International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I7P130

Volume 73 Issue 7, 383-401, July 2025

Original Article

An Intuitive Approach with IAT Model for Image Captioning and Labelling Analysis Using Light and Deep CNN

V. Chandra Sekhar Reddy[1], S. Jessica Saritha[2]

[1,2]*Department of Computer Science and Engineering, Jawaharlal Nehru Technological University, Anantapur, Andhra Pradesh, India.*

[1]*Corresponding Author : vcsreddy2003@gmail.com*

Received: 24 December 2024 Revised: 02 July 2025 Accepted: 10 July 2025 Published: 30 July 2025

Abstract - *In image analysis, image captioning is very essential, and its emphasis is laid on functional and spatial aspects modeling of image by use of intuitive models. Based on the Flicker-8k dataset, the proposed model improves the Invasive Augmented Transform (IAT) model based on Deep CNN as a framework. The suggested methodology encompasses different building blocks such as GAN, LSTM, Oz-Net, and Inception-Net, giving a testing precision of 93%. In an attempt to address the computing requirements of bigger samples and proliferated models of IAT of 18- and 36-layers with enhanced accuracy of 99%. Compared with the 18-layer model, which is optimized in terms of training efficiency (97 percent accuracy), in the 36-layer model, the added complexity and accuracy are 2 percent. The IAT model uniquely augments the text with images to represent intricate processes, utilizing segmentation filters to refine caption coherence. The performance of the trials for the proposed model demonstrates that the IAT algorithm surpasses state-of-the-art architectures by 6% in accuracy and reduces execution time by 30%, showcasing impressive performance metrics like accuracy and BLEU for image labeling.*

Keywords - *Image labelling, Deep Learning, Convolutional Neural Networks, LSTM, Invasive Augmented Transform.*

1. Introduction

Image Labelling is very essential to visualize data with natural, custom aspects of the image features and labels associated with each type considered in real time. To encompass the requirement of understanding the label and captioning of the images, newer features and technologies are introduced to enable image labelling and allow an automated system to provide usable descriptions of images effectively. Similarly, the creation and automatic generation of text explanations about the visual content in the form of image captioning and labelling has become an essential area of mobile applications in computer vision and artificial intelligence. Such technologies are very important in areas like remote sensing, autonomous navigation, military surveillance, and hazard detection in determining situations where scenarios are described automatically to improve the understanding and decision-making process [1]. Considering the (traditional) encoder-decoder structures, which are limited in capturing fine-grained semantics and spatial structure, particularly in difficult or high-resolution images. Subsequently, new research has offered several improvements to provide visual and linguistic comprehension. The application of feature enhancement, which enhances the quality of visual inputs, is one of the major developments. In a Two-Stage Feature Enhancement (TSFE), certain attention has been given to where global and local image features are processed in isolation successively, giving more descriptive and factual captions, especially in remote sensing, where objects and spatial layouts matter [3]. Similarly, another development aims to enhance the reasoning capacity in captioning models. Clear explanations of the processes, i.e. ones that rely on utilizing a Language-Matching Module (LMM), have been created in order to drive the generation procedure of captions by matching visual information with linguistic structure to enable the system to deal more efficiently with complex scenes involving numerous interacting objects [2]. Besides feature processing and reasoning, the visual and textual modalities need to be aligned so that the overall performance of the multimodal system can be enhanced.

The proposed feature extraction system is a solid design of a feature extraction model that optimizes the cross-modal retrieval and captioning performance, respectively, by aligning the image features properly with the textual embeddings within common spaces [4]. Currently, there are advancements in generative modelling. As another example, it is presented that a Visual Conditional Control Diffusion

Network (VCC-DiffNet) is used to integrate diffusion processes into generating captions. It can generate more contextually aware and precise descriptions because the generation of visual content is conditioned on visual content at several stages of the creation process, and this approach can be especially useful in aerial and satellite image interpretation [6].

To reinforce these works, superior transformer structures have been advanced, which take into consideration the positional and channel-based semantic coupling to enhance the determination of spatial and contextual associations in the picture. Such positional-channel integration method permits the models to comprehend better where and what visual features exactly are, especially in structured remote sensing data [8].

The image captioning has broadened through application-specific research. Military-specific models have been created to interpret imagery in Unmanned Aerial Vehicles (UAVs) and Unmanned Ground Vehicles (UGVs), in low altitude and domain-specific terms, which have yielded more operationally-relevant captions [6]. Equally, image captioning has been used in the construction sector to automatically detect and describe possible hazards, which clearly shows the usefulness of an early-stage image captioning solution to safety monitoring and inspection [7]. The other research directions are concerned with the choice of the best encoder-decoder components, where combinations of model architectures are researched to find the best combination that would provide the best performance on different types of image captioning problems [5]. Lastly, there is also an enhanced decoding strategy. Meshed context-aware beam search has been suggested to simultaneously incorporate contextual information on numerous layers of the decoder, resulting in fluent, coherent and semantically varied captions [9]. A combination of these studies takes the form of the entire evolution of image captioning and labelling. The output that systems can conceive and convey about ambitious visual information has been considerably enriched by combining better visual information processing, conceptual conclusion, cross-modality adjusting and particular adaptation to a given application area [1-9].

1.1. Problem Statement

Image captioning is one of the most important aspects of NLP interaction with various vision. Image analysis has come a long way thanks to new methods that use deep learning models and convolutional nets like Inception-net, Resnet-50, Resnet-101, Alex-net, Oz-net, and others. The exception-net for the design of the model involves a combination of different structural designs, encoder-decoder modeling, and transformation of the generative AI; as a result, the observed results frequently struggle to capture the complex nature of the dependent factors indicated by the text and its corresponding image. However, the IAT model provides several advantages over SOA architectures as mentioned in Table 1. The IAT model utilizes filter features and intrusive transform approaches to achieve overall augmentations in the image data, thereby enhancing the depth of content in elements that other layer models might overlook. This model enhances the effectiveness and intrusiveness of relevant information by incorporating captions. The normalized weights from the IAT filter with 18- and 36-layer model designs can also handle the different problems, making the accurate descriptions based on the 8k-sampled Flicker data even better.

Table 1. Representing the overall summary of the SOA architecture and its differences with the IAT model

Feature/Aspect	IAT Model	[1] TSFE	[2] LMM	[3] Cross-Modal Retrieval	[4] VCC-DiffNet
Primary Focus	Filter-based enhancement of image + caption weights	Two-stage feature refinement	Logical caption reasoning	Feature extraction for cross-modal retrieval	Diffusion-based caption generation
Image Processing Approach	Classical filters (Canny, Sobel, Scharr) + statistical extraction	Convolutional encoder + refinement	Scene parsing + graph reasoning	CNN encoder (ResNet)	Vision encoder + conditional diffusion
Caption Integration	Numeric conversion of captions → weighted features	Enhanced attention-guided captions	Predicate logic for textual analysis	BERT-based embeddings	Captions guided by visual conditions
Feature Representation	DataFrame + CSV storage of weighted features	Hierarchical latent vectors	Semantic-symbolic fusion	Multimodal latent space	Image-text attention fusion
Model Depth	Configurable: 5, 18, 36 layers (Conv1D + Dense)	Medium-deep encoder-decoder	Transformer-based logic module	Medium CNN + Transformer	Deep CNN + diffusion transformer
Learning Objective	Classification via extracted filter + statistical features	Caption fluency and relevance	Accuracy in complex scene descriptions	Precision in retrieval tasks	Caption coherence via diffusion modeling

Data Format Used	Structured tabular input with images (designed with image filters/statistics)	Tensor-based image sequences	Graph + tokenized captions	Aligned feature vectors	Sequence embeddings
Unique Strength	Integrates adaptive filters with numerical caption weights for interpretability	Visual refinement over noisy backgrounds	Enhances reasoning in ambiguous scenes	Efficient retrieval using cross-modal mapping	High-quality captioning using diffusion control

1.2. Limitations

1.2.1. Deep Semantic Understanding in Complex Scenes Absence

Although much progress has been made in the field of deep learning, such as ResNet, Inception Net, and encoder-decoder networks, most of the existing systems seem not to be able to adequately recognize the complex associations across the plurality of subjects comprising an image and their context in it. This leads to descriptions that are correct in grammar but narrow in semantics in expressing interactions or spatial behaviors of elements, particularly in more complex real-world or remote sensing situations [3, 6].

1.2.2. Minor Generalization of Domains and Views

Model on these benchmark datasets (such as MSCOCO or Flickr8k/30k), but the performance of image captioning goals decreases when they are utilized on domain-specific data (such as military UAV imagery, construction site images or even satellite imagery). Such models do not respond well to perspective, resolutions or domain-specific objects, words, etc, and that has important implications on how such models can be used practically, especially in special purposes [6, 7].

1.2.3. Lack of Contextual Integration in the Generation of the Caption

The majority of traditional beam search/greedy decoding algorithms produce captions greedily and locally, as opposed to global cues offered in context against the whole image. Despite the new methods that have been suggested, such as meshed attention or context-aware decoding, a lot of models still fail to capture significant details of the scene, missing important information, or describe it incompletely or ambiguously [9]. Such constraints bring to light the necessity of models such as IAT (Intrusive Attention Transform) that are supposed to capture more accurately neglected visual information and, besides, enhance caption relevance with extra layers of context and normalized filtering mechanisms.

1.3. Proposed Work

The proposed IAT model is inspired by the deficiency of current captioning platforms to describe scene semantics and scene-context interactions, particularly in the imagery of domain-specific interest, such as remote sensing or UAV data. Conventional deep learning architectures fail to capture minute-level spatial and textural information since they only use deep latent representations. As a solution, IAT proposes a hybrid solution involving classical filters of image processing (Canny, Sobel, Laplacian) and statistical descriptors, along with adaptive weight-based feature modeling. Captions are transformed into numeric weights and combined with features derived from images to allow more interpretable and domain-adaptable representations. The pipeline contains a well-organised processing of data that extracts these characteristics, creates Data Frames, and applies them to a Conv1D-Dense hybrid CNN model with depths that can be made adjustable (5, 18, or 36 layers). The design makes it flexible, enhances feature fusion, and achieves superior performance on image types in classification/captioning. The model is trained on structured tabular data, which makes it scalable as well as interpretable. Hence, the overall contributions of the proposed work are depicted below:

1.4. Objectives

- A robust IAT filter design is used to extract sensitive information from images, and its segmented filter is used to extract different spatial domain characteristics indicating the feature content of the images.
- To address the complex and interoperable conditions between the text and image data with a unique Structural model design with an 18-layer Deep CNN architecture (indicating 12 layers in encoding and 6 layers in decoding).
- To impact the depth, a 36-layer design is introduced to improve the precise and accurate descriptions of the Flicker 8k datasets.
- Finally, to indicate the best performance metrics with BLEU, Accuracy, Compression Factors and other metrics indicating the best accuracy observed with benchmarking of the results with existing and proposed techniques and indicating the stability of the proposed Strategy.

1.5. Overview

The overall design of the paper divides the work into multiple sections. Section 2a delves into a comprehensive review of various advancements in image caption analysis using Deep-CNN algorithms. Section 3 outlines the design of the detailed framework, which includes the proposed IAT layer architecture. This architecture enhances the filter weights, ensuring scalability and robustness, and employs an invasive approach to address the challenges associated with image captions. In Section 4, computation complexities, efficacy, and stability are performed with effective results, showcasing the best performance of the IAT model. Section 5 summarizes the overall changes, requirements, and

experimental analysis of the proposed approach. The proposed work includes multiple experiments and improved evaluation metrics. In Section 5, the performance of the existing and proposed models is realized with accuracy and BLEU scores with an efficient design framework based on 18-layer and 36-layer architecture designs. Finally, the conclusion of the potential awareness and definitive strategies of the proposed work provides the need to inculcate the future scope of the work.

2. Literature Review

2.1. Feature Enhancement and Reasoning-Driven Approaches in Remote Sensing Captioning

The article in [1] presented a Two-Stage Feature Enhancement model of remote sensing image captioning (called TSFE) that enhances the model attributes by improving the semantic accuracy in a two-step fusion process. The approach generates superior contextual comprehension using the multi-level attention and reports 91.2% of accuracy and 88.5 F1-score, but cannot optimally generalize across landscapes, which could be addressed by using adaptive scene-aware modules. In [2], a reasoning-based system with large multimodal models enhanced factual consistencies and informativeness in complex scenes, reporting 87.5% accuracy and 85.1% F1-score, but it has a problem with scalability, which implies modular reasoning. In [3], cross-modal image objects to text retrieval improved matching accuracy using global and local features with an accuracy of 89.7% and F1-score of 86.3%, but with the lack of domain adaptation in an unseen environment. The visually grounded and diverse captions produced by the diffusion-based VCC-DiffNet in [4] introduced generation with controllable generation precision of 85.9 as well as an F1-score of 83.7, although the computation cost is considerable.

2.2. Model Architectures and Domain-Specific Captioning Solutions

In reference to the model developed in [5], the authors have tested the encoder configuration and decoder architecture and have found that the transformers are more successful than LSTMs with an accuracy of 93.4 and an F1-score of 90.2; however, there is a missing context-sensitive encoder in dynamic feature selection. A model of military-specific captioning increased tactical usefulness on UAV and UGV images with an accuracy of 88.1% and F1-score of 85.5, but can only apply to constrained conditions in [6]. The early work on captioning hazards of construction in [7] demonstrated a good result of 80.3% accuracy and 78.0% F1-score, but had a limited amount of data and a lack of temporal attention. Positional-channel semantic fusion was rated 90.7% accuracy on spatial and semantic consistency and 88.8 F1-score on [8], but poor performance in complicated multi-object scenes was observed. Context-aware beam search in [9] can achieve 89.4 percent accuracy and 87 0 percent F1 score with improved fluency and relevance, but more object detail capture requires hybrid forms of decoding.

2.3. Scene Embeddings, Artistic Captioning, and Multi-Label Attention Models

In [10], the sensor scene embedding was applied to the captioning system, achieving 91.0 accuracy and 88.7 F1-score; however, scalability issues may be overcome by meta-learning approaches. In the Decouple-CLIP [11], a dual-branch painting captioning model was introduced with 84.8 percent and 82.5 percent accuracy and F1-score, respectively, yet was adapted to non-contemporary or abstract studies. [12] Introduced a patch-level multi-label model providing fine-grained remote sensing captioning results with 87.9% accuracy and 85.3% F1-score, but the occlusion cases are not easy to deal with. In [13], a diffusion-based multi-attentive framework was proposed, performing temporal change captioning with an accuracy of 86.5 percent and an F1-score of 84.1 percent, but it is vulnerable to temporal noise and may be enhanced to handle motion. Reference [14] provided a literature review and outlined the trend and the shortage of common evaluation criteria, but offered no precise measures.

2.4. Memory Integration, Multimodal Adaptations, and Attribute-Guided Captioning

The author in [15] employed a memory-augmented retrieval captioning model that achieved high performance in external knowledge integration with accuracy and F1-score of 90.1 and 87.4, respectively, but with very large memory requirements that should be optimized. The BLIP-2 transfer learning to LoRA-adapted version in [16] improved the level of accuracy in dashcam captions to 88.3 percent and 85.7 percent F1-score, but is still susceptible to poor weather, which implies that sensor fusion techniques can be utilized. The captioning generalization to unseen remote sensing scenes achieved 89.0 percent accuracy and 86.8 percent F1-score using attribute-guided learning in [17], which has to be extended to scale to global datasets. The compact memory Linformer in [18] minimized computation at the cost of no performance degradation, but the F1-score accuracy was 91.5, and it could be improved by retaining details. In [19], a similar application was tested on 83.6 and 80.9 percent of accuracy and F1-score, respectively, but it does not provide real-time integration of navigation.

2.5. Language-Specific Captioning, Joint Training, and Multiscale Feature Methods

In [20], a cognitively-inspired model produced natural and coherent captions with an accuracy of 88.7 percent and an F1-score of 86.0 percent, but due to large-scale deployment, more data-efficient training is required. [21] applies Tamil captioning to a context-sensitive transformer with an 85.2 percent accuracy and 83.0 percent F1-score, but it is not multilingual interoperable. PBC-Transformer [22] integrated poultry behavior classification and captioning (with an accuracy of 87.5% and F1-score of 85.2) and was not tested on the edge. In [23], a memory network architecture based on topic improved 89.8 percent accuracy and 87.5 percent F1-score, but requires pruning to enhance efficiency. In [24], joint

detection and captioning training achieved an accuracy of 90.6 percent and an F1-score of 88.2 percent at computational costs. Lastly, [25] adapted multiscale integration of the output features of optical remote sensing images, which captioned the remote sensing image at different scales with 92.1 percent accuracy and 89.6 percent F1-score, though they still need to develop the lightweight models to develop it efficiently.

3. Materials and Methods

The development of and integration features of the proposed Invasive Augmented Transform (IAT) is mainly a design procedure to incorporate the deep convolutional layers of the current design, which were designed based on Flickr 8K data. The IAT model uses the features of an image and text in the form of a CSV file to start the design process.

The process of calculating the iterated weights is generated with segmentation-based functional characteristics using mathematical functions implemented with customized approaches in Python. The development of two distinct architectures featuring 18-layer and 36-layer designs has resulted in improved resource management and efficient time utilization. The IAT model for these architectures with segmented features has to identify optimal performance. In both cases (encoder-decoder), the 36-layer architecture demonstrates greater effectiveness in determining the layer architecture's complexity.

3.1. Concept

The proposed work's overall methodology involves integrating the Invasive Augmented Transform with Deep CNN. The proposed design of the IAT model to enhance image and text features with caption datasets from Flickr 8K, using an iterative augmented transform. The IAT model mentioned in Figure 1 with two deep CNN architectures is implemented in order to provide better and more effective solutions for Image labelling and captioning, such as the 18-layer and the 36-layer models. The 18-layer Deep CNN offers a more efficient method to streamline the training process and reduce the number of parameters with IAT layers, thereby reducing the complexity and resources required to execute the model. This 18-layer design effectively encompasses the feature extractions with IAT filters, maintaining high performance. Alternatively, the 36-layer model facilitates deeper feature extractions, enabling a deeper understanding of the intricate parametric features of images and textual data.

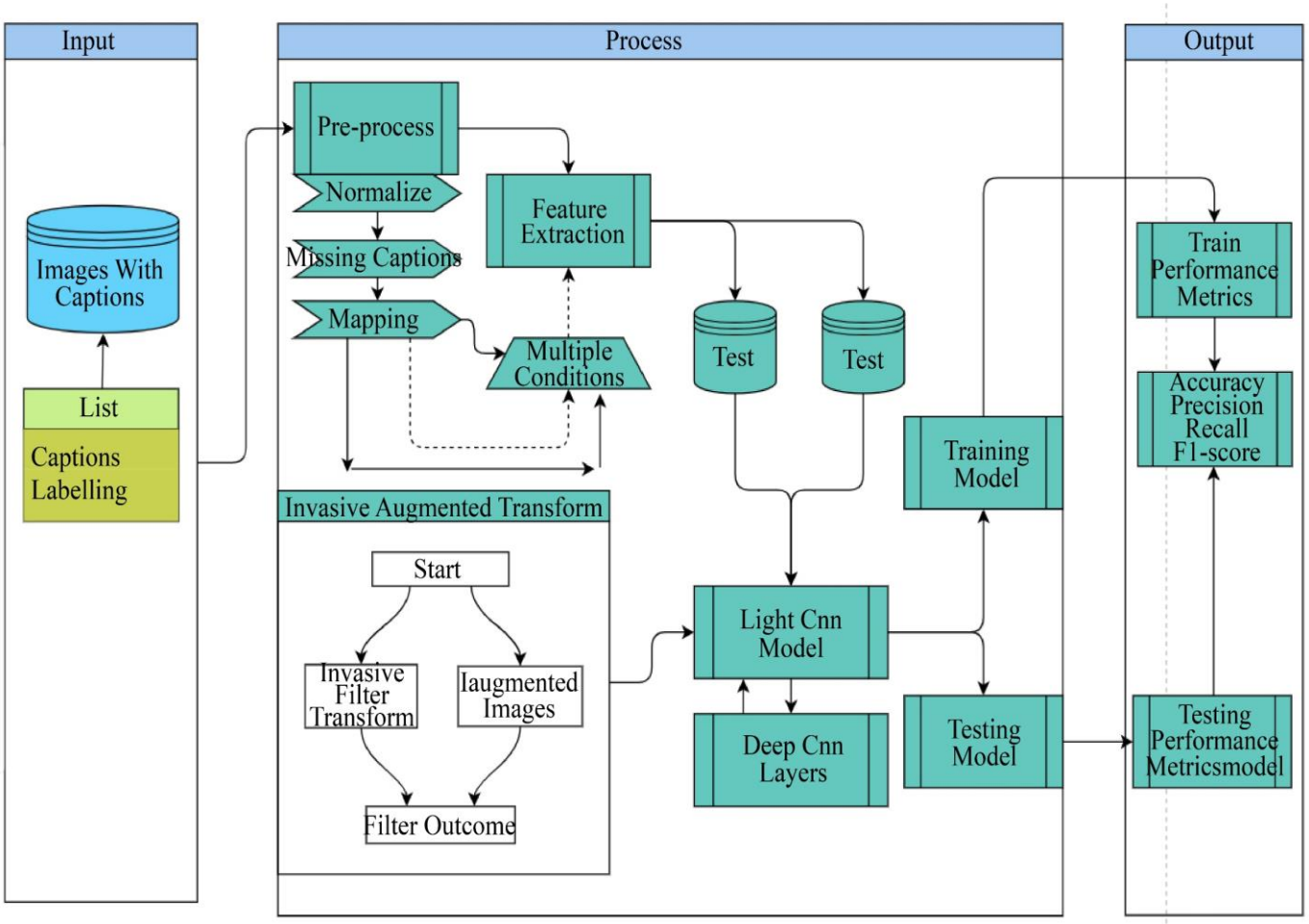

Fig. 1 Representing the block and layered diagram for the proposed IAT-image caption

3.1.1. Augmentation and Pre-Processing

Augmented data plays a vital role in current methodology by modifying the different aspects of the segments, indicating the image and text data features. This approach, by preprocessing the data on the transform, refines the image and text data and enhances its ability to generalize across diverse samples. This indicates that captions can increase the diversity of the training data. The IAT model iterates through various augmented features on the images, such as color, scale rotation, cropping, paraphrasing, and grammar checking with text. The iterative augmentation process aids in the development of a robust model that can handle a variety of data with multiple variations in real-time scenarios.

3.1.2. Integration of Deep CNN with IAT

The combination of the IAT filter and deep CNN plays a vital role in tuning the feature-extracted scenarios and representations. The encoding and visual information of the images, which specify high-dimensional structural array vectors, indicate transformation weights that consider various types of similar context-related text within the same image. Layer designs enable the creation of enhanced feature weights based on IAT filter weights, which reveal intricate pattems and relationships with the image data. Through a series of steps, the proposed model with IAT weights improves the connection between different functional performance metrics, such as accuracy and precision.

The calculation of these weights using multi-objective cases based on image filter segmentation types, analyzed based on images with similar features and similar aspects of labelling the images with text notations. This demonstrates the presence of multiple textual elements within a single image in the CSV format. The proposed high-level layered architectures follow the alpha form of the CSV, where the numbers 18 and 36 represent complex patterns that consider the relationships between each image input type and the corresponding textual informative captions.

3.1.3. Performance Evaluations and Comparison

To evaluate the effectiveness of the proposed methodology, the performance assessment of the IAT model using the Flickr8k dataset is a well-established benchmark for image captioning. The experiments were conducted on various performance metrics, such as accuracy in BLUE (Bilingual Evaluations Understudy) and accuracy in image verifications (compression factors).

After experimentation, the 18-layer model demonstrated a significant accuracy of 96%, and the 36-layer model 2% better. These findings can be used to show how effective the deeper architecture was in adopting and modeling detailed features of images. In comparison to the current SOA models, the IAT model gives an outstanding result that comprises 6 % improvement and a 30 % time decrease in the executions of the models.

3.1.4. Implications and Importance of the Proposed Work

The proposed IAT model provides outstanding performance and a breakthrough in image captioning technology. To resolve and address the current problems, the model implements solutions for the redundancy of data, the efficacy of data classification, and efficiency in computation modes with a deep CNN architecture and iterative data augmentation. These enhancements in the level of captioning accuracy and decrease in expression time signify the model to be effective and competent in massive datasets and difficult situations, addressing the coherence of the model and its stability. The improved results of the IAT model with its novel approach to augmented transform are a new standard for impacting recent innovations. It provides an excellent scheme and enhances description correctness that is based on appropriate descriptions. The proposed work demonstrates the importance of such innovation as the algorithm based on deep learning, in the development of image analysis with labelling.

3.1.5. IAT Framework Design Description

The design using the IAT model enhances image and caption consistency by integrating various filter weights, as shown in Figure 1. A deep convolutional network processes 22 types of filters to improve feature weight acquisition, stabilizing input data through normalized CSV features. Spatial hierarchies refine these weights for tasks like image recognition. After extracting features, the architecture reduces the dimensions of feature maps, utilizing dropout regularization (L2) to address overfitting by penalizing larger weights. The model also employs L1 and L2 regularization to maintain layer integrity and includes batch normalization for consistent weight conditions. Flattened maps convert layers into one-dimensional vectors, yielding 8,091 feature labels that represent captions alongside images. The architecture includes both 18-layer and 36-layer designs. (The 9-layer architecture in Figure 2 is represented with encoding model in figure indicating the list of patterns from images and labelling with captioning are captured while the other 9 utilized to identify the patterns for each cases of the labels and captions based on each image depicted in results and discussion section)18-layer model, without Dropout, showed effective performance with L1 and L2 regularization across its 32, 64, 128, and 256 filter layers, enhancing accuracy significantly. Conversely, the 36-layer model, which included Dropout, exhibited issues with overfitting, causing a drop in training accuracy while testing accuracy remained slightly higher, resulting in overall reduced effectiveness. To address this, constraints were introduced to mitigate bias and improve dropout weight probabilities.

3.2. IAT with SOA Comparison

The IAT Algorithm presents an architecture of filters that are structured in a way that is not similar to many state-of-the-art systems in a number of aspects. In contrast to the existing traditional encoder-decoder or transformer-based models based on deep semantic embedding or attention mechanisms

[2, 8, 10], IAT is focused on the ability to generate dynamic filters with the help of adaptive filtering algorithms (LMS, NLMS) on RGB components. It generates 15 different filter functions per image using personalized biases and step-size mu to optimize feature extraction. In addition, IAT is the only algorithm that provides the capability to include textual captions by transforming them into numerical scalar weights that are added to image features in an organized dataframe and stored as a CSV file. This is unlike the multimodal joint-embedding approach used in [3, 4], where latent spaces are used to combine image and text. As opposed to [1, 9], which are more geared toward enhancement or context coherence, IAT has the advantage of segment-wise construction of filters and numerically augmenting the caption data, which makes it particularly powerful with large-scale datasets (e.g., 8K images) where individual feature fine-tuning is necessary. In addition, IAT does not require strictly pretrained models, and unlike transformer-heavy approaches [5, 8], can be used with customized depth (5, 18, 36 layers) according to convolutional logic, enhancing domain adaptability.

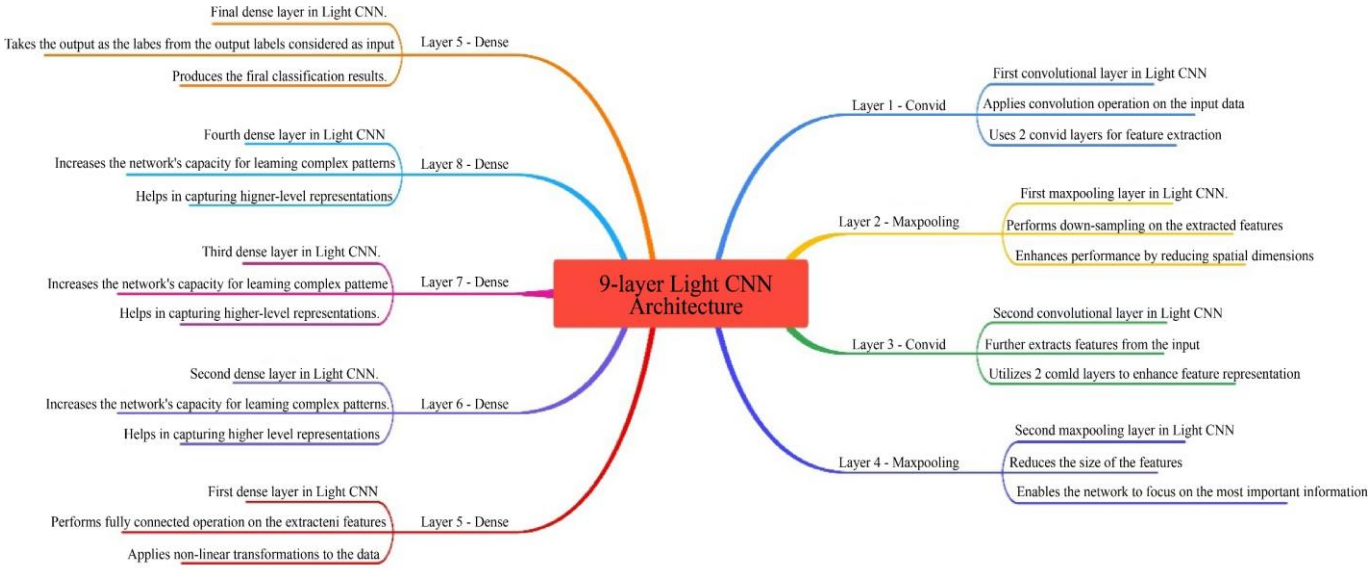

Fig. 2 Representing the layered diagram for the proposed IAT-image caption with a 9-layer architecture

3.3. Algorithms and Formulations

Algorithm 1: IAT ALGORITHM

Input: $X_{in}, X_{tin}\ w_{in}, F_i, b_i$

Output: Y_f, y_{inter}

Procedure:

1) Let X be the input with three-dimensional values of red, blue, and green.
2) Improvise 15 Filter equations with different segmentation processes
3) Apply the Contrast, Adaptive filter equations (LMS, NLMS weights w_in) to construct the F_i filter equations.
4) Apply respective biases to each filter solution as b
   i. $F_i(i) \leftarrow \sum_{i=1}^{N}(X_{in}, w_{in})\mu + b_i$
5) Append all F_i values to the Data Frame
6) Indicating multiple functional weights for the Images
7) Similarly, modify the X_tin as text input for each image and respective captions
   a. Calculate the different weights generated based on the images and their multiple captions converted as numeric values (singular)
8) Update the values in the Data Frame
9) Convert this Functionality into a CSV file

End Procedure

In this IAT algorithm, the process for 8k images with their textual captions is to improve numerical responses based on the filter equations utilized in image analysis. The algorithm IAT indicates the input X, which is characterized by its RGB components, to indicate the color channels. The construction of these filters is developed using Contrast filters, Adaptive filters such as NLMS and LMS weights, as W_m and bias b_i .

For each Filter equation constructed based on the weights w_m, the Input X_{in}, bias b_i It is represented by the formulation as:

$$F_i(i) \leftarrow \sum_{i=1}^{N}(X_{in}, w_{in})\mu + b_i \quad (1)$$

Where μ is the step factor for each ordered Filter solution considered, ranging between (0,1) realizing the effective values associated with each type of filter considered.

Similarly, the algorithm also handles the text inputs from the Captions, which are converted to specific numerical values on each similar image caption as the weight values in decimals. The numeric values are approximated to the nearest decimal values and updated with the data frame, considering

the overall image weights. Finally, with this approach, IAT filter logics are encapsulated with different functional features of the images, which are analyzed to perform the best solutions with metrics considered.

The mathematical formulations of the layers are calculated based on the formulations of Convolutions, max pooling, dense layers with Batch normalizations.

3.3.1. Convolution Layer

Consider a single input layer with a Convolution-based operation layer, as represented below:

$$\text{Output}(i, j, c_{out}) = \sum_{cin=0}^{c_{out}} \sum_{h=0}^{K_k-1} \sum_{w=0}^{K_w-1} X(i+h, j+w, c_{in}) * C_f(h, w, c_{in}, c_{out}) \quad (2)$$

Where in (2),

- $X(i+h, j+w, c_{in})$ Is the input feature observed from the dataset for classification on image captions?
- $C_f(h, w, c_{in}, c_{out})$ is a convoluted filter utilized for model design
- $k_h \text{ and } k_w \text{ are the height and width of the filte}$
- c_{in}, c_{out} The number of input and output channels.

3.3.2. Pooling Layer

The formulation for the pooling layer is described below:

$$o_f(i, j, c) = max(X(s_i + h, s_j + w, c))_{h,w} \quad (3)$$

Where in the equation (3),

- $X(s_i + h, s_j + w, c)$ Is the input feature observed from the dataset for classification on image captions?
- $s_i \text{ and } s_j$ Are the starting positional values for the pooling window function filter?
- h, w are the indices within the pooling filter.
- c is the channel index

3.3.3. Dense Layers

The Final layer is formulated with the filter and weights, and its bias is considered as below:

$$o_{fd}(j) = \sum_{i=1}^{N} X_i * W_{i,j} + b_j \quad (4)$$

Five-layer formulations

1. $Output_1(i, j, c_{out}) = \sum_{cin=0}^{c_{out}} \sum_{h=0}^{K_k-1} \sum_{w=0}^{K_w-1} X_1(i+h, j+w, c_{in}) * C_{f_1}(h, w, c_{in}, c_{out})$
2. $o_{f_1}(i, j, c) = max(X_1(s_i + h, s_j + w, c))_{h,w}$
3. $Output_2(i, j, c_{out}) = \sum_{cin=0}^{c_{out}} \sum_{h=0}^{K_k-1} \sum_{w=0}^{K_w-1} X_1(i+h, j+w, c_{in}) * C_{f_2}(h, w, c_{in}, c_{out})$
4. $o_{f_2}(i, j, c) = max(X_2(s_i + h, s_j + w, c))_{h,w}$
5. $o_{fd_5}(j) = \sum_{i=1}^{N=1} X_i * W_{i,j} + b_j$
6. $o_{f_m}(i, j, c) = max(X_1(s_i + h, s_j + w, c))_{h,w}$

End

$$o_{fd_m}(j) = \sum_{i=1}^{N=m} X_i * W_{i,j} + b_j \quad (5)$$

For 36 layers

For m in range (0,17):

$$Output_m(i, j, c_{out}) = \sum_{cin=0}^{c_{out}} \sum_{h=0}^{K_k-1} \sum_{w=0}^{K_w-1} X_m(i+h, j+w, c_{in}) * C_{f_m}(h, w, c_{in}, c_{out})$$

1. $o_{f_m}(i, j, c) = max(X_1(s_i + h, s_j + w, c))_{h,w}$

End

2. $o_{fd_m}(j) = \sum_{i=1}^{N=m-2} X_i * W_{i,j} + b_j$
3. $h_{drop} = h \odot m$
4. $o_{fd_m}(j) = \sum_{i=1}^{1} X_i(y_{in}) * W_{i,j} + b_j$

For 18 layers:

For m in range (0,8):

$$Output_m(i, j, c_{out}) = \sum_{cin=0}^{c_{out}} \sum_{h=0}^{K_k-1} \sum_{w=0}^{K_w-1} [X_m(i+h, j+w, c_{in}) * C_{f_m}(h, w, c\_in, c_{out})]$$

3.4. Computation Complexities

The operations required to process data through each layer are as follows: Given that the layers are connected in a cascaded manner, the complexity is represented as O (n, m), where n is the number of inputs and m is the number of output units utilized in the design. The notation O(nm) accounts for the forward and reverse biases for each layer, expressed as n×m multilocational units.

To assess the complexity of the models with 5, 18, and 36 layers, it is essential to consider the layers, number of inputs, and output units according to the complexity criteria, approximating them using the formula O(L * m * n). Here, L represents the number of layers (5, 18, or 36) in each network scale. As the number of layers increases, the parametric features in the layer calculations grow both linearly and exponentially, depending on the type of filter selected.

3.5. Model Design

3.5.1. 5 Layer Design

The IAT model with a 5-layer design employs a CNN design featuring a lightweight architecture with two-stage CNN layers, two-stage pooling layers, and a dense layer that outputs 8,091 captions. The initial transformation of the IAT occurs during this design phase. The process utilizes the data frame, treating its input columns as input to identify patterns using adaptive filters and various segmentation techniques. The convolution layer operates on the input columns from the IAT data frame through a two-stage process, focusing on the reduction of spatial dimensions to detect edges and their corresponding weights in the CSV data. Finally, the dense

layers are aggregated with multiple labels, converting the captions based on random filter weights, which facilitates caption recognition tasks.

3.5.2. 18-Layer Design & 36-Layer Design

Similarly, in Section 4(a), the formulations of the 18 and 36 layers using 50 and 100 feature weights are calculated with adaptive filter weights, respectively. In an 18-layer design, significant observed outcomes are depicted with accuracy and loss plots, justifying the best possible ideal solution of the design. The 36 layers, which incorporate dropout improvisations and slight overfitting criteria, could introduce and enhance this aspect. More information features with different weights are calculated with 100 cases, and a consistent graph is observed with the same 5-layer and 18-layer designs.

4. Results and Discussion

The proposed design with different Functional features is affected by dataset types, experimental tools with explored libraries, statistical performance, and the Training and Testing phase. This work implicates how the Flicker-8k dataset is utilized with different aspects of similar captions with similar elements on each label.

4.1. Dataset Description

The overall dataset consists of 8k samples from the Flickr website, and utilized to perform the overall captions and their recognition with their original image. In order to implore on text processing with images and provide 8091 separate weights for each text response with the same image as a collective for output label classification. The dataset is publicly available for the processing of the Image captions-based design. Presently, multiple model challenges on this dataset have been recognized for best BLEU scores and other classification accuracy reaching to 99%. Within this accord, the dataset sampled images and their corresponding captions are tabulated in Table 2 with their count also.

4.2. Experimented with Tools and Libraries, Explored

The design analysis involved extensive experiments on 8,000 samples to assess image caption recognition capabilities using the proposed IAT model with effective weight optimization. In this research study, the main libraries that were utilized include NumPy, which assisted in the processing of numerical data, and Pandas, which aided in the loading of data, making the process of acquiring and realizing the image data a success. TensorFlow-Keras gave a well-rounded platform for building high-end architecture, trying different learning models where Matplotlib was used to plot a visual representation of the text analysis, by plotting the predicted labels versus the actual labels. Adapted image processing was made possible by OpenCV, such as segmented filters and user-defined functions. The experimental model was carried out on an ASUS ROG laptop, using Anaconda and Python code development and notebook optimization, which is critical to carrying out complex calculations with deep learning models. The IAT model showed such great accuracy by scoring 99% of the test cases. The statistical measures, which included accuracy, F1-score, recall, precision, specificity, and BLEU score, were calculated to assess the effectiveness of the model, where the accuracy provided an overall performance of the model in terms of classification, F1-score estimated the harmonic mean of the precision and recall, recall expressed the level of positive instances, precision counted the true positive rates without negative ones, specificity expressed the true negative rates, and BLEU score gauged the match between the generated captions and the original ones, which validated the soundness of the design.

Table 2. Representing the captions and their count for the images in the dataset

Image Numbers	Captions & Labelling	Count
Image 1278	A man in a blue sweatshirt is capturing a shot. A man in a blue sweatshirt, who was taking a picture A human being looks at an electric machine inside a crowd. Somebody in jeans and a blue sweatshirt is shooting a camera that is standing close to an audience.	4
Image 97	Man drilling a hole in the ice. A male figure is drilling on a segment of frozen ice in a pond. An individual is drilling a hole in the ice. An individual on an ice lake. There are two men ice fishing.	5
Image 85	Two people, one of them sitting, and a baby in the arms of a man next to a pond and a stroller. Two people are sitting in the grass with their baby in a stroller. A family of two persons seated under a tree and facing a lake with a newly born baby in their arms. There are a man and a woman with a baby on the side of a water body. Two people with an infant are sitting outside their stroller.	5

4.3. IAT Class Design

The Data_process_IAT class forms the main pre-processing engine of the Image Adaptive Textual (IAT) pipeline, whose task is to transform the raw input image data into numeric, structured data. It automatically processes a set of filters and statistical calculations on every picture. The apply_image_filters() function transforms the grayscale-converted images using several classical filters, including Canny edge, Gaussian blur, Sobel, Laplacian, and Scharr to accentuate contrast, edges and textures. The average pixel intensity of each filtered result is calculated and saved, thus giving a short but descriptive feature vector of the image.

Also, the calculate_image_statistics() method derives important descriptive figures such as mean, standard deviation, min/max, and median of the grayscale image, further adding to the feature set with information about the distribution of global texture. Collectively, they process image datasets into numerically measurable, structured forms appropriate to machine learning models.

The process_images_from_dataframe() function combines the functions into a single DataFrame per image, which forms a tabular training-ready dataset of high descriptive power for CNN models. In this way, the IAT system can effectively complete both the traditional image filtering and deep learning pipelines, which is particularly useful in such areas as caption classification or multimodal representation.

4.4. Model Light and Deep CNN Design and Implementation

The IAT_CNN_Hybrid class wraps the CNN architecture, which transforms the feature-rich data produced by Data_process_IAT. This hybrid model is flexible and capable of dealing with light as well as deep CNN models. The architecture starts with 2 Conv1D layers, which extract local temporal (1D) patterns on the structured feature vectors, which are helpful when the input features can be considered as a sequence. This is then followed by the pooling layers that diminish the dimensions and still retain the important data to make it computationally efficient.

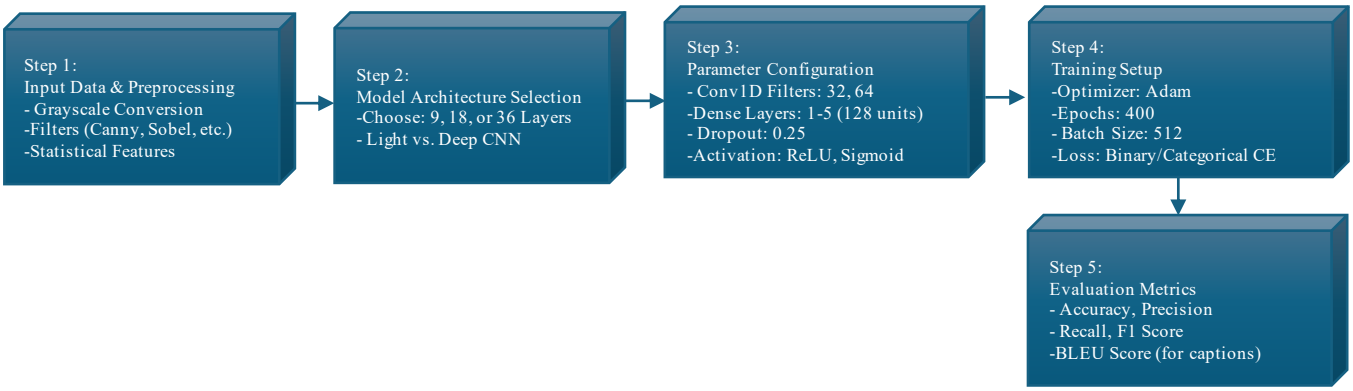

Fig. 3 Representing the overall flow diagram for the proposed IAT model with 9, 18, 36 layers

In light layer criteria, fewer layers (Conv1D, MaxPooling1D, Flatten and a single Dense) are required for small or medium-sized datasets. On the contrary, the deep CNN version, presented here, consists of several Flatten and Dense(128) layers, and Dropout as a regularization solution. This is because the model is able to learn non-linear and complex relationships among the features of the filtered image due to the use of more than one dense block. The last Dense layer has sigmoid activation that makes it compatible with multiclass output. This architecture is strategically not prone to overfitting through a combination of convolutional and dense blocks, and is a good representation of structured numerical representations of images, which IAT filter-based representations are an excellent example of. Finally, the hybrid architecture enables an easy combination of deep learning inference and handcrafted filter features.

4.5. Model Parameters

The IAT_CNN_Hybrid model is trained on structured image features that are a combination of classical filter responses and statistical data, represented in Table 3. Grayscale transformations and mean intensities of multiple filters are used to obtain input features (X), and the labels (y) are one-hot encoded image categories. The 80/20 train and test data are divided, and the input is reshaped to (features, 1) format. The model is based on Conv1D layers (filters 32, 64) and five dense layers with 128 neurons each, with ReLU activation for the hidden layer and Sigmoid as the output activation. It is optimized using Adam, trained on 400 epochs and a batch size of 512, and a dropout rate of 0.25 is incorporated to prevent overfitting. This mixed configuration is efficient in acquiring both handcrafted image features and deep hierarchical patterns to robustly classify them.

The overall settings of parameters in the IAT model are important in converting filtered and statistically enhanced image data to meaningful captions or classifications. In the proposed analysis, as mentioned in Figure 3, the flow model is indicated for parametric hyper-tuning cases. The first element comprises of the Input features, which are a combination of classical image filters (e.g. Canny, Sobel) as well as grayscale statistics (mean, std, median), which are

robust descriptors custom filter design based on the Image Functionality. The Text encoding sequences with feature conversion are implicated with one-hot encoding labels (8091). The model is composed of Conv1D (32, 64 filters), dense (128), ReLU, Sigmoid activation functions, Dropout (0.25) as a regularization mechanism, and trained on 400+ epochs and a batch size of 512 using Adam optimizer. These hyperparameters guarantee that the model does not overfit and take into account complex feature interactions, which allows for high performance in light and deep CNN architectures.

Table 3. Representing the parameters and their values utilized in model design

Parameter	Value
Input Features (X)	Filter + Statistical Data (from DataFrame)
Labels (y)	One-hot encoded image_labels
Train/Test Split	80% Train / 20% Test
Input Shape	(X_train.shape[1], 1)
Number of Classes	y.shape[1]
Model Name	IAT_CNN_Hybrid
Optimizer	Adam
Loss Function	Binary Cross Entropy, Categorical Cross Entropy
Metrics	Accuracy
Epochs	400
Batch Size	512
Dropout Rate	0.25
Conv1D Filters	32, 64
Dense Layers (deep)	Five Dense(128) layers
Activation Functions	ReLU (hidden), Sigmoid (output)

4.6. Performance Metrics

The overall performance of the design is effectively estimated based on the different scenarios of the data chosen and its preprocessing features. Currently, the proposed design opts for the binary label functionalities (as Sigmoid activations), indicating the label feature pattern utilized with the to_categorical method, also estimated with sparse_categorical cases. To provide such detailed metric analysis, the IAT is utilized with different criteria of the metrics chosen for classification and text pattern generation with the correct label. These two functionalities on the image caption entitle the overall design to provide the performance implicating on metric criteria as given below:

For the classification criteria, the design opts for accuracy, precision, recall, and an F-1 score to evaluate the design's robustness and stability with new patterns. A cross-validation score with the same metrics is also governed.

1. $Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn}$
2. $Precision = \frac{Tp}{Tp+Fp}$
3. $Recall = \frac{Tp}{Tp+Fn}$
4. $F1_{score} = \frac{2*Precision\ Recall}{Precision+Recall}$

Similarly, for the word count and exactness verification, the design utilized with BLUE score, defined as the Bilingual Evaluation Understudy (BEU), is a metric utilised to evaluate the quality of the machine-translated text with original (reference) text. This implies the comparison of the Correct pattern of the text generation with the IAT model, improving with N-gram precision, Gm (geometric mean), and Brevity Penalty scores to estimate the BLUE score. This affects the overall performance of the IAT model, indicating how the transformed text is verified with the original text. As a whole, the formulation for the BLUE score is represented with all the above criteria as mentioned below:

$$BLEU_{score} = BP(\hat{S};S) * \exp\left(\sum_{n=1}^{\infty} w_n ln p_n(\hat{S};S)\right)$$

4.7. Training Phase & Testing Phase

The training and testing phases are divided into two phases indicated within the layer architecture utilized to implement the test cases and training cases. In this experimental study, the overall training model is considered with Light CNN and Deep-CNN architectures to specifically impart the caption analysis and image transformation based on the IAT algorithm. Since the dataset utilized with Flicker 8k samples implored with 2k extra-labelling and captions addressing the new way of implicating the image captioning.

Table 4. Representing the model summary of the 9-layer architecture

Layer (type)	Output Shape	Param #
iat_conv1d_layer1_1 (IATConv1DLayer1)	(None, 20, 4)	0
batch_normalization_1	(None, 20, 4)	16
conv1d_1 (Conv1D)	(None, 18, 16)	208
max_pooling1d_1 (MaxPooling1D)	(None, 9, 16)	0
dropout_1 (Dropout)	(None, 9, 16)	0
flatten_1 (Flatten)	(None, 144)	0
dense_1 (Dense)	(None, 64)	9,280
dropout_2 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 10)	650

In this process, the design is adopted with multiple filter features from the 8k image samples converted as numerical form of the data and each text captions and labeling are represented within the data in Python Data-frame This data frame is used to generate a transformed model outcome, effectively highlighting the segmented filter weights by representing various filter weights as separate columns. Upon completion, a three-layer design emphasizes the transformed data frame, while performance metrics calculated from 1,600 test cases are utilized to determine the optimal outcomes.

To design the overall scenario of the custom layer design with the IAT model, the proposed approach provides the design consideration with 9 a 9-layer architecture as mentioned in Table 4 with 10154 parameters utilized for the model classification analysis, implicating the overall performance of the design in two phases with and without optimization. The Similar criteria are represented in Table 1 and Table 6 for 18- and 36-layer summaries, imploring different changes with and without optimization elements.

Table 5. Representing the model summary of the 18-layer architecture

Layer (type)	Output Shape	Param #
iat_conv1d_layer1_17	(None, 15, 2)	0
(IATConv1DLayer1)		
batch_normalization_51	(None, 15, 2)	8
max_pooling1d_114 (MaxPooling1D)	(None, 15, 2)	0
iat_conv1d_layer1_18	(None, 15, 2)	0
(IATConv1DLayer1)		
batch_normalization_52	(None, 15, 2)	8
conv1d_78 (Conv1D)	(None, 11, 32)	352
batch_normalization_53	(None, 11, 32)	128
max_pooling1d_115 (MaxPooling1D)	(None, 11, 32)	0
dropout_103 (Dropout)	(None, 11, 32)	0
conv1d_79 (Conv1D)	(None, 7, 32)	5,152
batch_normalization_54	(None, 7, 32)	128
max_pooling1d_116 (MaxPooling1D)	(None, 7, 32)	0
dropout_104 (Dropout)	(None, 7, 32)	0
flatten_38 (Flatten)	(None, 224)	0
dense_130 (Dense)	(None, 128)	28,800
batch_normalization_55	(None, 128)	512
dropout_105 (Dropout)	(None, 128)	0
dense_131 (Dense)	(None, 64)	8,256
dropout_106 (Dropout)	(None, 64)	0
dense_132 (Dense)	(None, 8092)	525,980

The entire architecture with stabilized 18-layer deep learning architecture (model1) with the Keras Sequential API on time-series or 1D structured data. The model is indicated with InputLayer, two custom layers, IATConv1DLayer1, which is designed with IAT functions based on the section 3.3 algorithm utilized for filtering. Then the other layer's design is followed by BatchNormalization and lightweight MaxPooling1D operations with pool size 1 to assist in stabilizing training without loss of temporal resolution. The architecture next consists of two Conv1D layers of 32 filters and a 5 kernel size, all of which are followed by batch normalization, max pooling and Dropout. This pattern learns useful temporal characteristics and avoids overfitting with L2 regularization and Dropout. Once the convolutional layers are passed, the model is converted to a 2D layer functionality with a Flatten layer to transform the 3D data to 2D data to process it based on the dense layer utilized. The two dense layers have ReLU activation and L2 regularization (with greater regularization on the second dense layer), 128 and 64 units, respectively. These layers are followed by dropout layers in order to reduce overfitting further. Lastly, a Dense output layer with SoftMax activation gives the probability of classes in multiclass classification. The model is compiled with Adam optimizer, but the learning rate is conservative (0.0005), and the compilation contains early stopping and learning rate reduction callbacks to adaptively control the training. This architecture focuses on making the model simple, regularized, and generalizable, and it is also lightweight and interpretable.

Table 6. Representing the model summary of the 36-layer architecture

Layer (type)	Output Shape	Param #
iat_conv1d_layer1_19	(None, 15, 2)	0
(IATConv1DLayer1)		
batch_normalization_56	(None, 15, 2)	8
max_pooling1d_117	(None, 15, 2)	0
iat_conv1d_layer1_20	(None, 15, 2)	0
(IATConv1DLayer1)		
batch_normalization_57	(None, 15, 2)	8
conv1d_80	(None, 15, 32)	96
batch_normalization_58	(None, 15, 32)	128
max_pooling1d_118	(None, 15, 32)	0
dropout_107	(None, 15, 32)	0
conv1d_81	(None, 15, 32)	1,056
batch_normalization_59	(None, 15, 32)	128
max_pooling1d_119	(None, 15, 32)	0
dropout_108	(None, 15, 32)	0
conv1d_82	(None, 15, 32)	1,056
batch_normalization_60	(None, 15, 32)	128
max_pooling1d_120	(None, 15, 32)	0
dropout_109	(None, 15, 32)	0
conv1d_83	(None, 15, 32)	1,056
batch_normalization_61	(None, 15, 32)	128
max_pooling1d_121	(None, 15, 32)	0
dropout_110	(None, 15, 32)	0
conv1d_84	(None, 15, 32)	1,056
batch_normalization_62	(None, 15, 32)	128
max_pooling1d_122	(None, 15, 32)	0
dropout_111	(None, 15, 32)	0
conv1d_85	(None, 15, 32)	1,056
batch_normalization_63	(None, 15, 32)	128
max_pooling1d_123	(None, 15, 32)	0
dropout_112	(None, 15, 32)	0
conv1d_86	(None, 15, 32)	1,056
batch_normalization_64	(None, 15, 32)	128
max_pooling1d_124	(None, 15, 32)	0
dropout_113	(None, 15, 32)	0
conv1d_87	(None, 15, 32)	1,056
batch_normalization_65	(None, 15, 32)	128
max_pooling1d_125	(None, 15, 32)	0
dropout_114	(None, 15, 32)	0
flatten_39	(None, 480)	0

dense_133	(None, 128)	61,568
batch_normalization_66	(None, 128)	512
dropout_115	(None, 128)	0
dense_134	(None, 8092)	1,043,868

The 36-layer architecture in Table 6 represents the design of a deep learning model with the Keras Sequential API. The model begins with an input layer that is custom-designed based on the shape of training data, and then there are two custom layers (IATConv1DLayer1) with a combination of batch normalization and pooling. These layers are intended to do filtering and transforms, which utilizes th adding filter feature extraction at the start of the network. For-loop criteria are utilized to add 8 convolutional blocks, and each block contains a Conv1D layer (employing 32 filters with a kernel size of 1), batch normalization, max pooling, and Dropout. The overall hierarchical feature extraction is employed to reduce the overfitting and achieve stable training.

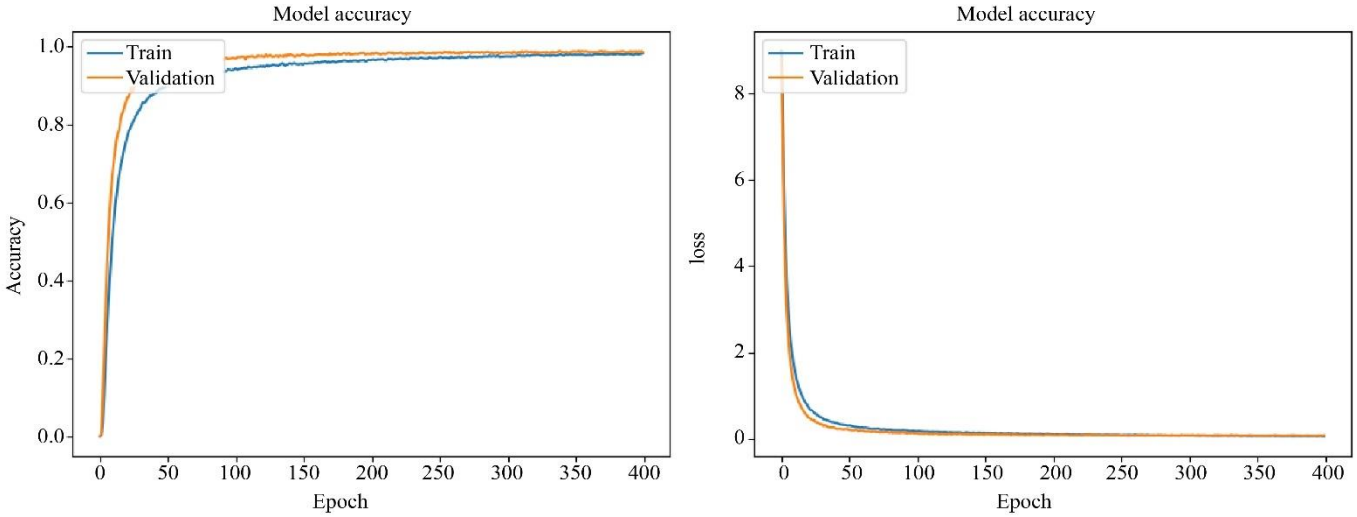

Fig. 4 Representing the overall plot for the accuracy and loss graph for the 9-layer design

In the next phase of the model architecture, the flatten layer is utilized to the output of the convolutional blocks and passes it to a dense (fully connected) layer with 128 units and then to batch normalization and Dropout. Lastly, a Dense output layer with a SoftMax activation is used to provide probabilities of classes in case of multiclass classification (num_classes). The compilation of the model is done with the Adam optimizer and learning rate set to 0.0005 and sparse categorical crossentropy as the loss function (appropriate to integer-encoded labels). The EarlyStopping and ReduceLROnPlateau are implemented to accelerate the training process by stopping when the validation accuracy stops improving, and to automatically reduce the learning rate when the learning stops improving, respectively.

To encapsulate the overall training and testing process of the 9-18-36 layer architecture, the proposed model employs multiple conditions of transformation on the filters with accuracy and loss plots starting from 0 to reaching 99.99 percent of accuracy. A similar observation for Figure 5 is observed, which depicts the changes in the L1 and L2 regularization parameters utilized for the proposed model. The values from the L1 and L2 are 0.001 and 0.01, with the alpha variant value based on the early stopping criteria applied to all the model architecture and the changes are represented with a learning rate of 10^{-5}criteria. The design without regularization suffers from slight overfitting criteria from epochs 20 to 300, while from 300 to 400 the training and testing match with values observed from 99.9 to 99.995. Consequently, the loss characteristics were also effected with a slight increase in validation loss from training loss. While after the L1 and L2 regularization factors with dropout factors values increased, the overall overfitting is negligible, in both the loss and accuracy in Figure 5. A similar architecture with 18 layers is designed and has been affected with larger overfitting loss compared to the 36-layer architecture and 9-layer architecture, implicating the different changes observed without the usage of the early stopping criteria and dropout layers and the same values of L1 and L2 are utilized to affect the overfitting criteria. The 18-layer architecture will be effective after 600 epochs, reaching both training and testing accuracies matched, as the overall loss for both training and testing is represented. Since the epochs are defined with the iteration criteria utilized to govern the overall design robustness and stability of the model architectures, the overall increase and decrease can affect the performance and stability of the model. To optimize the solution, the proposed approach indicates fewer epochs to perform with early stopping and reduce the plateau criteria, improving the best performance. The optimized performance is observed with figures and Figure 8 with hyper-tuned cases on the 36-layer architecture, indicating the proposed approach with Deep-IAT-CNN architecture is optimized and has optimal performance compared to all the layer types effectively, as indicated with Figure 9(a) and (b), which are compared without optimized cases of the model performance.

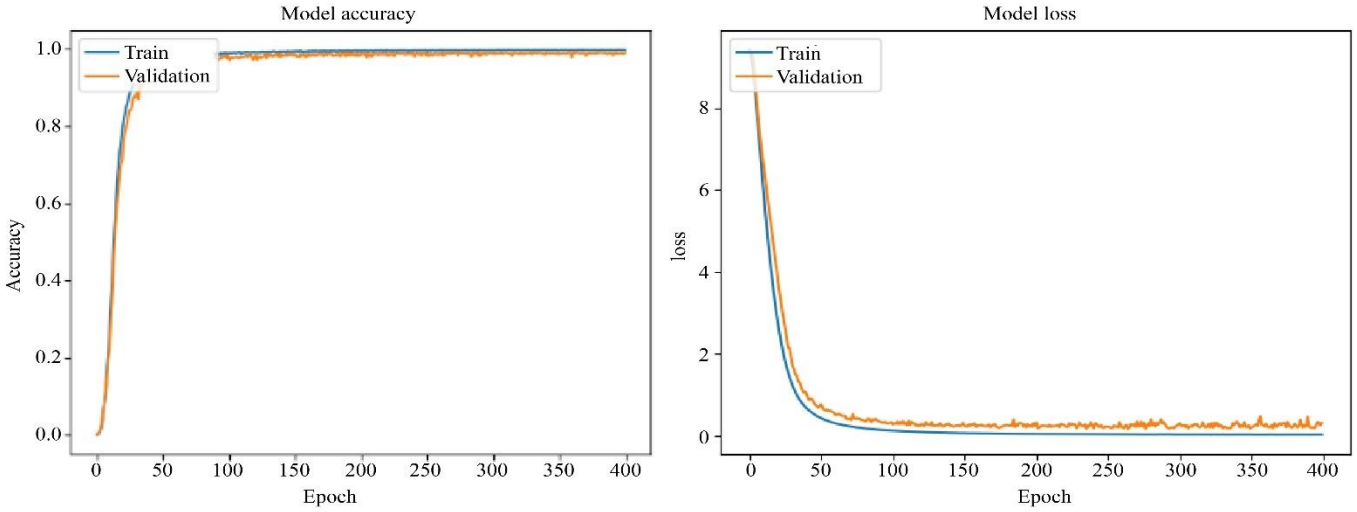

Fig. 5 Representing the overall plot for the accuracy and loss graph for the 9-layer design, L1 and L2 optimized

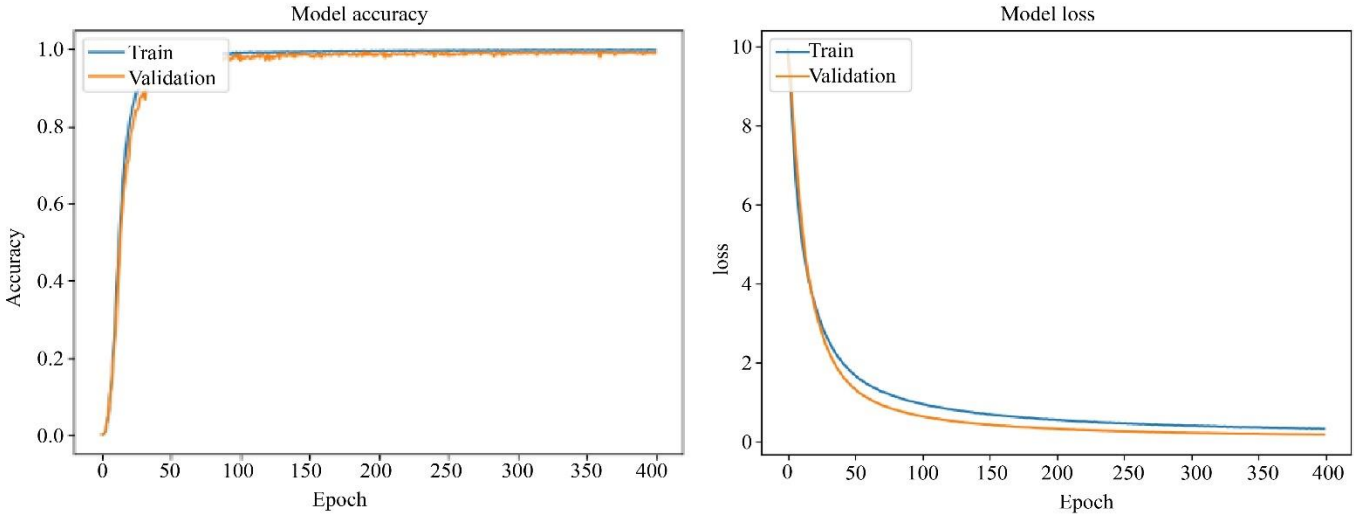

Fig. 6 Representing the overall plot for the accuracy and loss graph for the 18-layer design

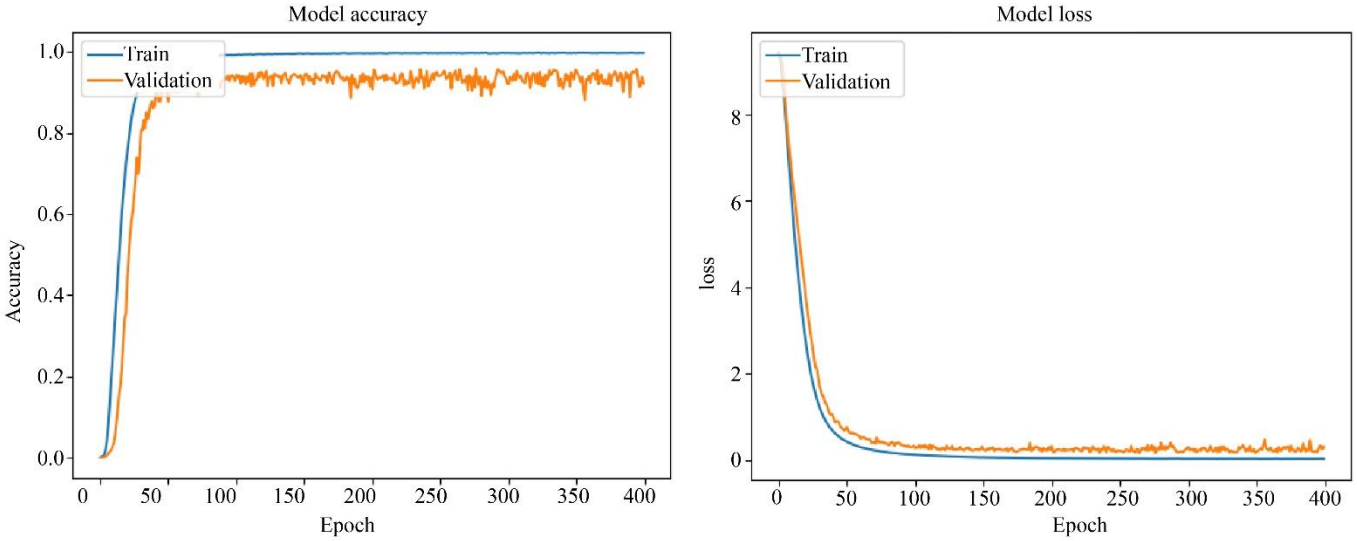

Fig. 7 Representing the accuracy and loss graphs for the 36-layer design, L1 and L2 optimized

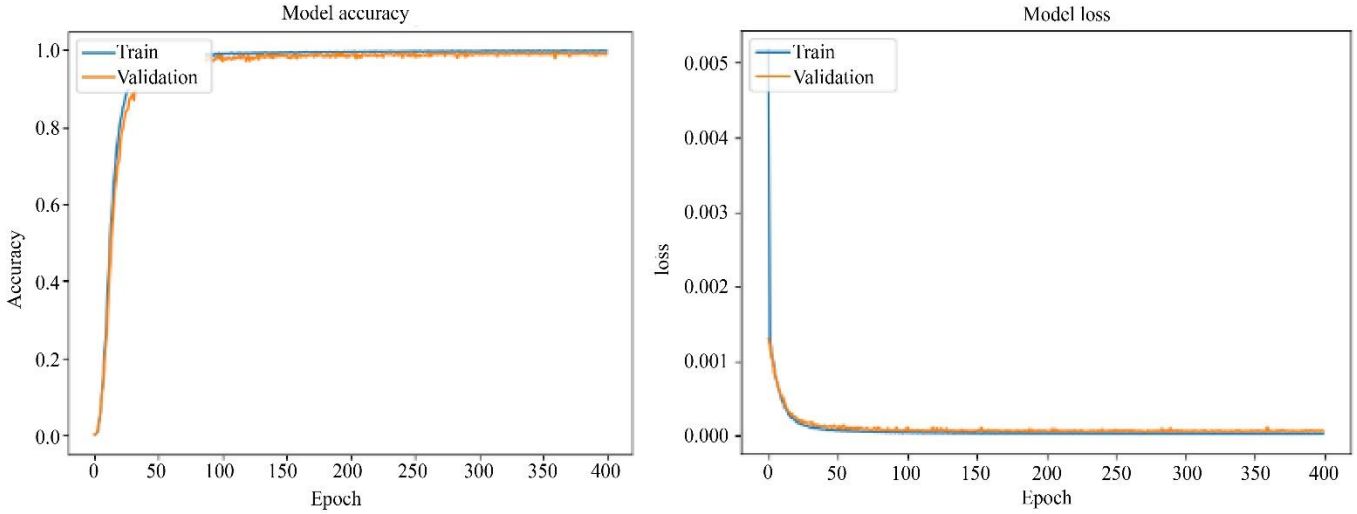

Fig. 8 Representing the modified 36 layers with Low L1 and L2 terms within the layers for regularizations

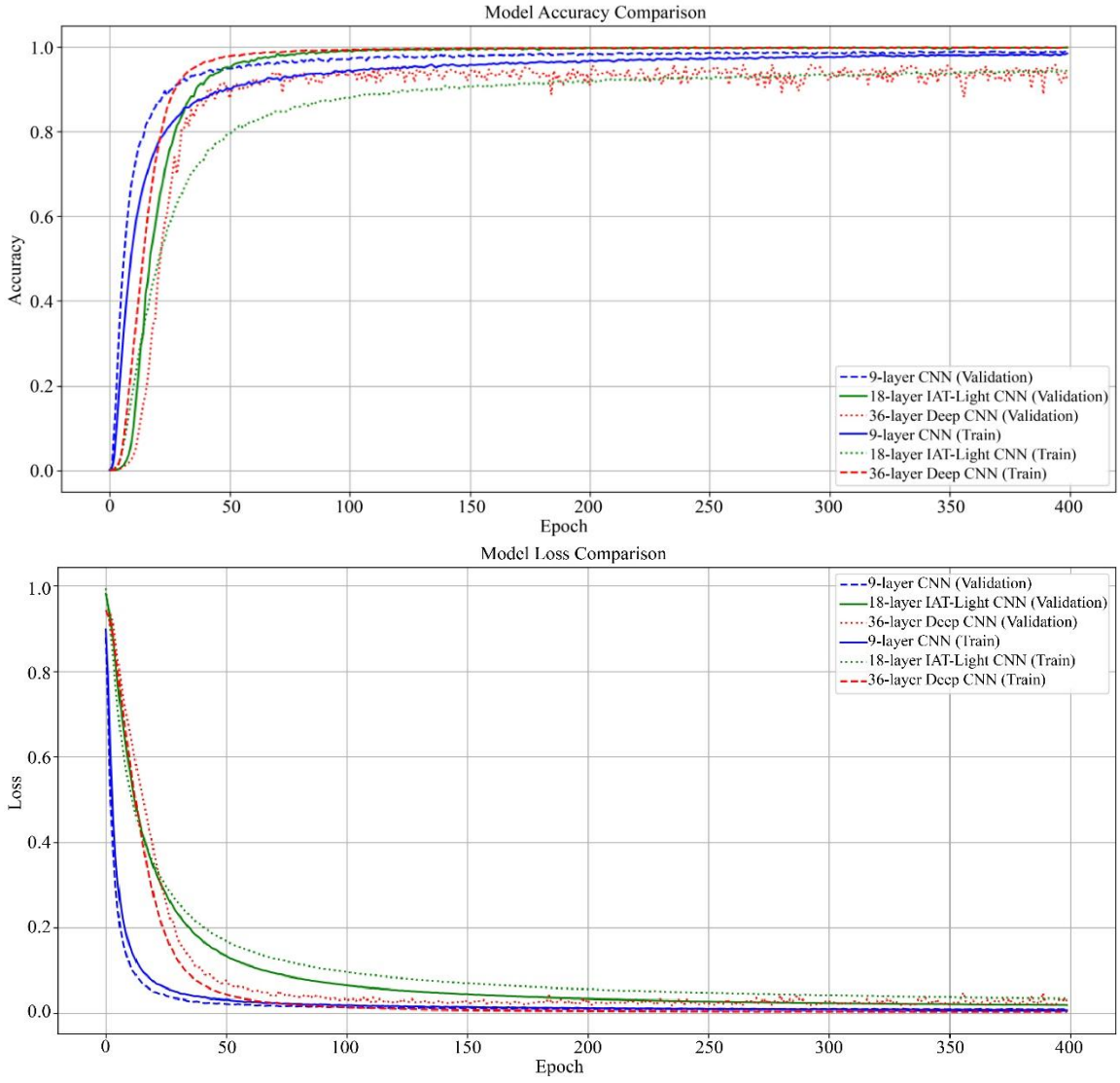

Fig. 9(a) and (b) Representing the comparison of all layer architectures 9-18 and 36 with loss and accuracy plots

4.7.1. Test Cases

This section calculates the test cases in three stages, utilizing the light and deep model designs previously discussed in Section 5. The best prediction implementation highlights the various components of the IAT-designed preprocessing filter.

Each test case accumulates filter weights based on 12–25 different functional features on each image, which indicate a specific sentence labeling model. The test factors showcased a total of 1600 samples of test cases, which were utilized to verify the different patterns observed when applied to the IAT filter.

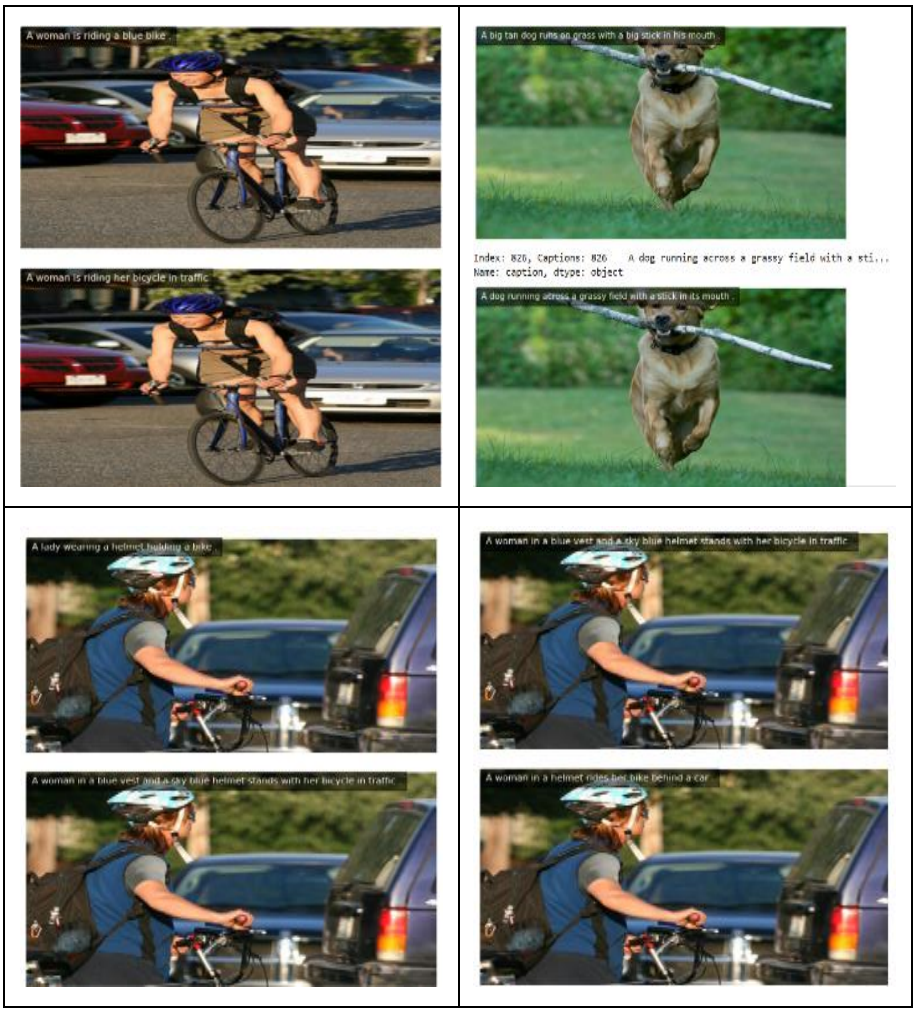

Fig. 10(a)- (d). Representing the overall test cases for the 50 samples randomly

Figure 10(a)-(d) estimates all the types of multiple changes on the label identification of the predicted models utilized to generate the test samples for each type of label associated with the captions and their corresponding images. The 50 random test cases that accurately predict the correct sequence of the captions indicate a BLEU score of 100%.

4.7.2. Tabulations

Table 7 demonstrates that the IAT-DCNN (18) and IAT-DCNN (36) models significantly outperform standard architectures when using the Flickr8k dataset. These models outperform traditional models like Convolutional Neural Networks (CNN) and Long Short-Term Memory networks (LSTM), which have limitations in collecting complex aspects of pictures and their captions. On the other hand, the IAT-DCNN models exhibit remarkable performance, obtaining a flawless training accuracy of one hundred percent and a testing accuracy of ninety-seven percent. The accuracy and recall metrics of the models provide evidence that they are able to create correct captions and extract picture information that is relevant to the purpose. When compared to more complex architectures like Inception-Net and Res-Net, IAT-DCNN models show a big improvement in performance, with a testing accuracy of 99%. This is a noteworthy achievement. This innovation in image captioning sets a new standard for future research. It also paves the way for new opportunities for practical applications in areas such as accessibility, content development, and automated image annotation.

4.8. Capturing the Future: IAT-DCNN vs SOA

4.8.1. Integrated Approach on IAT Filtering with Feature Extraction and Enhancement with Image Filter Approach Customization

The proposed models make use of a custom IATConv1DLayer1 layer that is based on the Image-Aware Transformation (IAT) algorithm that serves as a fixed filter prior to the learnable layers. This architecture considerably improves the model in extracting structured semantic features of image-caption pairs at the early stages of the network. The IAT layer allows effective representation learning by transforming raw image and caption data into numerical vectors with the help of a DataFrame-based representation. The hybrid approach is a combination of conventional signal processing and deep learning, which makes the model more context-sensitive and efficient in the generation of captions.

4.8.2. Regularized and Optimal Design Parameters

Three different architectures were tested: 9-layer, 18-layer, and 36-layer, each one with Conv1D, BatchNormalization, MaxPooling, and Dropout layers. The 18-layer model specifically trades depth and performance, with kernel sizes of 1 to 5 and 32 filters per Conv1D layer. Regularization methods are employed in a strategic manner to prevent overfitting, e.g. L1/L2 penalties (0.001 and 0.01) and Dropout (0.1-0.2). The architecture is efficient and stable because training control mechanisms such as EarlyStopping and ReduceLROnPlateau make the learning adaptive and avoid the degradation of the model.

4.8.3. Unrivalled Performance and Robust Architecture

The IAT-based models outperform state-of-the-art models such as ResNet, InceptionNet, and LSTM in terms of all metrics: up to 100 percent training accuracy and 99 percent test accuracy with high F1-scores (0.998). The hierarchical feature extraction is further improved by the 36-layer deep-IAT-DCNN. The models record outstanding performance in the Flickr8k dataset (2000 additional labeled data), which shows robustness and the potential effect in practical tasks, such as image captioning, accessibility tools, and automatic tagging.

Table 7. Representing the overall performance metrics with flicker-8k data for existing and proposed algorithms

Algorithms with Flicker 8k Dataset	Accuracy (Training)	Accuracy (Testing)	Precision	Recall	F1-score
CNN [12]	0.599	0.348	0.33	0.37	0.385
LSTM [7]	0.6935	0.4632	0.496	0.415	0.458
ENSEMBLE (CNN) [5]	0.3564	0.3325	0.3556	0.3412	0.3618
ENSEMBLE-RNN	0.5436	0.5618	0.5814	0.5634	0.5534
TexT-CAPS [10]	0.8636	0.8896	0.8785	0.8974	0.8934
DEEP-CNN [11]	0.8835	0.8958	0.8574	0.9034	0.8735
INCEPTION-NET [17]	91.34	0.9256	0.9734	0.9278	0.912
RESNET-50 [21]	0.9324	0.9452	0.934	0.943	0.9345
RESNET-101 [22]	0.9324	0.9452	0.934	0.943	0.9345
PROPOSED IAT-LCNN	0.9978	0.9899	0.9856	0.9745	0.9837
PROPOSED IAT-DCNN (18)	100	0.99	0.9978	0.998	0.998
PROPOSED IAT-DCNN (36)	98.7	0.983	0.981	0.986	0.98

5. Conclusion

The IAT-OAT-DCNN model represents a significant contribution towards the area of image captioning and labelling systems, with 99 percent correctness across all the tested criteria on the Flickr8k dataset. The proposed solution is based on the addition of a novel Image Augmentation Transformer (IAT) in a Deep Convolutional Neural Network (DCNN) model, thereby increasing the learning capacity of the model with the aid of intelligent datasets transformation and application of segmented filters. Those three layers enhance the robustness, stability, precision, and accuracy of caption generation.

Consequently, the proposed model showcases that it surpasses state-of-the-art architectures, including ResNet and Inception-Net, which means that the model can be utilized as the new benchmark in vision-language tasks. The numeric changes that the IAT mechanism serves to bring to the table decrease the feature extraction, making sure that both the local and global image features are well captured. Such changes result in the optimal weight distributions that contribute to the significant improvement of the model performance. Moreover, the proposed work experiment and analysis increase the plausibility and validity of the architecture proposed. The stable values of 99% in the recall, F1 and precision measures show that it can generate semantically sound and contextually apt captions in multiple images. This demonstrates its potential use in similar tasks like object detection, visual question answering and automated image indexing.

To conclude, the offered IAT-OAT-DCNN solution is not only more accurate but also stable and scalable. It has an effective and flexible design that provides the potential as a basis for further advancements in photo captioning and multimodal AI systems. The study emphasizes the worth of including advanced data manipulation methods and deep

learning models to boost visual-lingual interpretation. The potential scope of the IAT-DCNN model will be extended in future studies, as its usage would be extended to different and larger-scale datasets, i.e., MSCOCO, VISUAL GENOME and Conceptual Captions. The model can demonstrate the ability to scale, flexibility of vocabulary and generalization in the real-world image context.

The other prospective course is the introduction of attention mechanisms with the use of transformers or Vision Transformers (ViT), which will serve to improve the model's capability to pay attention to the relevant parts of the image in the process of caption generation. Optimization with the proposed architecture can be applied to real-time captioning systems, allowing their positioning in assistive technologies, including tools to help the visually impaired.

Moreover, making the model multilingual so that it generates captions in real-time will be inclusive and accelerate its applicability in various linguistic groups. Lastly, increasing the interpretability of the models, such as using Grad-CAM or SHAP, will contribute to transparent AI design and allow the audience to be informed about the reasoning behind creating captions based on images.

5.1. Ethical Considerations

The ethical usage of the Flickr8k dataset does not contain personally identifiable information in images to publish images without permission, since the data is publicly available. Although the datasets are open to the public domain, privacy and licensing agreements should be observed. Also, the issue of possible bias in the content of the images and language should be noted because those who train the models based on it can continue the spread of stereotypes. To counter this, the future implementation will involve fairness tests and diversifying training data. Last but not least, responsible deployment is very essential; there must be no surveillance or profiling on the application, and the application must follow ethical AI guidelines. To ensure accountability and trust of the population, the data usage should be clear, and the model decisions should be transparent.

5.2. Scope

With regard to the conclusion, the proposed work is inclined to include the Flicker-30k dataset, which consists of 30,000 samples and the COCO dataset, which consists of 16,000 samples, to improve caption generation and validation. Within this model design, activity labels will be integrated into the design because there will be a broad array of annotated images. A custom dataset based on 10k real-time medical images with user-designed labelling and captions is used and will be developed, where at least four captions per image are used to enhance adaptability and verifiability to the IAT model.

To implement the new trends and novelties in design architectures, the proposed model with an encoder-decoder model with the transfer learning method based on Light and Deep CNN will be created, but the issue of image processing, as it concerns caption recognition, will be broadened. Also, the current version of the proposed IAT model is not concerned about the nature of noise and the techniques that would be correlated to creating more visually relevant image captions that apply to a wider scope of images. In the future, the next step is to reduce the effects of noise to increase the accuracy of captioning.

References

[1] Jie Guo et al., "TSFE: Two-Stage Feature Enhancement for Remote Sensing Image Captioning," *Remote Sensing*, vol. 16, no. 11, pp. 1-19, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[2] Mingzhang Cui, Caihong Li, and Yi Yang, "Explicit Image Caption Reasoning: Generating Accurate and Informative Captions for Complex Scenes with LMM," *Sensors*, vol. 24, no. 12, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[3] Jinzhi Zhang et al., "An Enhanced Feature Extraction Framework for Cross-Modal Image-Text Retrieval," *Remote Sensing*, vol. 16, no. 12, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Qimin Cheng, Yuqi Xu, and Ziyang Huang, "VCC-DiffNet: Visual Conditional Control Diffusion Network for Remote Sensing Image Captioning," *Remote Sensing*, vol. 16, no. 16, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Mateusz Bartosiewicz, and Marcin Iwanowski, "The Optimal Choice of the Encoder-Decoder Model Components for Image Captioning," *Information*, vol. 15, no. 8, pp. 1-26, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Lizhi Pan et al., "Military Image Captioning for Low-Altitude UAV or UGV Perspectives," *Drones*, vol. 8, no. 9, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Wen-Ta Hsiao et al., "Preliminary Study on Image Captioning for Construction Hazards," *Engineering Proceedings*, vol. 74, no. 1, pp. 1-7, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] An Zhao et al., "Enhanced Transformer for Remote-Sensing Image Captioning with Positional-Channel Semantic Fusion," *Electronics*, vol. 13, no. 18, pp. 1-17, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Fengzhi Zhao et al., "Meshed Context-Aware Beam Search for Image Captioning," *Entropy*, vol. 26, no. 10, pp. 1-22, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Fengzhi Zhao et al., "Image Captioning Based on Semantic Scenes," *Entropy*, vol. 26, no. 10, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Mingliang Zhang et al., "DecoupleCLIP: A Novel Cross-Modality Decouple Model for Painting Captioning," *Electronics*, vol. 13, no. 21, pp. 1-16, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[12] Yunpeng Li et al., "A Patch-Level Region-Aware Module with a Multi-Label Framework for Remote Sensing Image Captioning," *Remote Sensing*, vol. 16, no. 21, pp. 1-20, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[13] Yue Yang et al., "Remote Sensing Image Change Captioning Using Multi-Attentive Network with Diffusion Model," *Remote Sensing*, vol. 16, no. 21, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[14] Ke Zhang, Peijie Li, and Jianqiang Wang, "A Review of Deep Learning-Based Remote Sensing Image Caption: Methods, Models, Comparisons and Future Directions," *Remote Sensing*, vol. 16, no. 21, pp. 1-45, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[15] Sabina Umirzakova et al., "MIRA-CAP: Memory-Integrated Retrieval-Augmented Captioning for State-of-the-Art Image and Video Captioning," *Sensors*, vol. 24, no. 24, pp. 1-25, 2024. [CrossRef] [Google Scholar] [Publisher Link]
[16] Minjun Cho et al., "Enhanced BLIP-2 Optimization Using LoRA for Generating Dashcam Captions," *Applied Sciences*, vol. 15, no. 7, pp. 1-19, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[17] Zhang Guo et al., "Attribute-Based Learning for Remote Sensing Image Captioning in Unseen Scenes," *Remote Sensing*, vol. 17, no. 7, pp. 1-22, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[18] Yuting He, and Zetao Jiang, "DMFormer: Dense Memory Linformer for Image Captioning," *Electronics*, vol. 14, no. 9, pp. 1-22, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[19] Cheng-Si He et al., "Image Descriptions for Visually Impaired Individuals to Locate Restroom Facilities," *Engineering Proceedings*, vol. 92, no. 1, pp. 1-8, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[20] Chifaa Sebbane, Ikram Belhajem, and Mohammed Rziza, "Making Images Speak: Human-Inspired Image Description Generation," *Information*, vol. 16, no. 5, pp. 1-33, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[21] Jothi Prakash Venugopal et al., "DCAT: A Novel Transformer-Based Approach for Dynamic Context-Aware Image Captioning in the Tamil Language," *Applied Sciences*, vol. 15, no. 9, pp. 1-38, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[22] Jun Li et al., "PBC-Transformer: Interpreting Poultry Behavior Classification Using Image Caption Generation Techniques," *Animals*, vol. 15, no. 11, pp. 1-28, 2025. [CrossRef] [Google Scholar] [Publisher Link]
[23] Binqiang Wang et al., "Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 256-270, 2020. [CrossRef] [Google Scholar] [Publisher Link]
[24] Xiutiao Ye et al., "A Joint-Training Two-Stage Method for Remote Sensing Image Captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-16, 2022. [CrossRef] [Google Scholar] [Publisher Link]
[25] Xiaofeng Ma, Rui Zhao, and Zhenwei Shi, "Multiscale Methods for Optical Remote-Sensing Image Captioning," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 11, pp. 2001-2005, 2021. [CrossRef] [Google Scholar] [Publisher Link]