

# A Unified Hybrid Encoder–Decoder Model for Vision–Language Image Captioning

Moloy Dhar<sup>1</sup>, Mrinmoy Sen<sup>2</sup>,  
Bidesh Chakraborty<sup>2</sup> and  
Suparna Biswas<sup>3,\*</sup>

<sup>1</sup>Department of Computer Science  
& Engineering, Guru Nanak Institute  
of Technology, Kolkata, India

<sup>2</sup>Department of Computer Science  
& Engineering, Haldia Institute of  
Technology, Kolkata, India

<sup>3</sup>Department of Computer Science &  
Engineering, Guru Nanak Institute of  
Technology, Kolkata, India

\*E-mail: suparna.biswas@gnit.  
ac.in

Received for publication  
December 11, 2025.

## Abstract

Image captioning is a combined activity that generates relevant text descriptions of images using natural language processing and computer vision. The outcomes are heavily influenced on the supplied image's quality and clarity. This research presents a hybrid system that combines ResNet50 and MobileNet into a single encoder in order to enhance captioning performance. The Integrated Model of Vision and Language Processing helps create captions that are more precise, logical, and context-aware by learning both language patterns and visual characteristics simultaneously. After testing the suggested model on the Flickr8k and Flickr30k datasets, it achieved a maximum accuracy of 98%.

## Keywords

deep learning, convolutional neural network, image captioning, natural language processing, recurrent neural network, encoder–decoder

## 1. Introduction

A cohesive method that connects visual comprehension with natural language production is the integrated model of vision and language processing. The capacity of an Integrated Model of Vision and Language Processing to combine linguistic generation and visual comprehension into a single, cohesive framework makes it novel for improving image captioning (IC). Conventional IC systems frequently produce superficial descriptions or semantic gaps because they approach image analysis and phrase generation as distinct activities. An integrated model, on the other hand, learns joint representations in which linguistic patterns and visual characteristics support one another, enabling more precise descriptions and better contextual understanding. This method captures fine-grained actions, emotions, and object relations that are often missed by traditional models. Furthermore, by using multimodal transformers and attention processes, the system may dynamically focus on the most pertinent visual regions while producing text that is both contextually meaningful and

syntactically valid. The model's versatility in a variety of fields, including digital media, healthcare, and surveillance, underscores its inventiveness and represents a major breakthrough in automated visual-linguistic intelligence. It uses sophisticated deep learning (DL) methods to decipher intricate visual characteristics and convert them into insightful textual descriptions. It makes caption generation more precise and context-aware by integrating visual encoders and language models into a unified framework. By overcoming the shortcomings of independent vision or language systems, this integrated framework seeks to improve the caliber and coherence of picture captions. The primary component of image posts on social media, captions convey a rich, semantic representation that can provide articles surrounding a photo and take into account the user's knowledge by indicating when, why and where the photo was taken at present, academics and researchers are giving the area of IC a lot of thought. The development of DL has brought attention to the field of producing IC.

Current IC techniques have advanced significantly, particularly with the application of encoder–decoder (E–D) architectures and DL. Nonetheless, a number of issues still exist that restrict their descriptive quality, generality, and real-world performance. The majority of models may overlook fine-grained details or contextual interactions in favor of extracting global data or focusing on certain regions. Semantic richness, fluency, and diversity are frequently lacking in captions. Models for captioning have trouble with items that are not visible during training. The capacity to describe new or uncommon items is necessary for real-world applications. Depending on the context, there are several legitimate ways to interpret a picture. Commonsense thinking is frequently not incorporated into models.

Building solutions that are not reliant on human annotation is becoming more and more important as they become more cost-effective. Examples of these solutions include assistive technology and autonomous image indexing. Here, we contrast six different IC architectures and introduce a hybrid architecture for IC. Using a hybrid design of two well-known convolutional neural networks (CNNs), ResNet50 and MobileNet, we will provide a novel approach to picture captioning in this study. That approach is highly applicable to problems where computing resources and performance are important since it combines the benefits of two models: efficiency and speed with MobileNet and deep feature extraction with ResNet50 for powerful feature extraction along with multilingual text-to-speech (TTS) module for accessibility. Pre-processing image data and creating an informative caption are two of the processes needed for the suggested method.

To balance representational capacity and computational performance, ResNet50 and MobileNet were used. Stronger backbones, such EfficientNet or Vision Transformers, raise training costs and increase sensitivity to data scale, but they can produce slight improvements. These contemporary buildings did not contribute enough to Flickr8k/30k to warrant the additional complexity. With its reliable convergence and complementary properties, the hybrid ResNet50–MobileNet encoder makes it possible to evaluate the suggested fusion technique more clearly.

Our paper’s contribution is as follows:

1. Our suggested approach builds a hybrid model for picture representation by fusing the potent visual feature eradication potentiality of MobileNet and ResNet50.

2. Subsequently being retrieved, the features are concatenated and put into an encoding-decoding architecture located on long short-term memory (LSTM).
3. With an accuracy of 98.42%, our hybrid model—which combines the MobileNet and ResNet50 models performs better than any other model that has been used with this hybrid architecture.
4. The final model performs better when it comes to producing thorough and contextually appropriate descriptions for input photos.

The composition of the following is our paper: Section II illustrate Literature Survey, Section III states Methodology, Section IV discusses feasible datasets, evaluation measures, and a preliminary study among the models. Section V provides a succinct summary of the classic issues in semantic segmentation (SS) and information classification, along with workable solutions and future possibilities, and Section VI highlights on Discussion.

## II. Literature Survey

Early end-to-end captioning models established the encoder–decoder paradigm: a CNN encodes the image and an Recurrent Neural Network (RNN) decodes text, learning to generate captions directly from image features. Showing that an image-to-sequence neural framework can produce fluent captions and strong baseline scores on Common Objects in Context (COCO) and other datasets [1].

CNN–RNN architecture that uses an LSTM-based recurrent network to produce descriptive captions and a convolutional network to extract visual features from remote-sensing pictures [2]. It shows that despite the lack of domain-specific datasets, deep vision-language models can be successfully applied to satellite imagery to generate relevant written descriptions. Overall, the study shows how DL can be used to understand Earth-observation data and offers a straightforward but efficient baseline for automated remote-sensing IC.

Region-level features and object-centric attention further advanced the field. The bottom–up and top–down attention framework replaced grid CNN features with object-region features (from an object detector), computing attention at the object/region level. This approach improved fine-grained descriptions and became standard input for many later models [3].

Optimization techniques also matter: reinforcement-learning (RL) based training such as self-critical

sequence training (SCST) directly optimizes non-differentiable metrics (e.g., Consensus-based Image Description Evaluation (CIDEr) and substantially improves evaluation scores by aligning training objectives with test-time metrics [4].

Attention mechanisms improved caption relevance by letting the decoder focus on salient image regions at each generation step [5]. Introduced spatial attention maps over CNN feature maps and demonstrated qualitative and quantitative gains, plus interpretable visualizations of where the model “looks” while generating words.

More recent work integrates fully attentive transformer architectures, memory meshes, and large-scale vision–language pretraining [6]. The

Meshed-Memory Transformer ( $M^2$ ) introduced mesh-like decoder connectivity and multi-level image representations to better fuse low- and high-level visual cues with language generation.

Large-scale pretraining methods (e.g., Object-Semantics Aligned Pre-training (OSCAR) and related vision–language pretraining models) use object tags or region features as anchors to learn stronger cross-modal embeddings that transfer well to captioning and other V + L tasks [7].

Collectively, these integrated models show a clear trajectory: stronger visual representations (object/region-level), dynamic attention or transformer-based fusion, and training objectives or pretraining strategies tailored to cross-modal alignment—all

## Representative works—key advantages & disadvantages

| Reference | Core idea                                                                                              | Advantages                                                                                                                                         | Limitations                                                                                                                             |
|-----------|--------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|
| [1]       | CNN encoder + RNN decoder end-to-end image→sentence                                                    | Simple, effective end-to-end learning; good baseline for fluency.                                                                                  | Limited visual grounding (global CNN features); often produces generic captions and can miss fine details.                              |
| [2]       | Deep vision–language architectures used in natural images can be effectively adapted to remote sensing | Effective feature extraction for remote sensing images. Improved natural-language description generation. Reduced dependence on manual annotation. | Lack of attention mechanisms. Dataset constraints. No multimodal fusion enhancements. Not benchmarking against state-of-the-art models. |
| [3]       | Uses object-level region features (bottom–up) + top–down attention controller                          | Stronger grounding to objects; better fine-grained descriptions and improved metrics; became standard input for many models.                       | Requires object detector pretraining (computational and annotation cost); detector biases affect captions.                              |
| [4]       | RL-based sequence-level training optimizing CIDEr and other metrics                                    | Directly optimizes evaluation metrics, leading to large empirical improvements; reduces train–test objective mismatch.                             | RL training can be unstable; can encourage metric-specific artifacts (over fitting to metric quirks).                                   |
| [5]       | Adds spatial attention over CNN feature maps during decoding                                           | Improved relevance; produces interpretable attention maps; better at localizing salient regions while generating words.                            | Attention over grid features is still coarse; struggles with object identity and fine relations.                                        |
| [6]       | Multi-level transformer encoder + mesh-like decoder memory for better cross-level fusion               | Exploits low- and high-level features simultaneously; strong performance with a fully attentive architecture; captures relations better.           | More complex architecture and heavier compute; may need careful tuning and larger data.                                                 |
| [7]       | Pretraining on image–text pairs using object tags as anchor tokens                                     | Strong cross-modal representations; excellent transfer to captioning and V + L tasks; reduces need for task-specific data.                         | Large-scale pretraining is compute- and data-intensive; it relies on detector-generated object tags (propagates detector errors).       |

CNNs, convolutional neural networks; RL, reinforcement learning.

of which reduce the semantic gap between image understanding and fluent, context-rich caption generation.

On common datasets such as Flickr8k and Flickr30k, transformer-based models routinely beat RNN and CNN-LSTM architectures regarding Bilingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit ORdering (METEOR), and CIDEr scores [8, 9, 10].

Not only is it difficult to classify images using hand-collected characteristics, but sufficient time is also required for the analysis of the characteristic data [11]. Simultaneously, standard machine learning (ML) cannot be used for the refining of large datasets, and this makes it challenging to comprehend how features are developed, resolution made, and models trained, all of which contribute to the bad model's classification performance [12]. As such, categorization of visual processes using standard ML is relevant in multiple application domains.

There has been a substantial evolution in the classic picture captioning problem, which focuses on producing descriptive text from single-modal visual input. The goal of these developments is to improve caption variety, generalizability, and practicality. Outlined the bottom–up and top–down attention model [13, 14, 15], in which an LSTM-based decoder receives object characteristics that were extracted using Faster R-CNN. This successfully integrated object semantics with visual saliency. On the other hand, [16, 17] presented scene graph-based captioning, which uses scene graphs to represent the visual relationships between objects, improving semantic reasoning in caption production.

SCST, first presented by [18], is among the first and most influential efforts. The resulting captions' variety and relevancy were enhanced by this policy gradient-based method. Furthermore, RL has been used with beam search optimization approaches, such as varied beam search, to enhance caption uniqueness and prevent generic or repeating results [19, 20]. By overcoming the drawbacks of conventional maximum likelihood estimation -based training, RL and optimization approaches have greatly improved picture captioning. Improved fluency, relevance, and diversity in produced captions have resulted from the move toward task-aligned learning objectives, reward-driven optimization, and hybrid training methodologies. Krishna et al. [21] provides the Visual Genome dataset, a comprehensive resource that connects language and visual data. It allows greater scene knowledge beyond basic object

identification because to its extensively annotated photos with objects, properties, and connections.

Presents the YOLOv10-MsA, an improved object identification model intended to detect insulator flaws in real time from Unmanned Aerial Vehicle (UAV) photos [26]. It enhances feature extraction and increases detection accuracy, especially for minor or subtle flaws, by including an attention mechanism into the YOLOv10 architecture. For power transmission systems, the method seeks to provide automated inspection that is quicker and more dependable.

Regional Mutual Learning (RML)-YOLO, an enhanced YOLO-based detection system created especially for detecting insulator flaws in UAV aerial photos, is proposed by [27]. To increase detection accuracy and robustness in the face of complicated backgrounds and variable imaging settings, it improves feature extraction and multi-scale learning. Using MobileNetV2-SSD, [28] suggests a real-time soccer ball identification system for wheeled robots that maximizes speed and accuracy for embedded deployment and exhibits dependable performance under changing lighting, distance, and dynamic match conditions. By incorporating a Convolutional Block Attention Module (CBAM) attention mechanism into the backbone, substituting a Bidirectional Feature Pyramid Network (BiFPN) for Path Aggregation Network (PANet) to improve multi-scale feature fusion, and optimizing anchor boxes using an enhanced K-means algorithm, [29] offers an improved insulator defect detection method that improves the YOLOv5 model. These improvements collectively increase accuracy in identifying intricate, small defects in aerial transmission line images. In order to achieve high precision and real-time performance in complex scenes, [30] proposes a cascaded detection technique that first locates transmission line insulators in UAV-captured aerial imagery using an improved YOLO-based model. Then, a lightweight YOLOv4-tiny network is applied to the identified regions to detect missing defects. In order to improve feature extraction and detection performance, [31] suggests modifying the network structure and training approach of the YOLOv8 framework (Ultralytics, United States), specifically designed for insulator defect detection. The goal of these enhancements is to improve the robustness and accuracy of detecting different types of insulator flaws in intricate inspection images.

An improved version of the YOLOv8x object detection model is presented in [32] with the goal of lowering computational costs while increasing detection accuracy. It improves the training approach and

network architecture to attain more rapid and effective performance without compromising accuracy.

In order to improve accuracy, speed, and robustness in comparison to other popular detectors, [33,34] examines the architecture and optimization techniques of the YOLOv8 object detection model, talking about enhancements in network structure, feature extraction and fusion, training, and inference (including sample imbalance and non-maximum suppression). The performance of YOLOv8 is also empirically assessed, and its benefits, drawbacks, and possible future uses in object detection tasks are highlighted. Assesses the YOLOv8 object identification model's ability to recognize drones in photos, showing that strategic data augmentation (such as rotation and blurring) and meticulous hyperparameter tuning greatly increase detection accuracy and resilience on a publically accessible UAV dataset [35, 36]. In terms of precision and recall for drone detection tasks, it demonstrates that the optimized YOLOv8 technique outperforms a number of other DL models.

The Mask Region-based Convolutional Neural Network (R-CNN) framework for object detection is examined in [37], which describes its design, segmentation capabilities, performance strengths, limits, and applications across many computer vision (CV) tasks. The Single Shot Detector (SSD) object identification framework is described in [38], along with its single-stage design, multi-scale feature maps, performance benefits, drawbacks, and applicability for real-time vision problems. [39, 40] analyzes the Fast R-CNN object detection technique, describing its region-based architecture, feature pooling effectiveness, performance advantages, and contribution to the development of more precise and quicker detection systems.

In order to improve accuracy and decrease training complexity on plant disease datasets, [41] refines Vision Transformer models using adapter modules to propose an effective tomato leaf disease classification technique. Examines how well several ML and DL models—such as LightGBM, CatBoost, ANN, DNN, and WDN—predict student outcomes while incorporating ElasticNet regularization to manage overfitting [42, 43]. It assesses and compares their predictive capabilities to determine which methods produce the most accurate and trustworthy outcomes on educational datasets.

In order to enable simultaneous object identification and instance segmentation, [44] presents Mask R-CNN, a system that builds upon Faster R-CNN by including a parallel branch for pixel-level mask prediction. For every object instance in an image, it

produces precise bounding box, class, and mask outputs by combining a lightweight ROIAlign layer with task-specific subnetworks. Traces the development of DL-based object detection techniques from the earliest convolutional models to contemporary single-stage and two-stage detectors, examining their architectures, advantages, and disadvantages [45, 46]. It also covers important issues and potential avenues for further study in the areas of precision, speed, and practicality. Li et al. [47], Wang et al. [48] look into useful methods, such as feature processing enhancements, training plans, and architectural modifications, to increase Mask R-CNN performance in visual detection tasks. On benchmark datasets, it shows that these modifications can increase both processing efficiency and detection accuracy.

Our primary focus is on the existing 2D and 3D image DL segmentation processes that use CNNs and more advanced hierarchical feature elimination processes as a crucial component of CV. Additionally, we compare the assessments of the induced models to analyze their achievements, highlighting the most convincing and suitable approaches for various and challenging tasks. Since it lends intuition to further creation of a realistic model drawing, we think this is informative. The ability to identify certain semantic concepts contained in images or videos and to apply this advanced semantic data to the CNN-LSTM architecture has significantly enhanced effectiveness.

### III. Methodology

In our hybrid model (ResNet-50 + MobileNet based IC framework), the input image dataset is considered as Flickr8k and Flickr30k. Generated output is optimized image captions. Split dataset into training (60%), validation (20%), and testing (20%). Resize all images to (224 × 224) pixels. Normalize pixel values to a range suitable for CNNs. Here, validation and test images were left unaltered, and only the training set was subjected to data augmentation. Shear changes were purposefully avoided to avoid distorting item relationships. Instead, geometric transformations included zooming in the range of 0.9–1.1, width and height shifts up to ±10%, random rotation within ±15°, and horizontal flipping with a chance of 0.5. Semantic coherence between images and captions was maintained by limiting photometric modifications to brightness variations within ±10% and avoiding contrast manipulation. The CNN input range was used to normalize all images, resize them to 224 × 224 pixels, and apply dynamic augmentations while training. Apply data augmentation

with rotation, flipping, and shifting. In Caption Preprocessing, convert all captions to lowercase and remove digits, punctuation, and redundant spaces. Special tokens are included like <pad>: padding token, <unk>: unknown word token, <start>: start of sequence, <end>: end of sequence. Using common tokenizers (Keras tokenizer), each caption is tokenized into separate words. Build vocabulary and remove words below the frequency threshold. Tokenize captions and pad sequences to a fixed length and then convert tokens into word embedding vectors. The vocabulary was created only from training captions with a 5-word minimum frequency criterion. An additional token was assigned to uncommon words. Based on the training length distribution, the maximum length of the captions (34 tokens) was either truncated or padded. Padding tokens were hidden during the computation of loss, and special tokens (<start>, <end>, <pad>, <unk>) were specifically included.

In order to preserve just the high-level residual feature representations, the hybrid feature extraction procedure starts by loading the pre-trained ResNet-50 model and eliminating its final classification layer. Similarly, to maintain its lightweight spatial features, a pre-trained MobileNet model is loaded with its top classification layer deleted. After that, each input image is sent via ResNet-50 to produce deep residual features. At the same time, MobileNet processes the same picture to fetch compact spatial information. The output feature maps of both networks are subjected to independent applications of global average pooling (GAP) in order to reduce dimensionality and preserve the most informative global descriptors. To do so generate a single hybrid feature vector  $F$  that successfully combines the depth of residual learning with the effectiveness of mobile-optimized spatial representation, the pooled feature vectors from ResNet-50 and MobileNet are concatenated. The classification heads are removed, and GAP is used to generate feature vectors of 2,048 and 1,024 dimensions, respectively, for ResNet50 and MobileNet. After these vectors are concatenated, a single 3072-dimensional representation is created, and it is then projected to 512 dimensions through a dense layer. The resultant vector is utilized to initialize the LSTM decoder's hidden and cell states, guaranteeing that visual context is included from the very first decoding step. This approach creates a more thorough feature representation by utilizing the advantages of both models, namely MobileNet's lightweight design and ResNet50's deep residual connections. The feature maps produced

by ResNet50 and MobileNet are global average pooled after exiting their respective architectures. GAP is calculated as follows for a feature map of size  $H \times W \times C$ :

$$GAP_c = \frac{1}{H * W} \sum_{i=1}^H \sum_{j=1}^W S(i, j, c) \quad (1)$$

where  $S(i, j, c)$  is the feature map result at location  $(i, j)$  for channel  $c$ . The fused feature vectors are sent through a thick layer using SoftMax activation, which generates probabilities over the lexicon of potential words in order to create captions. The odds of each vocabulary class are generated using the SoftMax function:

$$P(y = c|x) = \frac{e^{z_c}}{\sum_j^C e^{z_j}} \quad (2)$$

where  $z_c$  is what came out of the last dense layer for class  $c$ .  $C$  is the number of classes overall. LSTM produces a vector of raw scores at every time interval.

Initializing the LSTM's hidden state with the hybrid feature vector  $F$  is the first step in the caption creation method utilizing an LSTM decoder. This ensures that the visual context from the image is integrated right away. To start the sequence generation, the <start> token is then sent to the decoder. To direct the creation process, the embedding of the previously predicted word is input into the LSTM at each time step and mixed with the picture feature context. The LSTM creates a probability distribution across the language by updating its hidden state, which is then passed over a dense layer and a softmax function. The following output is the word with the highest likelihood. As far as the <end> token is formed or the maximum caption length is achieved, this iterative process keeps going. Ultimately, the entire caption  $S$  is created by successively combining all of the predicted words (Figure 1).

In order for the visual characteristics and language model to learn to cooperate well, the hybrid E-D architecture is jointly trained using the photos and captions from the training set. Cross-entropy loss, which measure the conflict between the ground reality captions and the estimated word probabilities, is used to tune the network parameters, and the Adam optimizer modifies the weights to boost convergence. The model's performance is assessed on a different validation split following each epoch to track generalization and identify any indications of overfitting. In order to ensure that learning stops at the ideal moment for optimal captioning performance, training

continues until the model reaches convergence or an early stopping condition is activated.

Using the trained hybrid E-D model, the assessment step starts by creating captions for each test image. The characteristic of these captions is then measured utilizing BLEU-1 to BLEU-4 scores to determine  $n$ -gram overlap with ground-truth descriptions. Apart from captioning measures, accuracy, precision, recall, and F1-score are recycled to evaluate the model's general predictive accomplishment, providing a deeper understanding of its error patterns and consistency. Plotting accuracy–loss curves and the ROC curve, which show how well the model separates accurate predictions across thresholds, helps visualize the learning behavior, and robustness. Lastly, the performance of the hybrid model is contrasted with outcomes from separate captioning systems based on ResNet-50 and MobileNet to highlight the enhancement brought about by unified feature representation.

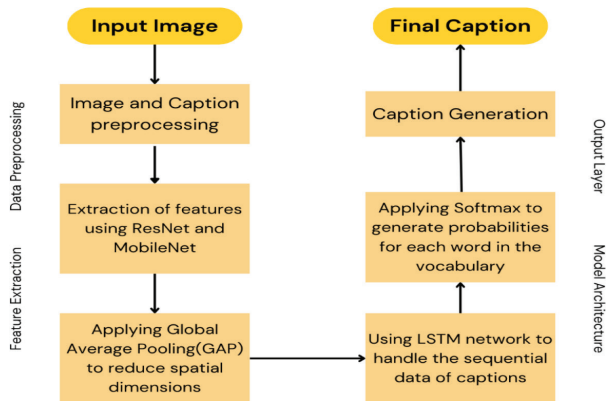


Figure 1: Proposed hybrid architecture model. GAP, global average pooling; LSTM, long short-term memory.

A multimodal IC network that combines CV and Natural Language Processing (NLP) into an E-D framework is shown in this diagram (Figure 2). After that, this fused feature vector is input into an LSTM-based language model, which works with the NLP component to produce meaningful and cohesive captions. The network's capacity to translate complicated visual content into structured natural language is demonstrated by the system's output of a descriptive sentence like "A young girl eating a sandwich at a table," which is further refined into a final caption that is grammatically correct. Two pretrained DL models are used to extract visual characteristics. A hybrid encoder then combines the outputs of these models to create a more robust visual representation. The created sentence's naturalness is gauged by the comparison. This serves as a loss of naturalness to guarantee that the phrase is cohesive and linguistically fluid. The created caption and a collection of pictures are sent to the listener network. Captions that are unable to distinguish the target image from other distractions are penalized in discriminability loss. Distinctiveness in relation to distract pictures is guaranteed by discriminability loss, whereas, fluency and resemblance to human language are guaranteed by naturalness loss. The captions are optimized by the algorithm to be human-like, natural, and detailed enough to differentiate identical photos. As can be seen in Figure 3, each section of the picture generates a phrase, which might then be combined to create a whole IC. Encoder uses CNN to fetch spatial and temporal characteristics from the video, and decoder uses LSTM to construct the caption word-by-word, form the foundation of a popular framework for video captioning that was first presented by [17, 18, 19].

The decoder lacks an explicit attention mechanism and is implemented as a vanilla LSTM. By

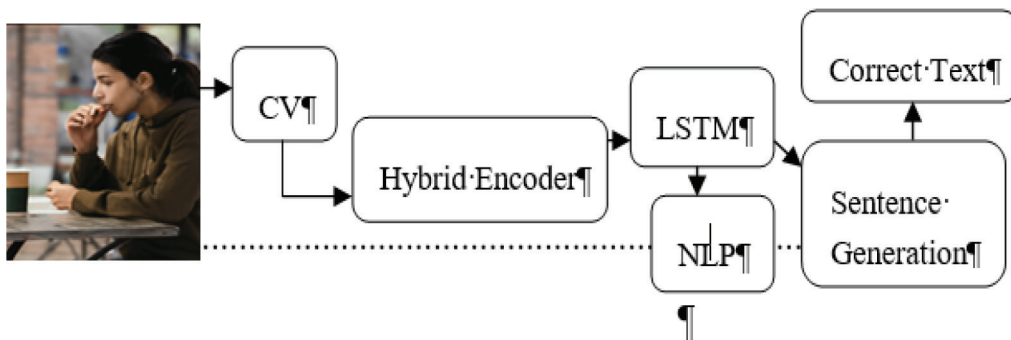


Figure 2: An illustration of a multimodal IC network. CV, computer vision; IC, image captioning; LSTM, long short-term memory.

using the fused CNN features to initialize the LSTM's hidden and cell states, visual context is added. Neither Luong nor Bahdanau attention modules are used.

A multimodal IC pipeline's whole forward pass is depicted in Figure 3, shows how an input image is gradually converted into a Natural Language (NL) description. A collection of visual features that capture significant aspects of the scene are produced by the encoder. After that, these features are sent into a hybrid encoder, which creates a richer fused output by fusing visual representations with more contextual information that has been learned. The Caption Generator evaluates the combined features and produces a descriptive sentence, finally receives this fused representation. The finished product shows how visual comprehension and language generation interact in a single forward pass, is a caption that describes an image of a child sitting at a table and working on an art or learning task.

After creating captions in English, we integrate the Google Translate Application Programming Interface (API) into the regular IC process to provide multilingual IC. The Google convert API is then used to convert the English captions into additional languages, such as Hindi, Bengali, and Spanish. Google TTS, or gTTS, is what we utilize to achieve this. This adaptability is necessary to enable the system to successfully meet the user requirements of the particular area. In addition to improving caption quality, the hybrid architecture allows for resource-constrained environments to function without seeing a decrease in speed.

With an initial learning rate of 0.0001, the Adam optimization method was used in the model training procedure. During the training phase, no decay mechanism was implemented and a fixed learning rate technique was used. Early halting was enabled and monitored using validation loss, with a patience threshold of 10 epochs, to avoid overfitting and needless computation 32 samples per iteration were used in the training process. Although 100 training epochs was the maximum, the early stopping criterion made sure that training ended as soon as performance on the validation set stopped getting better. In order to capture intricate temporal correlations in the data, the LSTM component of the model was set up using 512 hidden units. For the LSTM output layer, a dropout rate of 0.5 was used to enhance generalization and decrease overfitting. In order to further improve regularization, the fully connected fusion layer also included a dropout rate of 0.3 L2 regularization was applied using a weight decay factor of  $1 \times 10^{-5}$  in order to avoid the model parameters growing too much. With a maximum norm threshold of 5.0, gradient clipping was carried out to preserve stable training and prevent exploding gradients, which are frequently connected to recurrent networks.

To ensure consistent convergence and fair comparison, the same hyperparameter settings were used for both the Flickr8k and Flickr30k datasets in all tests. As demonstrated by tightly aligned training and validation loss curves, the reported setup produced consistent validation performance and avoided overfitting.

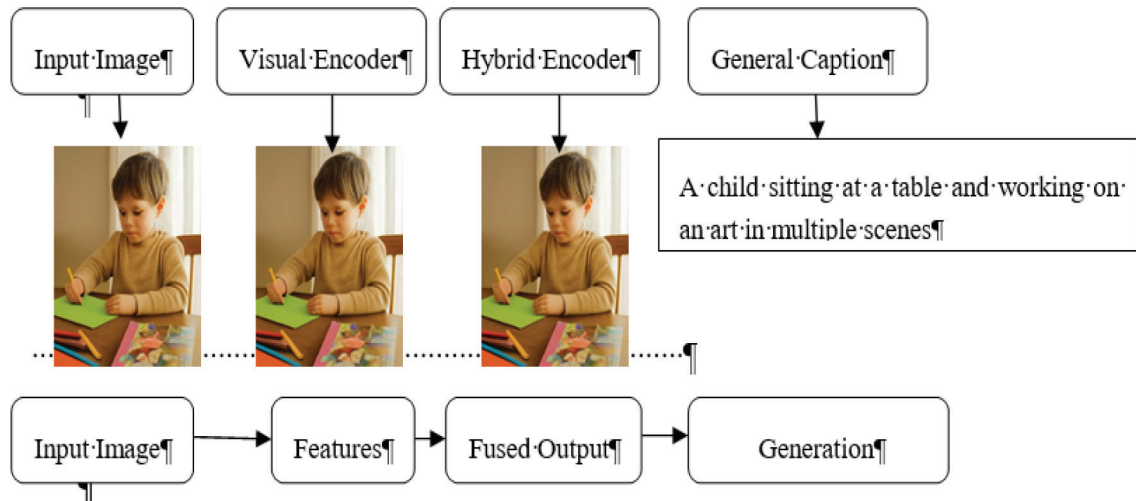


Figure 3: Dense captioning explanation numerous annotations with individual forward pass.

## IV. Dataset

To meticulously gather information designs and modify the quantity of parameters required for its gradient merging, DL need have access to lots of training information. Select a dataset is a crucial feature of a dataset that has a large number of photos (Flickr8k and Flickr30k dataset), each with a different caption that accurately describes the image's content. The current technique for extracting the attributes of an input picture CNN. Utilizing pre-trained image categorization model like the Visual Geometry Group (VGG) model (VGG-16 and VGG-19), Inception V3, MobileNet, ResNET50, and DenseNet model. We are precisely training the CNN on the pictures in the IC dataset.

Flickr8k dataset contains 8,092 photos in all, each with five informative captions. Here 4,855 photos, or 60% of the dataset, were selected for training from this collection. The remaining 1,619 photographs, or 20% of the total, were put aside for testing, while the validation set had 1,618 images. Every image kept its five captions across the corresponding splits, and care was taken to prevent image overlap between the test, validation, and training sets.

Flickr30k dataset contains 31,783 photos in all, each with five informative captions, make up the. 19,070 photos, or 60% of the dataset, were chosen for the training set in order to conduct the experimental evaluation. 6,357 photographs, or 20% of the total, made up the validation set; the remaining 6,356 images, or 20% of the total, were set aside for testing. To guarantee that no image featured in more than one split, the dataset was rigorously divided based

on image identifiers, and each image kept its five matching captions.

When preparing the dataset, no explicit class-based stratification was done because neither Flickr8k nor Flickr30k had ground-truth categorical labels, like object classes or scene categories, at the dataset level. Consequently, a 60/20/20 split was used to randomly divide the images for testing, validation, and training. These partitions were generated using a fixed random seed to guarantee reproducibility. Throughout the procedure, strict picture-level separation was upheld, guaranteeing that no image was included in more than one subset and that all captions for a particular image were limited to the same split.

Every experiment was conducted without the use of PyTorch in a standardized software environment with Ubuntu 20.04. LTS and Python 3.8.10, TensorFlow 2.9.1, and Keras 2.9.0 as the DL framework. NumPy 1.21.6, OpenCV 4.6.0, NLTK 3.7, Matplotlib 3.5.3, and Scikit-learn 1.1.2 were also used in the implementation to handle activities related to evaluation, visualization, image processing, natural language processing, and numerical computing.

## V. Results and Analysis

The models we considered in this article are described in summary form (Table 1). Here, we establish a comparative analysis of six IC architectures and introduce a new hybrid architecture model.

Our result of VGG-19 is better than [2, 8, 13], and VGG-16 is better than [5, 9]. Table 1 demonstrates

**Table 1: Obtained outcome from the individual DL model along with hybrid model**

| Model name                          | BLUE-1         | BLUE-2         | Trainable parameters | Non-trainable parameters |
|-------------------------------------|----------------|----------------|----------------------|--------------------------|
| VGG-19                              | 0.87524        | 0.81234        | 139,570,240          | 0                        |
| VGG-16                              | 0.85432        | 0.79648        | 118,684,640          | 0                        |
| ResNet50                            | 0.84657        | 0.78543        | 4,231,976            | 19,888                   |
| MobileNet                           | 0.83636        | 0.75032        | 705,330              | 3,314                    |
| Dense Net                           | 0.81564        | 0.75546        | 1,410,664            | 6,630                    |
| InceptionV3                         | 0.79654        | 0.71542        | 3,526,650            | 15,670                   |
| <b>Proposed hybrid architecture</b> | <b>0.98426</b> | <b>0.90558</b> | <b>5,204,864</b>     | <b>33,428</b>            |

Bold values indicate the performance of the proposed model.  
DL, deep learning.

that our proposed hybrid architecture model is better than [2, 5, 6, 8, 9, 12, 15, 18, 20, 22–25].

BLEU scores were calculated at the corpus level using the Flickr datasets and NLTK's corpus\_bleu, which included five reference captions per image. In contrast to higher-order BLEU scores (BLEU-3/4), which are significantly lower, the exceptionally high BLEU-1 score indicates extensive unigram overlap across numerous references and frequent object tokens. Without pointing to data leakage or caption memorization, the high BLEU-1 is explained by the usage of multiple references and unigram-level evaluation. It is also observed that by reducing overfitting, as seen by a smaller gap between training and validation loss and more steady convergence after roughly 50 epochs, the implemented technique improved

validation BLEU-1 by about 3.5% and validation BLEU-2 by about 3.8%.

Two line graphs from a 100-epoch training procedure are displayed in the picture (Figure 4). Training and validation loss are shown on the left graph, which shows multiple acute spikes and a general decline. Training and validation accuracy increases progressively from roughly 0.95 to almost 1.0 in the graph on the right. Both curves show that the model's performance is becoming better, with accuracy stabilizing near the end. Although recall is not quite as good as precision, it is still quite good in our hybrid model (Figure 5). This picture shows how well our hybrid model performs in categorization, particularly when making positive predictions. Exceptionally high F1 score due to nearly flawless recall and accuracy.

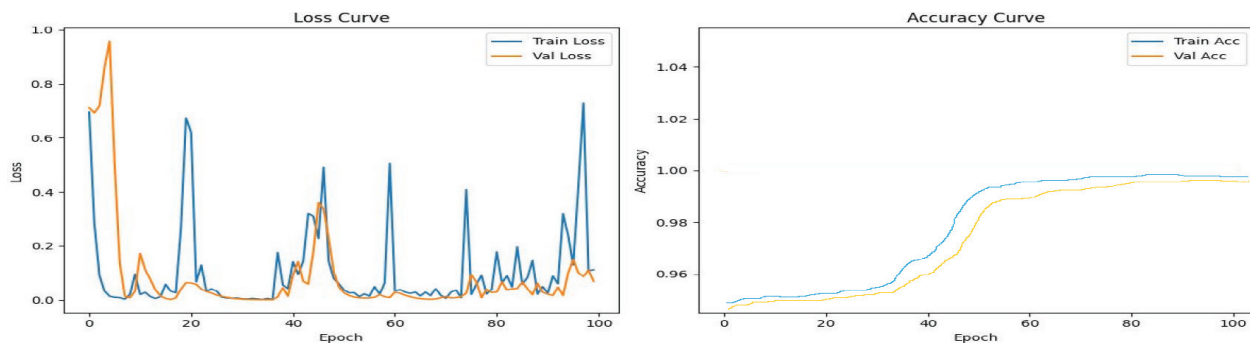


Figure 4: Accuracy & loss curve of proposed hybrid architecture model for epoch 100.

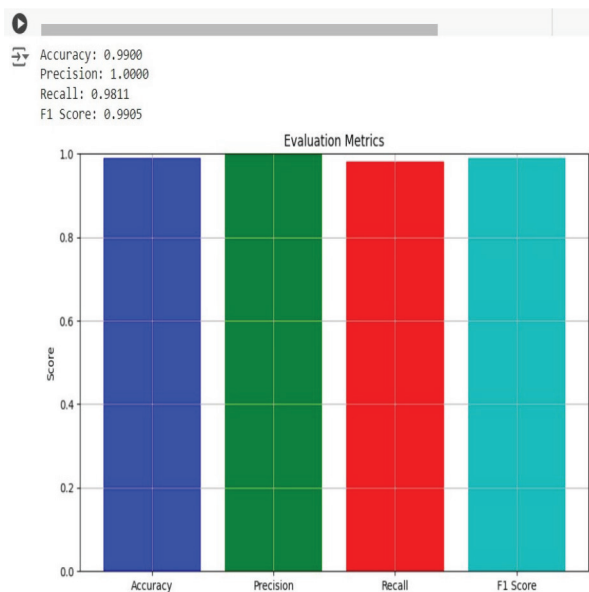


Figure 5: Evaluation metrics.

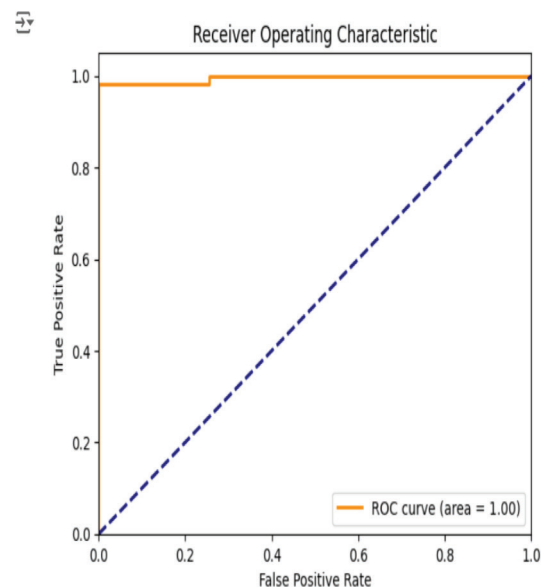


Figure 6: ROC curve. ROC, receiver operating characteristic.

Capability of a hybrid DL model is used to discriminate between two classes in a binary classification task is demonstrated by the model's receiver operating characteristic (ROC) curve (Figure 6). When a hybrid model performs well, the curve bows considerably in the upper-left corner, denoting a high true positive rate and low false positive rate.

As supplementary diagnostics, token-level categorization metrics were also supplied; however, BLEU scores are the main assessment for caption quality. Lexical prediction stability, convergence behavior, and confidence calibration are all revealed by these measures, which were calculated by comparing predicted and ground-truth words at each decoding step (not including padding). Sequence-level captioning metrics are not supposed to be replaced by them.

Because of its improved feature representations, the hybrid model continues to perform well even at more stringent thresholds.

Two scatter plots analyzing the relationship between sample size and two performance metrics are displayed in our hybrid model (Figure 7). Both are combined in the hybrid model, which improves performance but may have higher latency based on the fusion method and sample size. Due to fusion efficiency, the hybrid model usually has a latency that is greater than that of MobileNet alone but less than that of ResNet50 alone in big batches. For all sample sizes, the Matthews Correlation Coefficient (MCC) is flat at zero. This indicates that regardless of the number of samples used, the classifier did not discover any significant patterns (performance comparable to random guessing). As the sample size grows, the prediction latency drastically reduces. Delay is greater for small sample sizes (~5–20). As the sample size grows, the model gets quicker, most likely as a result of overhead being amortized by batching calculations. The internal state of the model becomes more stable. This is advantageous as predictions

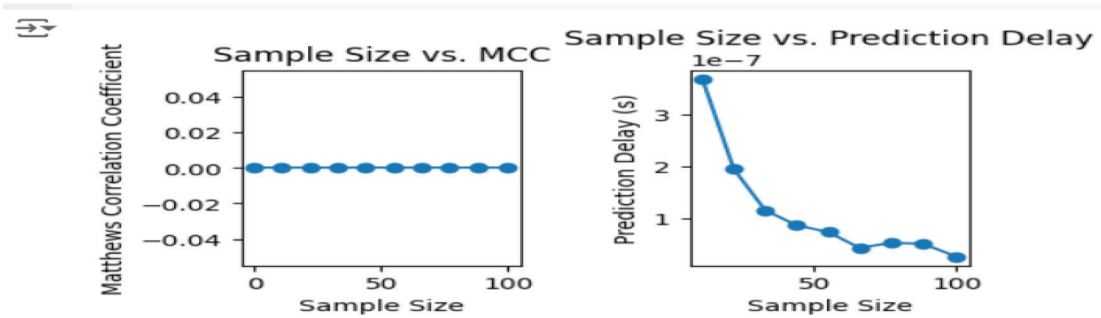


Figure 7: Sample size vs. MCC and sample size vs. prediction delay.

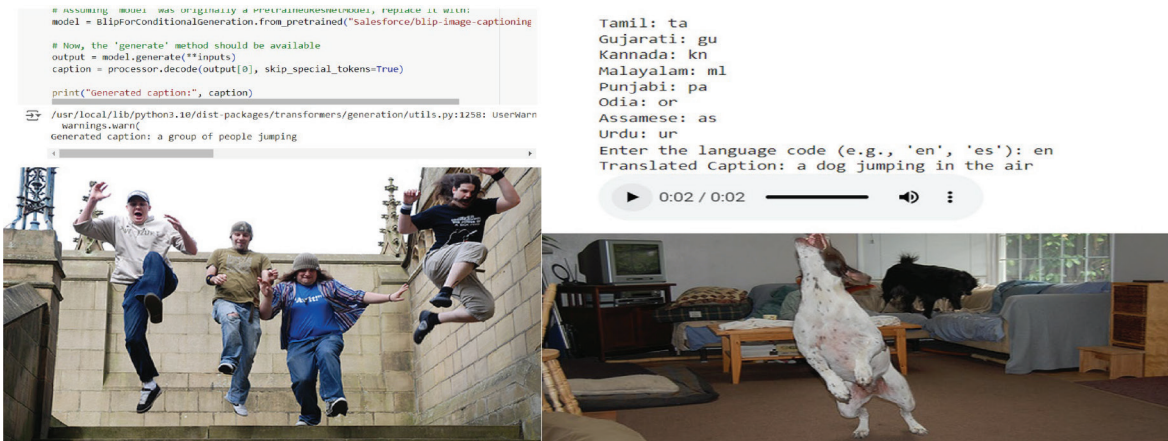


Figure 8: Sample output.

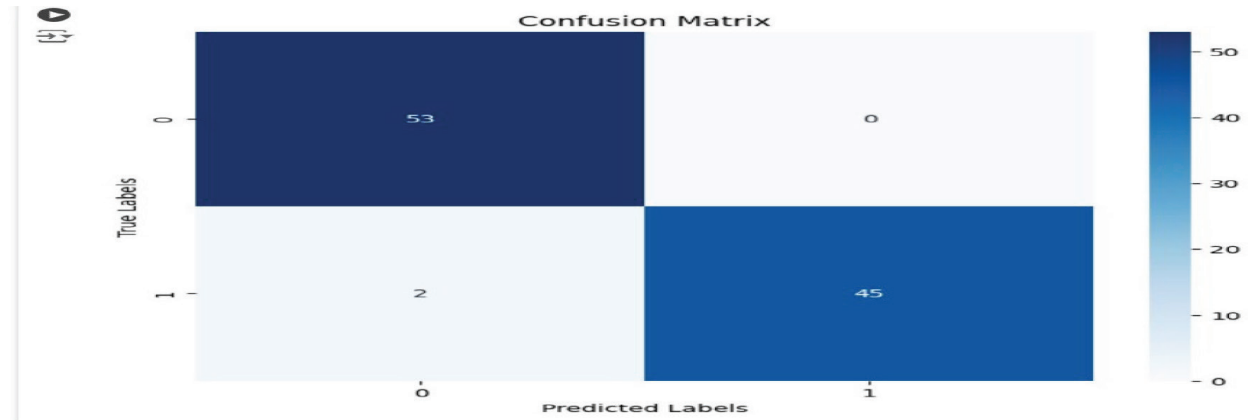


Figure 9: Confusion matrix.

Table 2: Paired samples statistics (*T*-test)

|        |                  | Mean   | <i>N</i> | Std. deviation | Std. error mean |
|--------|------------------|--------|----------|----------------|-----------------|
| Pair 1 | HybridModel_BLEU | 0.1927 | 1,999    | 0.18642        | 0.01874         |
|        | MobileNetV2_BLEU | 0.0000 | 1,999    | 0.00005        | 0.00000         |

Table 3: Paired samples correlations

|        |                    | <i>N</i> | Correlation | One-sided <i>p</i> | Two-sided <i>p</i> |
|--------|--------------------|----------|-------------|--------------------|--------------------|
| Pair 1 | HybridModel_BLEU & | 2        | 0.091       | 0.186              | 0.372              |
|        | MobileNetV2_BLEU   | 1,999    |             |                    |                    |

Table 4: Paired samples test

|        |                                        | Mean  | Std. deviation | Std. error mean | Lower   |
|--------|----------------------------------------|-------|----------------|-----------------|---------|
| Pair 1 | HybridModel_BLEU –<br>MobileNetV2_BLEU | 0.192 | 0.18641        | 0.01874         | 0.15553 |

become more accurate as data scales. More samples increase the computing efficiency, but the classifier's performance remains random and does not improve at all. If the model is supposed to learn significant patterns, this is a serious warning sign. Figure 8 shows sample output of our hybrid architecture model. Figure 5 shows an excellent accuracy rate of 98.42% of our hybrid architecture model. Recall is just below 100% as a result of a small number of false negatives. Overall, our model works well in almost flawless categorization. Recall could be somewhat improved

to collect all the positives, but precision and recall are well balanced. Figure 9 displays a heatmap of a confusion matrix with bright and dark blue squares, showing numerical values in each cell, labeled axes, a color bar, and expected versus actual classes.

*T*-test is utilized to determine if the means of two groups differ significantly from one another and can be used to compare the following in the context of a hybrid model. The hybrid model's performance measures in comparison to ResNet50 or MobileNet alone. Null Hypothesis, that is, the performance metric does

**Table 5: Paired samples test**

|        |                                        | Upper   | <i>t</i> | df | One-sided <i>p</i> | Two-sided <i>p</i> |
|--------|----------------------------------------|---------|----------|----|--------------------|--------------------|
| Pair 1 | HybridModel_BLEU –<br>MobileNetV2_BLEU | 0.22989 | 10.286   | 98 | <0.001             | <0.001             |

df, degrees of freedom.

**Table 6: Paired samples effect sizes**

|        |                                        |                    | Standardizer <sup>a</sup> | V point estimate | Lower |
|--------|----------------------------------------|--------------------|---------------------------|------------------|-------|
| Pair 1 | HybridModel_BLEU –<br>MobileNetV2_BLEU | Cohen's <i>d</i>   | 0.18641                   | 1.034            | 0.788 |
|        |                                        | Hedges' correction | 0.18786                   | 1.026            | 0.782 |

**Table 7: Paired samples effect sizes**

|        |                                     |                    | Upper |
|--------|-------------------------------------|--------------------|-------|
| Pair 1 | HybridModel_BLEU – MobileNetV2_BLEU | Cohen's <i>d</i>   | 1.276 |
|        |                                     | Hedges' correction | 1.267 |

**Table 8: Npar tests (Wilcoxon signed ranks test) ranks**

|                                        |                | <i>N</i>          | Mean rank | Sum of ranks |
|----------------------------------------|----------------|-------------------|-----------|--------------|
| HybridModel_BLEU –<br>MobileNetV2_BLEU | Negative ranks | 1797 <sup>a</sup> | 49.00     | 4753.00      |
|                                        | Positive ranks | 0 <sup>b</sup>    | 0.00      | 0.00         |
|                                        | Ties           | 202 <sup>c</sup>  |           |              |
|                                        | Total          | 1999              |           |              |

not differ significantly. Alternative Hypothesis that is, the performance metric differs significantly.

In order to perform an exploratory analysis, statistical tests were performed on sentence-level BLEU values and depicted in Table 2. Due to the weak independence assumptions and the known volatility of sentence-level BLEU, these results are presented purely for completeness and are not utilized to make key conclusions.

During inference, greedy decoding (beam width = 1) was employed for caption synthesis. Beam search was not used for BLEU evaluation.

The one-sided *p* value is used in paired samples correlations to test the directional hypothesis on the correlation between two related variables (Table 3). Paired samples correlation is used to determine

**Table 9: Test statistics<sup>A</sup> Mobilenetv2\_BLEU – Hybridmodel\_BLEU**

|                        |                     |
|------------------------|---------------------|
| Z                      | –8.551 <sup>b</sup> |
| Asymp. Sig. (2-tailed) | <0.001              |

whether two related (paired) variables, such as pre-test and post-test scores from the same group, have a statistically meaningful linear relationship.

The two-sided *p* value is used in paired samples correlations to determine whether two linked variables have a significant positive or negative correlation.

Table 4 shows the paired samples test. The means of two related groups are compared using a paired samples *t*-test to see if between them, there is

**Table 10: Ablation results (BLEU scores)**

| Model variant        | Fusion               | GAP          | BLEU-1       | BLEU-2       | BLEU-3       | BLEU-4       |
|----------------------|----------------------|--------------|--------------|--------------|--------------|--------------|
| MobileNet            | –                    | Yes          | 0.836        | 0.750        | 0.664        | 0.521        |
| ResNet50             | –                    | Yes          | 0.846        | 0.785        | 0.691        | 0.552        |
| ResNet50 + MobileNet | <b>Addition</b>      | Yes          | 0.912        | 0.843        | 0.731        | 0.584        |
| ResNet50 + MobileNet | <b>Concatenation</b> | <b>Yes</b>   | <b>0.984</b> | <b>0.872</b> | <b>0.721</b> | <b>0.592</b> |
| ResNet50 + MobileNet | <b>Concatenation</b> | No (Flatten) | 0.927        | 0.861        | 0.702        | 0.556        |

Bold values indicate the performance of the proposed model.  
GAP, global average pooling.

a statistically significant difference. Here, degrees of freedom ( $df$ ) =  $n - 1$ , that is, one less than the number of paired observations (Table 5).

The effect size measures the extent of the mean distinctness between the two related groups in the context of a Paired Samples  $t$ -test. Cohen’s  $d$  is a frequently used effect size.

The point estimate in paired samples effect sizes typically refers to Cohen’s  $d$ , a standardized effect size, which illustrates how much two paired scores differ from one another (Table 6). The effect sizes were estimated using the denominator. The sample standard deviation of the average variation is determined by Cohen’s  $d$ . Hedges’ correction mixes a correction factor using the sample standard deviation of the mean deviation (Table 7).

The “Negative Ranks mean” describes the average of the rankings given to instances in where the second variable is less than the first (i.e., a negative difference) in non-parametric tests (NPar tests), such as the Wilcoxon Signed Ranks Test (Table 8).

Here, ‘a’ means  $\text{MobileNetV2\_BLEU} < \text{HybridModel\_BLEU}$ , ‘b’ implies  $\text{MobileNetV2\_BLEU} > \text{HybridModel\_BLEU}$  and ‘c’ signifies  $\text{MobileNetV2\_BLEU} = \text{HybridModel\_BLEU}$ .

Here, ‘a’ denotes Wilcoxon signed ranks test and ‘b’ implies based on positive ranks (Table 9).

Richer feature extraction and stronger semantic understanding are made possible by ResNet50’s deeper residual architecture, which results in higher BLEU scores than MobileNet. Although MobileNet is competitive, it puts computational economy ahead of expressiveness and representational depth (Table 10). According to ablation experiments, the hybrid encoder routinely performs better than single-encoder baselines. The benefits of maintaining complementary representations are confirmed by the better BLEU scores that concatenation produces compared to element-wise addition. GAP generates

**Table 11: Comparative accuracy for considering different papers**

| Reference         | Database                                | Accuracy obtained                                                                |
|-------------------|-----------------------------------------|----------------------------------------------------------------------------------|
| [1]               | Flickr 30k                              | For BLEU-1: 0.72<br>For BLEU-2: 0.63<br>For BLEU-3: 0.51<br>For BLEU-4: 0.38     |
| [2]               | Remote sensing image captioning dataset | For BLEU-1: 0.661<br>For BLEU-2: 0.476<br>For BLEU-3: 0.375<br>For BLEU-4: 0.302 |
| [8]               | Flickr 8k                               | 0.45 (20 Epoch) and 0.73 (50 Epoch)                                              |
| [11]              | Flickr 8k                               | For BLEU-1: 0.28<br>For BLEU-2: 0.17<br>For BLEU-3: 0.08<br>For BLEU-4: 0.02     |
| [12]              | Flickr8k                                | For BLEU-1: 0.31<br>For BLEU-2: 0.10<br>For BLEU-3: 0.04<br>For BLEU-4: 0.02     |
| [18]              | Flickr 8k                               | For BLEU-1: 0.76<br>For BLEU-2: 0.74<br>For BLEU-3: 0.73<br>For BLEU-4: 0.72     |
| [19]              | Flickr 8k                               | For BLEU-1: 0.75<br>For BLEU-2: 0.61<br>For BLEU-3: 0.47<br>For BLEU-4: 0.32     |
| [22]              | Flickr 8k                               | For BLEU-1: 0.74<br>For BLEU-2: 0.65<br>For BLEU-3: 0.49<br>For BLEU-4: 0.31     |
| Proposed approach | Flickr 8k and Flickr30k                 | For BLEU-1: 0.98<br>For BLEU-2: 0.87<br>For BLEU-3: 0.72<br>For BLEU-4: 0.59     |

compact global descriptors, which further enhance stability and performance. With GAP and concatenation, ResNet50+ MobileNet provides the best performance. HybridModel performed significantly better than MobileNet ( $p < 0.001$ ). Average BLEU score difference = 0.193, effect size (Cohen's  $d \approx 1.03$ ) implies a large difference. Wilcoxon signed-ranks test purpose is to confirm T-Test results without assuming a normal distribution. 97 out of 99 samples had higher BLEU scores for HybridModel. Statistically significant difference ( $Z = -8.55$ ,  $p < 0.001$ ). Both tests strongly indicate that Hybrid Model outperforms MobileNet in BLEU score performance for IC (Table 11).

## VI. Conclusions

The focus of extroverted IC research is on DL-located processes, where adversarial learning, deep reinforcement, and live attention mechanisms are present. The exercise of the present processes is considered in our study. The best-performing model is GAN-based, and Up-Down has the most influence. The society will benefit from our review in a reciprocal manner, as it will include the present review articles for further traced research in IC. A DN sketch using a residual bidirectional LSTM is proposed in this paper. The novel network can stop the gradient from disappearing in space and time to raise recognition rates. Cutting-edge leads in SS and IC have been induced in this research. In order to complete both duties, the persuasiveness and characteristics of the essential strategies have been discussed. Our assumption is based on the difficult task of manually classifying photos into semantic groups and annotating complex human images by segmentation and captioning. Accessibility for visually impaired people can be greatly enhanced by an integrated vision-language model that produces accurate image captions, allowing them to engage more meaningfully with digital content and take part more completely in social, professional, and educational activities. By automatically producing thorough product descriptions, lowering manual labor costs, enhancing catalog quality, and making it easier for small enterprises to grow their online presence, enhanced picture captioning can benefit e-commerce platforms. Automated captioning of photos taken by drones or mobile devices can help with real-time decision-making in industries including agriculture, healthcare, and crisis management, increasing service effectiveness and financial results. Integrated vision-language systems enable faster data processing and cheaper operating costs for businesses by eliminating the need for human

annotation and content moderation, which ultimately boosts productivity and economic growth. Better caption creation encourages inclusivity in social media and digital communication by making information accessible to a wide range of people, regardless of their literacy or language skills. This promotes social justice and empowerment.

## VII. Future Scope

Later on, IC will be used to handle complicated scenarios, incorporate contextual knowledge, provide language assistance, and improve accuracy using sophisticated AI models. Better integration of textual and visual data, real-time captioning, and customization of captions based on user preferences are possible areas for future improvement. In order to obtain even more intriguing captioning results, the LSTM model may be repurposed as a bidirectional model to map between pictures and their sentence-located descriptions. Future studies will go toward an unsupervised environment in order to complete this task. This would guarantee that future efforts are focused primarily on creating intricate ML algorithms and mathematical models that possess the capacity to push the limits of the present-day status of the art. Transformer-based architectures, which provide better long-range dependency modeling and can enhance the precision and complexity of output captions, can be the focus of future research. Semantic comprehension can be further strengthened by improving multimodal fusion by incorporating additional data sources like scene graphs, object detectors, or contextual metadata. We recognize that gated fusion or attention-based methods are viable substitutes. To keep complete feature information while preserving architectural simplicity and training stability, we purposefully use simple concatenation in our study. Consistent increases were not obtained from preliminary comparisons with various fusion procedures, suggesting that concatenation is enough in the given experimental setup. Improved generalization across many visual domains will be aided by larger and more varied datasets. Although transformer-based decoders currently dominate picture captioning, we purposefully use an LSTM decoder to keep the architecture simple and concentrate on assessing the suggested hybrid visual encoder. LSTMs continue to offer robust training and competitive results on medium-scale datasets. We consider transformers to be a potential expansion and plan to continue this comparison in the future. To make captioning more human-like and context-aware,

attention-based processes and cross-modal alignment techniques can be improved. Real-time caption creation is still a major topic for advancement, especially for applications like autonomous systems and assistive technologies. Lastly, investigating domain-specific captioning—for surveillance, remote sensing, or healthcare—can result in specialized and extremely accurate captioning models that are suited to practical requirements.

## References

- [1] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, "Show and tell: A neural image caption generator," In CVPR, doi:10.1109/CVPR.2015.7298935, 2015.
- [2] G. Hoxha, F. Melgani, J. Slaghenauuffi, "A new CNN-RNN framework for remote sensing image captioning," In: 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Tunis, Tunisia, p. 1-4, June, 2020, doi:10.1109/M2GARSS47143.2020.9105191.
- [3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering", In CVPR, 2018, <https://doi.org/10.48550/arXiv.1707.07998>.
- [4] S.J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, "Self-critical sequence training for image captioning," In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- [5] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," Int. Conf. Mach. Learn., pp. 2048–2057, 2015.
- [6] M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, "Meshed-Memory Transformer for Image Captioning," In CVPR, 2020, <https://doi.org/10.48550/arXiv.1912.08226>.
- [7] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, et al., "OSCAR: object-semantics aligned pre-training for vision-language tasks," In: European Conference on Computer Vision (ECCV), p. 121-137, 2020.
- [8] J. Sudhakar, V.V. Iyer, S.T. Sharmila, "Image caption generation using deep neural networks," In: International Conference for Advancement in Technology (ICONAT), Goa, India. p. 1-3, March, 2022, doi:10.1109/ICONAT53423.2022.9726074.
- [9] P. Li, M. Zhang, P. Lin, J. Wan, M. Jiang, "Visual-text reference pre-training model for image captioning," Comput Intell Neurosci, 2022:9400999, 2022.
- [10] N. Li, Z. Chen, S. Liu, "Meta-learning for image captioning," Proc AAAI Conf Artif Intell, 2019, 33:8626-8633.
- [11] W. Kang, W. Hu, "A survey of image caption tasks," In: 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI), Nanjing, China, p. 71-74, November, 2022, doi:10.1109/CEI57409.2022.9950150.
- [12] H. Maru, T. Chandana, D. Naik, "Comparison of image encoder architectures for image captioning," In: 5th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, p. 740-744, April, 2021, doi:10.1109/ICCMC51019.2021.9418234.
- [13] S. Yıldız, A. Memiş, S. Varlı, "Automatic Turkish image captioning: the impact of deep machine translation," In: 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Türkiye, p. 414-419, October, 2023, doi:10.1109/UBMK59864.2023.10286693.
- [14] P. Tian, H. Mo, L. Jiang, "Improved image captioning via semantic feature update," In: 40th Chinese Control Conference (CCC), Shanghai, China, p. 7938-7943, July, 2021, doi:10.23919/CCC52363.2021.9549991.
- [15] A. P. Singh, M. Manoria and S. Joshi, "A Review on Automatic Image Caption Generation for Various Deep Learning Approaches," 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, pp. 1-5, doi: 10.1109/ICCCNT56998.2023.10308085, 2023.
- [16] Yu Z, Fu K, Jin H, Bai J, Zhang H, Li Y, "Local and global multimodal interaction for image caption," In: 4th International Conference on Electronic Communication and Artificial Intelligence (ICECAI), Guangzhou, China, p. 164-169, May, 2023, doi:10.1109/ICECAI58670.2023.10176671.
- [17] Krasin I, Duerig T, Alldrin N, Ferrari V, Abu-El-Haija S, Kuznetsova A, et al., "Open images: a public dataset for large-scale multi-label and multi-class image classification [dataset on the Internet],", June, 2017, Available from: <https://github.com/openimages>.
- [18] Kumari A, Chauhan A, Singhal A, "Vision 360: image caption generation using encoder-decoder model," In: 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, p. 312-317, January, 2022, doi:10.1109/Confluence52989.2022.9734167.
- [19] D. J. B. Saini, S. Kumar, K. Joshi, A. K. Pathak, S. Jain and A. Singh, "A Novel Approach of Image Caption Generator using Deep Learning,"

Third International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, 2023, pp. 24-29, 2023, doi: 10.1109/ICUIS60567.2023.00012.

[20] Z. U. Kamangar, G. M. Shaikh, S. Hassan, N. Mughal and U. A. Kamangar, "Image Caption Generation Related to Object Detection and Colour Recognition Using Transformer-Decoder," 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, pp. 1-5, 2023, doi: 10.1109/iCoMET57998.2023.10099161.

[21] Krishna, R., Zhu, Y., Groth, O. et al. Visual Genome: Connecting Language and Vision Utilizing Crowd sourced Dense Image Annotations. *Int J Comput Vision* 123, 32–73, 2017, <https://doi.org/10.1007/s11263-016-0981-7>.

[22] Z. U. Kamangar, G. M. Shaikh, S. Hassan, N. Mughal and U. A. Kamangar, "Image Caption Generation Related to Object Detection and Colour Recognition Using Transformer-Decoder," 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 2023, pp. 1-5, 2023, doi: 10.1109/iCoMET57998.2023.10099161.

[23] A. P. Singh, M. Manoria and S. Joshi, "A Review on Automatic Image Caption Generation for Various Deep Learning Approaches," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, pp. 1-5, 2023, doi: 10.1109/ICCCNT56998.2023.10308085.

[24] D. J. B. Saini, S. Kumar, K. Joshi, A. K. Pathak, S. Jain and A. Singh, "A Novel Approach of Image Caption Generator using Deep Learning," 2023 Third International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), Gobichettipalayam, India, pp. 24-29, 2023, doi: 10.1109/ICUIS60567.2023.00012.

[25] K. Fatema, S. Montaha, M. A. H. Rony, S. Azam, M. Z. Hasan, and M. Jonkman, "A robust framework combining image processing and deep learning hybrid model to classify cardiovascular diseases using a limited number of paper-based complex ECG images," *Biomedicines*, 10, 2835, 2022, <https://doi.org/10.3390/biomedicines10112835>.

[26] J. Yang and N. A. Mat Isa, "YOLOv10-MsA: Attention-Augmented Real-Time Insulator Defect Detection from UAV Imagery," *Emerging Science Journal*, vol. 9, no. 6, pp. 2899–2914, 2025, doi: 10.28991/ESJ-2025-09-06-02.

[27] Z. Deng, X. Li, R. Yang, "RML-YOLO: An Insulator Defect Detection Method for UAV Aerial

Images," *Applied Sciences*, 15(11):6117, 2025, <https://doi.org/10.3390/app15116117>.

[28] R.D. Puriyanto, I.D. Yunandha, H. Maghfiroh, A. Ma'arif, Furizal and I. Suwarno, "Ball Detection System for a Soccer on Wheeled Robot Using the MobileNetV2 SSD Method," *Emerging Science Journal* 9, 5, 2782–2796, 2025, doi: <https://doi.org/10.28991/ESJ-2025-09-05-028>.

[29] X. Lian, D. Wang, "Insulator defect detection algorithm based on improved YOLOv5," *Frontiers in Computing and Intelligent Systems*, 3(2), 44-47, 2023. <https://doi.org/10.54097/fcis.v3i2.7168>.

[30] J. Liu, C. Liu, Y. Wu, Z. Sun, H. Xu, "Insulators' Identification and Missing Defect Detection in Aerial Images Based on Cascaded YOLO Models," *Comput Intell Neurosci*, 2022:7113765, 2022, August, doi: 10.1155/2022/7113765. PMID: 36035858; PMCID: PMC9402330.

[31] X. Sun, S. Qi, "A Study on Defect Detection of YOLOv8 Insulators Based on Improvements," *Academic Journal of Computing & Information Science*, Vol. 7, Issue 5: 37-43, 2024, <https://doi.org/10.25236/AJCIS.2024.070505>.

[32] N. Fahad, M. J. Hossen, and M. S. Sayeed, "Efficient object detection with an optimized YOLOv8x model," *HighTech and Innovation Journal*, vol. 6, no. 3, Article 09, Sep. 2025, doi: 10.28991/HIJ-2025-06-03-09.

[33] C. Jia, D. Wang, J. Liu, W. Deng, "Performance Optimization and Application Research of YOLOv8 Model in Object Detection," *Academic Journal of Science and Technology*, 10(1), 325-329, 2024, <https://doi.org/10.54097/p9w3ax47>.

[34] Y. Zhao, N. C. Rodelas, "Enhanced Small Target Detection Methodology via Optimized YOLOv8 Framework," *Academic Journal of Science and Technology*, 11(3), 80-84, 2024, <https://doi.org/10.54097/7terh631>.

[35] B. Yilmaz, U. Kutbay, "YOLOv8-Based Drone Detection: Performance Analysis and Optimization," *Computers*, 13(9):234, 2024, <https://doi.org/10.3390/computers13090234>.

[36] U. Verma, "YOLOv8: An Enhanced Object Detection Model for Distance Estimation," *International Journal of Intelligent Systems and Applications in Engineering*, 12(21s), 3852, 2024, <https://www.ijisae.org/index.php/IJISAE/article/view/6156>.

[37] S. Erniwati, V. Afifah and B. Imran, "Mask Region-based Convolutional Neural Network in Object Detection: A Review," *IJACI: International Journal of Advanced Computing and Informatics*, vol. 1, no. 2, pp. 106–117, 2025, doi: 10.71129/ijaci.v1i2.pp106-117.

- [38] D. Juhartini, D. Arwidiyarti, and Desmiwati, “Single Shot Multibox Detector (SSD) in Object Detection: A Review,” *IJACI: International Journal of Advanced Computing and Informatics*, vol. 1, no. 2, pp. 118–127, 2025, doi: 10.71129/ijaci.v1i2.pp118-127.
- [39] R. Sastra, D. Hariyanto Dicky, and B. Apriyansyah, “Fast Region-based Convolutional Neural Network in Object Detection: A Review,” *IJACI: International Journal of Advanced Computing and Informatics*, vol. 2, no. 1, pp. 34–40, 2026, doi: 10.71129/ijaci.v2i1.pp34-40.
- [40] W. Ilham and A. Ahmad, “A comprehensive review of ConvNeXt architecture in image classification: Performance, applications, and prospects,” *IJACI: International Journal of Advanced Computing and Informatics*, vol. 2, no. 2, pp. 108–114, 2026, doi: 10.71129/ijaci.v2i2.pp108-114.
- [41] R. D. Ramadhan and Y. A. Adam, “Efficient tomato leaf disease classification using adapter based fine tuning of Vision Transformers,” *IJACI: International Journal of Advanced Computing and Informatics*, vol. 2, no. 2, pp. 127–138, 2026, doi: 10.71129/ijaci.v2i2.pp127-138.
- [42] V. Eriyandi and A. N. Ahmad, “A comparative machine learning and deep learning models with ElasticNet regularization for predicting student outcomes: LGBM, CatBoost, ANN, DNN, and WDNN,” *IJACI: International Journal of Advanced Computing and Informatics*, vol. 2, no. 2, pp. 139–150, 2026, doi: 10.71129/ijaci.v2i2.pp139-150.
- [43] N. S. Rahmawati and C. I. Amalia, “Lightweight ensemble models for static malware detection: Addressing deep learning trade-offs with the Kaggle PE dataset,” *IJACI: International Journal of Advanced Computing and Informatics*, vol. 2, no. 1, pp. 57–66, 2026, doi: 10.71129/ijaci.v2i1.pp57-66.
- [44] K. He, G. Gkioxari, P. Dollár and R. Girshick, “Mask R-CNN,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020, doi: 10.1109/TPAMI.2018.2844179.
- [45] Z.-Q. Zhao, P. Zheng, S.-T. Xu and X. Wu, “Object detection with deep learning: A review,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019, doi: 10.1109/TNNLS.2018.2876865.
- [46] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan and S. Belongie, “Feature Pyramid Networks for object detection,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 936–944, doi: 10.1109/CVPR.2017.106.
- [47] H. Li, X. Qi and G.-J. Qi, “A guide to convolutional neural networks for visual recognition,” *Foundations and Trends® in Computer Graphics and Vision*, vol. 15, no. 3, pp. 207–371, 2021, doi: 10.1561/06000000117.
- [48] X. Wang, D. Zhang, G. Wang, and Z. Cui, “Performance improvement techniques for Mask R-CNN in visual detection tasks,” *Journal of Visual Communication and Image Representation*, vol. 81, p. 103426, 2021, doi: 10.1016/j.jvcir.2021.103426.