Baghdad Science Journal

Volume 23 | Issue 4 Article 13

4-23-2026

A Multi-Model Ensemble Framework for Assistive Image Captioning with Voice Interaction for the Visually Impaired Users

Alaa Noori Mazher
Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq,
alaa.nouri2401p@sc.uobaghdad.edu.iq

Ghadah K. AL-Khafaji
Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq,
ghada.toma@sc.uobaghdad.edu.iq

How to Cite this Article

Mazher, Alaa Noori and AL-Khafaji, Ghadah K. (2026) "A Multi-Model Ensemble Framework for Assistive Image Captioning with Voice Interaction for the Visually Impaired Users," *Baghdad Science Journal*: Vol. 23: Iss. 4, Article 13.
DOI: https://doi.org/10.21123/2411-7986.5269

BAGHDAD SCIENCE JOURNAL 2026;23(4):1299–1315

Scan the QR to view the full-text article on the journal website

RESEARCH ARTICLE

A Multi-Model Ensemble Framework for Assistive Image Captioning with Voice Interaction for the Visually Impaired Users

Alaa Noori Mazher *, **Ghadah K. AL-Khafaji**

Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq

ABSTRACT

Visual impairment greatly restricts an individual's ability for perception and interaction with the environment, making everyday activities challenging without the availability of appropriate assistive technologies. One major barrier is the inability to perceive visual scenes, which are essential for spatial awareness, wayfinding, and situational knowledge. This study aims to address this gap through the proposal of an intelligent assistive image captioning system specifically designed for visually impaired individuals. The system employs the best available computer vision and natural language processing techniques in generating context-oriented text captions for an image infused with interactive voice responses. Here, four models of Convolutional Neural Networks, namely InceptionV3, InceptionResNetV2, Xception, and DenseNet201, with LSTM-based decoders, are used each to generate the initial captions. The additional captions were generated using a transformer-based ViT-GPT2 architecture. An ensemble method is used to choose the best caption, which is done using BLEU scores. For audio, Google Text-to-Speech is used, while for real-time voice queries, YOLOv8n is used to detect humans. We have tested our system using the Flickr8k dataset, and the results show that the ensemble method outperforms CNN-LSTM and ViT-GPT2 architectures. To be precise, the ensemble method achieved a BLEU-1 score of 0.7363, a BLEU-4 score of 0.2642, a METEOR score of 0.4545, and a ROUGE-L score of 0.5107. This shows that using multiple models is helpful to obtain better captions, thus increasing interactivity, which is a significant step towards real-time, accessible, and intelligent assistance for visually impaired individuals.

Keywords: CNN-LSTM models, Ensemble learning, Image captioning, Visually impaired, ViT-GPT2 transformer

Introduction

The ability to interpret and perceive visual scenes is important for human interaction with the world. In visually impaired or blind individuals, the lack of visual information significantly handicaps autonomy, spatial orientation, and situation awareness. The World Health Organization states that more than 285 million individuals worldwide have some form of visual impairment, making successful assistive technologies essential at the social level.[1] Traditional mobility aids such as guide canes or tactile maps address some issues but fail to provide dynamic, context-dependent information. Recent advances in artificial intelligence (AI), particularly in computer vision and natural language processing (NLP), have opened up new possibilities for smart systems that can interpret images and generate descriptive narratives to bridge this gap.[2,3]

A number of ways have been proposed to automatically image captioning, which primarily rely on visual-linguistic deep learning-based architectures. Early models employed convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs), particularly long short-term memory (LSTM) units, for generating sequential text.[4,5] These approaches could capture local and temporal dependencies but were prone to

Received 7 May 2025; revised 5 August 2025; accepted 27 August 2025.
Available online 23 April 2026

* Corresponding author.
E-mail addresses: alaa.nouri2401p@sc.uobaghdad.edu.iq (A. N. Mazher), ghada.toma@sc.uobaghdad.edu.iq (G. K. AL-Khafaji).

https://doi.org/10.21123/2411-7986.5269

failure in the event of long-range contextual reasoning as well as lacking attention mechanisms for identifying critical regions of the image.

To address these limitations, attention-based encoder-decoder models were introduced, enabling models to selectively pay attention to significant visual features while outputting words.[6,7] This framework significantly improved caption coherence and consistency with visual content. More recently, transformer-based models such as ViT-GPT2 and BLIP-2 have been used because they have superior ability to learn global dependencies between modalities by employing self-attention layers to simultaneously process image patches and text tokens.[8–10]

Despite improvements in architectural design, traditional models of captions have limited utility when used in assistive settings. Most traditional models provide pre-defined, generic captions, which do not capture essential perceptual details that are critical to visually impaired individuals, such as the number of individuals, colors, or spatial locations of objects.[11,12] Furthermore, traditional captions rarely offer voice-enabled questions, which limits their ability to offer contextual feedback. Therefore, the inclusion of expert modules for object recognition, color analysis, and multimodal interaction remains essential to propel captioning towards practical, aid-oriented applications.[13–16]

Visually impaired users' assistive captioning systems not only have requirements beyond general image description work. Voice interaction, query answering for individual users, and human language explanation of visual things need to be done by such systems. Application includes helping users perceive what is nearby, detect obstacles or barriers, detect people in the case of people, or detect things in living and public spaces.[17–19] Upcoming work also puts emphasis on modularity, light-weight architectures with object detection, scene classification, and attribute recognition combined into a unified pipeline to be executed on edge devices in real-time.[20]

The main contribution of this paper is the development of a unified, modular ensemble framework that combines four CNN-LSTM pipelines (InceptionV3, InceptionResNetV2, Xception, DenseNet201) with a ViT-GPT2 transformer to generate diverse caption candidates; the introduction of a BLEU-1–driven selection mechanism that dynamically chooses the most accurate caption at inference time without additional fusion training; the seamless integration of Google Text-to-Speech for real-time audio feedback and YOLOv8n for voice-driven object detection; the design of a lightweight, edge-compatible architecture for deployment on resource-constrained assistive devices; and the incorporation of an interactive voice query module enabling users to ask context-specific questions (e.g., "How many people are present?") and receive immediate spoken responses. While most image captioning systems today focus solely on optimizing for accuracy metrics, our work presents the first comprehensive assistive ecosystem designed for visually impaired people. This is a major step towards a paradigm shift from merely providing image captions to offering active and vocal visual assistance.

Related work

Numerous works have followed the endeavor of automatic image captioning, seeking to generate natural language descriptions from visual content. The early years used template- and retrieval-based methods, resulting in generic and rigid captions. With deep learning, the models began to leverage CNNs for feature extraction and RNNs for generating language, resulting in higher fluency and coherence. Attention mechanism integration into transformer models further enhanced the extent of correspondence among visual regions and words that had been created.

In,[21] the authors address the challenge of improving image captions' quality and contextuality by introducing an ensemble model using a transformer encoder-decoder architecture with attention. Their method combines several CNN-based feature extractors and transformer-based caption generators and a voting system to select the caption with the highest BLEU score. The study used Flickr8K and Flickr30K datasets for training and evaluation. The model achieved 0.728 and 0.798 for BLEU-1, respectively.

In,[22] the authors propose a hybrid model that integrates Neural Image Captioning (NIC) and k-Nearest Neighbor (kNN) methods to enhance caption generation performance. The NIC module employs an InceptionV3 CNN with LSTM, while the kNN employs visually similar images to obtain captions. A logistic regression classifier selects the best caption from features derived from both models. The model was trained and tested on the Flickr8K dataset and achieved 59.67 BLEU-1 and 18.20 BLEU-4.

In,[23] the authors propose a fusion-based architecture with pretrained Auxiliary Language Models (AuxLMs) such as BERT integrated into traditional encoder-decoder architectures for caption generation and error correction. The model uses a ResNet-101 CNN, LSTM decoder, and a range of fusion techniques (Simple, Cold, and Hierarchical Fusion) to improve the quality of the generated captions. Experiments on the Flickr8k dataset show that the Simple Fusion approach gives the highest BLEU-1 score of 64.7 and BLEU-4 score of 22.8.

In,[24] the authors address the challenge of training effective image captioning models from small datasets with limited computation resources. They introduce enhancements like multi-level attention based on object-level and convolutional features, language model rescoring during inference, and training caption augmentation by paraphrasing. On the Flickr8k dataset, their whole model achieves 68.6 BLEU-1, 48.5 BLEU-2, 34.7 BLEU-3, 24.5 BLEU-4, 23.2 METEOR, and 49.2 CIDEr-D.

In,[25] the authors investigate the employment of an image captioning model using VGG16 for feature extraction and LSTM with attention for generation. Training the model with the Flickr8k dataset (8092 images) on 6000 images and testing with the BLEU score obtained a highest BLEU-4 score of 18.2%.

In,[26] a hybrid model that combines an image feature vector and a partial caption vector is introduced to generate image descriptions. The method processes the text sequence, extracts features of images, and decodes the output through a combination of these two levels. With the Flickr8k dataset (6000 for training, 1000 for test, and 1000 for dev), the model was tested by generating multiple sentences with Beam Search and BLEU metrics evaluation. The optimal performance was achieved through Beam Search k=5 with a BLEU-1 of 66.9%, BLEU-2 of 46.3%, BLEU-3 of 29.2%, and a BLEU-4 of 23.2%.

In,[27] another enhanced image captioning model that utilizes a better visual attention mechanism was proposed. The model is designed with an encoder-decoder framework, where the encoder extracts visual features using a single CNN (ResNet-101, EfficientNet-B0, ResNeXt-101) or a Dual-CNN setting (ResNet-101, EfficientNet-B0) and the decoder utilizes LSTM to produce sequence words. The captioning generation was maximized using beam search with beam size k = 3. The experiments conducted on the Flickr8k dataset (6000 for training, 1000 for validation, and 1000 for testing) showed that ResNet-101 and EfficientNet-B0 with a Dual-CNN encoder, visual attention mechanism, and early stopping worked better. The proposed approach yielded BLEU-1 of 68.76%, BLEU-2 of 49.15%, BLEU-3 of 35.46%, and BLEU-4 of 24.71%.

In,[28] a comparative study was proposed to analyze the performance of image captioning based on two deep learning networks: ResNet-50 and VGG16. The model captures the image features using ResNet-50 or VGG16 encoders and employs an LSTM network to generate the related captions. Flickr8k dataset (6000 training images, 1000 for validation, and 1000 for test) was used for experimentation. Both models were trained for 20 epochs, and their outputs were tested against BLEU scores. The findings confirmed that ResNet-50 outperformed VGG16, with BLEU-1 at 61.9%, BLEU-2 at 45.2%, BLEU-3 at 36.8%, and BLEU-4 at 26.2%, while those of VGG16 were lower. This confirms the superior capability of ResNet-50 to generate accurate and contextual captions for the Flickr8k dataset.

Proposed methodology

The proposed assistive image captioning system is designed to benefit visually impaired people by generating descriptive captions and enabling interactive audio-based feedback. The system workflow begins by loading images from Flickr8k, which is preprocessed to resize into normalized sizes and formats suitable for deep learning models. Feature extraction is done using four pre-trained convolutional neural networks (CNNs): InceptionV3, InceptionResNetV2, Xception, and DenseNet201. Every CNN extracts high-level semantic features from the input images, producing different visual representations. These are then taken as inputs to LSTM-based decoder models, which are learned to generate similar natural language captions. In conjunction with this, a Vision Transformer (ViT) model in combination with a GPT-2 language model (ViT-GPT2) is employed to generate individual caption candidates one at a time.

For further increasing the accuracy and diversity of generated descriptions, an ensemble technique is used. Captions produced by the four CNN-LSTM models and the ViT-GPT2 model are scored with BLEU-1 against reference captions, and the caption that has the highest score is selected as the final output. Besides caption generation, the system includes a Google Text-to-Speech (gTTS) engine to speak the selected caption aloud to the user, which brings accessibility. Furthermore, to enhance engagement, the user can make voice queries, e.g., ask the number of people detected in the scene. In response, a YOLOv8n object detection algorithm is applied on the image by the system to detect the accurate number of people correctly and in real time.

This multi-component, modular architecture not only offers accurate and detailed visual descriptions but also offers interactive features specially targeted at fulfilling real-time perceptual requirements of visually impaired users. With the integration of feature extraction, caption generation, model ensembling, speech synthesis, and object detection, the proposed model offers a complete solution for accessible image understanding. Fig. 1 shows the proposed system.

Dataset

The Flickr8k dataset is a benchmark dataset for vision-language and image captioning research. It was introduced to facilitate the training and

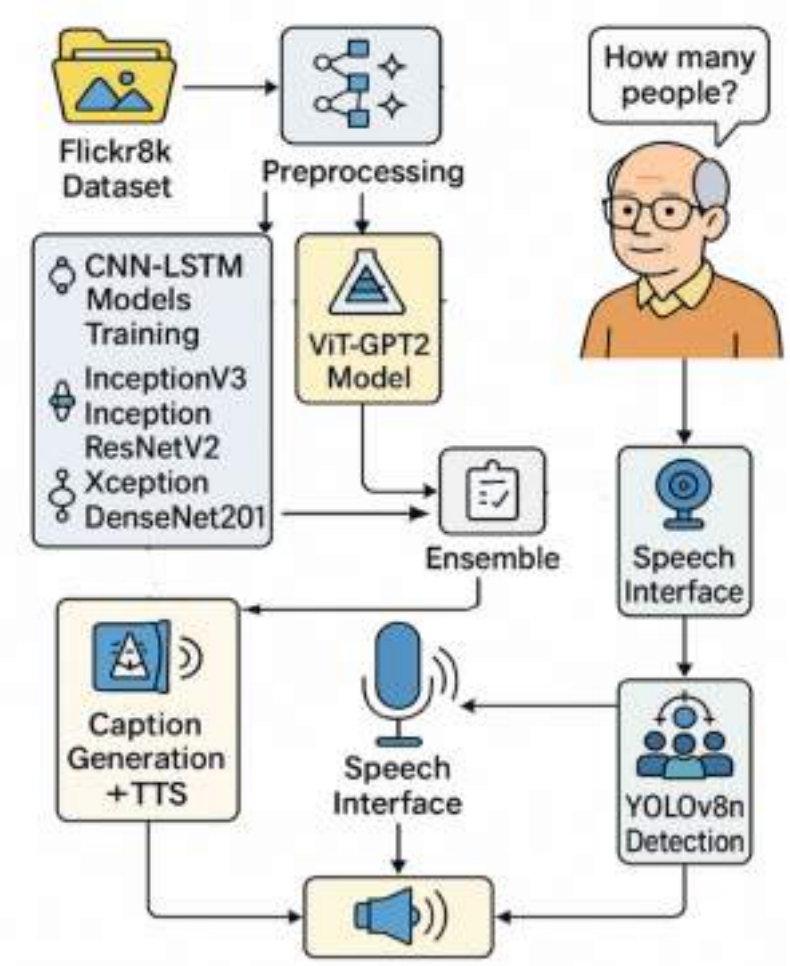

Fig. 1. Assistive image captioning and query system.

evaluation of models capable of describing images in natural language. The dataset consists of 8,092 photos crawled from the Flickr internet photo-sharing website. The photos contain general scenes with human beings, animals, or objects performing various actions or in various settings. To enable supervised learning for the generation of captions, each image in Flickr8k has five independent human-made captions that explain the visual content. The captions use simple, grammatically correct English and typically range between 8 and 20 words in length. a sample from the Flickr8k dataset including a human-annotated caption. Fig. 2 show sample from Flickr8k.

Preprocessing stage

Before model training, visual and textual data must be preprocessed in a manner such that they are properly formatted and standardized. Preprocessing within this research is divided into two general categories: image preprocessing and text preprocessing. The two-faceted preprocessing guarantees that the two modalities are aligned and optimized for effective feature extraction and sequence modeling.

Image preprocessing

The images in the Flickr8k dataset underwent several preprocessing activities so as to make them suitable for the convolutional neural networks (CNNs) used in the feature extraction process. All images are first normalized to fit the input dimensions required by the chosen CNN models. This means that each image is resized to 299 × 299 pixels or 224 × 224 pixels, depending on which CNN model is being used.

After that, each pixel value is divided by 255 to normalize the input data to a range between 0 and 1, replacing the original range of 0 to 255. This is to stabilize the model during training, as it ensures that each input to the neural network has a certain numeric characteristic.

After normalization, the images were then subjected to pre-trained CNN models to extract high-level semantic features from the images. The extracted feature vectors were then stored in a store for faster

Fig. 2. Sample from the Flickr dataset, including a human-annotated caption.

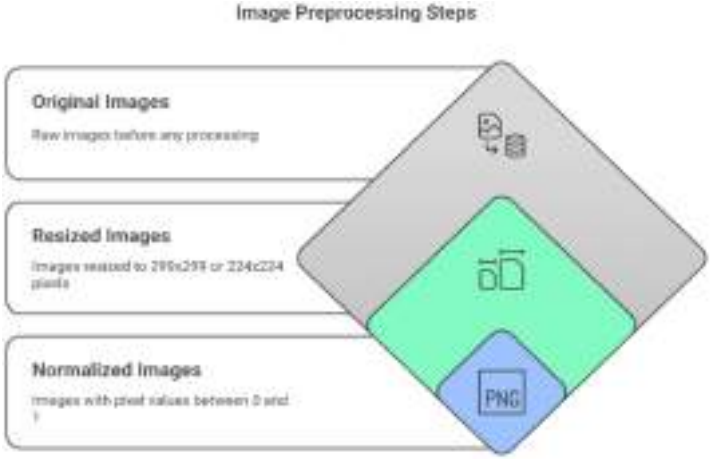

Fig. 3. Image preprocessing.

access during training and testing of the caption generation models. This avoids repeated extraction of features and saves computation time, thereby speeding up the training process. Fig. 3 shows the image preprocessing stage.

Text preprocessing

At the same time as the preprocessing of the images, the text associated with the images also went through the process of cleaning and structuring. The text associated with the images was cleaned by first removing

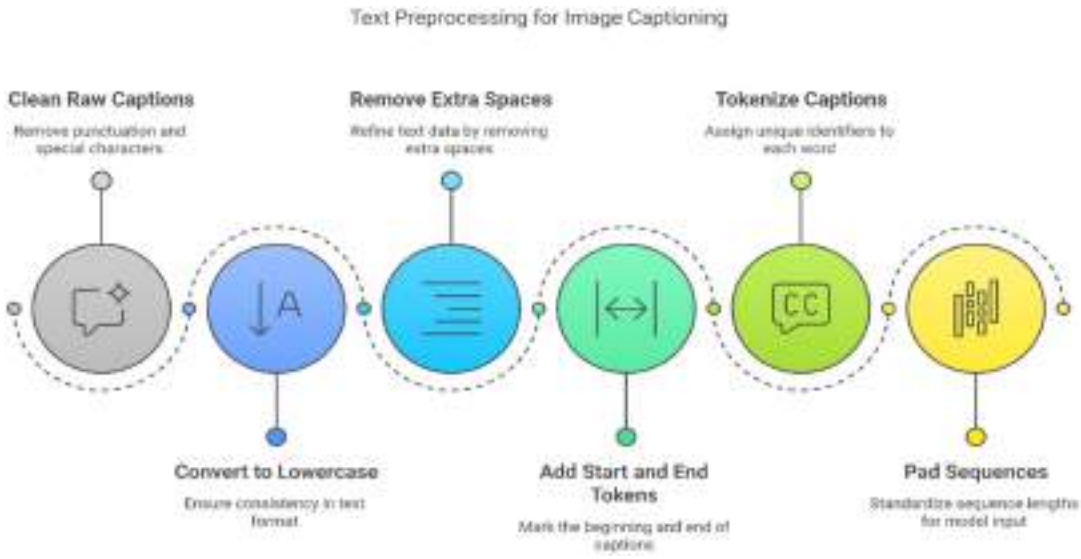

Fig. 4. Text preprocessing.

Algorithm 1: Text Preprocessing.

Input: Raw caption text
Output: Padded sequences ready for training

Step 1: Remove punctuation using regex: [^\w\s]
Step 2: Convert to lowercase using .lower()
Step 3: Remove numbers using regex: \d+
Step 4: Tokenize using split() method
Step 5: Add special tokens: ['<start>'] + tokens + ['<end>']
Step 6: Create vocabulary mapping: word → index
Step 7: Convert to sequences using vocabulary
Step 8: Convert to sequences using vocabulary

For instance, the raw caption 'A brown dog is running in the park!' undergoes the following transformation:
After cleaning: 'a brown dog is running in the park'
After tokenization: ['a', 'brown', 'dog', 'is', 'running', 'in', 'the', 'park']
After special tokens: ['<start>', 'a', 'brown', 'dog', 'is', 'running', 'in', 'the', 'park', '<end>']
After indexing: [1, 15, 267, 45, 12, 189, 8, 23, 156, 2]
After padding (max_length = 20): [1, 15, 267, 45, 12, 189, 8, 23, 156, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]"

all the punctuation marks, special characters, and numbers. Then the text was converted into lowercase format to maintain consistency with the data. Also, the spaces were removed.

In order to create the captions for the sequence modeling, the following changes were made: a "start" token was added at the beginning of each caption, and an "end" token was added at the end of each caption. This is to ensure the sequence generation models have a clear idea of the start and end points of each caption.

After token addition, the captions were tokenized by creating a vocabulary index with a unique integer identifier for each unique word. The padded tokenized data is created by finding the maximum length of the captions in the data and then padding to that length. The padding is necessary because it will ensure that all the input data for the models will be of the same dimension, and this is necessary for the batch processing and convergence of the models. Through the above processes, the images and the data were translated into a standard form that is compatible with deep learning, and this is a sound foundation for feature extraction and caption generation. Fig. 4 illustrates the text preprocessing stage.

Feature extraction

After this preprocessing step, the next important step is to extract important features from the images that will help in accurate and contextually relevant caption generation. In this study, four pre-trained convolutional neural network models were used for this purpose, and they are InceptionV3, Inception-ResNetV2, Xception, and DenseNet201. These models were chosen for this study because they have shown promising results in extracting important features from images and performing well in various computer vision-related tasks.

All resized and normalized images from the Flickr8k dataset were passed through the convolutional base of the selected CNNs. The classification layers typically added to the models were removed, and the feature vectors were extracted from the last global average pooling layers. This way, only the required abstract visual representations were retained and discarded classification-specific outputs that do not generalize well for caption generation tasks.

The feature vectors thus obtained were of varying dimensions depending on the specific architecture employed. For instance, InceptionV3 and Xception produced feature vectors of 2048 dimensions, InceptionResNetV2 produced feature vectors of dimension 1536, and DenseNet201 produced feature vectors of dimension 1920. Such dense feature embeddings preserve dense spatial and contextual information regarding the visual scenes depicted in the images that is useful to generate rich and coherent textual descriptions.

In order to enhance training efficiency and avoid redundant computation, all the feature vectors extracted were cached to disk. This method of caching is efficient in retrieving the features without the need to reprocess the images using the CNN models. The utilization of multiple CNNs in feature extraction allows the system to take advantage of the strengths of different CNN models, thus increasing the diversity in the caption generation system.

With the aid of an extended feature extraction process, the visual input data has been successfully transformed into a rich and informative representation, paving the way for the next stages of sequence modeling and caption generation. The detailed workflow of the feature extraction phase is depicted in Fig. 5. The sequential steps involved in the feature extraction phase include image resizing, pixel value normalization, feature extraction with the aid of pre-trained convolutional neural networks (InceptionV3, InceptionResNetV2, Xception, DenseNet201), and caching the feature vectors.

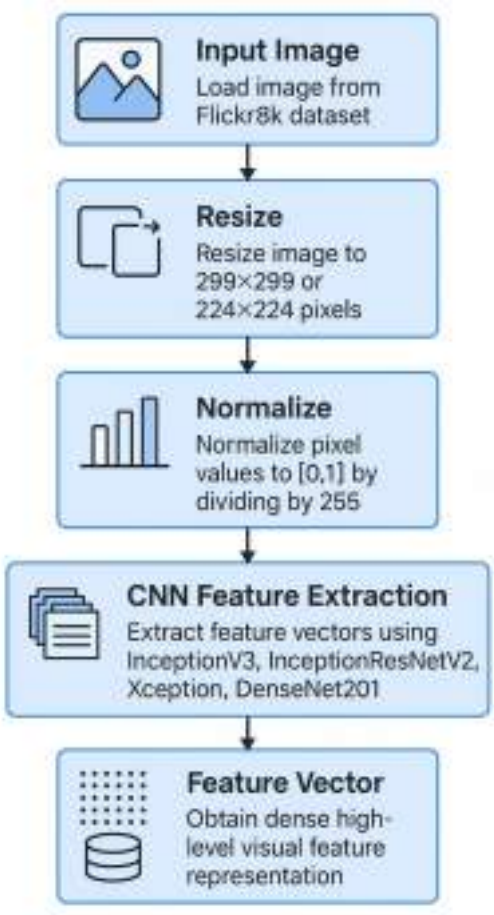

Fig. 5. Feature extraction stage.

Caption generation models

After extracting high-level visual features from images using pre-trained CNN models, the next crucial stage involves generating descriptive captions. Two distinct caption generation approaches were employed in this work: a classical sequence modeling architecture using Long Short-Term Memory (LSTM), and a transformer-based method using the Vision Transformer (ViT) and GPT-2. Both architectures take as input a visual representation of the image, but differ in how they process and generate the caption.

CNN-LSTM-Based captioning

In this model architecture, each image is modeled as a dense feature vector derived from one of the pre-trained CNN models (InceptionV3, Inception-ResNetV2, Xception, or DenseNet201), as shown in Fig. 6. The visual feature vector is first passed through a dropout layer with a dropout rate of 0.5, and then through a 256-neuron dense layer with ReLU activation. This prevents overfitting and projects

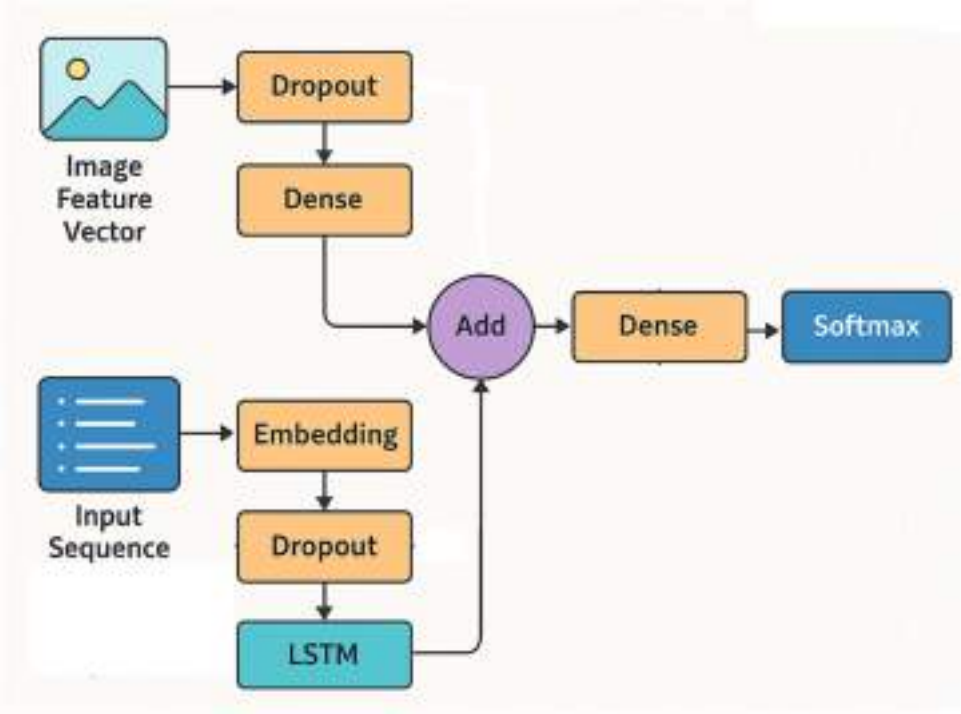

Fig. 6. CNN-LSTM-based captioning.

the input into a more compact and descriptive representation.

Simultaneously, the partial input caption, which is a sequence of word indices, is fed through an embedding layer with output size 256. It is then fed through a dropout layer (dropout rate = 0.5) and a single-layer LSTM with 256 hidden units to extract the temporal relationships in the sequence.

The visual and textual branch outputs are summed using an element-wise addition operation (add layer). The summed vector is fed through a fully connected dense layer of 256 units and ReLU activation. A SoftMax output layer ultimately predicts the next word in the vocabulary, with vocabulary size determined dynamically from the dataset tokenization process. The model is optimized with categorical cross-entropy loss using the Adam optimizer. This architecture is used uniformly for four CNN-based pipelines, which all utilize the same LSTM decoder. Uniform use of the decoder architecture allows an apples-to-apples comparison of how different CNN encoders impact caption quality.

Transformer-Based captioning with ViT-GPT2

In contrast to the sequential LSTM-based model, the transformer-based model uses a Vision Transformer (ViT) as the image encoder and a GPT-2 decoder to generate captions. The ViT module divides the input image into fixed-size patches, embeds them, and applies self-attention layers to produce a sequence of image tokens. These tokens are then passed into the GPT-2 model to generate the caption word by word using its multi-head attention and positional encoding mechanisms.

This method utilizes the parallel computing ability of transformers, as well as their ability to process longer-term dependencies compared to recurrent networks. Furthermore, the pre-training on text data, which is characteristic of GPT-2, greatly aids in increasing the variety and coherence of generated captions.

Using both transformer-based and CNN-LSTM captioning methods, the system enjoys the advantages of sequence modeling as well as attention-based architecture. The output captions of each individual model are subsequently evaluated individually, along with when collectively used using an ensemble method, in order to select the most contextually appropriate description of every image.

Caption selection using an ensemble model

Even if single captioning models can generate semantically accurate descriptions, their outputs will differ in terms of detail, phrasing, or accuracy. To counteract this variation and improve overall caption quality, an ensemble strategy was used to select the optimal caption per image. Instead of generating a

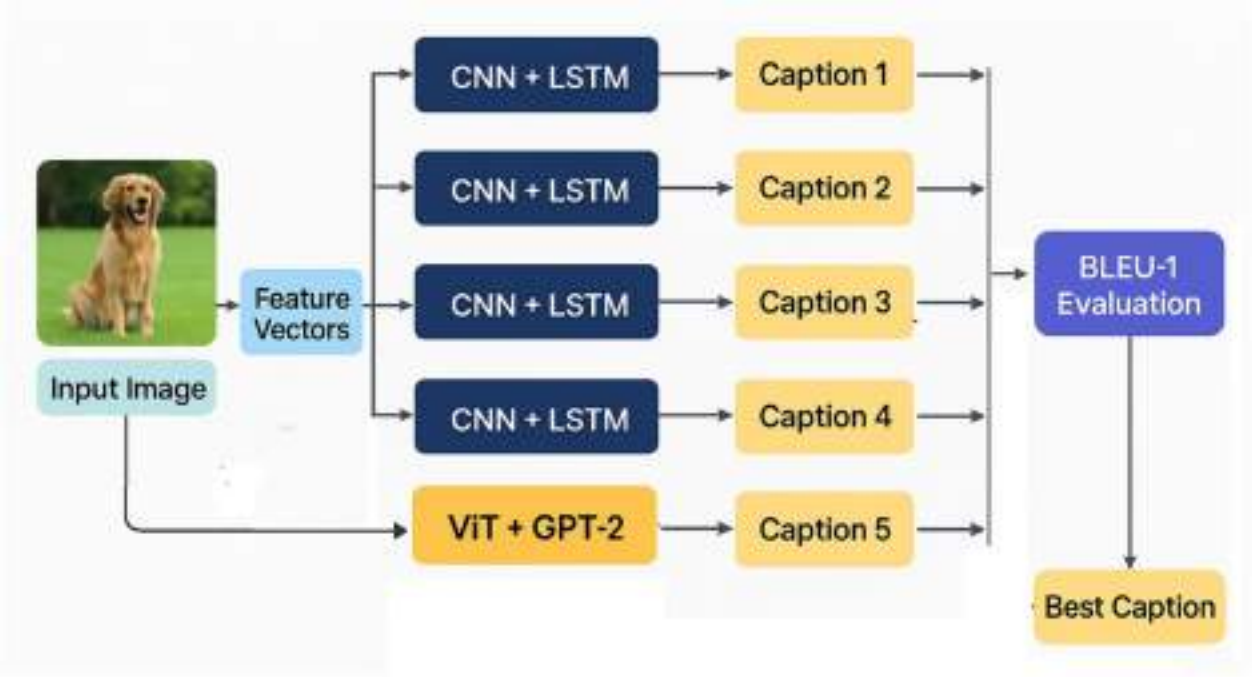

Fig. 7. Caption selection using ensemble model.

new caption from aggregated model responses, this method applies scoring to the candidate captions generated by multiple models and selects the top-scoring one.

The test dataset is composed of the output generated by five models. The models include four CNN-LSTM-based captioning pipelines: InceptionV3, InceptionResNetV2, Xception, and DenseNet201. Additionally, the test dataset incorporates a transformer-based captioning model named ViT-GPT2. Each test image is processed by all five models to generate captions. The candidate captions are then assessed using the BLEU-1 metric, which measures the n-gram overlap between the generated captions and the reference captions in the Flickr8k dataset.

The BLEU-1 is calculated for each candidate caption against its respective ground-truth references. The best caption is selected as the final output based on the maximum calculated BLEU-1 score. The voting procedure on the basis of metrics ensures that the selected caption is not only syntactically correct but also closer to semantically human-written descriptions. The ensemble model hence derives the strengths of the distinct captioning models and overcomes any weakness of a single model.

This efficient and simple selection technique makes it easier to produce more accurate and robust captions without injecting trainability into a fusion model to allow its functioning. The selection also maintains modularity for easy integration in the future of other assessment metrics or captioning models into the core design. Fig. 7 illustrates caption selection using the ensemble model.

Text-to-speech integration

To enable the visually impaired to use the captioning system, the selected caption generated by the ensemble model is converted into audible speech using a text-to-speech (TTS) engine. This capability provides real-time and user-focused access to image contents without any visual interaction, which aligns with the assistive objectives of the system.

Within this work, the generated captions have been spoken by the Google Text-to-Speech (gTTS) API. The gTTS delivers natural-sounding voice synthesis reliably across languages with the capacity for high-quality voices of English and others. The gTTS engine uses the best caption determined via the BLEU-based assessment as the input to transform into text to speech. It converts the given text and yields a respective MP3 file representing an audio version of it.

The resulting audio is output to the user by means of an embedded audio output interface. The module allows the user to hear the visual scene description instantly, thereby making the system interactive and functional for real-time assistive applications. Additionally, since the TTS process is light as far as processing demands are concerned, it is fitting for

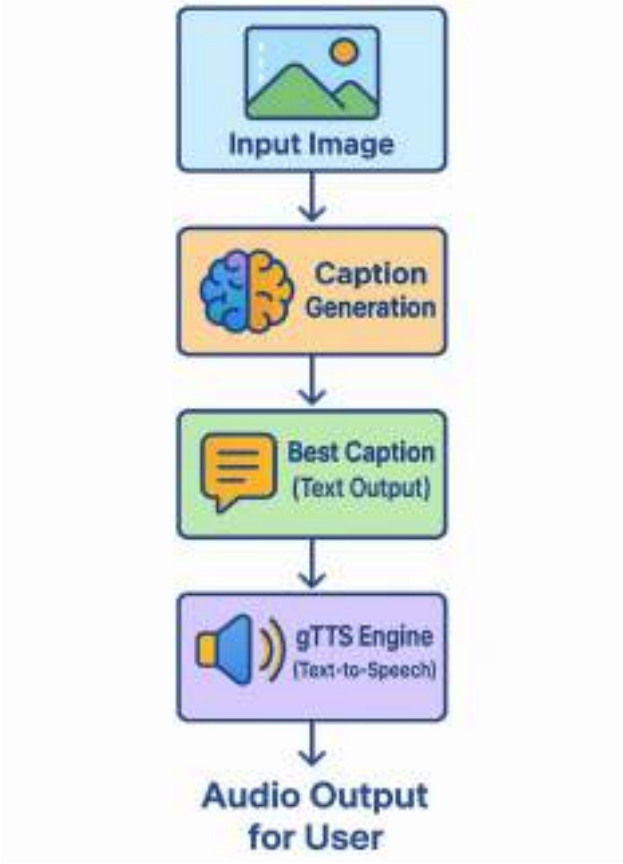

Fig. 8. Text-to-speech integration.

use on edge devices where there might be resource constraints.

With the inclusion of text-to-speech capability, the proposed framework covers the complete assistive loop: from visual input, to caption generation and selection, to audio output, for seamless visual scene interpretation for blind or visually impaired individuals. The process of converting the caption into audio using TTS is illustrated in Fig. 8.

Interactive voice query system

Aside from the display of the passive captions, the proposed system for the assistive tool also has an interactive voice query feature that aims to provide dynamic and interactive access for the user based on the provided image. This feature allows the blind user to interact with the system by asking questions based on the scene in the image after the initial caption has been provided.

After the appropriate caption has been generated and displayed by the system, the system goes into listening mode and waits for the user's input. The user may speak out questions such as "How many people are in the image?" or "Is there a dog in the photo?" in natural language. For the system to respond to the questions provided by the user, the system uses the feature of automatic speech recognition (ASR) that allows the system to convert the user's speech into text.

For quantitative queries about objects in the image, the system utilizes the YOLOv8n deep learning object detection model. YOLOv8n is a lightweight deep learning CNN designed for real-time object detection in an image. This allows for the efficient and effective identification of objects in an image. Once the objects are identified in the image, the system generates an appropriate response based on the user's query.

The response that is produced is then sent back to the same text-to-speech (TTS) system that was used for the playback of the captions. This creates an auditory response for the user's query. This type of feedback loop allows the user to have real-time access to more detailed scene information, thus helping to circumvent the limitations of the generic captions. This also allows the user to have a form of control within the description process.

The ability of this module to allow for interactive questions and responses completely redefines the system as no longer just a static tool for generating captions but as an intelligent tool for assisting the visually impaired. As shown in Fig. 9, the user is able to interact with the system by asking questions vocally in order to receive additional information.

System deployment and hardware integration

This system is characterized by a modular design that incorporates multiple inputs of images to ensure maximum adaptability to various environments of operation. Currently, the system is enabled to capture real-time images through a camera interface or by uploading them from storage devices. The system is accessible by screen readers, thus ensuring maximum accessibility, and is deployable on edge devices such as Raspberry Pi. This is enabled by the separation of the input and hardware-specific components from the main captioning system. Therefore, there is maximum adaptability to various environments without modifying the main AI system.

Results

The experiments were conducted on a machine with an Intel Core i7 processor, 32 GB RAM, and an NVIDIA RTX 3080 GPU with 10 GB VRAM. The deployment was done using Python 3.9, TensorFlow 2.x,

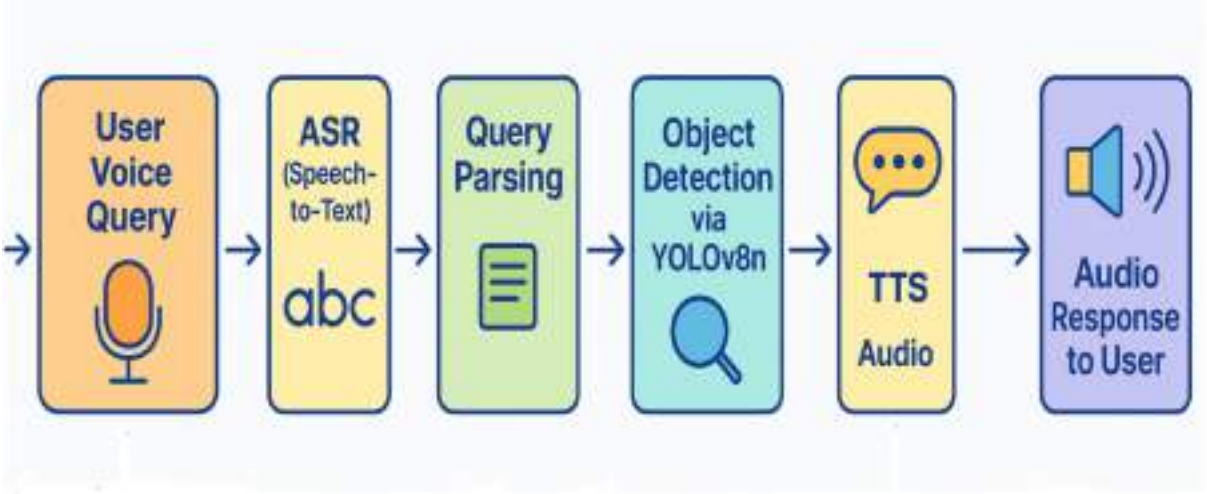

Fig. 9. Interactive voice query system.

Keras, and HuggingFace Transformers for ViT-GPT2 integration.

The Flickr8k dataset, with 8,092 images, each of which has five human captions, was used in all experiments. The dataset was randomly split into 80% training (6,474 images) and 20% testing (1,618 images). The training set was used to train the CNN-LSTM models, and the test set was reserved for evaluation and caption quality assessment. Images were preprocessed by resizing to the input size requirement of each CNN model (either 224 × 224 or 299 × 299), normalizing pixel values to the range [0, 1], and extracting deep features from pre-trained CNN models. Captions were preprocessed by removing punctuation and numbers, lowercasing, tokenizing, padding to a fixed number of tokens, and prepadded with "start" and "end" tokens to enable sequence generation.

The generated captions were evaluated with the following standard metrics:

- BLEU-1, BLEU-2, BLEU-3, and BLEU-4: Compute n-gram overlaps between reference and candidate captions. BLEU-1 score counts unigram matches, while BLEU-4 considers up to four-gram matches. More lexical matching with human descriptions is reflected by higher scores.
- METEOR: Computes unigram alignment considering stemming and synonym matching. METEOR is a combination of precision and recall and is generally more aligned with human preference than BLEU.
- ROUGE-L: Calculates the longest common subsequence (LCS) between the reference and generated captions, providing a metric that is concentrated on fluency and structural similarity.

Training times on Intel i7, 32GB RAM, NVIDIA RTX 3080: InceptionV3-LSTM (6.5h), InceptionResNetV2-LSTM (7.2h), Xception-LSTM (6.8h), DenseNet201-LSTM (7.5h), ViT-GPT2 fine-tuning (12h), with 2h for feature extraction, totaling ~40 hours.

Training progress and convergence

In order to compare the convergence behavior of the proposed CNN-LSTM models, all four models—InceptionV3, InceptionResNetV2, Xception, and DenseNet201—were trained on the training subset of the Flickr8k dataset (80% of data) for up to 50 epochs. Throughout training, training loss and validation loss were monitored to monitor optimization progress and potential overfitting.

In all models, a common trend of decreasing loss values was observed during the learning of visual-linguistic patterns. The InceptionV3-based model revealed a decrease in validation loss from 3.46 to 2.06 at epoch 40. The InceptionResNetV2 model was found to converge very well, achieving a minimum validation loss of 2.06 at epoch 38. The Xception model performed best with a minimum validation loss of 1.99, and DenseNet201 converged to a minimum of 2.05.

This reveals that all models had high feature representations and performed exceptionally well in terms of generalization over the entire training set. Furthermore, early stopping was incorporated to prevent overfitting, and the models retained the optimal weight for performance, as measured by minimal loss during validation. Fig. 10 represents the learning and validation loss for all models over the entire training period. This reveals the smooth learning of the framework.

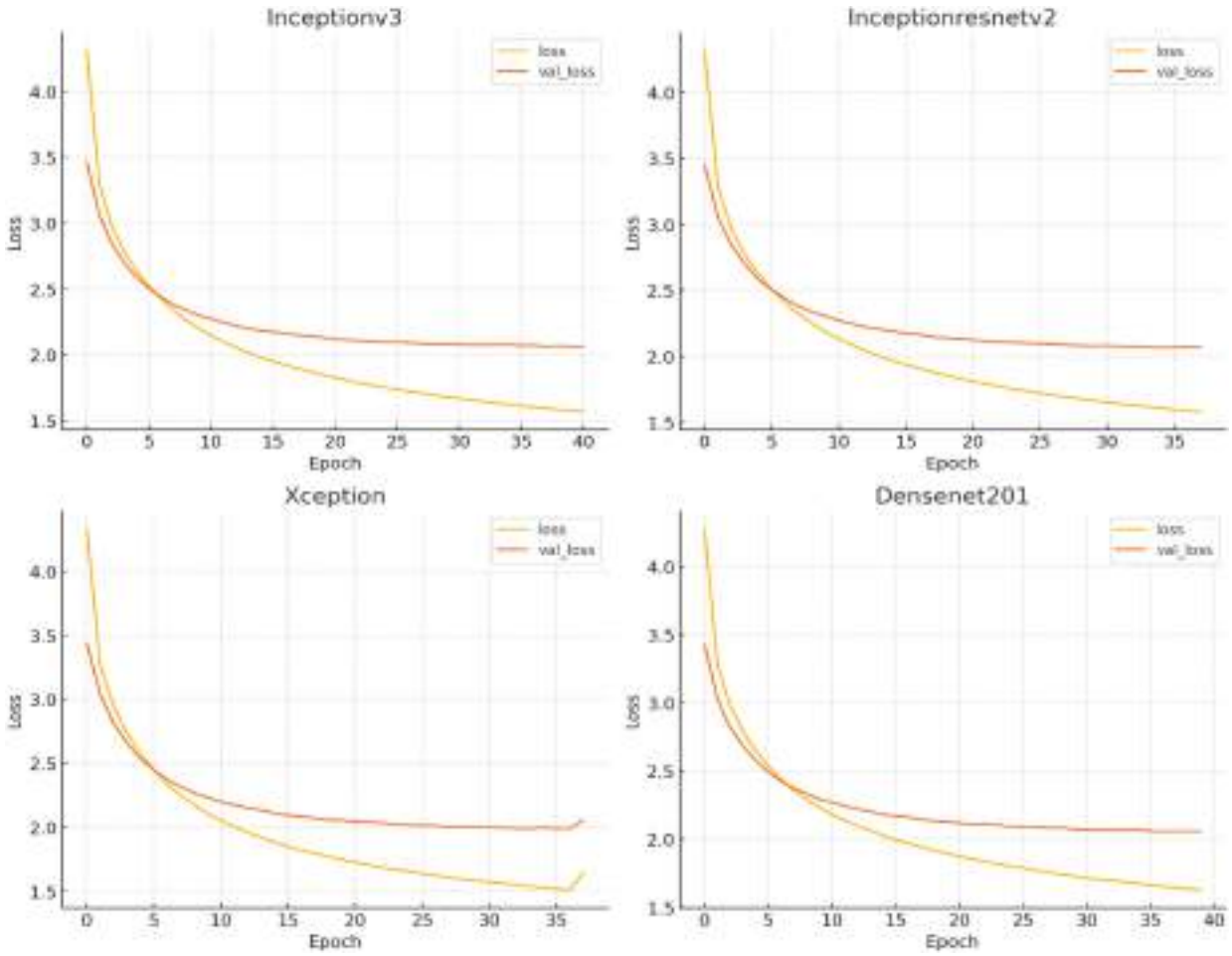

Fig. 10. Learning and validation loss system.

Evaluation of caption generation models

In order to critically evaluate the performance of the proposed models for caption generation, an experimental assessment is performed on a representative set of the Flickr8k dataset. The selected set of images contained high-quality human-annotated reference captions for comparison. The assessment is performed using a set of popular natural language generation evaluation metrics, namely BLEU-n (where n ranges from 1 to 4), METEOR, and ROUGE-L. These metrics offer complementary information about the generated captions in comparison to the reference captions.

The performance of each individual CNN-LSTM model was tested. The InceptionV3 model scored 0.4773 in BLEU-1, and 0.1235 in BLEU-4, with a METEOR score of 0.2999, and a ROUGE-L score of 0.3669. The InceptionResNetV2 model also scored the same, with BLEU-1 being 0.4783, BLEU-4 being 0.1363, METEOR being 0.3069, and ROUGE-L being 0.3685. The Xception-based model scored slightly higher than the other two with a BLEU-1 of 0.5024, BLEU-4 of 0.1366, METEOR of 0.3155, and ROUGE-L of 0.3912. DenseNet201 recorded the highest amongst the CNN-LSTM models with a BLEU-1 of 0.5096, BLEU-4 of 0.1391, METEOR of 0.3194, and ROUGE-L of 0.3983.

In contrast, the transformer version ViT-GPT2 outperformed all CNN-LSTM models by a large margin on all four evaluation metrics. It obtained the BLEU-1 score of 0.6929, BLEU-4 of 0.2192, METEOR score of 0.4274, and ROUGE-L value of 0.4895. This noticeable improvement stems from the model's ability to embed visual and text tokens simultaneously using self-attention, enabling a richer global contextual representation.

To further increase performance and caption variety, an ensemble strategy was employed that mixed the outputs of the four CNN-LSTM models and ViT-GPT2. The ensemble process selected the highest-ranked BLEU-1 caption from among the candidate captions generated by the individual component

Table 1. Evaluation of CNN-LSTM, ViT-GPT2, and ensemble models on standard metrics.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
InceptionV3	0.4773	0.2997	0.1908	0.1235	0.2999	0.3669
InceptionResNetV2	0.4783	0.3096	0.2004	0.1363	0.3069	0.3685
Xception	0.5024	0.3250	0.2078	0.1366	0.3155	0.3912
DenseNet201	0.5096	0.3282	0.2114	0.1391	0.3194	0.3983
ViT-GPT2	0.6929	0.4885	0.3275	0.2192	0.4274	0.4895
Ensemble	0.7363	0.5437	0.3823	0.2642	0.4545	0.5107

models. The ensemble model generated the best overall results with a BLEU-1 of 0.7363, BLEU-4 of 0.2642, METEOR of 0.4545, and ROUGE-L of 0.5107. These results clearly show that the combination of various model structures can produce more coherent, denser, semantically and contextually correct captions.

Thus, in summary, the evaluation process clearly proves the advantage of using the ensemble method, while also emphasizing the effectiveness of using CNN-based and transformer-based architectures together. This hybrid approach creates a balanced system that maintains fine-grained semantics while preserving linguistic relationships, making it highly appropriate for assistive visual captioning for the visually impaired. Table 1 shows the evaluation results of the CNN-LSTM models, ViT-GPT2, and their ensemble using BLEU (1-4), METEOR, ROUGE-L, and test images of the Flickr8k dataset.

Result & discussion

The experimental results in the form of Table 1 provide significant insights into the performance of the proposed models of image captioning, as illustrated in Fig. 11. Out of the four models based on the CNN-LSTM architecture, the DenseNet201-based image captioning model performed the best, achieving the following metrics: BLEU-1 = 0.5024, BLEU-4 = 0.1366, METEOR = 0.3155, and ROUGE-L = 0.3912. This result implies that the DenseNet201 architecture was successful in extracting dense features, which were interpreted as accurate image captions. The InceptionResNetV2 and Xception models performed slightly lower, while the InceptionV3 model performed slightly lower compared to the DenseNet201 architecture.

In contrast, the ViT-GPT2 architecture based on the transformer architecture for both vision and language tasks substantially outperformed the CNN-LSTM architectures in all four evaluation criteria. Specifically, the ViT-GPT2 architecture yielded a BLEU-1 score of 0.6929 and a BLEU-4 score of 0.2192, suggesting that the ViT-GPT2 architecture is able to capture long-range dependencies in the input image and generate captions that are smooth and context-aware. The ViT-GPT2 architecture also yielded a METEOR score of 0.4274 and a ROUGE-L score of 0.4895.

The ensemble model, which combined the output of CNN-LSTM and ViT-GPT2 using a selection mechanism based on BLEU-1 score, performed best in overall performance. This is because the ensemble method achieved the best score in BLEU-1 at 0.7363, BLEU-4 at 0.2642, METEOR at 0.4545, and ROUGE-L at 0.5107. This shows that the ensemble method is able to leverage the strengths of CNN and transformer models in providing captions that are not only accurate but also rich in context.

The model has been trained on the 8,092 images provided by the Flickr8k dataset and exhibits the ability for generalization beyond the particular data. Generalization beyond the data is accomplished by the application of pre-trained convolutional neural networks such as InceptionV3 and Densenet201 on large-scale data sets such as ImageNet. This allows the system to derive general visual and linguistic patterns and not merely rely on the memorization of data sets. The extensive vocabulary of captions also allows for the system's ability to derive accurate captions. Ensemble modeling also allows for the system's robustness and the prevention of overfitting. The system's accuracy is also demonstrated by the application of the system on unseen data.

The current system has some limitations. For instance, the computational requirements for ensemble processing may pose a challenge for use in devices with low processing capacity. Language support is also limited to English, with plans for the inclusion of Arabic in the subsequent studies. Another limitation is the privacy aspect with regard to the use of image and voice data. There are also chances of performance degradation with low-quality and culturally unfamiliar images.

In conclusion, the overall results have shown that the use of hybrid and ensemble methods improves the quality of generated image captions and produces output that is closer to human references.

Table 2. Performance comparison with state-of-the-art models.

Model	BLEU-1	BLEU-4	METEOR	ROUGE-L
Ensemble (Proposed)	0.7363	0.2642	0.4545	0.5107
ViT-GPT2	0.6929	0.2192	0.4274	0.4895
DenseNet201-LSTM	0.5024	0.1366	0.3155	0.3912
Xception-LSTM	0.4783	0.1363	0.3069	0.3685
InceptionResNetV2-LSTM	0.4773	0.1235	0.2999	0.3669
InceptionV3-LSTM	0.4773	0.1235	0.2999	0.3669
[21] Transformer Ensemble	0.728	0.728		
[22] NIC + kNN	0.5967	0.182		
[23] ResNet101 + BERT Fusion	0.647	0.228		
[24] Multi-level Attention	0.686	0.245	0.232	
[25] VGG16 + LSTM	0.582	0.182		
[26] Feature + Caption Fusion	0.669	0.232		
[27] Dual-CNN + Attention	0.6876	0.2471		
[28] ResNet50	0.619	0.262		
[28] VGG16	0.561	0.223		

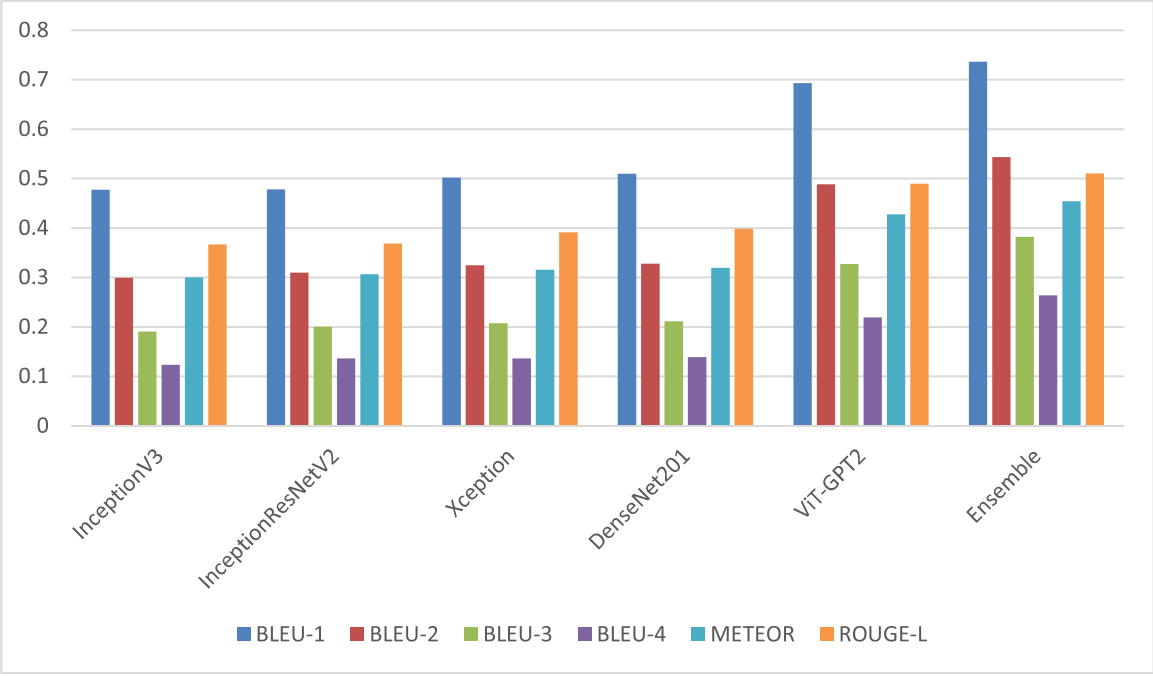

Fig. 11. Performance metrics for captioning models across evaluation scores.

Comparison with state-of-the-art models

To put the performance of the proposed captioning system into perspective, a comparison with existing state-of-the-art approaches in the image captioning literature was performed. Table 2 overviews the performance of a variety of models evaluated on the Flickr8k dataset, with BLEU-n, METEOR, and ROUGE-L scores as conventional benchmarks.

Among the compared methods,[21] employed an ensemble of CNN-based feature extractors and transformer-based generators with a voting scheme for selecting captions based on BLEU scores. Their model achieved 0.728 and 0.798 BLEU-1 on Flickr8k and Flickr30k, respectively. In,[22] a hybrid NIC + kNN system achieved 59.67 BLEU-1 and 18.20 BLEU-4 by combining deep features from InceptionV3 + LSTM and caption retrieval by nearest neighbors.

In,[23] the integration of pre-trained BERT as an Auxiliary Language Model (AuxLM) into a fusion-based encoder-decoder model obtained 64.7 and 22.8 BLEU-1 and BLEU-4 scores, respectively. Similarly,[24] put forward a multi-level attention-based model augmented with paraphrasing and language rescoring, obtaining 24.5 BLEU-4 and 23.2 METEOR. Other competing models are the VGG16 + LSTM with

attention in[25] (BLEU-4 = 18.2%) and the hybrid Beam Search system in[26] with a BLEU-4 of 23.2%.

The enhanced dual-CNN model in[27] paired ResNet-101 and EfficientNet-B0 with visual attention and beam search (k = 3) to achieve BLEU-4 of 24.71%. Lastly,[28] confirmed the superiority of ResNet-50 over VGG16 with BLEU-4 scores of 26.2% and 22.3%, respectively.

On the other hand, the proposed system outperformed all of these methods. Specifically, the ensemble model—aggregating outputs of four CNN-LSTM models (InceptionV3, InceptionResNetV2, Xception, DenseNet201) and ViT-GPT2—achieved BLEU-1 of 0.7363, BLEU-4 of 0.2642, and METEOR of 0.4545, all of which demonstrate considerable improvements in both lexical correctness and contextual relevance. These results display the power and synergy achieved by architectural diversity and aggregation in the proposed system. The comparative performance of the proposed model and existing methods is summarized in Table 2.

Conclusion

The proposed research aims to develop an effective and comprehensive system for assistive image captioning. The system is intended for the visually impaired. Four CNN-LSTM architectures are used: InceptionResNetV2, InceptionV3, DenseNet201, and Xception. Additionally, the ViT-GPT2 transformer architecture is used. The proposed system has competitive performance in generating effective captions. The best caption with the highest semantic accuracy is determined by BLEU-based ensemble scoring. The output has an order of magnitude improvement compared to individual models. The system also uses Google's Text-to-Speech (gTTS) for providing audio feedback. YOLOv8n is also used for human detection based on the user's voice query. Experimental results show that the proposed ensemble outperforms CNN-based and transformer-based approaches in terms of BLEU, METEOR, and ROUGE-L. Future research directions include the integration of effective scene graphs and user context. Also, the proposed system may be deployed on mobile devices for edge computing. Future research directions include testing the proposed system with more than 500 images of the Iraqi environment and testing the proposed system with 20 visually impaired people. Future research directions include the integration of the Arabic language and dialects. Future research directions include the integration of the Middle Eastern context and the use of spatial reasoning and safety questions.

Acknowledgment

The authors would like to express their sincere gratitude to the University of Technology – Iraq and the University of Baghdad for their academic advice and research facilities that greatly assisted in the completion of this work.

Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images that are not ours have been included with the necessary permission for republication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- Author(s) signed an ethical considerations approval.
- Ethical Clearance: The project was approved by the local ethical committee at University of Baghdad.

Authors' contributions statement

A.N.M. has contributed to the conception, design, acquisition of data, and implementation of the analysis of the manuscript. G.K.A. has contributed to the interpretation, verified the analytical methods, and revised the manuscript. Both authors have discussed the result analysis and contributed to the manuscript.

Data availability

The datasets generated and analyzed during the current study are available in the [Flickr8k] repository, [https://www.kaggle.com/datasets/adityajn105/flickr8k].

References

1. Vashist P, Senjam SS, Gupta V, Gupta N, Shamanna BR, Wadhwani M, *et al.* Blindness and visual impairment and their causes in India: Results of a nationally representative survey. PLOS ONE. 2022;17(7):e0271736. https://doi.org/10.1371/journal.pone.0271736.
2. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, *et al.* Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis. 2017; 123:32–73. https://doi.org/10.1007/s11263-016-0981-7.
3. Gurari D, Li Q, Stangl AJ, Guo A, Lin C, Grauman K, *et al.* VizWiz Grand Challenge: Answering Visual Questions from Blind People. CVPR; Salt Lake City, UT, USA. 2018;3608–3617. https://doi.org/10.1109/CVPR.2018.00380.

4. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: A neural image caption generator. CVPR. Boston, MA, USA. 2015;3156–3164. https://doi.org/10.1109/CVPR.2015.7298935.
5. Abdullah HS, Ali NH, Abdullah NA. Evaluating the Performance and Behavior of CNN, LSTM, and GRU for Classification and Prediction Tasks. Iraqi J. Sci. 2024;65(3):1741–1751. https://doi.org/10.24996/ijs.2024.65.3.43.
6. Jassem MD, Abdulrahman AA. Survey on distributed denial of service attack detection using deep learning: A review. Int J Nonlinear Anal Appl. 2022;13(2):753–762. https://doi.org/10.22075/IJNAA.2022.6458.
7. Lu J, Xiong C, Parikh D, Socher R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. CVPR. Honolulu, HI, USA. 2017;375–383. https://doi.org/10.1109/CVPR.2017.345.
8. Abdullah HS, Ali NH, Abdullah NAZ. Evaluating the Performance and Behavior of CNN, LSTM, and GRU for Classification and Prediction Tasks. Iraqi J. Sci. 2024;65(3):1741–1751. https://doi.org/10.24996/ijs.2024.65.3.43.
9. Chen Y-C, Li L, Yu L, El Kholy A, Ahmed F, Gan Z, *et al.* UNITER: UNiversal image-TExt representation learning.LNCS. 2020;104–120. https://doi.org/10.1007/978-3-030-58577-8_7.
10. Hossain MDZ, Sohel F, Shiratuddin MF, Laga H. A comprehensive survey of deep learning for image captioning. ACM Comput Surv. 2019;51(6):1–36. https://doi.org/10.1145/3295748.
11. Ibrahim V, Bakar JA, Harun NH, Abdulateef AF. A Word Cloud Model based on Hate Speech in an Online Social Media Environment. Baghdad Sci J. 2021;18(2):938–946. https://doi.org/10.21123/bsj.2021.18.2(Suppl.).0937.
12. Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. CVPR. Honolulu, HI, USA. 2017;1179–1195. https://doi.org/10.1109/CVPR.2017.131.
13. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, *et al.* Bottom-up and top-down attention for image captioning and visual question answering. CVPR. Salt Lake City, UT, USA. 2018;6077–6086. https://doi.org/10.1109/CVPR.2018.00636.
14. Okolo GI, Althobaiti T, Ramzan N. Assistive systems for visually impaired persons: Challenges and opportunities for navigation assistance. Sensors. 2024;24(11):3572. https://doi.org/10.3390/s24113572.
15. Ahsan H, Bhalla N, Bhatt D, Shah K. Multi-Modal Image Captioning for the Visually Impaired. NAACL 2021;53–60. https://doi.org/10.18653/v1/2021.naacl-srw.8.
16. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. CVPR. Las Vegas, USA. 2016;779–788. https://doi.org/10.1109/CVPR.2016.91.
17. Al-Jamali NAS. Convolutional Multi-Spike Neural Network as Intelligent System Prediction for Control Systems. Journal of Engineering. 2020;26(11):184–194. https://doi.org/10.31026/j.eng.2020.11.12.
18. Gu J, Stefani E, Wu Q, Thomason J, Wang X. Vision-and-language navigation: A survey of tasks, methods, and future directions. ACL; Dublin, Ireland. 2022;7606–7623. https://doi.org/10.18653/v1/2022.acl-long.524.
19. Khassaf NM, Ali NH. Improving Pre-trained CNN-LSTM Models for Image Captioning with Hyper-Parameter Optimization. ETASR. 2024;14(5):17337–17343. https://doi.org/10.48084/etasr.8455.
20. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, *et al.* An image is worth 16×16 words: Transformers for image recognition at scale. ICLR. 2021;1-21. https://doi.org/10.48550/arXiv.2010.11929.
21. Al-Badarneh I, Hammo B, Al-Kadi O. An ensemble model with attention based mechanism for image captioning. Comput. Electr. Eng. 2025;123(A):110077. https://doi.org/10.1016/j.compeleceng.2025.110077.
22. Arora K, Raj A, Goel A, Susan S. A hybrid model for combining neural image caption and k-nearest neighbor approach for image captioning. AISC. 2022;51–59. https://doi.org/10.1007/978-981-16-1249-7_6.
23. Kalimuthu M, Mogadala A, Mosbach M, Klakow D. Fusion models for improved image captioning. ICPR. 2021;12666: 381–395. https://doi.org/10.1007/978-3-030-68780-9_32.
24. du Plessis M, Brink W. Improving the performance of image captioning models trained on small datasets. CCIS. 2022;77–91. https://doi.org/10.1007/978-3-030-95070-5_6.
25. Yonia DL, Ariansyah I, Dina AB, Suciati N. Enhancing image captioning performance with VGG16 feature extraction and LSTM sequence processing. ICECOS. 2024;77–82. https://doi.org/10.1109/ICECOS63900.2024.10791213.
26. Nguyen DTT, Nguyen HT. Image caption generator with a combination between convolutional neural network and long short-term memory. Biomedical and Other Applications of Soft Computing. 2023;225–238. https://doi.org/10.1007/978-3-031-08580-2_21.
27. Zagon B, Praetawan J, Emmanuel O, Olarik S. Enhancing image caption performance with improved visual attention mechanism. ICIC. 2025;16(01):73. https://dx.doi.org/10.24507/icicelb.16.01.73.
28. Sri Neha V, Nikhila B, Deepika K, Subetha T. A comparative analysis on image caption generator using deep learning architecture—ResNet and VGG16. AISC. Springer Singapore. 2022;1420:209–218. https://doi.org/10.1007/978-981-16-9573-5_15.

إطار تجميعي متعدد النماذج لتوليد وصف صوتي تفاعلي للصور لمساعدة ذوي الإعاقة البصرية

علاء نوري مزه، غادة كاظم الخفاجي

قسم علوم الحاسوب، كلية العلوم، جامعة بغداد، بغداد، العراق.

الخلاصة

تُقيد الإعاقة البصرية قدرة الفرد على الإدراك والتفاعل مع البيئة، مما يجعل الأنشطة اليومية تحديًا دون توفر تقنيات مساعِدة مناسبة. أحد أهم المعوقات هو عدم القدرة على إدراك المشاهد البصرية، الضرورية للوعي المكاني والتنقل والمعرفة الظرفية. تهدف هذه الدراسة إلى سد هذه الفجوة عبر اقتراح نظام ذكي لتوليد وصف نصي للصور مدموج بتفاعلات صوتية موجهة لذوي الإعاقة البصرية. يوظف النظام أحدث تقنيات الرؤية الحاسوبية ومعالجة اللغة الطبيعية لتوليد تسميات نصية سياقية لكل صورة مصحوبة باستجابات صوتية تفاعلية. تم الاعتماد أولًا على أربعة نماذج من الشبكات العصبية التفافية (InceptionV3, InceptionResNetV2, Xception, DenseNet201) مع مفكِّكات تعتمد على LSTM لتوليد التسميات الأولية؛ بالإضافة إلى نموذج المحول المعتمد على ViT-GPT2 لتوليد تسميات إضافية. ثم يُطبق أسلوب التجميع لاختيار التسمية الأمثل استنادًا إلى تقييم BLEU. يستخدم النموذج خدمة تحويل النص إلى كلام من جوجل (gTTS) للإخراج الصوتي وYOLOv8n للكشف الفوري عن الأشخاص بناءً على استعلامات صوتية. جرى التدريب والاختبار باستخدام مجموعة بيانات Flickr8k، وأظهرت النتائج تفوّق النموذج التجميعي على النماذج الفردية (CNN-LSTM وViT-GPT2). فقد حقق النموذج التجميعي قيمة BLEU-1 بمقدار 0.7363، وBLEU-4 بمقدار 0.2642، وMETEOR بمقدار 0.4545، وROUGE-L بمقدار 0.5107، متفوقًا على الممارسات الحالية في توليد أوصاف متناسقة وغنية بالمعلومات. تؤكد النتائج فعالية المنهجية متعددة النماذج في تحسين جودة الوصف والتفاعل، مما يمثل خطوة هامة نحو تقديم مساعدة ذكية وفورية ومتاحة لذوي الإعاقة البصرية.

الكلمات المفتاحية: نماذج CNN-LSTM، التعلم التجميعي، توليد وصف الصور، ذوو الإعاقة البصرية، محول ViT-GPT2.