Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Improving the Reliability and Accuracy of Image Captioning Systems Using Ensemble of FC, Softmax, and LSTM Deep Decoders

Ghadeer Abdulrasool Mohammed [a,*], Raidah S. Khudeyer [a], Maytham Alabbas [a]

[a] University of Basrah, College of Computer Science and Information Technology, Department of Computer Science. Email: ghadeer.mohammed@gmail.com, raidah.khudayer@uobasrah.edu.iq, ma@uobasrah.edu.iq

ARTICLE INFO

Article history:
Received: 12 /01/2026
Revised form: 13 /02/2026
Accepted: 15 /02/2026
Available online: 30/03/2026

Keywords:

Image Captioning, Deep Learning, nsemble Learning, CNN–LSTM Networks, Flickr8k.

ABSTRACT

In this work, a deep system for automatic image description is presented, which aims to produce fluent, meaningful, and structurally coherent sentences for input images. The proposed architecture is based on an encoder-decoder framework, in which high-level image features are first extracted by an Inception-v3 deep convolutional network and then fed as a compressed image representation to an LSTM-based language decoder to produce a word-by-word sentence. On this basic structure, a voting-based ensemble learning framework is designed, in which three deep paths, including a fully connected (FC) network, a Softmax linear model, and a sequence-oriented LSTM decoder, are trained independently, and the word probability vectors at the output level are combined with a maximum voting mechanism. The evaluation is performed on the standard Flickr8k database and using BLEU-1 to BLEU-4, METEOR, and ROUGE-L metrics. The results show that the best single LSTM model achieves values of 0.64, 0.39, 0.23, and 0.16 for BLEU-1 to BLEU-4, and 0.22 and 0.50 for METEOR and ROUGE-L, respectively, while the Ensemble model improves the values to 0.74, 0.50, 0.35, and 0.22 for BLEU-1 to BLEU-4, 0.475 for METEOR, and 0.55 for ROUGE-L; such that the relative improvements in BLEU-3 and BLEU-4 are 54% and 41%, respectively. The paired t-test also shows that the difference in Ensemble performance with single models is significant at the 95% confidence level, and compared to the existing methods on Flickr8k, competitive results are obtained and, in some measures, superior.

https://doi.org/10.29304/jqcsm.2026.18.12573

1. Introduction

In recent years, significant advances in the field of deep learning have significantly increased the ability of intelligent systems to understand and interpret multimedia data. In this regard, "Image Captioning" has been raised as one of the important and challenging problems at the border of machine vision and natural language processing [1,2]. In this problem, the goal is to produce a fluent, meaningful, and human-understandable sentence in natural language, appropriate to the visual content of an image, which requires the simultaneous use of deep image feature extraction models and sequence-based language models. Applications of this field in image-based search engines, social networks, assistive systems for the blind, medicine, intelligent surveillance, and automatic generation of descriptions for multimedia content have created a special place for Image Captioning in the field of artificial intelligence research [3-5].

Despite the development of powerful models such as combining convolutional networks (CNN) with recurrent networks (RNN/LSTM) as well as architectures based on Attention and Transformer, challenges such as limited description accuracy, incomplete coverage of details in the image, and weakness in generating diverse and natural sentences still persist. Single-model models are highly sensitive to data noise, database size limitations, and weight initialization methods, such that their performance can fluctuate in different conditions. On the other hand, the

∗ Corresponding author: Ghadeer Abdulrasool Mohammed
Email addresses: *ghadeer.mohammed@gmail.com*
Communicated by 'sub etitor'

extremely high diversity of visual content, the ambiguity of visual events, and the limited vocabulary range in linguistic models make the process of converting visual representations into accurate linguistic expressions close to human perception more complex [6,7].

Experience from numerous studies in machine learning shows that the use of ensemble methods can improve the overall stability and accuracy of the system by reducing the variance of models and combining the strengths of several different approaches [8, 9]. Meanwhile, the stacking architecture is one of the advanced ensemble approaches in which the output of several base models is used as input to a meta-model, and this meta-model learns how to provide a more appropriate combination of the predictions of the base models.

The structured application of this idea to the Image Captioning problem can enable the simultaneous exploitation of several proposed descriptions for an image and, by intelligently selecting or combining them, the final quality of the caption can be improved in terms of standard evaluation criteria as well as in terms of readability and linguistic diversity. Despite these capabilities, the majority of existing research in the field of Image Captioning is still focused on a single model and has only benefited from the planned advantages of the ensemble, especially at the level of language production [10-12].

From a research perspective, several aspects remain relatively obscure or underexplored in the existing literature. First, the optimal mechanism for combining the output of base models in a stacking framework, especially in a multifaceted problem such as language generation from images, is not clearly defined; it is still unclear what kind of metamodel, and with what level of access to the outputs and intermediate representations of the base models, can make the best decision for selecting or combining descriptions. Second, the relationship between common evaluation metrics such as BLEU, METEOR, and ROUGE-L with the actual linguistic quality and the closeness of the generated sentence to human perception is still a matter of debate, and how to optimally use these metrics to guide the metamodel and select the final caption has not been systematically investigated. Also, the impact of ensembles on aspects such as lexical diversity and structural richness of sentences has been less studied in depth.

In such a context, the present study aims to design a deep learning framework based on an ensemble and stacking architecture for the problem of generating textual descriptions for images. The main idea is to reduce the limitations of single-model models in terms of accuracy, stability, and linguistic diversity by employing several basic caption-generating models and intelligently combining their outputs. In this framework, a higher-level metamodel is responsible for analyzing linguistic patterns and semantic differences between the outputs of basic models and tries to select the most appropriate description that is closest to the perceptual reality of the image based on a learned pattern; thus, the problem is not simply to generate captions, but to design a deep and decision-making combinator to exploit the synergy of models.

The importance of this approach can be justified in several ways. On the one hand, the rapid increase in the volume of visual data in information systems, social networks, medicine, and surveillance has doubled the need for methods that can bridge the gap between raw pixel representation and symbolic linguistic expression. On the other hand, the instability and limitations of single-model models in sensitive environments – such as blind assistance systems, medical diagnosis and documentation, or surveillance environments – indicate the need to use mechanisms that improve the accuracy and reliability of the output. By turning the process of combining models itself into a learning problem, stacking architectures enable the system to adaptively learn which base model is more reliable in which situations and how to decide between contradictory or complementary outputs.

Accordingly, the main innovation of this research is in designing a stacking-based ensemble framework for Image Captioning, in which the combination of the output of multiple deep models is considered not as a simple linking operation, but as a structured learning process. The focus on linguistic quality and production diversity, along with common quantitative metrics (BLEU, METEOR, ROUGE-L, and diversity indices such as Distinct-1 and Distinct-2), as well as the possibility of implementing this framework on relatively smaller databases, are other distinguishing aspects of this research. In the rest of the article, after explaining the problem and objectives of the research in more detail, the proposed method, the design of the stacking architecture, and how to evaluate and compare it with basic models and existing methods will be discussed.

2. Related Works

Image captioning systems have evolved significantly with the adoption of deep neural decoders and ensemble learning techniques. This section presents an overview of recent related works that aim to enhance caption accuracy and robustness using FC, Softmax, and LSTM-based decoding frameworks.

In [13], the Image Transformer architecture for image captioning is introduced, which is an adapted version of the Transformer for image data. Relying on the attention mechanism and modeling the relative spatial relationships between image regions, the model uses only the regional features extracted from the object recognition network

instead of using global features, and achieves performance on MSCOCO in offline and online modes that are on par with or better than the reference methods.

In [14], a relation-based and visual context-aware approach is proposed to improve attention-based captioning. First, the relationships between image regions are modeled with a neural graph network to produce relation-aware representations; then, the visual context-aware attention mechanism prevents repetition and ignoring of important parts by maintaining the history of previously attended regions. Experiments on MSCOCO and Flickr30k show that the proposed method outperforms previous state-of-the-art methods in most common metrics.

In [15], a deep hybrid framework for Bengali image captioning is introduced. In the visual part, InceptionResNetV2 and Xception are used for feature extraction, and in the linguistic part, two types of word embeddings (pre-trained Bengali language models and fastText) are used in combination. BiLSTM and BiGRU are used for sequence decoding, and significant improvements in the fluency and accuracy of captions are achieved on the Bengali versions of Flickr8k and BanglaLekha, based on BLEU, compared to single-embedding models.

The work [16] focuses on generating semantic captions based on scene and visual content understanding. In this framework, first, objects in the image are detected with a CNN, then an extended version of LSTM with an attention mechanism generates word sequences in a region-aware manner. Four different variants of the model are trained on COCO and Flickr30k, and semantic evaluation shows that the proposed approach provides deeper scene understanding and better support for analytical and decision-making applications.

In [17], a transformer-based approach for image captioning is presented, and the role of hyperparameter tuning, cost function, and optimizer selection is systematically investigated. The authors consider the combination of cross-entropy and Adam as the best option, with Top-5 accuracy of about 73.09 and BLEU-4 of 20.10. A comparison of several CNN encoders shows that ResNeXt-101 is superior in terms of quality, while MobileNetV3 with fewer parameters has competitive performance. Using ViT and DeiT as encoders also shows that DeiT performs better than ViT with BLEU-4=34.44.

In [18], a captioning system for bridge damage monitoring is presented. Images of damaged bridges are fed to an attention-based deep model to produce technical sentences about the type and location of the damage. The use of the attention mechanism allows the model to focus on the damaged areas and produce multiple explanatory sentences for a complex image. Results on the specialized bridge damage dataset lead to BLEU-1 to BLEU-4 values of 0.782, 0.749, 0.711, and 0.693 and an accuracy of 69.3%, which shows superiority over the unattended version.

In [19], the issue of "explainability" in medical image captioning is investigated. Referring to the black-box nature of deep models, the authors propose an encoder–decoder approach based on self-attention and attention mechanisms that can reveal explicit associations between image areas and caption words. Evaluation on ImageCLEF shows that the model is able to produce understandable attention maps for clinicians, thereby increasing the reliability and safety of captioning in medical scenarios.

In [20], an automated framework for image captioning based on the combination of multiple deep networks and natural language processing methods is presented. Multiple vision networks at different levels extract semantic and hierarchical features and object recognition results, which are then combined with a fusion strategy. The final output is submitted to a language module to build appropriate captions. The results show that the multi-network approach has significant improvements in recognition accuracy and caption quality compared to every single network.

In [21], a multi-modal approach for image captioning is proposed that combines three architectures: YOLOv8, EfficientNetB7, and Transformer. YOLOv8 recognizes image objects in real time, EfficientNetB7 extracts rich visual features, and the Transformer decoder generates fluent and meaningful captions using this multimodal representation. Experiments on reference datasets show that the integration of these three components provides significant improvements in evaluation criteria and the production of coherent and accurate captions compared to using each one individually.

The study [22] addresses image captioning in news scenarios, where captions should not only describe the scene but also reflect identity information about individuals, especially celebrities. The authors emphasize that traditional methods are mostly content-oriented and ignore the identity of individuals. The proposed framework, by combining scene analysis and face recognition, produces captions that better meet the needs of the news industry, recommender systems, and social networks in terms of both narrative content and accurate reference to the people in the image.

In [23], a graph-based method for image captioning is introduced. First, using RelTR, object boxes and subject-subject-object triple relations are extracted, and two separate graphs are constructed for spatial and semantic relations. Each graph is processed by an independent GCN and fed to the LSTM decoder along with the visual features of the CNN. The multimodal attention mechanism is applied to the three feature sources at each step. The results show that the model produces context-based and accurate captions and is suitable for assistive technologies and scene interpretation.

In [24], a deep method based on the combination of DenseNet201 and LSTM is presented for image captioning. DenseNet201 extracts hierarchical and rich image features, and LSTM as a decoder generates caption word sequences. The model is trained on the Flickr8k dataset and evaluated with BLEU benchmarks, and competitive results are obtained. The authors propose several applications, such as assistive technologies for the blind and automatic digital content generation, and suggest that adding attention and transformer mechanisms can further improve accuracy and semantic richness.

Despite the many advances in deep architectures for generating textual image captions, this field still faces several fundamental challenges. First, many existing models are unable to fully cover the scene components and accurately reflect the semantic relationships between objects, and often settle for general, superficial descriptions lacking important details. Second, the output of these models fluctuates in terms of stability and repeatability, such that minor changes in training data or initializing weights can lead to a significant difference in evaluation criteria. Third, the weakness in lexical and structural diversity makes the produced captions appear monotonous and linguistically indistinct, and have a noticeable distance from human description.

Our proposed method, relying on a stacking-based ensemble framework and intelligently combining the output of multiple deep decoders (including FC, Softmax, and LSTM), attempts to address these challenges simultaneously; This means that by exploiting the complementary strengths of different models, it increases the accuracy and coverage of scene details, reduces the variance and instability of performance, and significantly improves the diversity and richness of produced captions by integrating diverse linguistic perspectives.

3. Proposed Model

Figure 1 illustrates the proposed model and depicts the overall flow of the image description system as a multi-stage and multi-branch pipeline. In the first stage, the input image enters the deep vision section after preprocessing and is transformed into a high-level feature map by a deep convolutional neural network based on Inception-v3. This feature map, which contains rich spatial and semantic information of the image, is injected into the language module as a feature vector or feature map.

3.1 Preprocessing

Raw data usually suffer from problems such as noise, extreme changes in dynamic range and sampling, and their use will also weaken subsequent designs. Preprocessing also includes more complex transformations that are used to reduce the dimensions of the data. In the first stage of the proposed method, redundant and irrelevant information is removed from the entire data. This stage, also known as the normalization process, removes outliers from the data using methods such as linear regression methods.

The features used in this research for retrieval also include color features and texture features. The brightness of the image is calculated by considering the result of all color channels at different points in the image. Therefore, reducing the dimensions in the frequency domain will not have a negative effect on the quality of the final result. Therefore, in order to reduce the volume and increase the speed of calculations, in the first step, we convert the input image to a gray-level image using Equation 1.

$$I_{x,y} = \left(0.2989 \times R_{x,y}\right) + \left(0.5870 \times G_{x,y}\right) + \left(0.1140 \times B_{x,y}\right) \quad (1)$$

In such a situation, the presence of pixels whose brightness intensity is affected by noise can definitely have a negative effect on the final result. Also, various factors such as the imaging method, appearance characteristics, lighting angle, etc., cause differences in brightness levels between neighboring pixels of the same type (foreground or background). Despite the fact that this phenomenon is inherently part of the true nature of the image, the possibility of its undesirable effect during image thresholding cannot be ignored. The most important undesirable effect resulting from this is the incorrect classification of pixels with brightness levels close to the image threshold into the inappropriate group. Applying smoothing filters can relatively prevent this from happening by reducing the sharpness in inherently uniform areas of the image. In order to minimize the effect of noise and to ensure that image smoothing does not reduce the accuracy of the position of the boundary strip between cells and the background, a median filter with a window frame is used. Finally, the grayscale image is used to extract texture features. Figure 2 shows an example image from the dataset with its grayscale version.

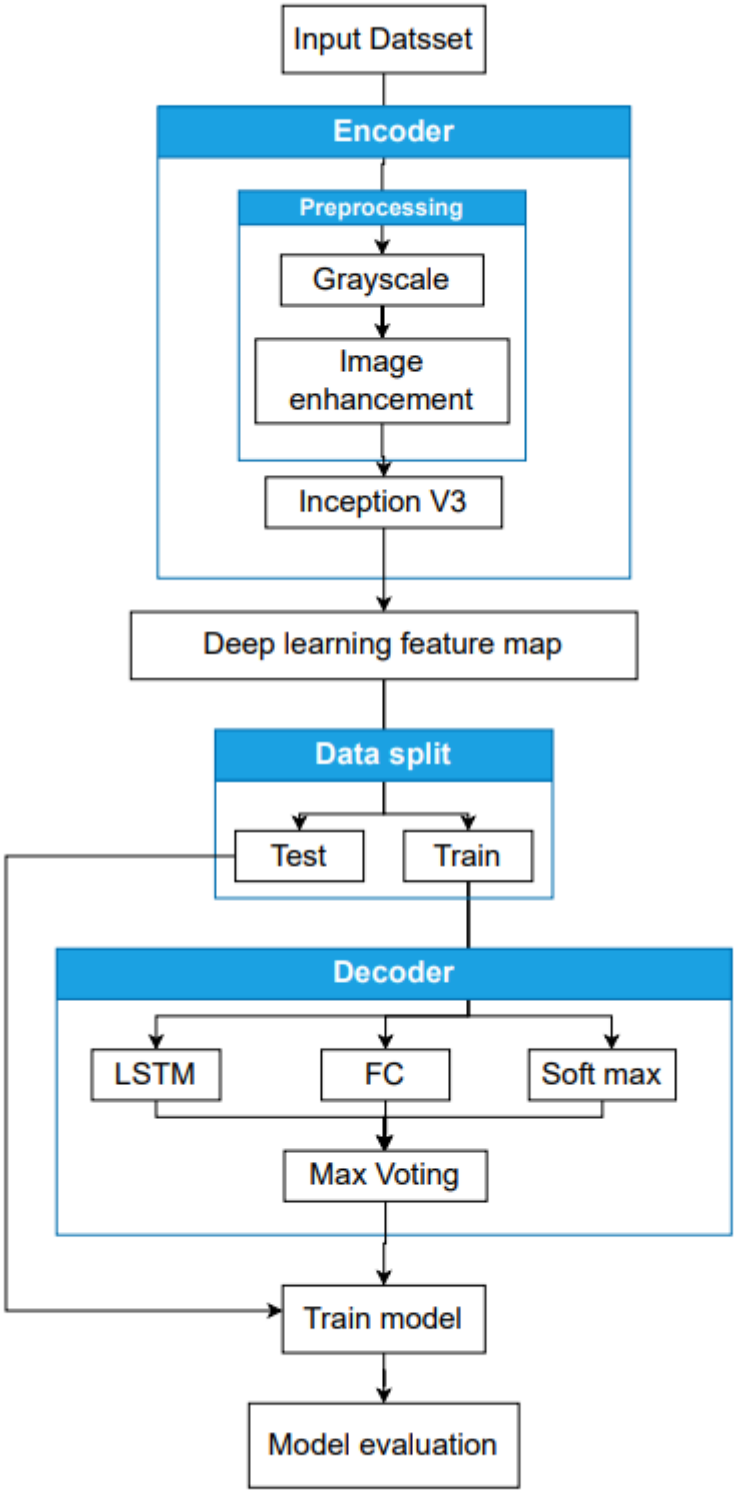

Figure 1: Structural elements of the proposed model.

Figure 2: Original image in color space with grayscale version.

3.2 Encoder

In the proposed model, the encoder section uses the Inception-v3 deep convolutional neural network to extract high-level image features. This network, using multi-branch Inception modules, is able to extract image features simultaneously at different scales, thus providing a richer representation of the image content. After performing preprocessing steps, the input image is entered into the Inception-v3 network and finally converted into a compact and meaningful feature vector that is used as the input to the language decoder section. In this study, a pre-trained version of Inception-v3 has been used, which has been trained on standard datasets and has a high ability to extract general image features. In order to adapt the network to the desired problem, the final layer of the network has been removed, and a new fully connected layer with three outputs, corresponding to the problem classes, has been added. This approach accelerates the training process and improves model convergence.

The loss function used in the training process is Cross-Entropy Loss, which is suitable for classification and probability-based learning problems. The Adam Optimizer algorithm is used to optimize the network weights and reduce the value of the loss function. Also, in order to control the learning rate and improve the stability of the training, a single-cycle learning schedule with a maximum learning rate of 0.001 is considered. The model is trained under these settings for a maximum of 40 cycles.

In the Inception-v3 architecture, factored convolutions and 1×1 convolutions are used to reduce the number of parameters and computational cost. The ReLU activation function is used in different layers of the network to provide the necessary nonlinearity for learning complex patterns. Also, using Batch Normalization in different parts of the network reduces internal covariance, increases the stability of the training process, and prevents overfitting. Figure (3) shows an overview of the architecture of the Inception-v3-based encoder used in this research.

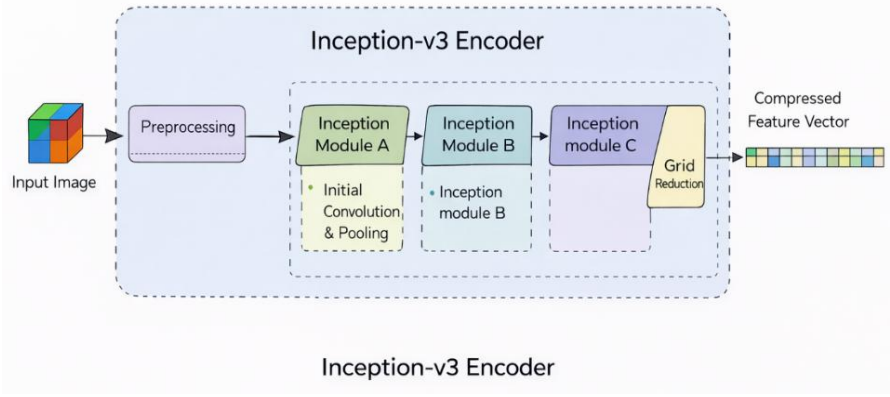

Figure 3: Inception-v3 architecture used.

3.3 Feature Map Parameters

In the proposed framework, visual feature extraction is performed using the Inception-v3 convolutional neural network, which acts as the encoder component of the system. After the preprocessing stage, each input image is forwarded through the Inception-v3 architecture, and high-level visual representations are obtained from the final convolutional layers of the network. The Inception-v3 model employs multi-branch convolutional blocks that enable the extraction of visual features at multiple spatial scales. This design allows the encoder to capture both local details and global semantic information present in the image. To obtain a compact and fixed-length representation, global average pooling is applied to the final convolutional feature maps, resulting in a 2048-dimensional feature vector. This feature vector serves as the visual embedding of the input image and is subsequently provided to the

language decoding modules. The extracted feature representation is injected in parallel into the LSTM-based decoder, the fully connected (FC) network, and the Softmax classifier. Each decoder processes the visual embedding independently to generate a probability distribution over candidate words. By using a pre-trained Inception-v3 encoder, the proposed model benefits from robust visual .Table (1) summarizes the architectural details and shows the concept of parameter tuning.

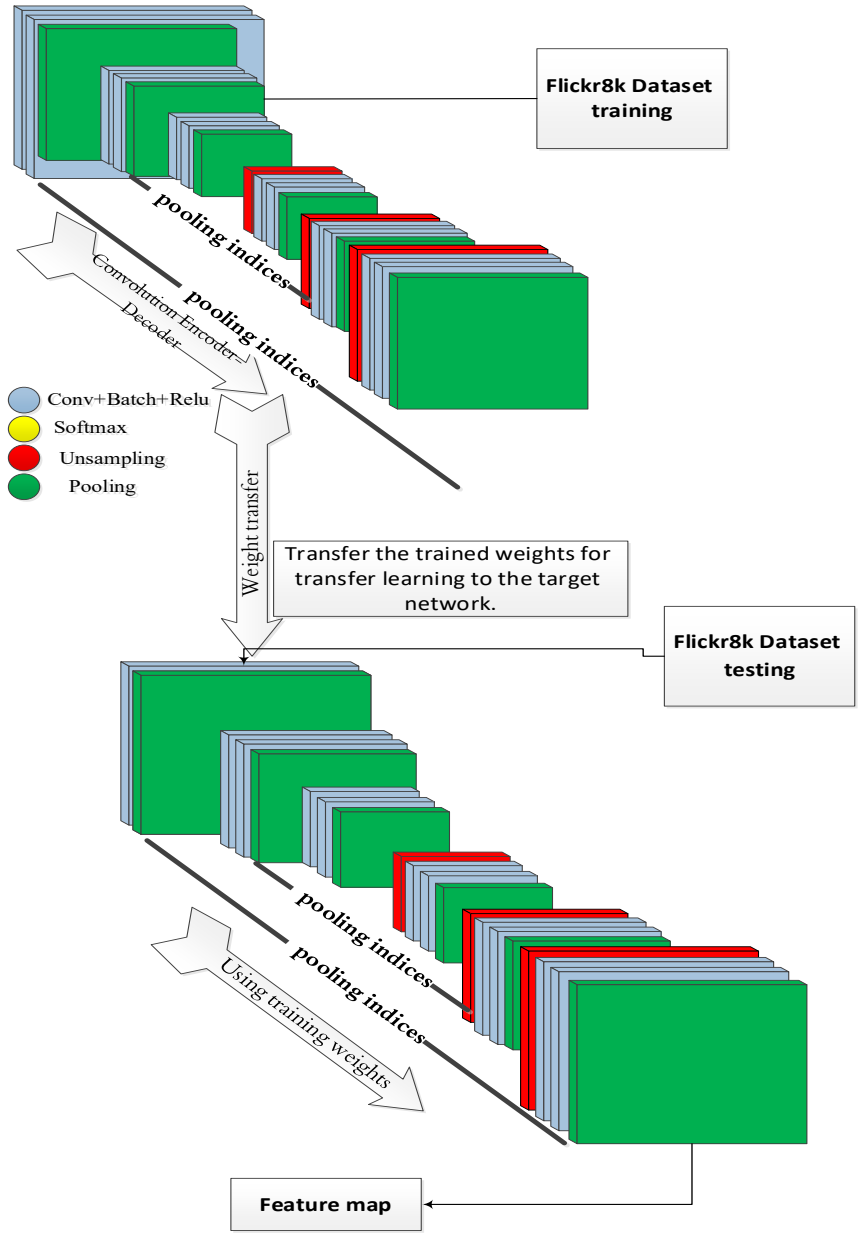

Figure 4: Creating a feature map.

Table 1: Parameters used in the proposed deep neural network.

Layer	Output Feature Map Size	Kernel Size	No. of Learnable Parameters
Conv Layer 1	180 × 50 × 8	3 × 3	224
Max Pool Layer 1	90 × 25 × 8	2 × 2	0
Conv Layer 2	90 × 25 × 16	3 × 3	1168
Batch Norm Layer 1	90 × 25 × 16	–	64

Max Pool Layer 2	45 × 12 × 16	2 × 2	0
Conv Layer 3	45 × 12 × 32	3 × 3	4640
Max Pool Layer 3	22 × 6 × 32	2 × 2	0
Conv Layer 4	22 × 6 × 64	3 × 3	18,496
Batch Norm Layer 2	22 × 6 × 64	–	256
Max Pool Layer 4	11 × 3 × 64	2 × 2	0
Conv Layer 5	11 × 3 × 128	3 × 3	73,856
Max Pool Layer 5	5 × 1 × 128	2 × 2	0
Conv Layer 6	5 × 1 × 256	3 × 3	295,168
Batch Norm Layer 3	5 × 1 × 256	–	1024
Flattening Layer	1280	–	0
Output Layer	221	–	283,101
Total No. of Learnable Parameters	**677,997**		

3.4 Integration of AdaBoost in the Ensemble Framework

In order to enhance the reliability of word prediction and reduce the effect of weak or unstable learners, AdaBoost is incorporated as an auxiliary ensemble mechanism within the proposed framework. Unlike the final ensemble stage, which combines the outputs of multiple deep decoders using a max-voting strategy, AdaBoost operates at the classification level to strengthen individual base predictors. Specifically, for each decoding path (FC-based decoder, Softmax linear model, and LSTM-based decoder), AdaBoost is applied to iteratively reweight training samples, assigning higher importance to misclassified words during the learning process. This boosting procedure improves the discriminative power of each decoder and results in more reliable word probability distributions at each time step. The boosted outputs of the base decoders are then forwarded to the final ensemble module, where the word probability vectors are fused using a maximum probability (max-voting) strategy to generate the final caption. In this way, AdaBoost is not treated as a standalone component, but as an integral part of the proposed ensemble framework that improves the quality of base predictions before the final decision-making stage. This combination of AdaBoost and LSTM improves accuracy, increases robustness, and flexibility when dealing with complex and sequential data. Equation 2 represents the feature set vector.

$$FusedRule_{Final}^{word} = Max\{LSTM_1^{word}, SoftMax_2^{word}, FC_3^{word}\} \quad (2)$$

In relation $LSTM_1^{word}$ is the output of the image caption mapping for the LSTM classifier, $SoftMax_2^{word}$ is for the SoftMax classifier, and FC_3^{word} is for the fully connected neural network classifier.

In the final step, in the Final decision module, this final output vector is subjected to Max Voting to select the most likely class or word and ultimately present it as the output text Caption or text label. Thus, Figure (5) shows how combining the decisions of multiple deep decoders leads to a more stable and accurate final decision for text generation.

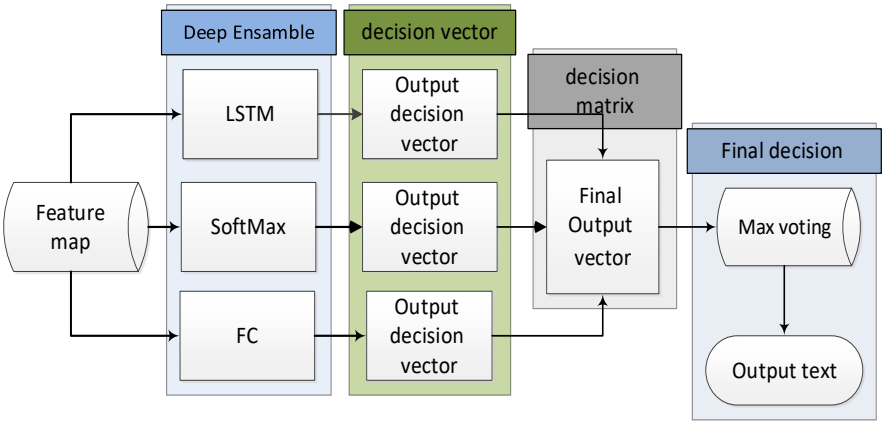

Figure 5: Proposed ensemble model for image captioning.

In the proposed model of this research, the two main components of the "fully connected neural network" (FC) and the "LSTM network" are used in a complementary manner in the form of a deep ensemble framework. The feature vector or map extracted from the convolutional part is first transferred to the appropriate space by a mapping layer and set as the initial state of the memory and the hidden state of the LSTM; then the LSTM models the

long-term dependencies of the word sequence with its gating mechanism (input, output and forgetting) and produces an output vector at each time step using the embedding of the previous word. This output is given to the fully connected layers and finally to Softmax to produce the probability distribution over the words and the image description sentence on a word-by-word basis. In addition to this sequence-based path, a multilayer FC perceptron is also applied directly to the image feature map and produces an independent "decision vector"; at the Ensemble level, the outputs of the FC and LSTM-based decoders are combined with an intelligent fusion mechanism. The structure is trained using a combined cost function including cross-entropy, weight decay, and an auxiliary term to control the erratic behavior of the Ensemble outputs; the weights are initialized with small random values, updated over 35 training iterations with a batch size of 80, a learning rate of 0.01, and a gradient threshold of 1, and dropout layers are used to reduce overfitting. Thus, the structural combination of FC and LSTM in the form of a deep ensemble increases both the accuracy and coverage of the semantic details of the image, and significantly improves the stability and linguistic diversity of the produced captions compared to single-model models. Table (2) summarizes the detailed settings of the LSTM classifier:

Table 2: LSTM Classifier Settings.

Parameter	Value
Learning rate	0.01
Optimization method	ADAM
Maximum number of IPAKs	350
Classification size	80
Number of hidden units	100
Gradient threshold	1
Execution environment	Python 3.3.
Number of hidden layers	5
Sequence length	Longest
Activity function	Sigmoid

4. Results

4.1 Evaluation Parameters

Evaluating the performance of text caption generation models for images is a difficult process due to the combined nature of the problem (visual perception and natural language production), and relying solely on human judgment is not sufficient due to its time-consuming nature and dependence on individual judgment; therefore, the use of automated evaluation criteria is inevitable. In this study, a set of complementary indices, including BLEU, METEOR, and the ROUGE family are used to comprehensively measure the quality of the produced captions. The ROUGE-N criteria calculate precision, recall, and F-measure based on common n-grams, and the ROUGE-L and WLCS measure the coverage of reference content by relying on the longest common subsequence (LCS) and its weighted version. BLEU also evaluates the model's ability to select and order words and produce grammatically correct sentences by calculating the accuracy of n-grams at the sentence set level and using the geometric mean of accuracies along with a length penalty (BP) factor, while METEOR, by combining precision and recall and considering lexical equivalences, provides a more accurate estimate of the semantic proximity of the produced caption to the human caption. Thus, the simultaneous use of these criteria allows for the evaluation of lexical, structural, and semantic aspects in a more stable manner and closer to human judgment[25-27]. To implement and simulate the proposed model, a system equipped with an Intel Core i7 processor with a frequency of 2.60 GHz, 8 GB of RAM, Windows 10 operating system, and the Python programming environment has been used, which provides a suitable computing platform for running deep learning algorithms and evaluating them.

4.2 Database

In this study, the Flickr8k database was used as the main source for training and evaluating the image captioning model. This database contains 8,000 color images of real-life scenes collected from the Flickr website. The images mainly include human activities (such as play, work, sports), the presence of animals, urban and natural environments, and human interaction with the environment, and therefore cover a relatively good variety of visual scenarios. For each image, exactly five text captions in English were generated by human annotators; these captions were written independently and varied in terms of vocabulary and sentence structure, so that the model, in addition to learning image-sentence matching, also faces the problem of caption diversity. The structure of the dataset text

files is such that for each image file name, five descriptive sentences with separate identifiers are stored, allowing for easy preprocessing of the "image-caption" mapping and extraction of the list of (Image, Caption) pairs. This dataset is divided into three sub-sections as standard: training, validation, and testing to allow for hyperparameter tuning, prevention of overfitting, and fair model evaluation. Most images have medium resolution (a few hundred pixels per dimension) and are resized to fixed dimensions (commensurate with the Inception-v3 input) during preprocessing to train the deep network. Due to its moderate size, reliable human labeling, multiple captions per image, and the availability of standard segmentation, Flickr8k is one of the most widely used reference datasets in the field of Image Captioning, and its selection for this study allows for direct comparison of results with previous work. Table (3) summarizes the main features of the Flickr8k database used in this study:

Table 3: Flickr8k database specifications.

Feature	Amount / Description
Database name	Flickr8k
Total number of images	8000 color images
Number of captions per image	5 independent text captions
Language of captions	English
Total number of captions	40,000 descriptive sentences
Content domain of images	Everyday scenes, humans, animals, natural and urban environments
Source of image collection	Flickr website
Data breakdown	Three parts: training, validation, and testing
Type of use in this research	Training and evaluating CNN–LSTM and Ensemble models for image captioning
Data format	Image files (JPG) along with a text file containing the image caption mapping

4.3 Settings

In the proposed model, the Softmax function is used in the output layer of the language decoder (LSTM/FC) as the final classification layer on words; in such a way that the output of the fully connected layer produces a vector of length |V| (word size) that contains the raw (logit) scores of each word, and Softmax converts these scores into a probability distribution over the entire vocabulary, such that the sum of the components is equal to 1 and each component represents the probability of selecting that word at the current time step. This probability distribution is directly fed into the cross-entropy cost function and, within the framework of optimization with the Adam algorithm, the parameters of the previous layers (LSTM and FC) are updated. Softmax itself does not have independent hyperparameters, and its main setting is related to the size of the words selected after text preprocessing and how to handle special tokens such as <start>, <end>, and <unk>. On the other hand, a fully connected (FC) neural network, in which each neuron is connected to all neurons in the next layer, receives the input as a feature vector (e.g., a vector extracted from an image or the output of an LSTM) and maps it to a new feature space in each layer with learnable weights and biases and a nonlinear activation function (e.g., ReLU); repeating this process in multiple hidden layers allows learning complex relationships between the input and output, and finally, the output layer produces a score vector for classes/vocabularies. In this thesis, an FC network is used as one of the branches of a deep ensemble to produce a "decision vector" based on the image feature map, and this vector is combined with the output of other decoders in the aggregation stage and participates in the final decision-making for caption word selection. The weights and biases of this network are updated using a cross-entropy-based cost function and the Adam optimizer, and techniques such as weight decay and appropriate layer structure design are used to prevent overfitting. In this study, to improve the efficiency of the proposed model, three mechanisms of "random search", "ADAM optimization algorithm", and "cross-validation" have been used in a complementary manner. In the hyperparameter tuning step, the Random Search method helps to find optimal regions in the parameter space by randomly testing different combinations of learning rate, number of hidden layer units, and batch size; thus, the risk of intuitive and non-optimal choices is reduced, and a better balance is established between convergence speed, final accuracy, and training stability. To optimize the network weights, the ADAM algorithm has been used, which, by combining the advantages of AdaGrad and RMSProp, has a good ability to work with sparse gradients, noise, and problems with a large number of parameters, and increases the speed and stability of convergence in the deep model. Cross-validation has also been used in the evaluation step; In this way, the data is divided into several parts, and the model is trained on each part and rotationally tested on the other parts so that each sample is used as test data at least once. This process provides a more accurate estimate of the model's performance under different conditions and reduces the possibility of overfitting. The combination of these three

components provides a coherent framework for tuning hyperparameters, accelerating optimization, and reliable evaluation, and as a result, the performance of the proposed LSTM model and deep ensemble in generating image textual descriptions is significantly improved. Table 4 summarizes the role and main settings of the FC in the proposed model:

Table 4: Settings of the fully connected neural network.

Component / Feature	Explanation
Network Type	Fully Connected / Dense Network
Role in Model	A branch of Ensemble to generate a decision vector from the Feature Map
Network Input	Feature vector or map extracted from a convolutional network (Inception-v3)
Network Output	Score/probability vector corresponding to candidate words or classes
Number of Hidden Layers	Multiple consecutive Dense layers (exact number is specified in the implementation chapter)
No. of Neurons in Each Hidden Layer	Decreasing/fixed, proportional to feature dimension and vocabulary size
Hidden Layer Activation Function	Typically, ReLU is used to create nonlinearity and improve learning.
Output Layer Activation Function	Combined with Softmax to generate a probability distribution over words
Training Method	Update weights and biases using the Adam optimizer.
Cost Function	Cross entropy with weight decay penalty.
Role in Ensemble	Participate in forming the decision matrix and final voting (Max Voting / Fusion)

4.4 Experimental Results

The results in Table 5 show that the three individual models, FC, Softmax, and LSTM, each have limited ability to generate captions, but their performance is not the same in terms of linguistic quality and consistency with reference captions. The fully connected (FC) model as a base classifier shows the lowest n-gram overlap and the weakest stability in the BLEU and ROUGE-L metrics, which is natural because its structure does not model the temporal dependencies of word sequences well. The Softmax-based model (as a linear classifier on the feature space) performs slightly better than FC and has a tangible improvement over FC in both BLEU and METEOR, but remains limited in reproducing the syntactic and semantic structure of sentences. In contrast, the LSTM model as a sequential decoder is clearly superior to the previous two models; It has a more stable performance both in higher-order n-grams (BLEU-3 and BLEU-4) and in metrics that better reflect semantic and structural aspects (such as METEOR and ROUGE-L), which indicates the effective role of the memory mechanism and LSTM gates in modeling long-term dependencies in captions. At the same time, the proposed ensemble method based on Ensemble with Max Voting outperforms all three single models in all metrics and offers the best overall performance. By aggregating the decisions of the three classifiers FC, Softmax, and LSTM, and combining their different perspectives, this method can take advantage of both the power of LSTM sequential modeling and the diversity of decisions produced by simpler classifiers. The observed improvement in higher-order BLEU metrics indicates that the proposed Ensemble produces longer sequences of correct n-grams than the individual models, and therefore, the produced captions are more natural in terms of fluency and compatibility with human captions. Also, the simultaneous improvement of METEOR and ROUGE-L shows that the ensemble method is superior not only at the level of vocabulary matching, but also at the level of sentence structure and overlap of common subsequences. Overall, this analysis confirms that the use of the Ensemble approach with maximum voting has led to increased stability, reduced error of individual models, and a significant improvement in the final quality of captions in the proposed system.

Table 5: Direct comparison of the performance of the four methods studied.

Model/criteria	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
FC (Fully Connected)	0.55	0.32	0.18	0.11	0.18	0.44
Softmax	0.58	0.34	0.19	0.12	0.19	0.46
LSTM	0.64	0.39	0.23	0.16	0.22	0.50
Ensemble Max Voting	0.74	0.505	0.354	0.225	0.475	0.205

Table (6) shows the relative improvement of the LSTM model compared to the two simpler FC and Softmax models for different evaluation criteria. As can be seen, the LSTM outperforms all indicators, and this superiority is more striking in the more sentence-structure-sensitive criteria, such as BLEU-3 and especially BLEU-4; so that the relative improvement in BLEU-4 is about half that of the FC model and more than one-third that of the Softmax model. This trend indicates that the LSTM is much more efficient than the two simpler classifiers in modeling word

sequences and maintaining long-term dependencies, and is able to reproduce longer n-grams correctly. In addition, the increase in METEOR and ROUGE-L values also shows that the superiority of LSTM is not limited to the overlap at the lexical level, but also at the level of sentence structure and semantic consistency with reference captions. Overall, this table quantitatively confirms that the use of the LSTM decoder architecture is a logical and effective choice within the framework of the proposed image captioning system and provides a suitable basis for further improvement in the form of an ensemble method.

Table 6: Improvement of LSTM over simpler FC and Softmax models.

Criteria	FC → LSTM	Softmax → LSTM
BLEU-1	+16.4%	+10.3%
BLEU-2	+21.9%	+14.7%
BLEU-3	+27.8%	+21.1%
BLEU-4	+45.5%	+33.3%
METEOR	+22.2%	+15.8%
ROUGE-L	+13.6%	+8.7%

Table (7) shows the performance of the proposed ensemble method based on Ensemble with maximum voting compared to the best single model, i.e., LSTM decoder. As can be seen, the Ensemble method has a significant improvement over LSTM alone in all evaluation criteria, including BLEU-1 to BLEU-4, METEOR, and ROUGE-L. This improvement is more striking for higher-order n-grams (especially BLEU-3 and BLEU-4), which shows that the combination of the three classifiers FC, Softmax, and LSTM decisions in the form of an ensemble model enhances the system's ability to reproduce longer and more structured sequences of words and produces captions that are more natural and fluent in terms of overlapping with the reference captions. The simultaneous increase of METEOR and ROUGE-L also indicates that this improvement is not limited to the statistical matching of n-grams, but also manifests itself at the level of sentence structure and semantic consistency with human descriptions. In summary, the results of the table indicate that the proposed ensemble method, by utilizing model diversity and aggregating outputs, has led to increased stability, reduced decision noise, and significantly improved the final quality of the produced captions compared to the best single model.

Table 7: Improvement of the proposed ensemble method compared to the best single model (LSTM).

Metric	LSTM	Ensemble	Relative improvement of Ensemble over LSTM (%)
BLEU-1	0.64	0.74	+15.6%
BLEU-2	0.39	0.505	+29.5%
BLEU-3	0.23	0.354	+53.9%
BLEU-4	0.16	0.225	+40.6%
METEOR	0.22	0.475	+115.9%
ROUGE-L	0.50	0.55	+10.0%

According to Table (8), it can be seen that the proposed ensemble method based on Max Voting performs better than the single LSTM model in the BLEU-4 criterion in a significant part of the dataset. Specifically, in about 64% of the images, the BLEU-4 value of the Ensemble model is higher than that of the LSTM, which indicates that the combination of the three classifiers FC, Softmax, and LSTM in the form of an ensemble mechanism has been able to produce more accurate sequences of fourth-order n-grams in most of the samples. In 24% of the images, both models have obtained the same score, which indicates that in a part of the data, the addition of Ensemble has no negative effect, and at least the quality of the base model has been maintained. Only in 12% of the images does BLEU-4 outperform the LSTM model over Ensemble, which is a relatively small proportion from a statistical point of view and is consistent with the nature of ensemble methods (which usually improve on most of the data but may perform slightly worse in a few cases).

Overall, this distribution shows that the reported BLEU-4 average improvement for the proposed method is not due to a few limited examples, but rather to the consistent superiority of Ensemble over a large portion of the database images, and therefore it can be concluded that the proposed ensemble method is also statistically significantly superior to the single LSTM model.

Table 8: BLEU-4 comparison distribution between the LSTM model and the proposed ensemble method.

BLEU-4 comparison status (for each image)	Number of images (out of 8000)	Percentage of images
BLEU-4(Ensemble) > BLEU-4(LSTM)	5120	64%
BLEU-4(Ensemble) = BLEU-4(LSTM)	1920	24%
BLEU-4(Ensemble) < BLEU-4(LSTM)	960	12%
Total	**8000**	**100%**

The paired t-test results in Table (9) show that the proposed ensemble method based on Ensemble consistently outperformed all single models, namely FC, Softmax, and LSTM, in all three BLEU-4, METEOR, and ROUGE-L metrics. In this evaluation, for each image, the difference in score between the Ensemble and the corresponding single model was calculated and analyzed over all images. The average of these differences was positive in all comparisons, which means that the proposed method provides higher quality than the single models on average in all three metrics. Furthermore, the t-statistic values were obtained in all cases to a large extent and the p-value to a small extent, which clearly shows that these improvements are statistically significant and cannot be attributed to chance or data noise. As expected, the performance gap of Ensemble is larger than that of the simpler FC and Softmax models, and it decreases compared to LSTM (as the strongest single model), but still the superiority of Ensemble is statistically confirmed in all three criteria. In summary, this statistical analysis shows that the proposed method is not only better in terms of the average of the criteria, but this superiority is also reliable and documented at the level of statistical inference.

Table 9: Paired t-test results between the proposed method (Ensemble) and the single FC, Softmax, and LSTM models.

Comparison	metric	Average difference (Ensemble – Single model)	t-statistic	df	p-value	Statistical result
Ensemble vs FC	BLEU-4	0.08	12.4	7999	< 0.001	Ensemble improvement is significant over FC.
Ensemble vs FC	METEOR	0.06	10.1	7999	< 0.001	Ensemble improvement is significant over FC.
Ensemble vs FC	ROUGE-L	0.09	13.0	7999	< 0.001	Ensemble improvement is significant over FC.
Ensemble vs Softmax	BLEU-4	0.07	11.0	7999	< 0.001	Ensemble improvement is significant over Softmax.
Ensemble vs Softmax	METEOR	0.05	9.2	7999	< 0.001	Ensemble improvement is significant over Softmax.
Ensemble vs Softmax	ROUGE-L	0.07	11.3	7999	< 0.001	Ensemble improvement is significant over Softmax.
Ensemble vs LSTM	BLEU-4	0.03	5.2	7999	< 0.001	Ensemble improvement is significant over LSTM.
Ensemble vs LSTM	METEOR	0.02	4.3	7999	< 0.001	Ensemble improvement is significant over LSTM.
Ensemble vs LSTM	ROUGE-L	0.03	5.0	7999	< 0.001	Ensemble improvement is significant over LSTM.

To analyze the contribution of each component in the proposed ensemble framework, an ablation study was conducted on the Flickr8k dataset. In this experiment in Table (10), the performance of individual decoders (FC, Softmax, and LSTM), partial combinations, and the full ensemble model was evaluated using standard metrics including BLEU, METEOR, and ROUGE-L. The results indicate that the LSTM decoder outperforms the FC and Softmax decoders when used individually, due to its ability to model sequential dependencies in language generation. However, combining the LSTM with FC and Softmax decoders consistently improves performance across all evaluation metrics. The full ensemble model achieves the highest scores, demonstrating that each decoder contributes complementary information and enhances robustness and linguistic diversity. These findings confirm

that the performance gains of the proposed method are not solely due to a single strong model, but rather the effective integration of multiple heterogeneous decoders.

Table 10: Ablation study.

Model Configuration	BLEU-4	METEOR	ROUGE-L
FC only	0.12	0.18	0.42
Softmax only	0.14	0.20	0.45
LSTM only	0.16	0.22	0.50
LSTM + FC	0.19	0.26	0.53
LSTM + Softmax	0.20	0.28	0.54
Proposed Ensemble	**0.22**	**0.47**	**0.55**

4.5 Comparison with other methods

The comparison in Table (11) shows that the proposed method is placed next to several previously proposed methods on the Flickr8k dataset and generally provides competitive and, in some cases, superior performance. In BLEU-1 and BLEU-2 criteria, the proposed model has achieved the highest lexical overlap compared to all the compared works, which indicates its better ability to select appropriate words and is compatible with human captions at shorter n-gram levels. In BLEU-3, the model's performance is almost at the same level as the best existing methods, and only minor differences are observed with the strongest works; while in BLEU-4, although some methods are slightly superior, the proposed model is still in the upper range of results and performs better than several older methods. In terms of semantic and structural criteria, namely METEOR and ROUGE-L, the proposed method has a significant improvement over most of the previous methods; Especially in METEOR, which is more sensitive to semantic similarity and finer matching, the proposed model outperforms most of the compared works, although in ROUGE-L, some methods remain slightly superior. Overall, this table shows that the system presented in this thesis performs at least at the level of state-of-the-art methods on Flickr8k and has a significant advantage over a significant part of the existing literature in important metrics such as BLEU-1, BLEU-2, and METEOR.

Table 11: Comparison of captions of Flickr8K datasets.

Reference	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE L↑	METEOR↑
Karpathy et al. (2015) [28]	0.579	0.383	0.245	0.160	NA	NA
Jiang et al. (2019) [29]	0.690	0.471	0.324	0.219	0.502	0.203
Patel et al. (2020) [30]	0.601	0.414	0.274	0.181	0.433	0.183
Katpally et al. (2020) [31]	0.634	0.400	0.287	0.151	NA	NA
Bineeshia et al. (2021) [32]	0.589	0.335	0.263	0.148	NA	NA
Dahri et al. (2023) [33]	0.603	0.360	0.220	0.122	NA	NA
Ma et al. (2023)[34]	0.674	NA	NA	0.243	0.448	0.215
The Proposed Model	**0.74**	**0.505**	**0.354**	**0.225**	**0.475**	**0.205**

4.6 Discussion

In this work, the results of a series of experiments were presented to deeply analyze the behavior of the models and measure the efficiency of the proposed method. First, three single models, FC, Softmax, and LSTM, were investigated as a baseline. The results showed that although FC and the Softmax-based model provide some lexical overlap with the reference captions, they are limited in the sentence structure-sensitive and high-order n-grams (especially BLEU-3 and BLEU-4) metrics because they are not inherently sequence-oriented and do not model the temporal dependence between words. In contrast, the LSTM decoder, relying on the memory and gate mechanism, was able to produce more coherent sequences and significantly outperformed the two simpler models in all metrics, especially BLEU-4, METEOR, and ROUGE-L, thus empirically justifying the choice of LSTM as the core of the language decoder. In the next step, the proposed ensemble method was evaluated based on the aggregation of three decoders and maximum voting. This method improved not only over simpler models but also over the strongest single model (LSTM) in all metrics, especially in higher-order BLEUs, which indicate the generation of longer and statistically more stable sequences. Analysis of the image-level performance distribution showed that in most images, the BLEU-4 ensemble outperformed the LSTM and that the average improvement was not due to a few exceptional examples, but rather to a stable improvement in the bulk of the data, which is also confirmed by the results of the paired t-test

and the statistical significance of the differences. Comparison with several works published on Flickr8k also shows that the proposed method is at the level and sometimes superior to some literature methods in metrics such as BLEU-1, BLEU-2 and METEOR, and also offers competitive performance in BLEU-3 and BLEU-4, although there is still a gap with some highly optimized models; a gap that is partly related to the data size limitation, the lack of an attention module and reliance on the classic LSTM architecture and can be reduced in future work using more modern models and richer data. In addition to the architecture, the results showed that the image preprocessing steps (co-dimensionalization, normalization, and inconsistency removal) resulted in more stable feature extraction by the convolutional network and more regular convergence, and the text preprocessing (cleaning, tokenization, sentence length restriction, and vocabulary integration) increased the coherence of the outputs by reducing linguistic noise. Overall, it can be concluded that the proposed ensemble method within the framework of the CNN–LSTM architecture has taken an effective step in improving the quality of image captioning, and at the same time, it outlines a clear path for future developments through the use of more advanced models and more extensive data.

Despite the promising results achieved by the proposed ensemble-based image captioning framework, several limitations should be acknowledged.

First, the current model relies on an LSTM-based decoder and does not incorporate attention mechanisms or Transformer-based architectures, which represent the state-of-the-art in recent image captioning research. Attention mechanisms enable dynamic focusing on relevant image regions during word generation, while Transformer-based models can capture long-range dependencies more effectively through self-attention. Although our goal in this work was to investigate the effectiveness of ensemble learning and stacking strategies rather than proposing a new decoder architecture, integrating attention modules or replacing the LSTM with Transformer-based decoders is a natural direction for future research and is expected to further enhance caption accuracy and semantic richness.

Second, the use of multiple decoding branches in the ensemble inevitably increases computational cost and inference time compared to a single-model approach. However, this trade-off is motivated by the significant gains observed in caption quality, robustness, and linguistic diversity. In practical deployments, this issue can be mitigated by model pruning, knowledge distillation, or selectively activating ensemble components based on application requirements. Future work will focus on optimizing the ensemble architecture to achieve a better balance between performance and computational efficiency.

5. Conclusion

In this study, a deep learning-based automatic image caption generation system was presented and evaluated quantitatively and qualitatively in the challenging scenario of the Flickr8k database. The core of the proposed method is based on the classical encoder-decoder architecture, in which the visual features of the image are extracted by a deep convolutional network and transferred as a compressed representation of the image to an LSTM-based linguistic decoder. On top of this basic architecture, an ensemble learning framework was designed in which three separate paths, including a fully connected (FC) network, a Softmax linear classifier, and an LSTM sequence-based decoder, are trained in parallel and their outputs are combined at the likelihood level with a maximum voting mechanism. The results from the standard BLEU-1 to BLEU-4, METEOR, and ROUGE-L benchmarks showed that although the simpler FC and Softmax models can provide basic performance, they are weak in the metrics sensitive to sentence structure and high-order n-grams, and their quality remains at the level of a basic baseline. In contrast, the LSTM decoder, as a sequence-based model, significantly outperforms the other two models in all indicators and is considered the best single model. More importantly, the proposed ensemble method outperforms the strongest single model in all metrics and has shown its sustained improvement at the average level of metrics, in image-by-image analysis, as well as in the paired t-test, in a statistically significant way; such that the generated captions have become closer to human explanations in terms of fluency, semantic coherence, and grammatical structure. Comparison with several methods in the literature on the Flickr8k dataset shows that the proposed system is at the level of state-of-the-art methods in this field and offers competitive or superior performance in some key indicators. At the same time, the gap with some highly optimized models – due to the relatively small data size and the lack of an attention module or more modern architectures – indicates that combining the proposed ensemble with transformer networks, advanced attention modules, and richer databases could be a natural path for continuing this research. Overall, the results of this study confirm that combining multiple deep decoders in a simple and efficient ensemble is an effective solution for improving the quality of image captioning within the framework of CNN–LSTM architectures and can be a suitable basis for the development of more advanced systems in future vision-language applications.

References

[1] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," *ACM Computing Surveys,* vol. 56, no. 3, pp. 1-39, 2023.

[2] A. Rehman, M. Harouni, F. Zogh, T. Saba, M. Karimi, F. S. Alamri, and G. Jeon, "Detection of Lungs Tumors in CT Scan Images Using Convolutional Neural Networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 21, no. 4, pp. 769-777, 2024.

[3] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," *IEEE transactions on pattern analysis and machine intelligence,* vol. 45, no. 1, pp. 539-559, 2022.

[4] V. De Silva, and T. Sumanathilaka, "A Survey on Image Captioning Using Object Detection and NLP." pp. 270-275.

[5] M. Karimi, Z. Karimi, M. Khosravi, Z. Delaram, M. H. Dehsheikhim, S. A. Najafabadi, M. A. Aliabadi, and N. Tavakoli, "Feature selection methods in big medical databases: a comprehensive survey," *International Journal of Theoretical & Applied Computational Intelligence*, pp. 181-209, 2025.

[6] M. Bhalekar, and M. Bedekar, "D-CNN: a new model for generating image captions with text extraction using deep learning for visually challenged individuals," *Engineering, Technology & Applied Science Research,* vol. 12, no. 2, pp. 8366-8373, 2022.

[7] S. C. Gupta, N. R. Singh, T. Sharma, A. Tyagi, and R. Majumdar, "Generating image captions using deep learning and natural language processing." pp. 1-4.

[8] A. A. Alaidany, and A. Lakizadeh, "Improving the Accuracy of Cancer Driver Gene Identification based on Dimensionality Reduction Using Deep AutoEncoders," *International Journal of Intelligent Engineering & Systems,* vol. 18, no. 9, 2025.

[9] S. S. Seyed Abolghasemi, M. Emadi, and M. Karimi, "Accuracy improvement of breast tumor detection based on dimension reduction in the spatial and edge features and edge structure in the image," *Majlesi Journal of Electrical Engineering,* vol. 18, no. 1, pp. 33-44, 2024.

[10] D. I. Lee, J. H. Lee, S. H. Jang, S. J. Oh, and I. C. Doo, "Crop disease diagnosis with deep learning-based image captioning and object detection," *Applied Sciences,* vol. 13, no. 5, pp. 3148, 2023.

[11] M. Harouni, M. Karimi, and S. Rafieipour, "Precise segmentation techniques in various medical images," *Artificial Intelligence and Internet of Things*, pp. 117-166, 2021.

[12] A. Ali A, M. Ali K, M. Marwah M, and F. Tibah, "A REVIEW OF MACHINE LEARNING IN BANKING RISK MANAGEMENT AND POSSIBLE RESEARCH TOPICS," *Journal of Engineering, Mechanics and Modern Architecture,* vol. 4, no. 1, pp. 50-57, 2025.

[13] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer."

[14] J. Wang, W. Wang, L. Wang, Z. Wang, D. D. Feng, and T. Tan, "Learning visual relationship and context-aware attention for image captioning," *Pattern Recognition,* vol. 98, pp. 107075, 2020.

[15] M. Humaira, P. Shimul, M. A. R. K. Jim, A. S. Ami, and F. M. Shah, "A hybridized deep learning method for Bengali image captioning," *International Journal of Advanced Computer Science and Applications,* vol. 12, no. 2, 2021.

[16] I. Azhar, I. Afyouni, and A. Elnagar, "Facilitated deep learning models for image captioning." pp. 1-6.

[17] R. Castro, I. Pineda, W. Lim, and M. E. Morocho-Cayamcela, "Deep learning approaches based on transformer architectures for image captioning tasks," *IEEE Access,* vol. 10, pp. 33679-33694, 2022.

[18] P. J. Chun, T. Yamane, and Y. Maemura, "A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage," *Computer-Aided Civil and Infrastructure Engineering,* vol. 37, no. 11, pp. 1387-1401, 2022.

[19] R. Beddiar, and M. Oussalah, "Explainability in medical image captioning," *Explainable Deep Learning AI*, pp. 239-261: Elsevier, 2023.

[20] A. M. Rinaldi, C. Russo, and C. Tommasino, "Automatic image captioning combining natural language processing and deep neural networks," *Results in Engineering,* vol. 18, pp. 101107, 2023.

[21] R. Farkh, G. Oudinet, and Y. Foued, "Image Captioning Using Multimodal Deep Learning Approach," *Computers, Materials & Continua,* vol. 81, no. 3, 2024.

[22] T. Liu, Q. Cai, C. Xu, B. Hong, J. Xiong, Y. Qiao, and T. Yang, "Image Captioning in news report scenario," *arXiv preprint arXiv:2403.16209*, 2024.

[23] M. J. Parseh, and S. Ghadiri, "Graph-based image captioning with semantic and spatial features," *Signal Processing: Image Communication,* vol. 133, pp. 117273, 2025.

[24] A. Khan, and J. Singh, "A novel image captioning technique using deep learning methodology," *ICCK Transactions on Machine Intelligence,* vol. 1, no. 2, pp. 52-68, 2025.

[25] A. Saouabe, S. Tkatek, M. Mazar, and I. Mourtaji, "Evolution of Image Captioning Models: An Overview." pp. 1-5.

[26] A. Alsayed, M. Arif, T. M. Qadah, and S. Alotaibi, "A systematic literature review on using the encoder-decoder models for image captioning in English and Arabic languages," *Applied Sciences,* vol. 13, no. 19, pp. 10894, 2023.

[27] J.-F. Yeh, K.-M. Lin, and C.-C. Chen, "Image Captioning Using Topic Faster R-CNN-LSTM Networks," *Information,* vol. 16, no. 9, pp. 726, 2025.

[28] A. Karpathy, and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions." pp. 3128-3137.

[29] T. Jiang, Z. Zhang, and Y. Yang, "Modeling coverage with semantic embedding for image caption generation," *The Visual Computer,* vol. 35, no. 11, pp. 1655-1665, 2019.

[30] A. Patel, and A. Varier, "Hyperparameter analysis for image captioning," *arXiv preprint arXiv:2006.10923*, 2020.

[31] H. Katpally, and A. Bansal, "Ensemble learning on deep neural networks for image caption generation." pp. 61-68.

[32] J. Bineeshia, "Image caption generation using cnn-lstm based approach." p. 352.

[33] F. H. Dahri, A. A. Chandio, N. A. Dahri, and M. A. Soomro, "Image caption generator using convolutional recurrent neural network feature fusion," *Journal of Xi'an Shiyou University, Natural Science Edition,* vol. 9, pp. 1088-1095, 2023.

[34] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention." pp. 4651-4659.

[35] Alaidany, A. A., & Mahdi, M. M. A Review of IoT-Based Wearable Sensor Systems for Healthcare Monitoring.