An Ensemble Model with Attention Based Mechanism for Image Captioning

Israa Al Badarneh[1], Bassam H. Hammo[1,2], Omar Al-Kadi[1]

[1]King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan.

[2]King Hussein School of Computing Sciences, Princess Sumaya University for Technology, Amman, Jordan.

Abstract

Image captioning creates informative text from an input image by creating a relationship between the words and the actual content of an image. Recently, deep learning models that utilize transformers have been the most successful in automatically generating image captions. The capabilities of transformer networks have led to notable progress in several activities related to vision. In this paper, we thoroughly examine transformer models, emphasizing the critical role that attention mechanisms play. The proposed model uses a transformer encoder-decoder architecture to create textual captions and a deep learning convolutional neural network to extract features from the images. To create the captions, we present a novel ensemble learning framework that improves the richness of the generated captions by utilizing several deep neural network architectures based on a voting mechanism that chooses the caption with the highest bilingual evaluation understudy (BLEU) score. The proposed model was evaluated using publicly available datasets. Using the Flickr8K dataset, the proposed model achieved the highest BLEU-[1-3] scores with rates of 0.728, 0.495, and 0.323, respectively. The suggested model outperformed the latest methods in Flickr30k datasets, determined by BLEU-[1-4] scores with rates of 0.798, 0.561, 0.387, and 0.269, respectively. The model efficacy was also obtained by the Semantic propositional image caption evaluation (SPICE) metric with a scoring rate of 0.164 for the Flicker8k dataset and 0.387 for the Flicker30k. Finally, ensemble learning significantly advances the process of image captioning and, hence, can be leveraged in various applications across different domains.

Keywords: Image captioning, ensemble learning, convolutional neural network, attention-based transformer

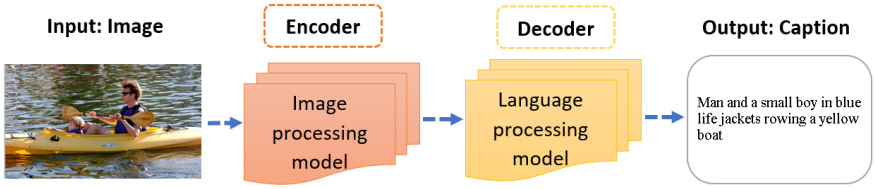


Figure 1: General architecture of image captioning model.

1 Introduction

Identifying key components in an image, understanding their relationships, and creating syntactically and semantically consistent descriptions of the visual content are all necessary to create an image caption. This is one of the hardest tasks in artificial intelligence because it requires the integration of two very different research communities: natural language processing and computer vision [1]. An overview of the standard architecture of the image captioning model is given in Fig.1. The general architecture of an image captioning system typically consists of several key components. It begins with an image input processed by a Convolutional Neural Network (CNN) for feature extraction, utilizing pre-trained models like ResNet or Inception to capture the essential visual elements. These extracted features are fed into a generation caption model such as Long-Short-Term Memory (LSTM) units or transformers. An optional attention mechanism can enhance this process by allowing the model to focus on specific image areas while forming each caption word. Finally, the system produces an output caption that represents the generated description of the image.

Recent developments in deep learning models, made possible by cutting-edge computational capabilities, have significantly advanced this discipline [2, 3]. Image captioning is challenging in artificial intelligence since it combines computer vision and natural language processing research. A captioning model aims to represent the text and scene, as this is essentially what the human brain does. Humans can automatically describe much information about any given image with a glance. One of the many difficulties and unsolved problems inherent in image captioning is the parallax error. It may be difficult for the human eye to identify an object, even at certain angles, where its appearance varies to the point of being undetected. An object class may include several objects of various forms and angles. Additionally, the visual assistant could have difficulty correctly identifying objects hidden by other objects. Object recognition is negatively affected by scene clutter [4]. There are many industries in which image captioning research can find practical applications. Examples include medical imaging for analysis and diagnostics [5–8], improving education for students [9, 10], supporting visually impaired people [11], helping virtual assistants [12], facilitating information retrieval [13], aiding video surveillance [14], improving social media content [15], and even assisting automated self-driving cars [16]. Additionally, it is essential to improve the quality of image search [17]. Template-based, retrieval-based, and deep learning-based approaches are the three primary categories of image captioning techniques. Template-based approaches create captions using predefined templates with

blank spaces; this results in grammatically correct statements but is restricted. With retrieval-based techniques, general but sometimes inaccurate semantic descriptions are generated by extracting captions from an existing set. Using deep neural networks for visual and linguistic modeling in deep learning-based approaches is a discovery that improves image captioning systems and offers useful solutions. [18] introduced an attribute node to provide a more detailed description of objects and to model high-level relationships within a visual semantic graph. The proposed method of [19] offers a novel approach to news image captioning, aiming to preserve semantic information, enhance style coherence with the news articles, and enable entity-aware, controllable caption generation. Before deep learning became popular, traditional machine learning approaches handled most image captioning tasks [1, 20]. Among these were feature extraction approaches such as the histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Scale-Invariant Feature Transform (SIFT). The items were classified using a classifier after extracting features [21, 22]. Deep learning-based techniques automatically find features and are more popular than traditional methods since feature extraction from huge amounts of data is challenging [4]. The use of deep machine learning for captioning images has received a lot of interest recently [1]. Deep learning algorithms can efficiently manage the difficulties and complexity of captioning images. Ensemble learning is a growing field of interest that addresses this issue by merging concepts from data fusion, modeling, and mining into a unified approach. It starts by extracting features using multiple algorithms that make predictions based on these characteristics. The ensemble learning then combines these insights to improve the general accuracy of the prediction through various voting mechanisms to achieve better results than any algorithm could provide [23]. Ensemble learning reduces the biases associated with individual models and enhances caption production overall by combining predictions from several models. This technique is particularly useful for improving the performance of complex architectures with several types of ensemble models, such as bagging, boosting, stacking, and voting.

The main contributions of this research are: (a) exploring the detailed design of transformer models, focusing on the effectiveness of different attention mechanisms,(b) introducing an innovative ensemble learning framework that leverages multiple deep neural network architectures to enhance the accuracy and richness of generated captions, and (c) enhancing the robustness and reducing overfitting of image captioning models by experimenting with different publicly available datasets.

The research is organized as follows. Section 2 describes related works on image captioning. Section 3 details the research methodology. Section 4 covers experiments and results. Section 5 provides limitations and future opportunities. Finally, Section 6 offers the conclusion and implications of the research.

2 Related works

Several papers have recently employed deep-learning techniques to generate captions for images. This section provides an overview and discussion of related works.

2.1 Neural Network based model

In the automatic generation of image captions[24], the encoder-decoder architecture was used in the suggested model. The encoder processed the input image to extract the relevant data, while the decoder used these features to generate the caption. The image was encoded into a feature vector with a specified length. Long Short-Term Memory (LSTM) cells were used to implement the decoder.Utilizing pre-trained models and deep learning approaches, the suggested method showed encouraging results.The work in [25] The suggested model performed better when creating informative captions for images. Authors of [26] provided a model that creates natural language descriptions of images for generating image descriptions. The method consisted of bidirectional RNNs over phrases, CNNs over image areas, and a structured objective using multimodal embedding to align the two modalities. The alignment model obtained state-of-the-art results.

2.2 Attention based model

The challenging task of automatically producing meaningful captions for images was discussed in [27], and suggested a collaborative model known as AICRL (Automatic Image Captioning based on ResNet50 and LSTM with Soft Attention). The encoder used a CNN called ResNet50 to represent the input image comprehensively. The model gathered the quality of generated captions by focusing on relevant locations. The experiment results showed that the AICRL model is useful for producing image captions. It offers a promising means of bridging the gap between natural language descriptions and visual content, making it applicable to various computer vision applications and beyond. It is remarkable that the aligned attention method is model-independent and may be quickly added to current innovative image captioning models to enhance their captioning capabilities. [28] presented a transformer-based model for image captioning; their strategy used a mask operation to automatically assess the influence of image region features and use the results as supervised information to direct attention alignment. This work provided a useful reference for self-supervised learning. The transformer-based framework LATGeO was proposed in [29] to caption images, and it includes multi-level geometrically coherent and visual recommendations to relate objects based on their localized ratios. LATGeO used object proposals to find coherence and connected its embeddings with less significant surrounds. A brand-new label-attention module (LAM), an extension of the traditional transformer, was developed to bridge the gap between the visual and linguistic worlds. Although normalization has traditionally only been used outside of self-attention, the work of [30] provided a unique normalization method and showed that doing so in hidden activation within self-attention is feasible and advantageous. They provide a class of geometry-aware self-attention (GSA) that extends self-attention to explicitly and efficiently consider the relative geometry relations between the objects in the image to model the geometry structure of the input objects for feature extraction. Faster-RCNN was used. The inputs to the transformer encoder are region-based visuals, and the transformer decoder predicts the subsequent word recursively using the attended senses and the embedding of the preceding words.

Motivated by the relationships between image features, [31] presented a new transformer-based model. The proposed model considered three types of spatial relationships in the image regions. The query region could be a parent, neighbor, or child. The decoder consists of an LSTM layer and an implicit transformer layer. The transformer was used parallel to decode different image regions in the decoder part. The results showed that the proposed model was better than others based on several evaluation metrics.The work of [3] illustrated the limitations of current methods, such as neglecting the interaction between a word and an object and the undiscovered relationship between objects. To solve these problems, they presented a multi-transformer (MT) for image captioning. The MT model can understand three types of relations: word-to-word, object-to-object, and word-to-object. The caption decoder took the encoder output and generated the caption using word embedding and a layer of LSTM.

2.3 Ensemble based model

Ensemble learning aims to increase generalizability and robustness over a single model by combining the predictions of various base models. Modern techniques for detecting hate speech in multimodal memes [32] applied the majority voting technique, also known as the hard voting or voting classifier, which combines many classifiers and voting classifiers. As a result, it performs better than any individual model utilized in the ensemble. Textual and visual hybrid methods are combined using the max voting technique to classify a fake or real news instance. [33], in this work, the maximum voting method was used. The proposed system consists of four independent parallel streams capable of detecting specific forgeries. All four streams handled each input instance. These independent predictions are finally combined using the maximum voting ensemble method.

In [34], an image captioning method was presented using a set of weighted multichannel fusion optimization enhancements to optimize the encoder and decoder. In the model that is being described, a multichannel encoder was suggested that can combine different models and algorithms to extract different information from the same image, researchers suggested combining separate decoders of the same type using the voting weight technique for decoder fusion to improve the description produced by the decoder. For the concept detection task, [35] considered an image retrieval approach using an ensemble of five different CNNs, where the top N photos most similar to the training set and their related CUIs were used to assign a set of CUIs to each query image. The top N images that look the most like a query image, determined by the cosine similarity between image embeddings, were extracted using CNN as the image encoder; then, an aggregation step was carried out to choose the set of CUIs to link to each query image. This involved soft majority voting. A recent work [36] proposes a soft voting-based ensemble model that benefits from the efficient operation of various classifiers on various modalities. Deep feature extraction from multimodal datasets was performed for the proposed model using deep learning methods (BiLSTM, CNN). The final feature sets were classified using the soft voting-based ensemble learning model after completing the feature selection process for the features that combine text and image features.In [37], an effective deep-set medical image captioning network (DCNet) was suggested to give doctors and patients explanations. Three well-known

pre-trained models, including VGG16, ResNet152V2, and DenseNet201, are combined into DCNet. Assembling these models leads to better outcomes, as it avoids an overfitting problem. A classifier was created according to the research conducted by [38] using a soft voting ensemble combining the common CNN models Predictions in the soft voting ensemble are combined and weighted according to the relevance of the classifier to produce the total of weighted probabilities.

2.4 Insights from previous research and our solution

The discussion above revealed that contemporary captioning models rely on RNN and LSTM as language models. However, one key issue with these approaches is the occurrence of vanishing gradients, limiting their effectiveness. Moreover, the RNN and LSTM models are not hardware-friendly and require additional computational resources. An alternative approach explored in the literature is using Generative Adversarial Networks (GAN) for image captioning. However, GANs come with challenges due to their discrete nature, making training such systems a difficult task [39, 40]. Using a hybrid approach, combining LSTM with transformer models introduces specific limitations and drawbacks. For example, it can increase the complexity of the model, attributed to architectural differences, resulting in a higher demand for resources and extended training times [41]. Consequently, this complexity can affect the interpretation of model decisions, hindering a clear understanding of the underlying reasoning. Image captioning is an attractive task that involves understanding visual and textual information. The need for image captioning arises from the need to make visual content accessible to individuals. Therefore, developing and implementing dedicated image captioning systems is essential to address this need. Therefore, this research aims to bridge this gap by introducing a hybrid approach that combines a transformer with an attention mechanism to help the model capture complex details in images and generate more contextually relevant captions. The rationale behind this combination is that transformers are great for capturing long-range dependencies in data, while attention mechanisms help them focus on relevant parts. Ensemble learning, on the other hand, can boost overall performance by combining multiple models. The subsequent section will explore the details of this approach.

3 Methodology

This work followed a methodology incorporating four stages: data description and data preprocessing, model development, experimentation, and performance evaluation. The following subsections discuss each stage in more detail.

3.1 Dataset

This section will introduce the commonly used datasets in image captioning. details of these datasets.

Flickr8K[42]: it was published for public use in 2013. The photographs in the dataset, which total 8000, are all from the photo and image-sharing website Flickr. The image content is mostly human and animal. The description for the label was

also crowd-sourced through Amazon's manual labeling program. Each image contains a description of five sentences. This dataset offered a comprehensive and diverse set of images comprising 6,000 training images, 1,000 validation images, and 1,000 test images. Flickr8k is a standard dataset for training and evaluating image captioning models, covering a wide range of scenes, objects, and activities characteristic of daily photography. Researchers use its rich diversity in images and textual descriptions to develop algorithms capable of generating accurate and contextually relevant captions.

Flickr30k[43]: Flickr8k dataset has been expanded to build Flickr30k, it contains 31,783 captioned images. The split dataset available to the public uses 29,000, 1,000, and 1,000 images for training, validation, and testing, respectively. Each image has five sentences that were written specifically for it. The photos in this dataset mostly show people participating in ordinary activities and events. Flickr30k is used to understand visual media (images) that match a language expression (an image description). This dataset is frequently used as a reference standard for sentence-based image descriptions. A research paper emphasizes the importance of the Flickr30K dataset in analyzing human descriptions of visual content, providing a comprehensive review of its features. Each image is richly annotated with contextually relevant descriptions that offer multiple viewpoints on its content. The dataset captures the diversity of human experience and includes various aspects of human actions, objects, scenes, and environments. This variety makes it particularly suitable for exploring how people interpret and describe visual scenes [44].

3.2 Data preprocessing

Because of raw textual data challenges, cleaning and preprocessing datasets before they are used in ML models have become essential. The approach we applied to text preprocessing was comprehensive and systematic. Several procedures were employed in the data preprocessing, including the following:

(a) Text normalization: Typically, actions are taken to reduce the number of extracted terms. They include eliminating special and non-letter characters ($, &, %,..).

(b) Text tokenization: In this step, a linguistic analysis of the text is performed. Separates words, character strings, and punctuation marks into tokens during indexing. This process aims to divide the text into a stream of discrete tokens, or words, by identifying the sentences' borders and eliminating any unnecessary punctuation.

(c) Adding start and end tokens: Finally, distinctive start and end tokens were appended to determine the beginning and end of each caption, adding a layer of structural clarity to the dataset. A unique padding token was introduced to address the variability and standardize the length of captions.

3.3 The proposed model for image captioning

The following are the steps applied through the model, and the following subsections discuss each stage in more detail. Algorithm 1 provides the pseudo-code outlining the operations of the model.

S1: Image feature extraction: A pre-trained CNN network like ResNet or EfficientNet extracts features from the input image. These features serve as a rich representation of the visual content.
S2: Text generation with a transformer: A transformer-based model generates textual descriptions by taking as input the image features and producing a sequence of words that form the caption.
S3: Attention Mechanism: Attention mechanisms are implemented within the transformer, allowing the model to focus on different parts of the image when generating each word in the caption. It enhances the model's ability to align visual and textual information.
S4: The Beam Search Algorithm: The beam search algorithm was applied with a width of $k = 10$.
S5: Ensemble learning: To get a more robust and accurate caption, the ensemble learning model trains multiple instances of the transformer with different random initializations or hyperparameters and then combines their output, either by averaging or voting.
S6: Training and fine-tuning: Train the combined model on a large dataset of image-caption pairs, then fine-tune the model on a specific dataset.
S7: Evaluation: Evaluate the performance of the ensemble model using metrics like BLEU, METEOR, and CIDEr.

3.3.1 Image feature extraction

a) Convolutional neural network (CNN): Popular deep learning models include recurrent neural networks (RNNs), convolutional neural networks (CNNs), deep belief networks (DBNs), and deep Boltzmann machines (DBMs). Using shared weight filters and hierarchical learning, CNNs are highly effective in understanding visual data [45, 46]. CNN-based encoders on ImageNet that have been pre-trained are frequently used in image captioning to convert images into visual vectors. Selective focus during generation is made possible by preserving fine-grained correspondence using sets from lower convolution layers [47, 48].

CNNs provide the following benefits over conventional neural networks when used in computer vision applications: 1) The main reason to consider CNN is its weight-sharing feature, which reduces the number of trainable network parameters, allowing the network to increase generalization and preventing overfitting. 2) Learning both the classification layer and the feature extraction layers simultaneously produces a well-structured model output that depends on the features that were extracted. 3) CNN facilitates large-scale network installation more easily than other neural networks [49]. See Fig. 2 that presents the architecture of the CNN model.

b) Transfer learning: Applying a previously learned model to a modified environment is known as transfer learning. Due to its ability to train deep neural networks on tiny datasets, it is particularly well preferred in the deep learning field. This is particularly helpful in data science because most real-world scenarios do not require millions of labeled data sets to train complicated models. To apply transfer learning to image captioning, the model was first trained on a standard dataset under supervision, and then its knowledge was transferred to a new dataset consisting of unpaired

Algorithm 1 Attention-based transformer model using ensemble learning

Input: dataset=[Set of images (S), corresponding set of captions (CI)]
Output: The final output caption for the tested image (o_c)

Evaluation metrics: (BLEU-[1-4], ROUGE-L, METEOR, CIDEr, & SPICE)
Step1: Dataset prepossessing
for each caption set CI of an image I **do**
    Normalization (CI)
    Text tokenization (CI)
    Adding start and end tokens (CI) <start> (CI) <end>
end for
for each image $I \in S$ **do**
    Augment (I)
end for
Step2: Feature extraction
M= [ResNet50, ResNet101, EfficientNetV2, VGG16, VGG19, EfficientNetB4, ResNet152, RegNetX120]
for each image $I \in S$ **do**
    for each pre-trained model $m \in M$ **do**
        f_i = extract feature map f_i of image I
    end for
end for
Step3: Caption generation
for each feature map f_i **do**
    g_c= generatedCaptionbyTransformer
    BestKCaption= Beam Search(10)
end for
Step4: Ensemble learning
for each g_c **do**
    o_c= voting-on (the generated caption from all models g_c)
end for

phrases and images [45]. Residual CNNs, such as ResNet-50, use identity mapping and shortcut connections to address overfitting and optimization issues. A pre-trained ResNet-50, trained on ImageNet, is utilized in image feature extraction by removing its final output layer [50]. The ResNet-101 image captioning model uses bottom-up attention to encode images as a baseline. The effectiveness of bottom-up attention to the baseline ResNet encoding is evaluated to evaluate the performance of the model [51]. Shorter connections between layers in DenseNet's architecture improve training efficiency and the depth of deep learning networks. Strong information flow is ensured by interlayer connection, which improves learning [52]. However, VGGNet is a popular image feature extractor that is frequently used in research applications because of its resilience and simplicity. ResNet, however, outperforms VGG in terms of efficiency, providing better accuracy with fewer parameters [53]. Compound-scaling EfficientNet models have recently proven superior to other CNNs' accuracy and efficiency when used with transfer learning datasets. They show promise in various fields, such as

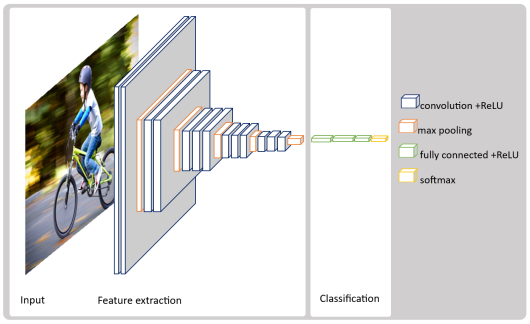


Figure 2: Architecture of CNN Model

the classification of COVID-19 [54, 55]. MobileNet is another architecture that maximizes computational effectiveness while maintaining good accuracy. The effectiveness of representation is improved by channel separation and reintegration [56, 57]. Lastly, Inception-v3, often used for transfer learning, has less computational overhead when used as an encoder. It gathers key information and adds it to a feature matrix that captures the essence of an image [3, 31, 58]. In the proposed methodology's workflow, the initial step toward image processing involves passing the image through a CNN to generate image features. Existing work studied various versions of CNN as feature extractors for image captioning. Feature extraction is based on eight CNN models discussed in Section. These models include: ResNet50, ResNet101, EfficientNetV2, VGG16, VGG19, EfficientNetB4, ResNet152, and RegNetX120. These features serve as input for the subsequent language processing model. Fig. 2 visually depicts the CNN model architecture. The convolution layer plays a vital role in downsampling the image into features and incorporating information from nearby pixels. The prediction layers then become active, using multiple convolution filters or kernels that pass over the image, each extracting unique aspects. To prevent overfitting and reduce the spatial size of the convolved features, a max pooling layer is used to provide an abstract representation of the convolved features. ReLU is the most widely used among various activation functions due to its ease of training and superior performance attributed to its linear behavior, as highlighted by [49].

3.3.2 Text generation with a transformer

One kind of neural network architecture is a transformer. Transformer was first introduced in the publication "Attention is all you need" [59]. Text data is well handled by the Transformer architecture, which is sequential by design. After receiving one text sequence as input, they create another one with a stack of encoder and decoder layers. The encoder and decoder stacks contain matching embedding layers for their respective inputs. There is an output layer at the end to create the final result. The encoder and a feedforward layer contain the crucial self-attention layer, which determines the connections between the words in the sequence. The decoder consists of the feedforward layer, the self-attention layer, and a second encoder-decoder attention

layer. There is a distinct set of weights for each encoder and decoder. Current image captioning algorithms get an excellent score by intuitively connecting informative parts of the image with transformer designs and attention. Some earlier transformer-based image captioning models, however, are limited in their ability to use the basic machine translation architecture of the transformer. A word in a text can be located to the left or right of another word, depending on how far apart they are. The degree of freedom in the relative spatial relationship between areas in images is more significant than in phrases [31]. This is because images are two- or three-dimensional, meaning a region might be anywhere besides the left or right of another region.

An encoder and a decoder are the two primary components of the transformer, as depicted in Fig. 3. Similar to parallel heads of self-attention, multi-head attention functions. The transformer uses self-attention to incorporate its understanding of other relevant terms into the word it is currently processing. The fully connected feedforward network is an additional component that consists of two linear transformations with different parameters at different layers, but that is the same across positions. To help with word position determination, the transformer adds a vector to each input embedding. Position embedding is a technique that considers the order of words in an input sequence. The vector generated by the decoder's stack is transformed into a larger vector known as a logit vector by the linear layer, a straightforward, fully connected neural network. The probabilities are supplied by SoftMax. The term related to the cell of highest probability is generated as output [59].

To obtain the attention scores, Fig.4 shows the scaled attention of the dot product in the left block, in which the self-attention computes the dot product of the query with all keys, which is then normalized using the SoftMax operator. The attention scores determine the weights, and each entity then becomes the weighted sum of all the entities in the sequence. On the other hand, on the right block, multihead attention consists of numerous self-attention blocks (h = 8 in the original Transformer model) to capture multiple complex interactions between various items in the sequence.

The language processing model encompasses three components: the transformer, the attention mechanism, and the ensemble learning model. A transformer-based model generates textual descriptions by taking the image features as input and producing a sequence of words that form the caption. In this work, the proposed language processing model uses the transformer with two key components: the encoder and the decoder (refer to Fig. 3). The image transformer utilized for image captioning will decode diverse information within image regions [60]. To establish the position of each word, the transformer introduces a vector added to each input embedding. Position embedding accounts for the sequential order of words in an input sequence. The linear layer, a straightforward, fully connected neural network, transforms the vector generated by the stack of decoders into a substantially larger vector referred to as a logit vector. Subsequently, SoftMax is applied to derive probabilities. The cell with the highest probability is selected, and the associated word becomes the output [59]. The transformer model addresses issues inherent in RNN and LSTM, facilitating increased parallelization and enhancing translation quality. Unlike LSTMs or RNNs, which process sentences one word at a time, transformer models are attention-based, capable of handling entire sentences [61].

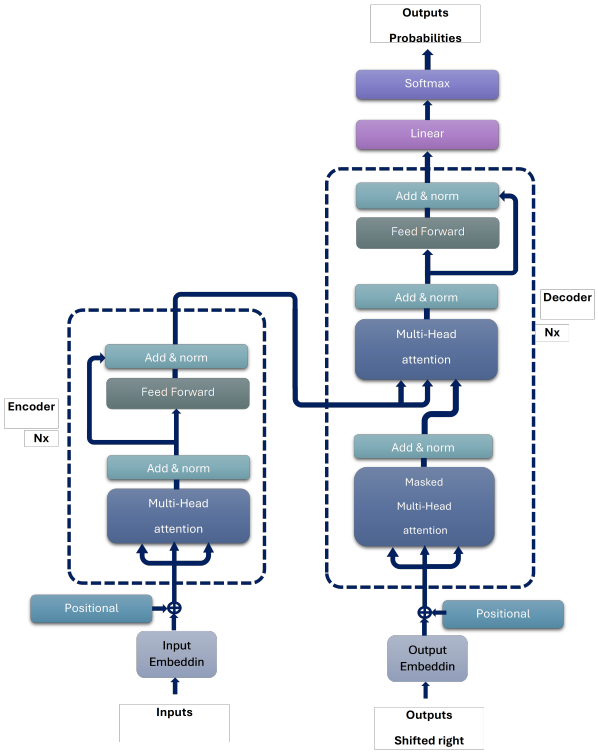


Figure 3: General Architecture of Transformer

3.3.3 The attention mechanism

Attention mechanisms focus on the most relevant features extracted by CNNs, which is crucial for tasks such as image captioning, where context is key. Attention in image processing mimics human attention patterns. Its strength lies in establishing meaningful connections between features and enhancing the models' ability to prioritize important features while filtering out noise. This aligns with the attention mechanisms that guide the focus of the model during training [4]. Despite the richness of the image data, not all features require explicit attention in captioning. When attention is integrated into the encoder-decoder picture captioning framework, sentence creation becomes contingent on hidden states computed using the attention method. The attention mechanism is a fundamental component of the encoder-decoder architecture within this framework. Using various types of input image patterns to guide the decoding process, ensuring that attention is focused on specific features of the input image at each time step. This composed focus on attention facilitates the generation of a descriptive caption for the input image [62].

Attention guides computations on significant regions to improve caption quality in image annotation. This is achieved by using soft and hard attention mechanisms to estimate the focus of attention. Soft attention, trainable via standard backpropagation,

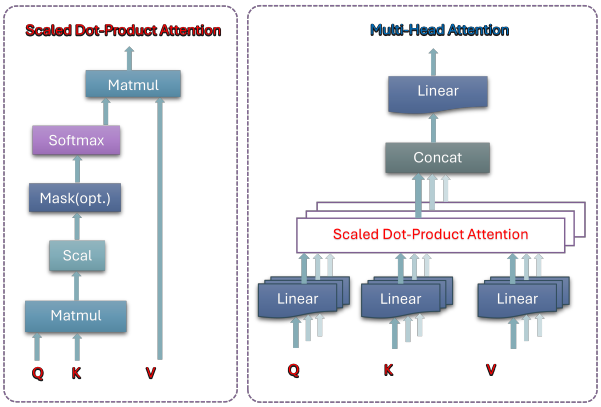


Figure 4: Scaled Dot Product Attention (left), Multi-Head Attention (right)

involves weighting the annotated vector of picture features when salient features are identified. On the other hand, stochastic hard attention is trained by maximizing a variation lower limit [45]. Recent studies have explored top-down and bottom-up attention theories, with recent experiments favoring top-down attention mechanisms [53]. Attentive encoder-decoder models lack global modeling skills. To address this, a reviewer module reviews encoder hidden states, producing a thought vector at each step. The attention mechanism plays a vital role in assigning weights to hidden states. These thought vectors capture global input aspects and effectively review and learn the encoded information from the encoder. Subsequently, the decoder uses these thought vectors to predict the next word in the sequence [62]. Visual attention in multimodal coverage mechanisms bridges the gap between encoder and decoder, improving data understanding [47, 63]. Scaled *Dot Product Attention*, introduced by [59], computes the dot products of the queries, the dimensions keys d_k, and the dimensions d_v values that make up the input; after that, the dot products of the query with all the calculated keys, divided by $\sqrt{d_k}$, and then a SoftMax function was applied to obtain the weights of the values. The attention function was continuously computed on a group of queries gathered into a matrix Q. The keys and values are also compacted in matrices K and V. The attention function is mathematically formalized in (1).

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

The mechanism of multi-head attention involves parallel passes through the attention mechanism. The formula expressed in (2) produces concatenated outputs to be transformed into the desired dimension [59].

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}\left(\text{head}_1, \ldots, \text{head}_\text{h}\right) W^O \\ \text{where head} &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{2}$$

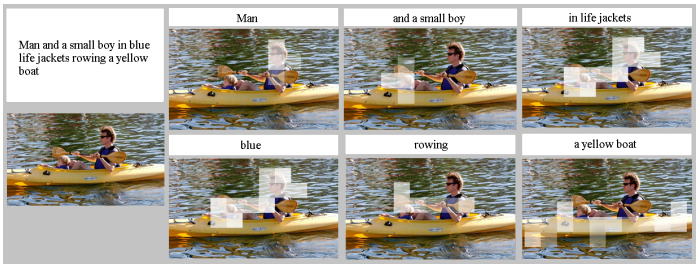


Figure 5: The visual architecture of the convolutional neural network.

The attention mechanism is implemented within the transformer, which allows the model to focus on different parts of the image when generating each word in the caption. It enhances the model's ability to align visual and textual information. Generally, individuals selectively attend to information, focusing on secondary data while disregarding certain primary data. This attention mechanism is essential for generation-based models within the encoder-decoder architecture, mirroring human visual focus in image captioning. In cognitive neurology, attention is identified as a shared higher cognitive skill that allows intentional oversight of received data. Originally proposed for image categorization, attention is widely used in NLP experiments, including machine translation, speech recognition, text understanding, and visual captioning [3, 64–66].

Fig. 5 visually depicts attention over time, illustrating how the model's focus shifts with the generation of each word to highlight relevant parts of the image. CNN highlights different areas of the image to draw attention to key components that are essential for correct interpretation. This detailed analysis demonstrates how CNNs analyze visual data to properly interpret and caption images. CNN starts by analyzing the image to find fundamental elements like borders and color spots. In this instance, it could pick up on the boy's and man's shapes and the boat's recognizable yellow color. More intricate characteristics are found; the network could identify the forms of a boat and a life jacket. As it gets farther into the network, CNN combines basic properties to detect things. The little child and the man are depicted as distinct individuals, each with a blue life jacket. The actions included in the image, through further analysis, acknowledge that the individuals are taking part in an activity (rowing). Combining all recognized details allowed for a comprehensive understanding of the scene about a "man and a small boy in blue life jackets rowing a yellow boat".

3.3.4 The Beam search algorithm

The greedy decoding technique outputs the word with the highest probability. However, it quickly accumulates potential errors. To solve this problem, the beam search

algorithm with a width of $k = 10$ was applied, maintaining k sequence candidates and selecting the most likely one at each step [67]. This approach generates a diverse group of captions. Previous studies supported beam search as the preferred algorithm for caption generation [68].

3.3.5 Ensemble learning

Typical learning techniques may not produce sufficient results because various features and the underlying structure of data are difficult for these methods to capture. So, building an effective model becomes a crucial problem in the data mining industry. An area of research that is gaining interest is ensemble learning, which tries to combine data fusion, data modeling, and data mining into a single framework. In which a set of features is first extracted using multiple learning algorithms to provide predictions based on these learned properties. Then, ensemble learning combines useful information to improve prediction accuracy across a variety of voting processes to outperform the results of any individual algorithm. Through the use of multiple machine learning algorithms, ensemble learning techniques generate weak predictions based on features extracted from a variety of data projections. The results are then fused with different voting mechanisms to produce performances that are better than those of any one of the constituent algorithms alone. Ensemble learning is used to improve architecture performance [23]. Several ensemble models exist[69], including bagging, boosting, stacking, voting:

a) Bagging: Breiman [70] created bootstrap aggregation, often known as bagging, to improve the classification performance of machine learning models by aggregating the predictions from randomly generated training sets. It was argued that bagging can increase accuracy because varying the learning set can result in appreciable changes to the predictor that is produced. In addition, diversity is achieved in bagging through the creation of bootstrapped copies of the input data, in which a number of randomly selected subsets are selected with replacements from the initial training set. As a result, the different training sets are considered distinct and are used to train different base learners for the same machine-learning algorithm.

b) Boosting: A machine learning method called "boosting" can turn a weak classifier into a powerful one. It is a kind of ensemble meta-algorithm that lowers variance and bias. A classifier that performs marginally better than random guessing is considered weak, whereas classifiers that achieve considerable accuracy are considered strong, and it is upon these classifiers that the boosting ensemble methods are based. [71] addressed the boosting algorithm regarding the possibility of producing a single strong learner from a group of weak learners.

c) Stacking: An ensemble learning framework called stacked generalization, or stacking, trains a different machine learning algorithm to aggregate the predictions of two or more ensemble members. Wolpert [72] first proposed an effort to reduce the generalization error in machine learning issues. When many machine learning models are particularly skilled at a specific position, stacking can be helpful. In this case,

the stacking strategy uses a different ML model to determine when to employ the predictions from the different models. It entails training a meta-learning algorithm to train a new model that combines the predictions from the base models with numerous base algorithms, the so-called level-0 models.

d) Voting: In problems involving regression and classification, the majority vote is the most widely used and logical combination approach [73]. The class with the majority vote is returned as the ensemble prediction in classification problems when the predictions for each class are added together. Meanwhile, regression tasks determine the majority vote by averaging the predictions made by each base learner. Let us assume that the t classifier's decision is $d_{t,c} \in \{0, 1\}$, $t = 1, \ldots, T$ and $c = 1, \ldots, C$, where T and C represent the number of classes and classifiers, respectively. Next, class ω_c* is chosen as the ensemble forecast by majority voting if

$$\sum_{t=1}^{T} d_{t,c} = \max_{c} \sum_{t=1}^{T} d_{t,c} \tag{3}$$

Several separate models are combined in ensemble learning to improve generalization performance. Deep learning architectures are now performing better than standard or shallow models. To improve the generalization performance of the final model, deep-enhanced learning models integrate the benefits of ensemble learning and deep learning. Using ensemble learning, architecture can operate effectively by combining its various parts to achieve a single objective. Numerous ensemble models exist, including voting, bagging, boosting, and stacking [33]. Voting combines predictions from multiple models, making the overall system more robust; even if individual models fail or make errors, the ensemble can still provide reliable results, and it allows the combination of different model architectures or pre-trained embeddings to enhance the overall understanding of image content. In addition, ensemble methods such as voting reduce overfitting by averaging out model biases and generalizing better to unseen data [74] [37]. We employ the voting approach to aggregate the predictions made by each technique. To obtain a more robust and accurate caption, the ensemble learning model trains multiple instances of the transformer with different random initializations or hyper-parameters and then combines their output through voting. The voting model is presented in Fig. 6 to combine the results of each of the eight transformer models. The BLEU score-1 was considered for this purpose, and the prediction result will be accepted from the model that gains the highest BLEU score.

4 Experimental results

This section presents the results obtained from the proposed model and compares them with the latest models.

4.1 Environment setup

To assess the performance of the proposed model, a set of experiments was conducted using the Google Colab Pro+ framework, equipped with 52 GB of RAM and 1 TB of

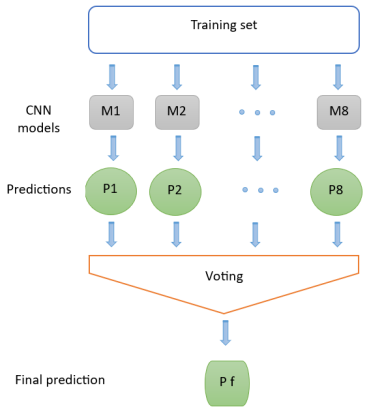


Figure 6: Voting model architecture.

storage capacity for implementation purposes. The proposed model was trained with a batch size of 64, employing the Adam optimizer [75], a learning rate set at 0.00001, 30 epochs with early stopping, and the ReLU activation function was utilized.

Our model employs a structured approach to training an image captioning system, focusing on effectively managing the learning rate. The loss function is defined using cross-entropy, which calculates the loss between predicted and true labels without reduction. To prevent overfitting, early stopping is implemented to monitor validation loss and halt training if no improvement is observed after a predetermined number of epochs, restoring the best model weights. A custom learning rate scheduler is utilized to dynamically adjust the learning rate throughout the training process. It starts at a low rate of 0.00001 and gradually increases, facilitating a stable training experience that enhances convergence and overall model performance. This scheduler works with the Adamax optimizer during model compilation, ensuring effective optimization while reducing the risk of overfitting. This approach promotes stable convergence during training and improves the model's ability to generalize to unseen data, which is essential for generating accurate and contextually relevant captions for images. The chosen parameters are based on empirical observations and established best practices in neural network training, with the aim of balancing efficient convergence and stable training while minimizing the risk of overfitting.

4.2 Evaluation metrics

While direct human judgment is the simplest way to evaluate text generated for images, scalability is challenging due to nonreusable human effort and subjective nature. To overcome these challenges, various evaluation metrics assess the performance of image captioning systems. These metrics measure the systems' ability to generate linguistically acceptable and semantically valid phrases. However, the choice of the most significant metric depends on the specific objectives of the image captioning task. BLEU and ROUGE are often considered standard. However, recent research has shown the value of incorporating diverse metrics such as METEOR, CIDEr, and SPICE to

provide a more comprehensive evaluation and performance results. The evaluation metrics applied in this study include BLEU, ROUGE, METEOR, CIDEr, and SPICE. Table1 provides a summary of common assessment metrics in image captioning, while the following section discusses them in more detail.

Table 1: Performance assessment metrics in image captioning

Metric	Evaluation task	Methodology
BLEU[76]	Machine translation	n-gram precision
ROUGE[77]	Document summarization	n-gram recall
METEOR [78]	Machine translation	n-gram with synonym matching
CIDEr [79]	Image captioning	tf-idf weighted n-gram similarity
SPICE[80]	Image captioning	Scene-graph synonym matching

4.2.1 Bilingual evaluation understudy (BLEU)

BLEU is a metric that evaluates the quality of machine-generated text by comparing individual segments to a set of reference texts [76]. Its approach varies with the number of references and the length of the text. BLEU scores are higher for short autogenerated text and range from 0 to 1. The comparisons of the gram and the bigram determine BLEU-1 and BLEU-2, with an empirically determined maximum order of four for optimal correlation with human judgments. BLEU assesses adequacy through unigram scores and fluency through higher n-gram scores. Although widely used and language-independent, BLEU has drawbacks. It favors brief output texts, and a high score does not guarantee higher quality, making it imperfect for certain evaluations [4].

4.2.2 Recall-oriented understudy for gisting evaluation (ROUGE)

ROUGE is a set of measures that evaluate text summaries by comparing word sequences and pairs to a database of human-written reference summaries [81]. Originally designed for machine translation accuracy and fluency assessment, it quantifies sentence-level similarity using the longest common subsequence between candidate and reference sentences. Similarly to BLEU, ROUGE is also computed by varying the n-gram count. However, unlike BLEU, which is based on precision, ROUGE is based on recall values. It captures sentence-level structure with in-sequence word matches, allowing non-sequential matching. ROUGE-L is the version that is used in the evaluation of image and video captioning. It calculates the recall and precision scores of the longest common subsequences (LCS) between each generated sentence and its corresponding reference sentence.

4.2.3 Metric for explicit ordering translation evaluation (METEOR)

METEOR is designed for machine translation evaluation and is considered more valuable than BLEU, with a stronger link to human evaluations [78]. It calculates scores based on generalized unigram matches between a candidate sentence and human-written reference sentences. The precision, recall, and alignment of the matched words

contribute to the score computation. In cases with multiple reference sentences, the candidate's final evaluation considers the best score among independently computed ones. METEOR considers unigram overlap and incorporates additional features like stemming and synonymy matching. It aims to address some limitations of BLEU and ROUGE by providing a more comprehensive evaluation [4].

4.2.4 Consensus-based image description evaluation (CIDEr)

CIDEr is an image caption quality assessment paradigm that relies on human consensus [79]. Assesses the similarity of a generated sentence to a set of human-written ground-truth sentences. Using the TF-IDF weighting for each n-gram in the candidate phrase, CIDEr encodes their frequency in reference sentences. CIDEr evaluates the grammar, significance, and accuracy of image captions and descriptions. Unlike metrics that work with a limited number of captions per image, CIDEr employs consensus utilization, making it suitable for analyzing the agreement between generated captions and human assessments [4].

4.2.5 Semantic propositional image caption evaluation (SPICE)

SPICE is a semantic concept-based image caption evaluation metric based on semantic scene graphs [80]. It uses a graph-based semantic representation extracted from image descriptions [1]. Generated and ground-truth captions are converted into an intermediate scene graph representation through semantic parsing to calculate the SPICE score. The F1 score derived from precision and recall measures the similarity between the generated and ground-truth caption scene graphs.

4.3 Quantitative analysis

Table 2 compares the results obtained by the proposed model with the latest methods and Fig. 7. As will be discussed soon, the proposed model exhibits superior performance, with the highest scores highlighted in bold. The result includes the research with models based on the Flickr8k dataset. The proposed model achieved the highest scores in BLEU-1, BLEU-2, and BLEU-3: 0.728, 0.495, and 0.323, respectively. These scores indicate how well our system's predictions align with reference captions on n-gram overlap. Our model obtained the highest result of the METEOR score, 0.604, which evaluates the semantic similarity between the generated and reference captions. The SPICE score, which focuses on semantic content overlap, was used to assess the quality of image captions. Our model achieved the highest SPICE value of 0.164, indicating strong semantic alignment. We also get competitive results for ROUGE L (0.432) and CIDEr (0.604). ROUGE L measures the longest common subsequence between generated and reference captions, CIDEr considers word frequency and diversity.

To demonstrate the effectiveness of the suggested model, we contrasted its results with those of the most advanced models on Flickr30k datasets, as indicated in Table 3 and Fig. 8. The suggested model outperformed the latest methods in Flickr30k datasets, as demonstrated by the table, as determined by the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores. ROUGE L METEOR CIDEr showed that the model

performs similarly to the state-of-the-art. It indicates that the proposed approach can provide meaningful and understandable human captions. The results derived from alternative metrics, including SPICE (0.387), confirm the efficacy of the model. It is important to remember that the majority of other methods do not share their results on this metric. SPICE scores provide a unique evaluation metric for image captioning by focusing on the semantic content of generated captions rather than just word overlap. In addition, SPICE scores correlate better with human judgments of caption quality, making them a more reliable measure of how well captions align with human expectations.

Table 2: Comparison of Flicker8K dataset image captioning

Reference	B1↑	B2↑	B3↑	B4↑	ROUGE L↑	METEOR↑	CIDEr↑	SPICE↑
[26] Karpathy et al.(2015)	0.579	0.383	0.245	0.160	NA	NA	NA	NA
[82] Jiang et al.(2019)	0.690	0.471	0.324	0.219	0.502	0.203	0.507	NA
[83] Patel et al.(2020)	0.601	0.414	0.274	0.181	0.433	0.183	0.452	NA
[84] Katpally et al.(2020)	0.634	0.400	0.287	0.151	NA	NA	NA	NA
[85] Bineeshia et al.(2021)	0.589	0.335	0.263	0.148	NA	NA	NA	NA
[25] Dahri et al.(2023)	0.603	0.360	0.220	0.122	NA	NA	NA	NA
[86] Ma et al.(2023)	0.674	NA	NA	0.243	0.448	0.215	0.636	NA
The Proposed Model	**0.728**	**0.495**	**0.323**	0.208	0.432	**0.235**	0.604	**0.164**

Table 3: Comparison of Flicker30K dataset image captioning

Reference	B1↑	B2↑	B3↑	B4↑	ROUGE L↑	METEOR↑	CIDEr↑	SPICE↑
[26] Karpathy et al.(2015)	0.573	0.369	0.240	0.157	NA	NA	NA	NA
[87] You et al.(2016)	0.647	0.460	0.324	0.230	0.189	NA	NA	NA
[88] Fu et al.(2016)	0.649	0.462	0.324	0.224	0.451	0.194	0.472	NA
[89] Lu et al.(2017)	0.677	0.494	0.354	0.251	0.204	NA	0.531	NA
[90] He et al.(2019)	0.666	0.484	0.346	0.247	0.467	0.202	0.524	NA
[82] Jiang et al.(2019)	0.689	0.468	0.319	0.220	0.487	0.191	0.428	NA
[58] Do et al.(2020)	0.695	0.463	0.341	0.232	0.451	0.302	0.486	NA
[91] Kalimuthu et al.(2021)	0.647	0.456	0.320	0.224	0.449	0.197	0.467	0.136
[86] Ma et al.(2023)	0.671	NA	NA	0.233	0.443	0.204	0.645	NA
[92] Abdussalam et al.(2023)	0.694	0.498	0.355	0.254	0.538	0.251	0.469	NA
The Proposed Model	**0.798**	**0.561**	**0.387**	**0.269**	0.443	0.213	0.565	**0.387**

4.4 Qualitative analysis

As demonstrated in Fig. 7 and Fig. 8, we have provided several sentences produced by our caption method to validate the effectiveness of our model. In general, our model demonstrates proficiency in generating captions that are not only relevant but also accurate in describing the image content. Fig. 7 presents samples of nearly correct captions from the Flicker8K dataset. Green text is used to identify the generated captions. As noted in Fig. 7.b, where the man is not riding the bike in the position in the picture, it is revealed that he is performing a trick by "do the trick". In Fig. 7.c, despite the terms "forest" and "wood" appearing in the references, the model was able to accurately depict the appearance of wood in the image accurately; however,

since the image is not a green forest, the description produced by the system used the more correct word "wood". The description of the sites was available through the picture Fig. 7 .a in the sentence "dirt road", through the second picture Fig. 7.f in the sentence "over a snowy hill", and Fig. 7.g on a beach. The "uniform" was generated for both Fig. 7. d and Fig. 7.h to clarify that they belong to a certain outfit.

On the other hand, Fig.8 shows the correct samples from Flicker30k. It is clear in Fig.8.e, how the number of women was accurately identified for the first scenario, particularly in settings with complex backgrounds, counting the number of items is a higher level of artificial intelligence than object recognition [82], we can observe how the model determines the gender, (man) in figures: Fig.8.c, Fig.8.d, Fig.8.h, and Fig.8.g, and (woman) in Fig.8.b, Fig.8.e, and Fig.8.i. An illustration of how the model can represent the location of an object as "in front of" is provided in Fig.8.g. Furthermore, the example in Fig.8.b "is performing a trick" and Fig.8.a "in a white uniform" effectively conveys the setting. "Is eating" in Fig.8.h and "Is cooking" in Fig.8.i serve as an illustration of how to differentiate between the actions related to a meal. Objects such as "saxophone" Fig.8.f and "javelin" Fig.8.d were correctly recognized. However, in Fig.8.c it is evident that the model interprets the white area as the color of the shirt. A single thing might have several characteristics depending on the situation at hand. Learning to identify attributes in computer vision is still a challenging task [82].

However, incorrect captions are shown in Fig.9 from the Flicker8K dataset. The right figure mistakenly describes the signs in the man's shirt as "hold a sign", whereas the left figure was incorrectly described. There are errors in the generated caption of the Flicker30k dataset, examples displayed in Fig.10, and the created captions are identified by red text. The left figure (the obstacle on a red track) was incorrectly described as "soccer on a field." However, the number of players was correctly listed as "two". In the right figure, the model produced the place in the caption "standing on a rocky mountain," while the number (two men) was incorrectly provided as "a man." These inaccurate captions show how the existing image captioning model has difficulty recognizing actions, context, and complicated settings. The model may fail to recognize context, leading to erroneous interpretations. For future plans, it may be necessary to detect and distinguish items within images effectively; object recognition algorithms should be improved. Improved context analysis methods should also be used to understand the connections between actions and objects.

4.5 Ablation study

An ablation study was conducted to assess the contributions of individual components within the proposed ensemble model for image captioning. The Flicker8K dataset was utilized for this purpose. In this study, evaluation metrics such as BLEU, ROUGE, METEOR, CIDEr, and SPICE were applied. This section outlines the methodology used and presents the findings, highlighting the importance of each model within the architecture. The following ensemble configurations were analyzed:

1. **Baseline 1:** MobileNetV2, VGG16, VGG19, and ResNet50. These CNN models were used for feature extraction, capturing essential visual elements from the images before the transformer processes the extracted features for text generation.

2. **Baseline 2:** MobileNetV2, VGG16, VGG19, ResNet50, and ResNet101. Like Baseline 1, this configuration incorporates additional CNNs to enhance feature extraction, providing a richer representation for the transformer during caption generation.
3. **Baseline 3:** MobileNetV2, VGG16,VGG19, ResNet50, ResNet101, RegNetX120 and EfficientNetB4. This ensemble combines multiple CNN models to maximize feature extraction capabilities, allowing the transformer to generate more accurate and contextually relevant captions.
4. **Full Ensemble Model:** In the full model, we evaluated the performance of the complete ensemble model, which integrates the outputs of all CNNs and a transformer language model to generate captions.

The results are summarized and compared in Table 4.

Table 4: Comparison of ensemble models using Flicker8K dataset

Model	B1↑	B2↑	B3↑	B4↑	ROUGE L↑	METEOR↑	CIDEr↑	SPICE↑
Baseline 1	0.656	0.428	0.269	0.165	0.395	0.208	0.446	0.132
Baseline 2	0.697	0.470	0.304	0.192	0.419	0.226	0.544	0.154
Baseline 3	0.705	0.476	0.309	0.197	0.421	0.227	0.545	0.156
Full ensemble model	**0.728**	**0.495**	**0.323**	**0.208**	**0.432**	**0.235**	**0.604**	**0.164**

The results indicate that the ensemble model consistently outperforms each base model in all metrics, validating the hypothesis that the combination of multiple models enhances feature representation. The complete ensemble demonstrated the highest scores on all metrics, indicating that the diversity of features captured by multiple CNNs leads to improved caption quality. This ensemble approach mitigates the limitations inherent in individual models by leveraging their unique strengths.

4.6 Discussion

The results demonstrated how our model differs from other methods in feature text extraction by focusing on salient image regions and characteristics through attention mechanisms. Table 2 and Table 3 highlight the important distinctions between our suggested model and the other models and emphasize the research contributions in the following areas:(1) enhanced prediction robustness: In contrast to previous methods, our model uses an ensemble learning strategy, which effectively combines eight CNN models via a voting process, to fine-tune the ideal caption for every image. This increases the architecture's robustness and generalizability, while greatly improving its efficiency. Our model efficiently reduces overfitting by combining predictions from many base models, resulting in a more robust and flexible solution. (2) comprehensive evaluation metrics: to gain a deeper understanding of the model's capabilities, we used a methodology in this research work that took a variety of indicators into account. A more realistic description of the overall performance of a recently suggested model in this growing field will come from a comprehensive evaluation that considers multiple factors.

To further validate our findings, we conducted paired t-tests between the full ensemble and each ablation variant. The results revealed that the performance differences were statistically significant ($p < 0.05$), reinforcing the necessity of using an ensemble model.

Several models continue to show difficulties, such as explosions in gradients and inaccurate sentence construction, which impact the image encoding and description process. Most modern captioning models use LSTM and RNNs as language models. However, long-term information must go through every cell before arriving at the present processing cell, since RNN and LSTM operate sequentially. As a result, it is readily tainted by repeatedly multiplying it by small values smaller than zero, leading to vanishing gradients that delay updating the network weights and the learning process. Some, but not all, of the issues with the disappearing gradient can be resolved via LSTM. Moreover, LSTM and RNN require additional resources and are not hardware-friendly. While LSTM-based image descriptions yield remarkable results, putting them into images requires more time, work, and parameters. An innovative caption creation framework was introduced by [93], the EnsCaption framework, combining caption generation and retrieval with a re-ranking procedure and adversarial network for improved accuracy. While EnsCaption shows strong performance it has limitation in recognition of fine-grained features.

Using an ensemble of CNN models in our image captioning framework significantly enhances performance, improving accuracy, and robustness. Ensembles generalize more effectively to unseen data, which is crucial for applications requiring high model reliability. This is especially significant in domains with high variability, such as medical imaging or environmental monitoring. However, this enhancement increases computational demands, leading to longer training times. In critical fields such as medical imaging, even a slight improvement in accuracy can impact patient outcomes, as improved detection rates can reduce false negatives and ensure timely treatment. The significance of these trade-offs is also evident in security applications, where accurate image analysis improves resource allocation, and in disaster response, where precise facial recognition can identify potential threats. Therefore, despite the considerable computational requirements, the significant advantages in essential applications justify adopting ensemble methods.

4.7 Real-World applications of the proposed image captioning model

Our proposed image captioning model could be applied in real-world scenarios in enhanced search engines, a search engine integrates an image captioning model to improve the search experience by delivering more informative and relevant image results. When a user enters a search query for images, the model analyzes each image in the database and generates detailed captions that accurately describe the content. These captions are indexed alongside the images, enabling the search engine to retrieve results that align more closely with the user's query. Another compelling use case is assistive technology for the visually impaired. In this scenario, visually impaired individuals use a mobile application to capture photos of their environment. The image captioning system analyzes these images and generates descriptive audio captions that

highlight key elements. The application significantly improves the user's quality of life by facilitating greater interaction with their surroundings.

5 Limitation and future work

The proposed Attention-Based Transformer Model for Image Captioning encounters several challenges. First, the model sometimes misinterprets the color of certain areas as corresponding to different areas or clothing items, highlighting the difficulty of recognizing multiple attributes associated with a single factor in computer vision. Second, there are instances where the model does not accurately count the number of elements in the target image; this task involves a higher level of artificial intelligence than simple object recognition. In addition, the model may struggle to understand complex settings, leading to incorrect interpretations. Lastly, a key limitation that affects model performance in image captioning is the presence of noisy or ambiguous images. Although this issue falls outside the scope of the current research, it is important to note that such images, characterized by distracting elements, low resolution, or unclear subjects, can hinder the model's ability to accurately interpret visual content, resulting in incorrect or irrelevant captions.

Future advances in image captioning can effectively address existing limitations through focused research initiatives. First, improving object recognition algorithms will enhance the model's ability to detect and distinguish items within images accurately. Second, improving the robustness of the model against noisy or ambiguous input to improve the quality of the caption and the overall performance in various real-world scenarios. Developing more comprehensive evaluation metrics can also offer a deeper assessment of caption quality. Furthermore, exploring how to handle diverse datasets effectively, ensuring that the model can generalize well across different scenarios. Finally, enhancing the visualization of image captioning models will allow insight into how the model focuses on specific areas of an image when generating captions.

6 Conclusions

Converting an input image into an explanation in words is known as image captioning. It can be used in various situations, including social networking, smart travel, assisting the blind, medical image captioning for healthcare, education, and training of children. Competition among researchers is leading to an increase in the number of unique models. In this research, we thoroughly investigated the transformer model and provided many helpful ways to attract attention. We have shown the potential for significant advancement in this field by implementing an attention mechanism in a transformer-based design. We introduce a novel ensemble learning framework to generate captions based on a voting mechanism that selects the caption with the highest bilingual evaluation understudy (BLEU) score. This framework enhances the richness of the captions generated by utilizing multiple deep neural network architectures. The robustness and efficacy of the proposed approach have been demonstrated by a comprehensive analysis of the Flickr8K and Flickr30K datasets combined with common

metrics and measurements. We believe that our approach will encourage future research to explore transformer-based designs further and push the limits of what is possible in multilingual image captioning.

7 Declaration

Data Availability: The datasets used in this study are publicly available and can be downloaded by the following links: Flickr8k (https://www.kaggle.com/datasets/adityajn105/flickr8k), Flickr30K https://www.kaggle.com/datasets/eeshawn/flickr30k)

References

[1] Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR) **51**(6), 1–36 (2019)

[2] Cheikh, M., Zrigui, M.: Active learning based framework for image captioning corpus creation. In: International Conference on Learning and Intelligent Optimization, pp. 128–142 (2020). Springer

[3] Yu, J., Li, J., Yu, Z., Huang, Q.: Multimodal transformer with multi-view visual representation for image captioning. IEEE transactions on circuits and systems for video technology **30**(12), 4467–4480 (2019)

[4] Ghandi, T., Pourreza, H., Mahyar, H.: Deep learning approaches on image captioning: A review. ACM Comput. Surv. **56**(3) (2023) https://doi.org/10.1145/3617592

[5] Ayesha, H., Iqbal, S., Tariq, M., Abrar, M., Sanaullah, M., Abbas, I., Rehman, A., Niazi, M.F.K., Hussain, S.: Automatic medical image interpretation: State of the art and future directions. Pattern Recognition, 107856 (2021)

[6] Ogura, A., Hayashi, N., Negishi, T., Watanabe, H.: Effectiveness of an e-learning platform for image interpretation education of medical staff and students. Journal of digital imaging **31**(5), 622–627 (2018)

[7] Depeursinge, A., Al-Kadi, O.S., Mitchell, J.R.: Biomedical Texture Analysis: Fundamentals, Tools and Challenges. Academic Press, ??? (2017)

[8] Al-Kadi, O.S.: Tumour grading and discrimination based on class assignment and quantitative texture analysis techniques. PhD thesis, University of Sussex (2010)

[9] Chendake, P., Korpal, P., Bhor, S., Bansal, R., Patil, S., Deshpande, D.: Learning system for kids. International Journal of Recent Advances in Multidisciplinary Topics **2**(6), 71–75 (2021)

[10] Ayyoub, H.Y., Al-Kadi, O.S.: Learning style identification using semi-supervised self-taught labeling. IEEE Transactions on Learning Technologies (2024)

[11] Ahsan, H., Bhatt, D., Shah, K., Bhalla, N.: Multi-modal image captioning for the visually impaired. In: Durmus, E., Gupta, V., Liu, N., Peng, N., Su, Y. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 53–60. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.naacl-srw.8 . https://aclanthology.org/2021.naacl-srw.8

[12] Nivedita, M., Chandrashekar, P., Mahapatra, S., Phamila, Y.A.V., Selvaperumal, S.K.: Image captioning for video surveillance system using neural networks. International Journal of Image and Graphics, 2150044 (2021)

[13] Wang, Z., Huang, Z., Luo, Y.: Paic: Parallelised attentive image captioning. In: Australasian Database Conference, pp. 16–28 (2020). Springer

[14] Saleem, S., Dilawari, A., Khan, U.G., Iqbal, R., Wan, S., Umer, T.: Stateful human-centered visual captioning system to aid video surveillance. Computers & Electrical Engineering **78**, 108–119 (2019) https://doi.org/10.1016/j.compeleceng.2019.07.009

[15] Shuster, K., Humeau, S., Hu, H., Bordes, A., Weston, J.: Engaging image captioning via personality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12516–12526 (2019)

[16] Fujiyoshi, H., Hirakawa, T., Yamashita, T.: Deep learning-based image recognition for autonomous driving. IATSS research **43**(4), 244–252 (2019)

[17] Guinness, D., Cutrell, E., Morris, M.R.: Caption crawler: Enabling reusable alternative text descriptions using reverse image search. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1–11 (2018)

[18] Xiao, F., Zhang, N., Xue, W., Gao, X.: Sentinel mechanism for visual semantic graph-based image captioning. Computers & Electrical Engineering **119**, 109626 (2024) https://doi.org/10.1016/j.compeleceng.2024.109626

[19] Zhang, Z., Zhang, H., Wang, J., Sun, Z., Yang, Z.: Generating news image captions with semantic discourse extraction and contrastive style-coherent learning. Computers and Electrical Engineering **104**, 108429 (2022) https://doi.org/10.1016/j.compeleceng.2022.108429

[20] Bai, S., An, S.: A survey on automatic image caption generation. Neurocomputing **311**, 291–304 (2018)

[21] Al-Kadi, O.S.: Combined statistical and model based texture features for improved image classification. In: 4th IET International Conference on Advances in

Medical, Signal and Information Processing-MEDSIP 2008, pp. 1–4 (2008). IET

[22] Al-Kadi, O.S.: Supervised texture segmentation: a comparative study. In: 2011 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), pp. 1–5 (2011). IEEE

[23] Dong, X., Yu, Z., Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Frontiers of Computer Science **14**, 241–258 (2020)

[24] Verma, A., Yadav, A.K., Kumar, M., Yadav, D.: Automatic image caption generation using deep learning. Multimedia Tools and Applications **83**(2), 5309–5325 (2024)

[25] Dahri, F.H., Chandio, A.A., Dahri, N.A., Soomro, M.A.: Image caption generator using convolutional recurrent neural network feature fusion. Journal of Xi'an Shiyou University, Natural Science Edition **9**, 1088–1095 (2023)

[26] Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)

[27] Chu, Y., Yue, X., Yu, L., Sergei, M., Wang, Z.: Automatic image captioning based on resnet50 and lstm with soft attention. Wireless Communications and Mobile Computing **2020**, 1–7 (2020)

[28] Fei, Z.: Attention-aligned transformer for image captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 607–615 (2022)

[29] Dubey, S., Olimov, F., Rafique, M.A., Kim, J., Jeon, M.: Label-attention transformer with geometrically coherent objects for image captioning. Information Sciences **623**, 812–831 (2023)

[30] Guo, L., Liu, J., Zhu, X., Yao, P., Lu, S., Lu, H.: Normalized and geometry-aware self-attention network for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10327–10336 (2020)

[31] He, S., Liao, W., Tavakoli, H.R., Yang, M., Rosenhahn, B., Pugeault, N.: Image captioning through image transformer. In: Proceedings of the Asian Conference on Computer Vision (2020)

[32] Velioglu, R., Rose, J.: Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. arXiv preprint arXiv:2012.12975 (2020)

[33] Meel, P., Vishwakarma, D.K.: Han, image captioning, and forensics ensemble multimodal fake news detection. Information Sciences **567**, 23–41 (2021)

[34] Zhong, J., Cao, Y., Zhu, Y., Gong, J., Chen, Q.: Multi-channel weighted fusion for image captioning. The Visual Computer, 1–18 (2022)

[35] Dalla Serra, F., Deligianni, F., Dalton, J., O'Neil, A.Q.: Cmre-uog team at imageclefmedical caption 2022: Concept detection and image captioning (2022)

[36] Salur, M.U., Aydın, İ.: A soft voting ensemble learning-based approach for multimodal sentiment analysis. Neural Computing and Applications **34**(21), 18391–18406 (2022)

[37] Singh, D., Kaur, M., Alanazi, J.M., AlZubi, A.A., Lee, H.-N.: Efficient evolving deep ensemble medical image captioning network. IEEE Journal of Biomedical and Health Informatics (2022)

[38] Kim, B.C., Kim, H.C., Han, S., Park, D.K.: Inspection of underwater hull surface condition using the soft voting ensemble of the transfer-learned models. Sensors **22**(12), 4392 (2022)

[39] Abu-Srhan, A., Abushariah, M.A., Al-Kadi, O.S.: The effect of loss function on conditional generative adversarial networks. Journal of King Saud University-Computer and Information Sciences **34**(9), 6977–6988 (2022)

[40] Abu-Srhan, A., Almallahi, I., Abushariah, M.A., Mahafza, W., Al-Kadi, O.S.: Paired-unpaired unsupervised attention guided gan with transfer learning for bidirectional brain mr-ct synthesis. Computers in Biology and Medicine **136**, 104763 (2021)

[41] Alkadi, O., Moustafa, N., Turnbull, B., Choo, K.-K.R.: A deep blockchain framework-enabled collaborative intrusion detection for protecting iot and cloud networks. IEEE Internet of Things Journal **8**(12), 9463–9472 (2020)

[42] Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research **47**, 853–899 (2013)

[43] Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)

[44] van Miltenburg, E.: Pragmatic factors in (automatic) image description. PhD thesis, Vrije Universiteit Amsterdam (October 2019)

[45] Oluwasammi, A., Aftab, M.U., Qin, Z., Ngo, S.T., Doan, T.V., Nguyen, S.B., Nguyen, S.H., Nguyen, G.H.: Features to text: a comprehensive survey of deep learning on semantic segmentation and image captioning. Complexity **2021** (2021)

[46] Almanaseer, W., Alshraideh, M., Alkadi, O.: A deep belief network classification approach for automatic diacritization of arabic text. Applied Sciences **11**(11), 5228 (2021)

[47] Biswas, R., Barz, M., Sonntag, D.: Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking. KI-Künstliche Intelligenz **34**(4), 571–584 (2020)

[48] Chen, J., Zhuge, H.: A news image captioning approach based on multimodal pointer-generator network. Concurrency and Computation: Practice and Experience **34**(7), 5721 (2022)

[49] Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L.: Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. Journal of big Data **8**(1), 1–74 (2021)

[50] Faiyaz Khan, M., Sadiq-Ur-Rahman, S., Islam, S., *et al.*: Improved bengali image captioning via deep convolutional neural network based encoder-decoder model. In: Proceedings of International Joint Conference on Advances in Computational Intelligence, pp. 217–229 (2021). Springer

[51] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)

[52] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)

[53] Staniūtė, R., Šešok, D.: A systematic literature review on image captioning. Applied Sciences **9**(10), 2024 (2019)

[54] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR

[55] Marques, G., Agarwal, D., Torre Díez, I.: Automated medical diagnosis of covid-19 through efficientnet convolutional neural network. Applied soft computing **96**, 106691 (2020)

[56] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520 (2018)

[57] Indraswari, R., Rokhana, R., Herulambang, W.: Melanoma image classification based on mobilenetv2 network. Procedia Computer Science **197**, 198–207 (2022) https://doi.org/10.1016/j.procs.2021.12.132 . Sixth Information Systems International Conference (ISICO 2021)

[58] Carmo Nogueira, T., Vinhal, C.D.N., Cruz Júnior, G., Ullmann, M.R.D.: Reference-based model using multimodal gated recurrent units for image captioning. Multimedia Tools and Applications **79**(41), 30615–30635 (2020)

[59] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, ., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

[60] Gong, L., Crego, J.M., Senellart, J.: Enhanced transformer model for data-to-text generation. In: Proceedings of the 3rd Workshop on Neural Generation and Translation, pp. 148–156 (2019)

[61] Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S., *et al.*: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)

[62] Bai, S., An, S.: A survey on automatic image caption generation. Neurocomputing **311**, 291–304 (2018) https://doi.org/10.1016/j.neucom.2018.05.080

[63] Chen, J., Zhuge, H.: A news image captioning approach based on multimodal pointer-generator network. Concurrency and Computation: Practice and Experience, 5721 (2019)

[64] Pedersoli, M., Lucas, T., Schmid, C., Verbeek, J.: Areas of attention for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1242–1250 (2017)

[65] Mishra, S.K., Dhir, R., Saha, S., Bhattacharyya, P., Singh, A.K.: Image captioning in hindi language using transformer networks. Computers & Electrical Engineering **92**, 107114 (2021) https://doi.org/10.1016/j.compeleceng.2021.107114

[66] Wang, H., Zhang, Y., Yu, X.: An overview of image caption generation methods. Computational intelligence and neuroscience **2020** (2020)

[67] Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From show to tell: A survey on deep learning-based image captioning. IEEE transactions on pattern analysis and machine intelligence **45**(1), 539–559 (2022)

[68] Li, J., Monroe, W., Jurafsky, D.: A simple, fast diverse decoding algorithm for neural generation. ArXiv **abs/1611.08562** (2016)

[69] Mienye, I.D., Sun, Y.: A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access **10**, 99129–99149 (2022)

[70] Breiman, L.: Bagging predictors. Machine learning **24**, 123–140 (1996)

[71] Schapire, R.E.: The strength of weak learnability. Machine learning **5**, 197–227 (1990)

[72] Wolpert, D.H.: Stacked generalization. Neural Networks **5**(2), 241–259 (1992) https://doi.org/10.1016/S0893-6080(05)80023-1

[73] Ballabio, D., Todeschini, R., Consonni, V.: Chapter 5 - recent advances in high-level fusion methods to classify multiple analytical chemical data. In: Cocchi, M. (ed.) Data Fusion Methodology and Applications. Data Handling in Science and Technology, vol. 31, pp. 129–155. Elsevier, ??? (2019). https://doi.org/10.1016/B978-0-444-63984-4.00005-3 . https://www.sciencedirect.com/science/article/pii/B9780444639844000053

[74] Ahad, A.: Vote-based: Ensemble approach. Sakarya University Journal of Science **25** (2021) https://doi.org/10.16984/saufenbilder.901960

[75] Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2014)

[76] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002). https://doi.org/10.3115/1073083.1073135 . https://aclanthology.org/P02-1040

[77] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)

[78] Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization, pp. 65–72 (2005)

[79] Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575 (2015)

[80] Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: European Conference on Computer Vision, pp. 382–398 (2016). Springer

[81] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries, p. 10 (2004)

[82] Jiang, T., Zhang, Z., Yang, Y.: Modeling coverage with semantic embedding for image caption generation. The Visual Computer **35**(11), 1655–1665 (2019)

[83] Patel, A., Varier, A.: Hyperparameter analysis for image captioning. arXiv preprint arXiv:2006.10923 (2020)

[84] Katpally, H., Bansal, A.: Ensemble learning on deep neural networks for image caption generation. In: 2020 IEEE 14th International Conference on Semantic Computing (ICSC), pp. 61–68 (2020). IEEE

[85] Bineeshia, J.: Image caption generation using cnn-lstm based approach. In: Proceedings of the First International Conference on Combinatorial and Optimization, ICCAP 2021, December 7-8 2021, Chennai, India (2021)

[86] Ma, Y., Ji, J., Sun, X., Zhou, Y., Ji, R.: Towards local visual modeling for image captioning. Pattern Recognition **138**, 109420 (2023)

[87] You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4651–4659 (2016). https://doi.org/10.1109/CVPR.2016.503

[88] Fu, K., Jin, J., Cui, R., Sha, F., Zhang, C.: Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2321–2334 (2016)

[89] Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 375–383 (2017)

[90] He, C., Hu, H.: Image captioning with text-based visual attention. Neural Processing Letters **49**, 177–185 (2019)

[91] Kalimuthu, M., Mogadala, A., Mosbach, M., Klakow, D.: Fusion models for improved image captioning. In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VI, pp. 381–395 (2021). Springer

[92] Abdussalam, A., Ye, Z., Hawbani, A., Al-Qatf, M., Khan, R.: Numcap: a number-controlled multi-caption image captioning network. ACM Transactions on Multimedia Computing, Communications and Applications **19**(4), 1–24 (2023)

[93] Yang, M., Liu, J., Shen, Y., Zhao, Z., Chen, X., Wu, Q., Li, C.: An ensemble of generation-and retrieval-based image captioning with dual generator generative adversarial network. IEEE Transactions on Image Processing **29**, 9627–9640 (2020)

Figure 7: Samples of correct captions (green text) generated by the proposed model from Flicker8K dataset

Figure 8: Samples of correct captions (green text) generated by the proposed model from Flicker30K dataset

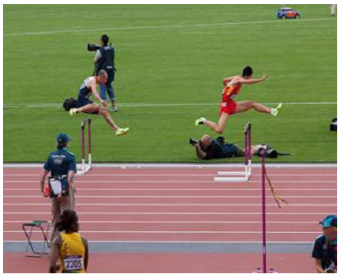
2860372882_e0ef4131d4.jpg
'<start> A man in bright orange short be skateboard along a course with cone as two person observe in the background . <end>', '<start> A man in orange short and knee pad be in a blurry picture . <end>', '<start> A man in orange short , knee pad and a helmet , move quick on a street course . <end>', '<start> a man in orange short move really fast <end>', '<start> A skater go through a course . <end>'.

a man in a red jacket be hold a sign

2848977044_446a31d86e.jpg
'<start> A cowboy be hang upside down from a horse . <end>', '<start> A man at a rodeo be ride his horse while upside down and not on a saddle <end>', '<start> A man be ride upside down on the side of a horse at a rodeo . <end>', '<start> A rodeo rider be throw headfirst from a horse . <end>', '<start> A rodeo rider do a vertical trick on a brown horse as an audience look on <end>'.

a man be jump on a bicycle of a street of person

Figure 9: Samples of incorrect captions (red text) generated by the proposed model from Flicker8K dataset



7711989162.jpg
'<start> two athlete one with red short and shirt and the other with blue short jumping through hurdle with cameraman behind them snapping photo <end>', '<start> two men are jumping over hurdle on a red track while other people are watching and photographing them <end>', '<start> olympic hurdler are jumping over hurdle on a track while others watch <end>', '<start> two men are competing in a track and field event involving hurdle <end>', '<start> two hurdler running a race <end>'

two men are playing soccer on a field

(a)

771366843.jpg
'<start> these two people in hat are standing on rocky terrain with their arm around each other <end>', '<start> two men posing for a piece in front of snowcapped mountain <end>', '<start> two men pose on a rocky area in front of a snowy mountain <end>', '<start> two men are standing in front of a snowcapped mountain <end>', '<start> the men are in a rocky mountain area <end>'

a man in standing on a rocky mountain

(b)

Figure 10: Samples of incorrect captions (red text) generated by the proposed model from Flicker30K dataset