveillance applications have been revealed to have a risk of misidentifying an object, risk to operation, and systems outputs being interpreted too liberally [8]. Affective captions bring forth ethical concerns about interpretation of the subjective, cultural awareness, and framing of emotional scene [1][17]. All these factors illustrate that the technical advancement must coincide with responsibly deploying by human-cantered evaluation.

VI. EXPERIMENTAL SETUP

The design and implementation of each component of the framework, all the hyperparameters, the design and implementation of the baselines, the evaluation metrics and protocols, and measures taken to ensure the validity and reproducibility of the experiments. The information provided here should be sufficient to reproduce the experiments

Table 6.1: Full Hyperparameter Configuration.

Hyperparameter	Value	Justification
Visual backbone	EfficientNet-B4 (ImageNet-21K)	Best accuracy/parameter on engagement scale
Input resolution	380 × 380	EfficientNet-B4 native resolution
Frame sampling rate	3 fps (30 frames/clip)	Optimal accuracy-efficiency trade-off
Caption strategy	Keyframe (3 frames/clip)	Best accuracy-efficiency balance
Text encoder	RoBERTa-base (fine-tuned)	Strong contextual representations
Max token length	256 tokens	Covers compound 3-caption descriptions
Sentiment classes	5	Captures full affective range
Fusion embedding dim d	512	Matches RoBERTa projection capacity
Self-attention layers	2	Sufficient for spatial context modeling
Cross-attention layers	4	Optimal by ablation (Section 8.3)
Attention heads	8	Multi-aspect attention coverage
FFN inner dimension	2048 (4×d)	Standard transformer scaling
Dropout rate	0.3 (fusion), 0.5 (classifier)	Stronger regularization at head
Batch size	16 per GPU (32 effective)	GPU memory optimized
Gradient accumulation	2 steps	Effective batch 32
Optimizer	AdamW	Decoupled weight decay
LR — backbone	5×10^{-5}	Slow fine-tuning of pretrained weights

LR — sentiment enc	5×10^{-5}	Consistent with backbone
LR — fusion/head	2×10^{-4}	Faster learning for new parameters
Weight decay	1×10^{-4}	Standard regularization
LR schedule	Cosine annealing + warmup	Stable convergence
Warmup epochs	5	Gradual LR ramp-up
Max epochs	50	Training budget
Early stop patience	7 epochs	Macro F1 on validation
Loss function	Weighted CE + label smoothing (ε=0.1)	Class imbalance + calibration
Alignment loss weight λ	0.1	Auxiliary supervision
Mixed precision	FP16 + GradScaler	Memory and speed efficiency
Random seed	42 (primary), 123, 456 (repeats)	Reproducibility
Gradient clipping	Max norm = 1.0	Prevent gradient explosion

VII. EVALUATION METRICS AND PROTOCOL

7.1 Primary Metrics

We use the following metrics computed on test sets to evaluate all systems:

Macro F1-Score (primary metric):

$$F1_{macro} = \frac{1}{C}\sum_{c=1}^{C} F1_c = \frac{1}{C}\sum_{c=1}^{C} \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}$$

Macro F1 is used as the main metric and stopping criterion as it treats all classes equally regardless of their class frequencies, which is important for the evaluation of class-imbalance problems.

Overall Accuracy:

$$Acc = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C}(TP_c + FP_c + FN_c + TN_c)}$$

Macro Precision and Recall:

$$P_{macro} = \frac{1}{C}\sum_{c=1}^{C} P_c\,, \qquad R_{macro} = \frac{1}{C}\sum_{c=1}^{C} R_c$$

AUC-ROC (one-vs-rest macro-averaged):

$$AUC_{macro} = \frac{1}{C}\sum_{c=1}^{C} AUC\ (c\ vs. rest)$$

7.2 Statistical Significance Testing

To verify that differences in performance between the proposed model and baselines are statistically significant (and not caused by random chance), McNemar's test is applied to prediction vectors on the test set. We use a p-value of 0.05. Moreover, 95% confidence intervals for macro F1 are obtained through bootstrapping (10,000 samples).

VIII. APPLICATIONS AND FUTURE SCOPE

Generation of image captions enables an extensive range of applications that process visual content into structured textual representations. The reviewed literature demonstrates that captioning models advance modalities in accessibility, automation, retrieval, surveillance, communication, robotics, and analytical decision support. Deep learning architectures have cultivated the application of captioning systems through mechanisms that enhance accuracy on an interpretive or semantic alignment basis and responsiveness to context. Emotional captioning has revealed that systems can assess fact-based cognition and at the same time include interpretive and stylistic cues, which can extend further applications in image analysis; i.e., art analysis, cultural interpretation, mood in multimedia, and so forth [1][4][17]. Reinforcement learning framework has contributed to improve caption performance depending on metric-based or user aligned objectives and improved fluency and reliability in the deployments of captions in situational cases [2]. Transformer based and context aware modelling has improved on the interpretive scope and delves into advanced applications that require some type of global visual cognitions or dynamic region of emphasis [5][6].

Assistive technologies designed for users with visual impairment highlight one of the more notable applications of image captioning. Captioning systems are capable of contextualizing scenes, objects, and interactions encountered in everyday contexts as descriptive summaries using screen readers or audio streams. Research that utilizes efficient visual

encoders built with stable sequence decoders highlights that any description for the sake of descriptions relies on the importance of dependable, grounded, and contextualized descriptions for the sake of user safety and awareness in navigation [14][19]. Methods that optimize emotional captioning possibly include affective nuances and expressive descriptions which might assist users in their understanding of emotional or artistic atmospheres. However, captioning systems in infographic, textual, or audio format will need to consider possible protection against inaccuracies in feelings, as one subjective experience may constitute over-exaggeration or hallucination [1][4][17]. The next iteration of research and development in this specific space will be focused on captioning with the ability to possibly identify uncertainty, clarity, and organize captions around factual reporting that draw from user feedback in an adaptive captioning system.

Captioning models that distil complex visual input into structured language concepts are advantageous to autonomous systems and surveillance platforms. Research initiatives aimed at military functions, and perhaps low altitude aerial imaging, have found that captioning systems can enhance reconnaissance, situational awareness, or general visual inspections to locate critical objects [8]. In these operational contexts, models are expected to recognize objects that are small, work through clutter, and convey operationally relevant descriptions in succinct language. Styles of model architectures based on vision transformers or hybrid transformer-GAN architectures, allow for improved global reasoning and stylistic uniformity meaning that they would be well suited to support summarizing large or dense scenes [5][7]. Nevertheless, captioning for these types of applications would require low latency inference, reliability, and adaptations to the domain, all of which would mitigate operational risks.

Systems that retrieve relevant images, otherwise known as content-based image retrieval (CBIR) systems, and in particular, those that achieve this task using captioning models, are becoming increasingly prevalent. These newly available multimodal systems have applied visual features and text embeddings to improve retrieval to align conceptual representations across modalities [13] [15]. Advancements in reward-based and adversarially trained captioning systems will also further enhance the quality of semantic queries resulting in richer and more expressive captions [2] [7]. The continual scale and diversity of new multimedia content, suggests that captioning systems will play a vital role in helping to structure content to facilitate retrieval, or even to augment retrieval with automated annotations.

A further issue regards the large number of languages available to capture captions, descriptions, narratives, and other visually expressed elements. Hybrid CNN–RNN and efficient attention-based decoders can produce human like captions appropriate for routine communication tasks [10][16]. Transformer based systems further expand on this capability to provide richer context and flexibility of style [5][20]. As systems become more effective in their expressive ability to capture elements of storytelling, captions can assist with the generation of automated alt text, contextual labelling, and overall accessibility in online platforms. Additional future systems can seek to incorporate the model of user preferences, or an adjustment for emotional tone for captions that better suit the audience or the communication situation.

Robotics and human–robot interaction are additional emerging application areas where captioning models are important to interpretability, safety, and task collaboration. Robots that incorporate captioning systems can share states of the environment, properties of objects, and act on intentions with humans in a natural language conversation creating a better understanding of interactions. Studies of context aware encoders have shown that enriching spatial reasoning and high resolution improves the accuracy of descriptions which are important for tasks of manipulation and navigational planning [6][9]. As robots gain the capacity for greater autonomy, captioning will become an important connection of communication that access the perceptual model with the human decision making.

Medical imaging analysis is a potential domain for embedding captioning models in the future. Not all of the reviewed studies are exactly reflected in medical workflows, however the architectural advances from attention based and transformer-based caption to structured diagnostic descriptions can be readily applied. Vision transformers and context aware attention approaches are ideally suited for capturing local abnormalities while considering other contextual global features, an important aspect of radiology reporting [5][6]. The direction of future research in the

area of Image Caption Generation is expected to include the inclusion of clinical datasets, more rigorous evaluative processes, and the development of greater interpretability with the implementation of interpretive model frameworks.

The future direction of Image Caption Generation is likely to explore several significant research paths. Improved grounding will be the most important of these paths. Reduction of hallucinations from creating captions and increase accuracy of the generated captions through the use of Object Detectors, Scene Graph Reasoning, or by using Cross Modal Consistency Constraints [1] [9][18]. The area of Emotional Captioning provides an interesting potential to introduce subjectivity into caption interpretation. In the future, captioning systems will need to find equilibrium among rich interpretive Ness and factual accuracy [4] and [10]. Another promising development in developing real-time interactive captions for Augmented or Virtual Reality is developing efficient Encoder-Decoder architectures; applying Encoder-Decoder model Quantisation Methods; or pruning with Optimised Attention Mechanisms in low latency on interactive environments [10][14]. Another area providing future research potential is Multilingual and Low-Resource Captioning in this domain. The language structure of the various languages provides evidence that the linguistic structure influences the quality of Image Captioning [11][20]. Therefore, researchers must time Multilingual Pretraining, Adaptive Tokenisation Strategies, or Culturally-Informed Annotators for the global use of these systems.

Continuing to advance the Image Caption Generation Domain with an improved Knowledge Base using Evaluation Methods, also developed in this domain, is essential. Existing benchmarks do not recognize the accuracy of semantics, grounding performance, or user experience. Prior research has recommended hybrid evaluation approaches to include automatic metrics, human guided scoring, and grounding assessment [3][12][20]. Future datasets will require richer annotations, different domains of application, and scenario specific evaluations for models to undergo comprehensive testing. In short, the use cases for image caption generation extends to accessibility, automation, retrieval, communication, robotics and analytical tasks. The future of this documentary will be contingent on developments and grounding, multimodal reasoning, calculations efficiency, multilingual modelling, and evaluation methods. Ongoing research in these areas will enable the scientific community to advance captioning systems to build more reliable, expressive, and adaptable captioning systems for real world environments.

IX. CONCLUSION

Image captioning has evolved from the classic convolutional and recurrent models to more sophisticated designs that combine reconstruction, attention, multimodal reasoning, and transformer modelling. As demonstrated in the existing literature, advancements in deep learning have enhanced the ability for captioning systems to extract meaningful visual features, reason about situational context, and generate coherent natural language descriptions. Encoder-decoder models provided the initial basis of the field, whereas attention mechanisms and transformer-type designs have sought to expand models on their auxiliary ability to capture global dependencies, generate semantically coherent descriptions, and manage increasingly complex visual environments. Additionally, emotional and affective captioning methods demonstrate the potential of some models to respond to subjective or stylistic qualities of images which is an important emergence that extends beyond conventional object recognition and language generation.

Nonetheless, various limitations exist; dataset bias continues to affect model behaviour, and models whether narrow or dense in representation commonly show difficulty when applied in narrower distribution or specialized domains of visual perception. Hallucination continues to be a significant problem, where models create incorrect content or unsupported content in the instance of ambiguous or missing visual information. Evaluation techniques still have not adapted fully to capture the viability of semantics, depth of interpretation, and user cantered considerations. Many existing criteria and metrics still measure lexical overlap relative to grounded or factual accuracy, limiting their ability to evaluate real world performance. Finally, models are subject to compute constraints that limits scaling, especially in mobile applications, robotics, or interactive settings where resources may be limited for computational platforms. Future gains require that all of these interrelated issues

are addressed. Robust ground mechanisms, multimodal alignment, and better integration and use of representations of structured scenes, will be required in order to reduce the issue of hallucination, and increase factual accuracy. Multi lingual and low resource captioning pipelines should be developed to support applicability globally and to reduce dependency on English text centered datasets. More efficient model architectures and pruning and quantization methods will be required for real time captioning in augmented reality, robotics, and other time and latency sensitive contexts.

Furthermore, evaluation techniques evolving toward hybrid human aligned criteria will provide a multi-faceted measure of descriptive quality, contextual relevance, and user trust. In summary, the trajectory of image caption generation represents sustained momentum derived from continuous deep learning, multi-modal processing, and attention-based modeling. Validating research in grounding, efficiency, linguistic diversity, and evaluation will remain critical to achieving the success of captioning systems as trustworthy and adaptable in real world contexts.

REFERENCES

[1] S. Ishikawa and K. Sugiura, "Affective image captioning for visual artworks using emotion-based cross-attention mechanisms," IEEE Access, vol. 11, pp. 24527–24534, Jan. 2023, doi: 10.1109/ACCESS.2023.3255887.

[2] T. Bai, S. Zhou, Y. Pang, J. Luo, H. Wang, and Y. Du, "An image caption model based on attention mechanism and deep reinforcement learning," Frontiers in Neuroscience, vol. 17, Oct. 2023, doi: 10.3389/fnins.2023.1270850.

[3] A. Verma, A. K. Yadav, M. Kumar, and D. Yadav, "Automatic image caption generation using deep learning," Multimedia Tools and Applications, Jun. 2023, doi: 10.1007/s11042-023-15555-y.

[4] S. Kavitha, K. Pon Karthika, J. Kaliappan, S. Selvaraj, R. Nagalakshmi, and B. Molla, "Caption generation based on emotions using CSPDenseNet and BiLSTM with self-attention," Applied Computational Intelligence and Soft Computing, vol. 2022, pp. 1–13, Sep. 2022, doi: 10.1155/2022/2756396.

[5] M. Nabi, R. Pachauri, S. Ahmad, and K. Varshney, "Visual image captioning through transformer," International Journal for Research in Applied Science and Engineering Technology, 2023. [Online]. Available: https://www.ijraset.com/research-paper/visual-image-captioning-through-transformer. [Accessed: Nov. 18, 2025].

[6] A. Bhuiyan, E. Hossain, M. M. Hoque, and M. A. A. Dewan, "Enhancing image caption generation through context-aware attention mechanism," Heliyon, vol. 10, no. 17, p. e36272, Sep. 2024, doi: 10.1016/j.heliyon.2024.e36272.

[7] S. Tyagi et al., "Novel advance image caption generation utilizing vision transformer and generative adversarial networks," Computers, vol. 13, no. 12, Art. no. 305, Nov. 2024, doi: 10.3390/computers13120305.

[8] L. Pan, C. Song, X. Gan, K. Xu, and Y. Xie, "Military image captioning for low-altitude UAV or UGV perspectives," Drones, vol. 8, no. 9, Art. no. 421, Aug. 2024, doi: 10.3390/drones8090421.

[9] M. A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," Journal of Big Data, vol. 9, no. 1, Feb. 2022, doi: 10.1186/s40537-022-00571-w.

[10] P. V. Kavitha and V. Karpagam, "Image captioning deep learning model using ResNet50 encoder and hybrid LSTM–GRU decoder optimized with beam search," Automatika, vol. 66, no. 3, pp. 394–410, Apr. 2025, doi: 10.1080/00051144.2025.2485695.

[11] F. Hoseini and A. Yaghoobi Notash, "Image captioning using bidirectional LSTM neural network," Discover Artificial Intelligence, vol. 5, no. 1, May 2025, doi: 10.1007/s44163-025-00315-8.

[12] S. Suneetha, A. M. L. Priya, C. L. Narayana, C. Enosh, and T. S. V. Prasad, "Image captioning using deep learning," International Journal of Engineering Research & Technology, vol. 13, no. 2, Feb. 2024, doi: 10.17577/IJERTV13IS020047.

[13] R. Farkh, G. Oudinet, and Y. Foued, "Image captioning using multimodal deep learning approach," Computers, Materials & Continua, vol. 81, no. 3, pp. 3951–3968, 2024, doi: 10.32604/cmc.2024.053245.

[14] S. R. E. and M.-P. T. B., “Image captioning web application using deep learning algorithms,” International Journal of Computer Applications, vol. 185, no. 40, pp. 29–33, Nov. 2023, doi: 10.5120/ijca2023923204.

[15] P. Naveen, Y. Saikrishna, M. Z. Hussain, and K. Rajesh, “Image captioning with machine learning algorithms,” International Journal of Engineering Research & Technology, vol. 14, no. 4, Apr. 2025, doi: 10.17577/IJERTV14IS040242.

[16] A. Singh Sengar, K. Pandey, and P. Tewari, “Image to caption generator using machine learning and deep learning models,” SSRN Electronic Journal, 2025, doi: 10.2139/ssrn.5190843.

[17] Z. Zhou, Z. Zhai, X. Gao, and J. Zhu, “Improved IEC performance via emotional stimuli-aware captioning,” Scientific Reports, vol. 15, no. 1, Jul. 2025, doi: 10.1038/s41598-025-06094-7.

[18] N. Kharate et al., “Unveiling the resilience of image captioning models and the influence of pre-trained models on deep learning performance,” International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 9s, pp. 01–07, Jul. 2023. [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/3089.

[19] A. Pal and I. Pal, “Vision Speak: Automated image-to-caption generation using deep learning,” SSRN Electronic Journal, Jan. 2025, doi: 10.2139/ssrn.5268594.

[20] S. Jaiswal, H. Pallthadka, R. P. Chinchewadi, and T. Jaiswal, “Capturing the essence: An in-depth exploration of automatic image captioning techniques and advancements,” International Journal of Computer Engineering Research Trends, vol. 10, no. 12, pp. 22–30, Dec. 2023, doi: 10.22362/ijcert.v10i12.912.

[21] H. A. Jardawala, “Deep learning-based sugarcane leaf disease detection using Xception: A precision agriculture approach,” in Proc. Parul University International Conference on Engineering and Technology (PiCET 2025), Vadodara, India, 2025, pp. 1654–1660, doi: 10.1049/icp.2025.1680.

[22] Y. Suthar and C. Thacker, “Automated Monkeypox detection using deep learning: An inception-based approach,” in Proc. Parul University International Conference on Engineering and Technology (PiCET 2025), Vadodara, India, 2025, pp. 1078–1084, doi: 10.1049/icp.2025.1551.

[23] adityajn105, “Flickr30k Dataset,” Kaggle. [Online]. Available: https://www.kaggle.com/datasets/adityajn105/flickr30k/. [Accessed: Nov. 18, 2025].