Preprint accepted in Neural Computing and Applications. DOI: https://doi.org/10.1007/s00521-025-11199-1

Attention-based Transformer Model for Arabic Image Captioning

Israa Al Badarneh[1], Rana Husni AlMahmoud[2], Bassam H. Hammo[1,3], Omar Al-Kadi[1]

[1]King Abdullah II School of Information Technology, The University of Jordan, Amman, Jordan.
[2]School of Electrical Engineering and Information Technology, German Jordanian University, Amman, Jordan.
[3]King Hussein School of Computing Sciences, Princess Sumaya University for Technology, Amman, Jordan.

Abstract

Image captioning generates descriptive text from an input image, establishing a connection between the image content and words. Recently, the most successful approaches for automatically creating image captions have been based on transformer learning models. Arabic image captioning has gained importance due to the unique characteristics of the Arabic language. This paper introduces an Attention-Based Transformer Model for Arabic Image Captioning (ARTIC). ARTIC employs a deep learning convolutional neural network (CNN) for feature extraction from the images and a transformer encoder-decoder architecture for generating textual captions. ARTIC utilizes an ensemble learning approach based on a voting mechanism that selects the caption with the highest bilingual evaluation understudy (BLEU) score to produce the captions. To evaluate the effectiveness of the proposed model, the publicly available Flickr8k benchmark dataset was used for Arabic image captioning. Our results show that ARTIC achieved the best scores for BLEU-1, CIDEr, and ROUGE at rates of (0.626),(0.838), and (0.471) respectively. The other metrics, such as BLEU-2 and METEOR, achieved competitive rates of (0.381) and (0.332), respectively. The experiments with the Flickr30k English dataset demonstrated the generalizability of the proposed approach to other languages. These results indicate that the suggested model outperformed other models used for comparison.

Keywords: Arabic image captioning, Transformer, Computer vision, Natural language processing, Attention mechanism, Ensemble learning.

1 Introduction

Generating an image caption involves recognizing essential elements within an image, discerning their relationships, and crafting descriptions of the image content that are syntactically and semantically coherent. This task is challenging within artificial intelligence, primarily because it requires the combination of two distinct research communities: computer vision and natural language processing. Image captioning approaches include template-based, retrieval-based, and deep learning-based methods. Template-based solutions use predefined templates with a set number of empty slots to identify objects, characteristics, and behaviors. This method produces grammatically sound captions, often more accurate than retrieval-based methods. Retrieval-based systems use existing captions to create general, syntactically correct captions. However, they cannot provide precise semantically and image-specific captions. Techniques used for image captioning include template-based retrieval and retrieval-based retrieval, both relying on pre-existing captions from the training set or clearly defined linguistic structures [1]. Due to the challenges of the task using template-based and retrieval-based approaches, a third deep learning-based strategy has been proposed in light of recent advances in deep neural networks, which are widely used in computer vision and natural language processing. Deep neural networks can provide valuable answers for visual and linguistic modeling [2]. They have, therefore, been used to improve existing systems and create several new ones [3]. Image captioning was mainly done with traditional machine learning-based methods before the notable breakthrough in deep learning techniques. Among these were feature extraction approaches such as the Histogram of Oriented Gradients, Local Binary Patterns, and Scale-Invariant Feature Transform. The items were classified using a classifier after the extraction of features. Deep learning-based techniques automatically find features and are more popular than traditional methods since feature extraction from massive amounts of data is challenging [4].

The massive volume of images available on the Internet, often without accompanying explanations, has led to the automation of image captioning [5]. Recent advances in deep learning models, powered by state-of-the-art computing capabilities, have driven significant progress in this field [6, 7]. Despite the remarkable advances achieved in various computer vision tasks such as scene recognition, object recognition, image segmentation, and classification, generating a natural language description for an image remains one of the exceptionally challenging tasks exceeding the complexity of many other computer vision tasks [2, 8]. Research in image captioning has a broad spectrum of practical applications that span various fields. Examples might include medical imaging for diagnostics and analysis [9], improving student learning experiences [10], helping visually impaired individuals [11], powering artificial intelligence-driven platforms [12], assisting virtual assistants [13], allowing efficient image retrieval [14], facilitating information retrieval [15], improving social media content [16], and even supporting automated self-driving cars [17]. In addition, it plays a crucial role in describing CCTV footage [18], improving image search quality [19], and improving facial recognition systems [20].

In recent publications, the application of deep machine learning for image captioning has gained significant attention [1, 21]. Deep learning algorithms effectively

handle the challenges and complexities inherent in image captioning. Like other languages tackling this vital field of research, the importance of Arabic image captioning has grown, considering that Arabic and its various dialects are spoken by more than 422 million people, making it the sixth most spoken language in the world [22]. The richness of the Arabic language, with about 12 million words in its lexicon, adds a layer of complexity and significance to image captioning endeavors [6]. Most captioning models use Recurrent neural networks (RNN) and Long short-term memory (LSTM) as their language models. However, vanishing gradients create a significant challenge to these techniques, which reduces their effectiveness. Moreover, LSTM and RNN models are not hardware-friendly and demand higher processing power [8, 23]. An alternative technique for image captioning has been researched in the literature and published by the authors of [24], Generative Adversarial Networks (GAN). However, GANs present challenges due to their discrete nature, making training such systems difficult [25]. Many issues and concerns arise when using BERT models and hybrid LSTM-Transformer approaches for natural language processing jobs such as captioning images. Due to their bulk and lengthy training times, BERT models are challenging to install on low-powered devices. Although hybrid LSTM-Transformer models have potential, their increased complexity due to architectural differences necessitates more excellent resources and extended training periods, which may impair interpretability and complicate the model's decision-making process. Our study proposed a hybrid approach that combines a transformer with an attention mechanism to address these gaps. This method is particularly significant for Arabic image captioning, aiming to enhance the model's ability to understand complex image elements and generate contextually appropriate captions. Transformers effectively capture long-range dependencies, supported by attention mechanisms focusing on relevant input data. Additionally, decision-making can often be more interpretable when using ensemble models [26]. The combination of forecasts from several models provides information on the reliability and consistency of the model's outcomes. The suggested model can further enhance performance and address the challenges in Arabic image captioning.

1.1 Motivation for automatic image captioning for Arabic

Morphological richness, complicated grammar, and cursive writing in Arabic present a challenge compared to English. Multiple dialects, each with a unique style and syntax, add another layer of complexity. The inherent characteristics of Arabic make linguistic competence imperative for accurate resolution. The challenge is compounded by homographs, where many Arabic words share the same written form as others [27]. High-quality images with accurately maintained metadata and tags are necessary for Arabic image captioning [28]. However, existing data sets, particularly the one provided by [29], with its 8092 images, pose a risk of overfitting due to their relatively small size for training purposes. In addition, creating resources is a time-consuming and costly effort [6]. Additionally, previous attempts at Arabic image captioning should have considered the benefits of the attention mechanism, as highlighted by [30]. In response to these challenges, this research proposes a new model tailored for Arabic image captioning. The transformer's ability to allocate attention to specific image regions allows words to represent localized features rather than a

global context, promising improved results [31]. Furthermore, the model can describe the relationships between the features of the image [32].

The main contributions of this research are: first, to present an attention-based approach for Arabic captioning using a transformer. Second, to enhance the robustness of Arabic image captioning models and reduce overfitting by adopting ensemble learning to aggregate the performance of collective models, and third, to highlight the challenges related to the Arabic language in the context of image captioning. Finally, the model will be evaluated by employing comprehensive assessment metrics not fully explored in the previous research.

The rest of the paper is organized as follows: Section 2 provides the research background. Section 3 reviews related work. Section 4 details the research methodology. Section 5 encompasses experiments and results. Section 7 provides limitations and Research Opportunities. Finally, Section 8 offers the conclusion and implications of the investigation.

2 Preliminaries and background

2.1 Convolutional neural network

Recurrent neural networks (RNNs), convolutional neural networks (CNNs), deep belief networks (DBNs), and deep Boltzmann machines (DBMs) are popular deep learning models. CNNs excel in interpreting visual data through hierarchical learning, using shared weight filters [33]. In image captioning, pre-trained CNN-based encoders on ImageNet are standard, transforming images into visual vectors. Using sets from lower convolution layers preserves fine-grained correspondence and enables selective focus during generation [2, 34]. Despite variations, CNNs remain effective in object recognition, offering advantages like weight-sharing, simultaneous feature extraction and classification learning, and simplified large-scale deployment [33, 35].

2.2 Transfer learning

Transfer learning applies pre-existing models to different contexts, aiding deep neural network training with limited datasets. In image captioning, successful implementation involved training on a standard dataset, then transferring knowledge to a novel dataset with unpaired phrases and images [33].

Feature extraction uses eight CNN models: ResNet50, ResNet101, EfficientNetV2, VGG16, VGG19, EfficientNetB4, ResNet152, and RegNetX120. Residual CNNs, like ResNet-50, tackle overfitting and optimization challenges through identity mapping and shortcut connections. In image feature extraction, a pre-trained ResNet-50, trained on ImageNet, is used by discarding its final output layer [36]. ResNet-101 is a baseline in an image captioning model that encodes images without bottom-up attention. The model's performance is assessed, focusing on the impact of bottom-up attention compared to the baseline ResNet encoding [37]. On the other hand, VGGNet, known for its simplicity and robustness, is a popular image feature extractor

often chosen for research applications. However, regarding efficiency, ResNet outperforms VGG, offering higher accuracy with a lower parameter count [38]. Recently, EfficientNet models, with compound scaling, excel in transfer learning datasets, consistently outperforming other CNNs in accuracy and efficiency. They prove effective in diverse domains, including applications like COVID-19 categorization [39, 40].

2.3 Transformers

Text data is well handled by the Transformer architecture, which is sequential by design [41]. After receiving one text sequence as input, another text sequence is created 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 feed-forward layer contain the crucial self-attention layer, which determines the connections between the words in the sequence. The decoder consists of the feed-forward layer, the self-attention layer, and a second encoder-decoder attention layer. The encoder and decoder have distinct weights and LayerNorm layers with residual skip connections. It uses embedding and position encoding layers for data inputs. The encoder stack includes multiple encoders with feed-forward and multi-head attention layers, while the decoder stack has numerous decoders with feed-forward layers and multi-head attention. A general architecture of a transformer is illustrated in Fig. 1.

2.3.1 Self-head attention

Self-attention is a process in which each element in a set is related, allowing a more accurate representation using residual connections [41]. Self-attention uses the scaled dot product mechanism, which works with three vectors: n_k element-strong query vectors, n_k element-strong key vectors, and n_k element-strong value vectors. The operator computes a weighted sum of value vectors based on the similarity distribution [42].

2.3.2 Multi-head attention

Multihead attention is a module that performs self-attention multiple times in parallel, concatenates, and transposes the attention output into the desired dimension [43]. It helps pay attention to different sequence sections, such as longer-term vs. shorter-term dependencies. There are two types of attention: soft and hard attention. Soft attention uses weighted image features instead of an image as input, ignoring unimportant areas and ensuring high-attention areas maintain their original value. The gradient is computed using backpropagation, and the accuracy is based on the weighted average representing the focus region. Monte Carlo simulations are used to calculate the average of all sample results [4].

2.3.3 Add and Norm Layers

The Add and Norm layers carry out two tasks. The initial phase is the "add" portion, which controls flow via residual connections. Layer normalization is carried out in the second phase, called normalization "Norm". As a result, the following equation would represent the output of the layer:

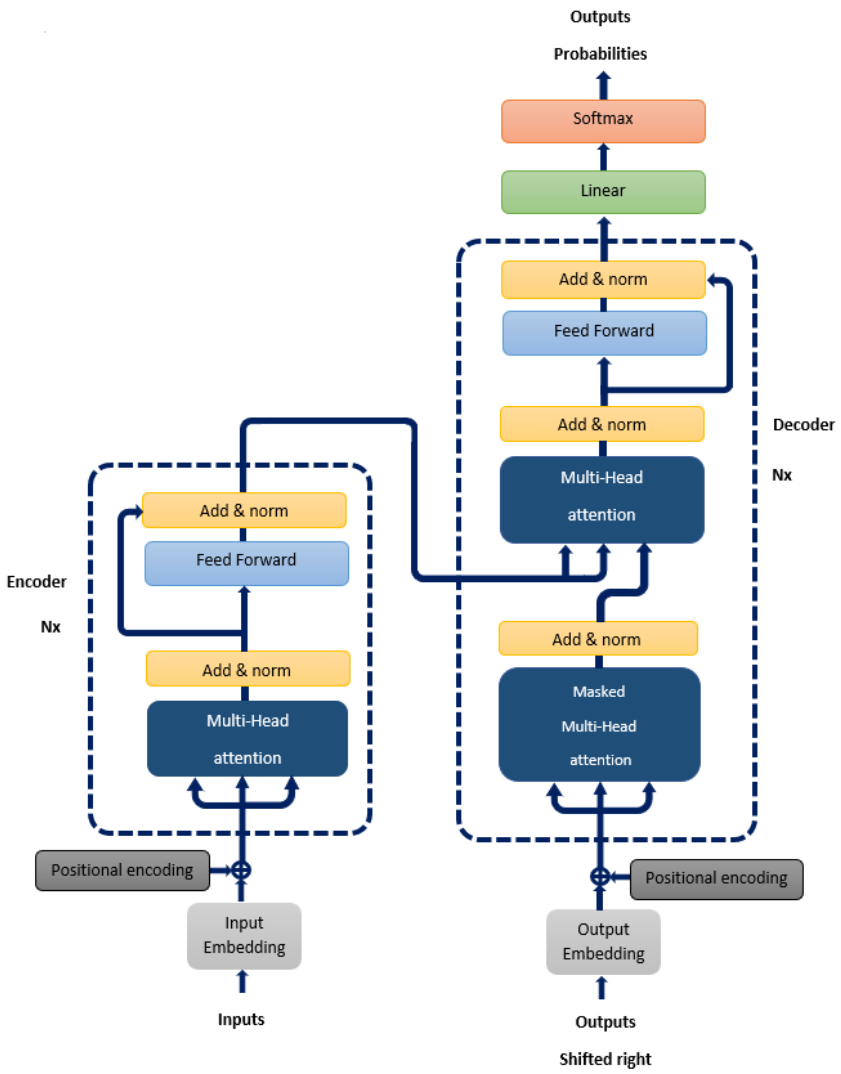


Figure 1: General architecture of a transformer.

$$\text{Add \& Norm} = \text{LayerNorm}\left(x + \text{Sublayer}\left(x\right)\right) \tag{1}$$

, where Sublayer (x) is the output and x is the input to any sublayer (MHA or Feed Forward).

2.3.4 Feed Forward Network

A completely connected point-wise feed-forward network is present in every layer. It performs two linear transformations using ReLU activation. This layer decides upon the weights used in training. It has the following numerical definition:

$$\begin{aligned} FF(x) &= \text{ReLU}\left(xW_1 + b_1\right)W_2 + b_2 \\ \text{ReLU}(x) &= \max(0, x) \end{aligned} \tag{2}$$

,Where b_1 and b_2 are biases and W_1 and W_2 are network weight matrices.

2.3.5 Positional Encoding

Transformers use positional encoding to introduce relative or absolute embedding positions into the model, maintaining the token sequence's format for parallel execution. Position-aware embeddings are created by combining language features with positional encodings.

2.3.6 Linear and SoftMax Layer

The decoder's output is projected as $n-$vocabulary size using a fully linked linear layer, where n is the expected result size and vocabulary size is determined by sentence length and vocabulary size. A SoftMax layer is applied for the probability distribution.

2.3.7 Encoder and Decoder Stacks

The encoder and decoder are two layers in a multi-headed attention model [41]. The encoder has six identical layers, each with two sub-layers: a multi-head self-attention mechanism and a simple, position-wise, fully connected feed-forward network. The decoder adds a third sub-layer, performing multi-head attention over the encoder stack's output. Residual connections are used around each sub-layer, followed by layer normalization. The self-attention sub-layer is changed to stop positions from paying attention to preceding positions. The predictions for location i depend on known outputs at positions less than i due to masking and offset by one position. The decoder is finalized with a linear layer as a classifier and a SoftMax to determine word probabilities.

2.3.8 Attention Function

An attention function maps a query, key-value pairs, and output vectors to one another. The output is a weighted sum of values determined by the query's compatibility function with its corresponding key. "Scaled Dot-Product Attention" is used for this process. The input consists of queries, keys, and dimensions d_k, with dot products computed and weights obtained using a SoftMax function. The attention function is continuously calculated on a group of queries, which are then compacted into matrices K and V. The output matrix is calculated as follows:

$$\text{Attention}(Q, K, V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

The work in [41] suggests that linearly projecting queries, keys, and values h times using distinct projections to d_q, d_k, and d_v dimensions can improve attention function performance. This approach allows the model to jointly attend to data from multiple representation subspaces at various places through multi-head attention. The study used $h = 8$ parallel attention layers, applying the formula $d_k = d_v = d_{model}/h = 64$ for each. The total computing cost is comparable to single-head attention with full dimensionality due to the lower dimension of each head.

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}\left(\text{head}_1, \ldots, \text{head}_\text{h}\right) W^O \\ \text{where head} &= \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \end{aligned} \tag{4}$$

2.3.9 Transformers-based Approaches for Image Captioning

Transformers have significantly advanced Arabic image captioning through their unique architectural features and attention mechanisms. Firstly, transformers provide an understanding of context by using attention mechanisms. In Arabic image captioning, attention mechanisms enable the model to focus on different parts of an image while generating descriptive text. This is crucial for Arabic due to its rich morphology and syntax, which requires a nuanced understanding of context, as the attention mechanism helps align visual contexts with linguistic contexts, allowing the model to produce detailed and accurate captions. For example, the sentence (رجل يرتدي قميصا حمراء يتزلج على لوح تزلج), "A man with a red shirt riding a surfing board", the attention mechanism helps the model focus on the "shirt" when generating the word (حمراء) "red" and on the "board" when generating the word (لوح) "board". Secondly, transformers handle long-distance dependencies and preserve word order by leveraging the self-attention mechanism to recognize relationships between words across long distances in a sentence. Traditional models like RNNs or LSTMs struggle with long-range dependencies due to vanishing or exploding gradients [41]. This is crucial for Arabic, a language with complex morphology and syntax, such as in sentences where descriptive phrases come after the main subject, example (طفلة صغيرة تتسلق الدرج إلى منزلها),(A little girl climbing the stairs to her house", in this case, the model needs to connect (تتسلق), "climbing", to (طفلة) "girl", transformers effectively manage these connections across the entire sentence or image features. Moreover, transformers enable multimodal learning where they integrate information from various sources, such as image features and textual descriptions; this integration is vital for generating accurate captions for images. Additionally, parallel processing and bidirectional [44] of transformers nature improve efficiency and context understanding compared to traditional RNNs and LSTMs [41] transformers generate the words of the caption as a full text with all words simultaneously [45]. Recent image captioning models augment transformer architectures for implicit region connections, achieving high scores. Some models, however, face limitations in adapting the transformer's internal architecture designed for machine translation to image captioning. Unlike text, images have multidimensional spatial relationships, posing challenges in spatial freedom [32].

2.4 Evaluation metrics

While direct human judgment is the simplest way to evaluate text generated for images, scalability is challenging due to non-reusable 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. Table1 summarizes 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[46]	Machine translation	n-gram precision
ROUGE[47]	Document summarization	n-gram recall
METEOR [48]	Machine translation	n-gram with synonym matching
CIDEr [49]	Image captioning	tf-idf weighted n-gram similarity
SPICE[50]	Image captioning	Scene-graph synonym matching

2.4.1 Bilingual evaluation understudy (BLEU)

BLEU is a metric evaluating the quality of machine-generated text by comparing individual segments to a set of reference texts [46]. Its approach varies with the number of references and text length. BLEU scores are higher for short, auto-generated text and range from 0 to 1. Unigram and bigram comparisons 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. While 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 specific evaluations [51].

2.4.2 Recall-oriented understudy for gisting evaluation (ROUGE)

ROUGE measures evaluate text summaries by comparing word sequences and pairs to a database of human-written reference summaries [52]. Initially designed for machine translation accuracy and fluency assessment, it quantifies sentence-level similarity using the longest common subsequence between candidate and reference sentences. Like 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. ROUGE-1, ROUGE-2, ROUGE-W, and ROUGE-SU4 serve diverse evaluation tasks, and their metrics range from 0 to 1.

2.4.3 Metric for explicit ordering translation evaluation (METEOR)

METEOR is designed for machine translation evaluation and is considered more valuable than BLEU, with a more vital link to human evaluations [48]. It calculates scores based on generalized unigram matches between a candidate sentence and human-written reference sentences. Precision, recall, and alignment of 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 [51].

2.4.4 Consensus-based image description evaluation (CIDEr)

CIDEr is an image captioning quality evaluation paradigm relying on human consensus [49]. It assesses the resemblance 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 grammar, significance, and accuracy for image captions and descriptions. Unlike metrics that work with a limited number of captions per image, CIDEr uses consensus utilization, making it suitable for analyzing the agreement between generated captions and human assessments [51].

2.4.5 Semantic propositional image caption evaluation (SPICE)

SPICE is a semantic concept-based image caption evaluation metric based on semantic scene graphs [50]. It utilizes 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.

3 Related works

Recent vision and language research advances have raised diverse image captioning models. This section presents the work from the literature explicitly addressing Arabic image captioning.

3.1 Root words RNN and DBN model

The first contribution to discuss is the three-stage root word-based method proposed by [53]. The approach begins by generating image fragments through a pre-trained deep neural network on ImageNet. These fragments are then linked to a set of root

words in Arabic. Subsequently, a deep belief network, pre-trained with Restricted Boltzmann Machines, determines different root words associated with image fragments and extracts the most contextually relevant words for the image [54]. The model employs a rank-based approach to go through image-sentence pairings, ultimately selecting the most fitting pair while discarding false associations. Finally, sentence captions are constructed based on dependency tree relations from the obtained words.

The evaluation of this model required two datasets. The first dataset, ImageNet, features manually written captions in Arabic by professional translators for 10,000 images. The second one has images from the Al-Jazeera news website, comprising 80,000 images for training and 30,000 for testing. The evaluation used BLEU-1 scores, with a unique approach of comparing results by directly generating captions in English and translating them to Arabic using Google Translate. The findings underscored the superiority of directly generating captions in Arabic, which produced significantly higher BLEU-1 scores compared to generating captions in English and subsequently translating them. The work proposed by [55] introduced an approach for directly generating image captions in Arabic. The authors extracted root words from images using a sophisticated blend of root-word-based Recurrent and Deep Neural Networks. Subsequently, these roots were translated into morphological inflections. Adding another layer of linguistic depth, the model used dependency tree relations to ensure the proper sequencing of words in Arabic sentences. Two diverse datasets served for experimentation. The first dataset comprised images from the Flickr8K dataset, accompanied by crafted captions in Arabic by professional Arabic translators. The second dataset has 405,000 images featuring captions collected from newspapers across various Middle Eastern countries. BLEU-[1-4] metrics systematically evaluated the model's performance. The results showed that generating Arabic captions directly in one stage produced superior results to a two-stage process that includes English captions in the Arabic translation process.

3.2 CNN-RNN encoder-decoder model

In [56], a generative merge model was introduced for Arabic image captioning. This model depends on the collaboration of two sub-networks: an RNN dedicated to sentences and a CNN tailored for images. The interaction between these sub-networks gave rise to the generation of captions. The model's architecture comprised three key components: (1) A robust RNN-LSTM-based language model deployed to encode varying-length linguistic sequences. (2) A fully convolutional network was employed to extract image features, drawing inspiration from the CNN VGG OxfordNet 16-layer. These features are demonstrated as a fixed-length vector, serving as input for the image encoder. (3) A decoder model that takes the outputted fixed vectors from the preceding models and makes the final predictions for the image captions. To test the efficacy of this model, the researchers constructed an Arabic dataset combining data from two English benchmark datasets, COCO and Flickr8k. The total dataset comprised 3427 images, divided into 2400 for training, 411 for development, and 616 for testing, maintaining a distribution of 70:12:18, respectively. Evaluation metrics included BLEU-[1-4]. The findings suggested that the merged model revealed promising

performance for Arabic image captioning, suggesting that even better results could be achieved with a more expansive corpus.

The Arabic Description Model (ADM), designed for the comprehensive generation of image descriptions in Arabic, was constructed as outlined in [57]. In a comparative analysis with its English-based predecessor, the ADM's foundation rested upon image features extracted from a CNN and a JSON file containing English image descriptions. The process involved translating the English JSON description file into Arabic, which was subsequently inputted into an LSTM network alongside the CNN-generated feature vector. This methodology facilitated the construction of a new JSON image description file tailored for the Arabic description model. The experimental phase employed a subset of the Flickr8k dataset, featuring 2000 images strategically partitioned into 1500 training images, 250 validation images, and 250 test images. BLEU-[1-4] metrics evaluated the model performance. The empirical findings revealed the superior efficacy of the English-based model over its Arabic counterpart. The authors highlighted the pitfalls of translating recognized English captions into Arabic, underscoring the inherent structural deficiencies in the resulting Arabic sentences.

In [29], the authors presented a dual-model framework for image captioning in Arabic. The first model, rooted in English image captioning, underwent a transformative process in which the English text was translated into Arabic. In contrast, the second model adopted an end-to-end approach, directly transcribing images into Arabic text. The image-centric model harnessed the power of a pre-trained CNN, VGG16, to map images to embeddings, a vector of substantial length, precisely 4096. This image embedding vector underwent further transformation through a fully connected layer with a Tanh activation function, ensuring output values within the bounded range of -1 to 1. For the linguistic component, a single hidden LSTM layer with 256 memory units formed the core of the language model. A critical aspect of their work involved the creation of a novel Arabic image captioning dataset. This dataset emerged from translating the well-established Flickr8K dataset, encompassing 8000 images, each accompanied by five distinct captions. These images, sourced from Flickr8K, mainly featured human and animal subjects. The translation process unfolded in two stages: all English captions were initially translated via the Google Translate API. The second stage involved editing and confirmation by a professional Arabic translator. The performance of the two models was evaluated with the new dataset using BLEU-[1-4] metrics. The results indicated the superiority of the end-to-end model, highlighting its superior efficacy in Arabic image captioning.

In multimedia transformation, [28] proposed an innovative Text-to-Picture system tailored for automatically converting simple Arabic children's stories into visually representative images. The attempt defined numerous challenges inherent in mapping natural text to multimedia, with an observed obstacle being the absence of captions and meaningful tags accompanying images sourced from the Google search engine. To overcome this limitation, the authors incorporated a deep-learning captioning model into their framework. This model unfolded in two integral stages: 1) Story text processing and image retrieval. 2) Image ranking using the automatic captioning process and sentence similarity. This approach comprised several key steps: 1) keyword extraction, in which preprocessing steps were made, and keywords for each sentence

were translated into English. 2) Query formulation: formulate queries to retrieve images for them. 3) Image selection: the retrieved images were prepared for captioning in the next step. 4) Image captioning, using deep CNN to represent an image and an RNN, especially LSTM, to provide the output sentence. 5) Sentence similarity is used to match the initial keywords and captions. 6) Sentences exhibiting higher similarity values with the initial keywords were prioritized and presented to the user, allowing for interactive engagement. 7) The final step involved image evaluation, in which users could assign a ranking of 1 to 5 for each relevant image, providing valuable feedback for refinement and improvement.

In Arabic image captioning, a recent advancement by [6] suggested an architecture rooted in the neural machine translation (NMT) paradigm, taking advantage of its superior performance compared to classical methods. The essence of the proposed model lay in an encoder-decoder architecture, with a CNN serving as an encoder to extract visual information from the input images. The decoder, an LSTM network, was crucial in generating a probability distribution over potential next steps to formulate the caption. To assess the model's efficacy, the researchers constructed a novel ArabicFlickr1K dataset containing 1095 images, each accompanied by three to five descriptive captions. The research introduced an active learning framework that recruited human annotators to collaboratively refine the automatic translation capabilities of the model. The model's performance underwent rigorous evaluation, using BLEU-[1-4] metrics, determining its proficiency in generating captions. The results obtained confirmed the potential of the proposed architecture in the field of Arabic image captioning. This ensured the model's ability in linguistic translation and highlighted the efficacy of the active learning approach in enhancing the system's overall performance.

In [58], the authors tested thirty-two combinations affecting caption generation. They include four preprocessing techniques, two deep learning techniques (LSTM, GRU), and two image feature extraction models (Inception V3, VGG16). The authors reviewed image captioning models for Arabic and English, focusing on the lack of available datasets. They fixed typos in the Arabic Flickr8k dataset and applied text preprocessing. The study revealed that using Arabic preprocessing and VGG16 image feature extraction improved the Arabic caption quality. However, the work did not observe any significant differences when using Dropout or LSTM compared to GRU. The Arabic dataset achieved the best BLEU-[1-4] results at [36.5, 21.4, 12, 6.6], respectively.

The study of [59] suggested an effective deep-learning model for Arabic image captioning, focusing on the impact of text preprocessing on attention weights and BLEU-N scores. The project aimed to provide meaningful, syntactically, and semantically accurate captions for computer vision and natural language processing, especially in Arabic and other languages with complex morphological structures. The authors investigated the impact of applying several text preprocessing techniques on the resulting BLEU-N scores, the quality of the created images, and the behavior of the attention mechanism before presenting an effective deep-learning model for Arabic image captioning. The model employed an LSTM to produce conventional Arabic captions, a translator-based approach for output root words, and a Region Convolutional Neural Network (RCNN) to map image objects to Arabic root words. RESNet-101 was the encoder,

while the LSTM network was the decoder. Preprocessing and tokenization of captions were performed with Pyarabic, which divided image captions into tokens by splitting them into spaces, and with FARASA, which separates Arabic words into their component clitics. The model was evaluated using the Flickr8k dataset. The model achieved its best BLEU-4 score using the FARASA segmenter with 200 randomly chosen images from the MSCOCO dataset that were utilized for quality testing of the model.

3.3 CNN-Transformer encoder-decoder model

Another progress in Arabic image captioning is the work of [60]. The model's architecture can be explained in two steps: The initial step forms the foundation for capturing rich visual information. It utilizes the power of a CNN encode to extract area features and object tags from the input image. The second step involves using a pre-trained transformer language model. The transformer takes the extracted region features and object tags to generate a coherent sentence. The fusion of visual and semantic elements in this process marks a crucial advancement in image captioning. After initialization, the models undergo fine-tuning through the Object Semantics Aligned Pre-training (OSCAR) learning method. This strategy is essential in simplifying the learning of semantic alignments between an image and text. It is achieved using object tags in images as anchor points, contributing to a more streamlined and effective learning process. The robustness of the proposed model was tested using the Flickr8K dataset, and the results showed that the model provided a new baseline for Arabic image captioning.

The work of [61] introduced an Arabic image captioning approach that utilized transformer models in both the encoder and decoder stages. In the decoder, they applied a pre-trained word embedding model, while in the encoder phase, they utilized feature extraction from images. The models experienced comprehensive training and testing on the Flickr8k Arabic dataset. A comparative analysis was conducted to assess the image feature extraction subsystem, integrating three vision models: SWIN, XCIT, and ConvNexT. Simultaneously, four different pre-trained language embedding models evaluated the caption-generated linguistic subsystem. The authors demonstrated that the optimal results emerged from combining three transformer-based models incorporating ConvNexT, SWIN, and XCIT as image feature extractors, along with the CamelBERT language embedding model.

The system proposed by [62] aimed to reduce some of the morphological complexity associated with the Arabic language using an improved text preprocessing pipeline with a word segmenter. Furthermore, they designed neural network topologies with transformers and attention processes. Three potential models with an encoder-decoder architecture were defined. All encoders in the three models were pre-trained on ImageNet. In the first model, the encoder was an LSTM-based model with an attention layer, while the decoder was MobileNetV2. In the second model, the decoder was a GRU-based model with an attention layer, while the encoder was MobileNetV2. In the third mode, a transformer-based model served as the decoder, while the encoder was EffeceintNet. The models were tested using the Arabick Flickr8k dataset. The transformers and the AraBERT segmenter produced the best BLUE scores with BLEU-1 = 44.3 and BLEU-4 = 15.6.

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, which requires additional computational resources [8, 23]. An alternative approach explored in the literature, as suggested by the authors of [24], is using GAN for image captioning. However, GANs come with challenges due to their discrete character, making training such systems a problematic task [25].

To generalize image captioning for multilingual support, a suggested model was provided by the study of [63]. The domain object dictionary approach was demonstrated, which generates image captions without processing additional learning data by adapting the object dictionary for each domain application. The default model was the OSCAR model, which is based on the BERT model, and the image COCO captioning data was learned. Instead of processing the learning data, the study's suggested strategy involves changing the object's dictionary to focus on the domain object dictionary, which produces different image captions by fully explaining the items needed for every domain. While maintaining the functionality of previous models, this filter captioning paradigm enabled the creation of image captions from various areas. This methodology can be used in several domains, including real-time traffic information commentary, sports commentary, art therapy, and image search. Generalization is a crucial challenge in image captioning that requires models to handle diverse images and scenarios beyond those encountered during training. Technique such as transfer learning was employed [64] to improve a model's ability to generalize across different domains. Transfer learning leverages a relatively limited amount of data to enable the development of high-performance models that perform better when applied to other domains. Compared to the model that was extensively trained on the target source, the fashion image captioning model achieved competitive performance and high-quality captions by executing the last adaptation stage of the pre-trained model using a relatively restricted collection of target samples. This adaptation stage is significantly less expensive than starting from scratch to train a fashion image captioning model. Additionally, this model can enhance performance over previously untested data distributions, improving the model's ability to generalize. However, it is worth mentioning that BERT models pose a challenge due to the lengthy training periods needed for fine-tuning specific downstream tasks. The deployment of BERT on devices with limited processing capabilities is further complicated by its substantial size and an extensive array of parameters [65]. Utilizing a hybrid approach, combining LSTM with Transformer models introduces specific limitations and drawbacks. For example, it can escalate model complexity, attributed to architectural differences, resulting in a higher demand for resources and extended training times. Consequently, this complexity can affect the interpretation of the model's 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 Arabic image captioning emerges from the necessity to make visual content accessible to Arabic-speaking individuals. Therefore, developing and implementing dedicated Arabic 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. In the following section, we would like to explore the details of this approach.

4 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.

4.1 Dataset

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

English Flickr30k dataset[66]: The Flickr8k dataset[67] has been expanded to create the Flickr30k, which now contains 31,783 images with five captions for each. The split dataset available to the public uses 29,000, 1,000, and 1,000 images for training, validation, and testing, respectively. Most of the images in this dataset depict individuals engaging in daily activities and situations. Flickr30k is used to comprehend visual information "images" that correspond to expressed statements "descriptions of images". This dataset is commonly used as a benchmark for sentence-based image descriptions. The research paper[68] highlights the importance of the Flickr30k dataset in understanding human descriptions of visual material by thoroughly reviewing it. Every image has comprehensive, contextually rich annotations that present multiple perspectives on the image's content. The dataset, which reflects the diversity of human experience as captured via photography, categorically includes various features of human actions, objects, scenes, and surroundings. This diversity ensures that the dataset is well-suited for studying how different people interpret and describe visual scenes.

Arabic Flickr8K[29]: The work in [67] revealed the English Flickr8K dataset, a valuable resource comprising 8,000 images from the prolific photo-sharing platform Flickr. Primarily known for its human and animal-centric content, this dataset served as a vital contribution to the public domain. The label creation involved crowdsourcing descriptions through Amazon's manual labeling program, each image annotated with five sentences. Flickr8k is a standard dataset for training and assessing image captioning models, covering a broad range of scenes, objects, and activities characteristic of daily photography. Researchers leverage its rich diversity in images and textual descriptions to develop algorithms capable of generating accurate and contextually relevant captions. [29] translated the English Flickr8K dataset into Arabic through a dual-phase approach. Initially, Google Translate API was employed, followed by a

thorough review by qualified Arabic translators. The rigorous selection process concluded with identifying the top three translated captions for each image from an initial pool of five. To evaluate the proposed method in this study, we utilized the Arabic version of the Flickr8K dataset. This dataset offered a comprehensive and diverse set of images comprising 6,000 training images, 1,000 validation images, and 1,000 test images. Each image within this dataset boasted three distinct Arabic captions. Exemplary instances from the Arabic Flickr8K dataset are illustrated in Fig. 2.

Figure 2: Exemplary instances extracted from Arabic Flickr8k dataset

4.2 Data preprocessing

Because of raw textual data's inherent challenges, cleaning and preprocessing datasets before they are used in ML models become essential. Following the best practices outlined by [29], the approach we applied to Arabic text preprocessing was comprehensive and systematic.

Several procedures were employed in the data preprocessing, including tokenizing words, stemming, and normalizing a few Arabic letterforms. The preprocessing tasks outlined in this work were borrowed from [69]. They incorporate the following:

1. Text normalization: Typically, actions are taken to reduce the number of extracted terms. They include eliminating special characters and non-letter characters ($, &, %,..), taking out the non-Arabic characters, replacing the initial (أإآ) or (آ) with bare alif (ا), using (ha) (ه) instead of knotted (ta marbuta) (ة), eliminating the word's initial "al" (ال), and swapping out the last (ya'a) (ي) with (ى).
2. Text tokenization: In this step, a linguistic analysis of the text is conducted. It separates words, character strings, and punctuation marks into tokens during the indexing process. 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.

3. 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.

How we split our dataset ensures a balanced distribution for training, validation, and testing. First, we divide the dataset into two parts: the first one has 90 % of the data, while the second has 10%. Next, the large part was split into two subsets: 80% was used for training and 10% for validation. The remaining 10% forms the testing set. This method is similar to [29] splitting but distinguishes itself by incorporating a dedicated validation set. This validation subset enhances the evaluation process during model fine-tuning, facilitating effective hyperparameter optimization and ensuring robust generalization to new data. This structured approach supports reliable model development and promotes transparency and comparability with prior research findings.

4.3 ARTIC: The attention-based transformer model for Arabic image captioning

Fig. 3 shows the schematic diagram of the proposed model. In contrast, for a better understanding of ARTIC, Algorithm 1 provides the pseudocode outlining the operations of the model. The following are the steps applied through the model, and the following subsections discuss each stage in more detail.

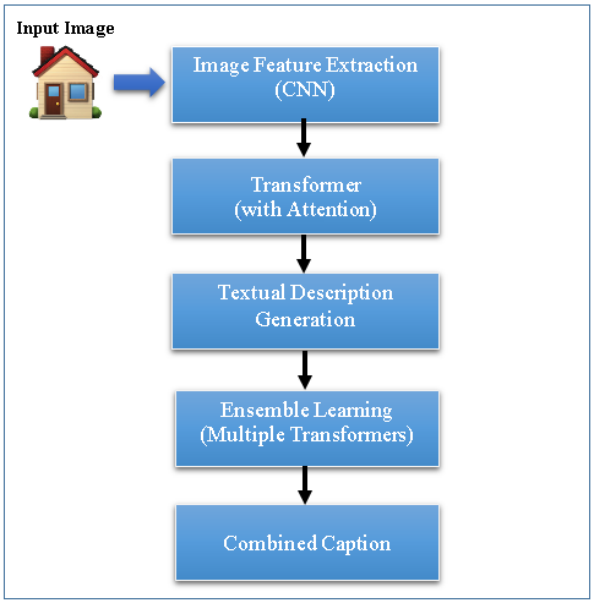


Figure 3: The schematic diagram of the proposed approach

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, 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.

S4: 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 outputs, either through averaging or voting.

S5: 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.

S6: Evaluation: Evaluate the performance of the ensemble model using metrics like BLEU, METEOR, and CIDEr.

Contemporary image captioning models have largely incorporated a flexible and effective encoder-decoder architecture, often called a CNN+RNN structure. The architecture of the proposed model consists of two primary models: the image-processing model and the language-processing model. In this configuration, the encoder typically employs a CNN image model to extract high-level feature vectors from input images and effectively "reads" them. Meanwhile, the decoder, often implemented as RNN, generates words based on the image representation acquired from the encoder. Its task is to produce a sequence of words that form a coherent, grammatically correct, and stylistically accurate phrase, effectively encapsulating the image's content [38].

The model proposed in this work adopts an encoder-decoder approach enriched with attention mechanisms, as recommended by [70]. The attention mechanism focuses on relevant sections of the image vital for the caption creation process, potentially leading to superior outcomes. Fig. 4 provides an overview of the typical architecture of the Arabic image captioning model.

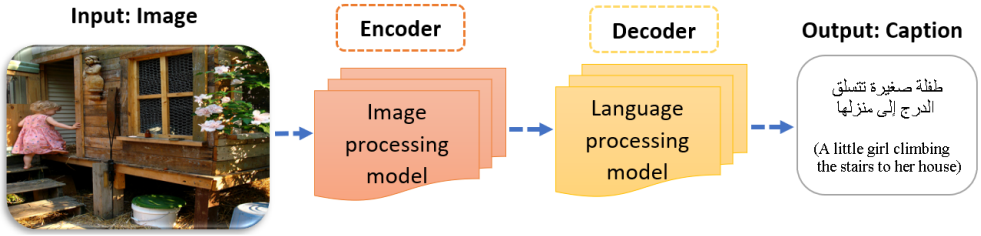


Figure 4: General architecture of the Arabic image captioning model.

Algorithm 1 Attention-based transformer model using ensemble learning

Input: Arabic Flickr8k 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

4.3.1 Image feature extraction

In the workflow of the proposed methodology, 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. 5 visually depicts the CNN model architecture. The convolution layer is vital 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.

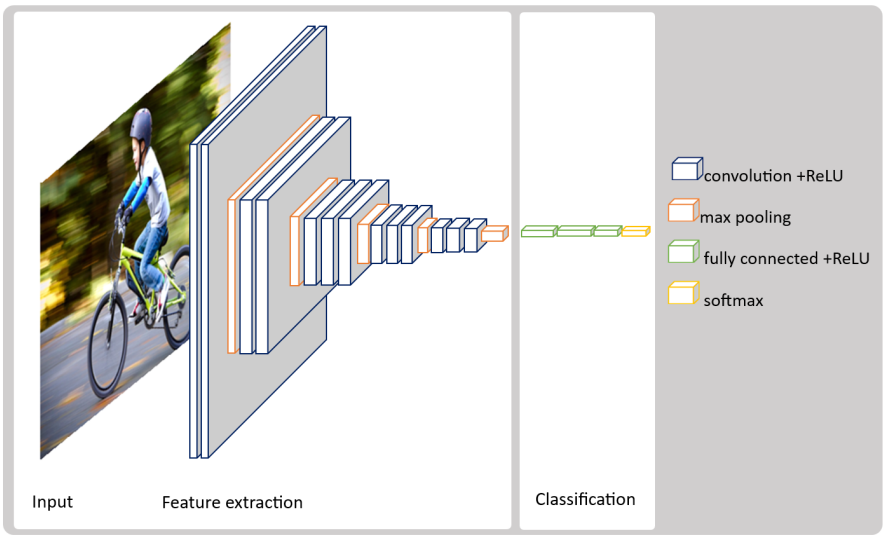


Figure 5: The visual architecture of the CNN

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 [35].

4.3.2 Text generation with a transformer

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. 1). The image transformer utilized for image captioning will decode various information within the image regions [71]. 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 [41]. The transformer model addresses the 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 [72].

4.3.3 The attention mechanism

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 specific 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 allowing intentional oversight of received data. Initially proposed for image categorization, attention is widely used in NLP experiments, including machine translation, speech recognition, text understanding, and visual captioning [7, 31, 73]. Fig. 6 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.

Figure 6: The visual architecture of the attention mechanism

Attention in image processing mimics human attention patterns. Its strength lies in establishing meaningful connections between features and enhancing the models' ability to prioritize essential features while filtering out noise. This aligns with the attention mechanisms that guide the model's focus during training [51]. 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 attentional focus facilitates the generation of a descriptive caption for the input image [74].

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,

involves weighting the annotated vector of picture features when salient features are identified. On the other hand, stochastic intricate attention is trained by maximizing a variation lower limit [33]. Recent studies have explored top-down and bottom-up attention theories, with recent experiments favoring top-down attention mechanisms [38]. Attentive encoder-decoder models lack global modeling skills. A reviewer module reviews encoder hidden states to address this, 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 [74]. Visual attention in multi-modal coverage mechanisms bridges the gap between encoder and decoder, enhancing data understanding [2, 75].

4.3.4 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 was applied with a width of $k = 10$, maintaining k sequence candidates and selecting the most likely one at each step [42]. This approach generates a diverse group of captions. Previous studies supported beam search as the preferred algorithm for caption generation [76].

4.3.5 Ensemble learning

Learning using typical techniques may be inadequate due to the complexity of data features and structures. Instead, ensemble learning (discussed in Section 2) integrates data fusion, modeling, and mining into a unified framework. In this essence, multiple learning algorithms extract features, while ensemble learning combines this knowledge, improving prediction accuracy through various voting processes [77]. To obtain 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 outputs by averaging or voting. Ensemble models like bagging, boosting, stacking, and voting aim to increase model performance by combining predictions from diverse base models. For example, a majority voting technique combines multiple classifiers' predictions in multimodal memes' hate speech detection, outperforming individual models [78].

This work uses a voting model, presented in Fig. 7, to ensemble the results obtained from each of the eight transformer models, discussed in Section 2. 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. Putting all the pieces together, Fig. 8 shows a comprehensive overview of the workflow of the proposed Arabic image captioning model.

5 Experimental results

This section presents the results obtained from the proposed model and compares them with the state-of-the-art models.

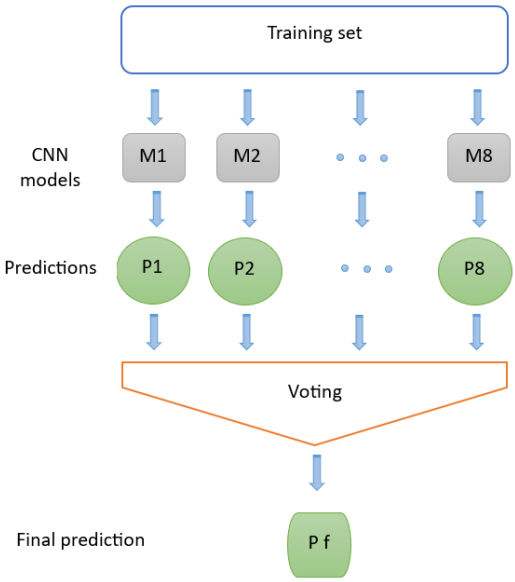


Figure 7: Voting model.

5.1 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 storage capacity for implementation purposes. The proposed model was trained with a batch size of 64, employing the Adam optimizer [79], a learning rate set at 0.00001, 30 epochs with early stopping, and the ReLU activation function was utilized.

Our model demonstrates a structured approach to training an image captioning model, focusing on effectively managing the learning rate. The loss function is initially defined using cross-entropy, which computes the cross-entropy loss between predicted and true labels without reduction. Early stopping is implemented to monitor validation loss and halt training if no improvement is seen after a set number of epochs, restoring the best weights. A custom learning rate scheduler adjusts the learning rate dynamically throughout training; it begins with a low rate, equivalent to 0.00001; this scheduler gradually raises the learning rate, providing a steady training process that improves convergence and overall model performance. This scheduler is integrated with the Adamax optimizer during model compilation, ensuring the model optimizes effectively while mitigating overfitting risks. The approach facilitates stable convergence during training and enhances the model's ability to generalize to unseen data, which is crucial for tasks like generating accurate and contextually relevant image captions. These numbers are chosen based on empirical observations and best practices in training neural networks, aiming to balance efficient convergence and stable training without risking overfitting.

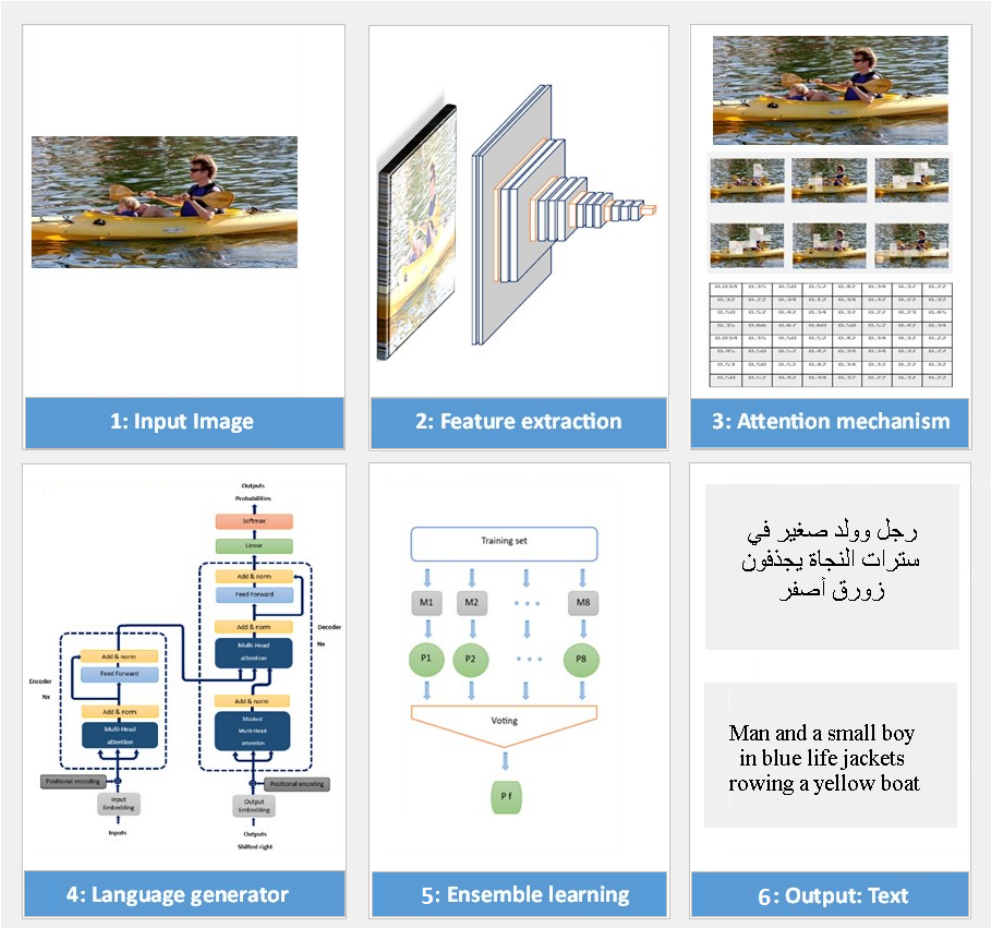


Figure 8: Workflow of the proposed Arabic image captioning model (ARTIC).

5.2 Evaluation metrics

While direct human judgment is the simplest way to evaluate text generated for images, scalability is challenging due to non-reusable 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. The evaluation metrics applied in this study, discussed in Section 2, include BLEU, CIDEr, METEOR, ROUGE, and SPICE.

5.3 Experimenting with Arabic Flickr8K dataset

Fig. 9 depicts a visual representation of the results for the Arabic Flickr8k dataset. The first row displays images and their corresponding names, and the second row

presents reference captions from the Arabic Flickr8k dataset. The third row exhibits captions generated by the proposed ARTIC model.

Figure 9: Image samples extracted from the Arabic Flickr8k dataset: (a) Samples from Flickr8k dataset, (b) Reference captions from the Arabic Flickr8K dataset, and (c) The generated captions from ARTIC, the proposed model

Table 2 compares the results of ARTIC against the state-of-the-art methods. The results of ARTIC are listed in two rows. The first row demonstrated the results without text pre-processing, while the second row shows the results after applying the pre-processing steps described in section 4.2. As shown in Table 2, ARTIC exhibits superior performance, with the highest scores highlighted in bold. Before we discuss the results, it is worth mentioning that most comparative studies did not report the data splits they used; others used different splits, and others used parts of the Flickr8k dataset. For instance, in the work of [29, 60, 62], the authors used the same dataset and test split, while in [61], they applied a different one. The test split in the remaining models was undisclosed.

Experimental results show a remarkable advancement of ARTIC over previous efforts on the Arabic Flickr8k dataset based on current image captioning evaluation metrics. This can be observable when compared with the models of [29, 58–60, 62], across metrics: BLEU-[1-4] (for simplicity, it is indicated as 'B,' CIDEr, and ROUGE in Table 2). For example, [62] got a better result for BLUE-4. However, ARTIC achieved a BLEU-1 score at the rate of (0.626), surpassing [80] at (0.489) and [61] at (0.598) for the same metric. Meanwhile, [61] reported higher outcomes across BLEU-[2-4] scores. Furthermore, ARTIC exhibited superior performance compared to the research findings of [61] and [80], demonstrating notable results using CIDEr with a score of (0.838), METEOR with a score of (0.332) and ROUGE with a score of (0.471). While [80] reported a METEOR score of (0.334), which is slightly higher than ARTIC (0.332). Although the SPICE metric was not reported in previous work, ARTIC achieved a rate of (0.110). It's worth mentioning that [61] achieved their highly tuned scores by

Table 2: Comparison of Arabic image captioning models

Reference	Dataset	B1	B2	B3	B4	CIDEr	METEOR	ROUGE	SPICE
[53]	Al-Jazeera news^{m}	0.348	NA	NA	NA	NA	NA	NA	NA
[55]	Flickr8k^{m}	0.658	0.559	0.404	0.223	NA	0.209	NA	NA
[56]	Flickr616	0.460	0.260	0.190	0.800	NA	NA	NA	NA
[57]	Flickr8k^{x}	0.344	0.154	0.760	0.350	NA	NA	NA	NA
[29]	Arabic Flickr8k	0.330	0.190	0.100	0.060	NA	NA	NA	NA
[58]	Arabic Flickr8k	0.365	0.214	0.120	0.066	NA	NA	NA	NA
[62]	Arabic Flickr8k	0.443	NA	NA	0.157	NA	NA	NA	NA
[60]	Arabic Flickr8k	0.391	0.246	0.151	0.093	0.428	0.317	0.334	NA
[59]	Arabic Flickr8k	0.391	0.251	0.140	0.083	NA	NA	NA	NA
[80]	Arabic Flickr8k	0.489	0.317	0.213	0.145	0.472	0.334	0.398	NA
[61]	Arabic Flickr8k	0.598	0.400	0.306	0.165	0.469	0.260	0.385	NA
ARTIC	Arabic Flickr8k	0.490	0.293	0.169	0.093	0.545	0.308	0.379	0.056
ARTIC^{+}	Arabic Flickr8k	**0.626**	0.381	0.224	0.134	**0.838**	**0.332**	**0.471**	**0.110**

^{m}Manual extraction ^{x}Subset of Flickr8k (2000 images) $^{+}$ARTIC with pre-processing

utilizing concatenated models featuring an embedding layer. Furthermore, a difference was observed in the data set-splitting strategy. This divergence in the dataset approach explains the differences in the reported results of the BLEU scores compared to our findings. Since the datasets and the code are private, confirming the findings or comparing this method to other models utilizing the same Flickr8K dataset is difficult.

5.4 Experimenting with English Flickr30K datasets

To illustrate how well the proposed model performed in another dataset, we compared its results to the best-performing models based on the English Flickr30k dataset, as shown in Table 3 and Fig. 11. The table shows that the suggested approach outperforms state-of-the-art techniques on Flickr30k datasets as measured by BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores, which are 0.798, 0.561, 0.387, and 0.269 respectively. The model's performance is comparable to that of the state-of-the-art, as demonstrated by ROUGE L METEOR CIDEr. It indicates the proposed approach is capable of producing captions that are clear and meaningful. The results from other metrics, such as SPICE (0.387), validate the methodology's effectiveness. It is critical to know that the results of most other approaches are not shared on this metric.

Table 3: Comparison of English image captioning - Flickr30K dataset

Reference	B1	B2	B3	B4	CIDEr	METEOR	ROUGE L	SPICE
[81]	0.671	NA	NA	0.233	0.645	0.204	0.443	NA
[82]	0.677	0.494	0.354	0.251	0.531	0.204	0.467	0.145
[83]	0.647	0.456	0.320	0.224	0.467	0.197	0.449	0.136
[84]	0.689	0.468	0.319	0.220	0.428	0.191	0.487	NA
[85]	0.694	0.498	0.355	0.254	0.469	0.251	0.538	NA
[86]	0.674	0.495	0.360	0.260	0.520	0.201	0.470	NA
[87]	0.690	0.493	0.347	0.241	0.528	0.195	0.465	NA
ARTIC	**0.798**	**0.561**	**0.387**	**0.269**	0.565	0.213	0.443	**0.387**

6 Discussion

6.1 Arabic image captioning interpretation and analysis

In general, our model demonstrates proficiency in generating captions that are not only relevant but also accurate in describing the image content. Fig. 10 presents samples of nearly correct captions, highlighted in yellow. In addition, ARTIC excels at producing more accurate captions for specific images. For instance, in Fig. 10 (c), the model accurately generates the phrase "blue dress" (زي أزرق) and "street" (شارع) elements that were not present in the reference captions. Conversely, the word "ice cream" (ايس كريم) is not detected in this example.

Fig. 10 (f) presents another example that illustrates the model's ability to generate a caption that refers to the location where the man stands, using the phrase "In front of a white building" (امام مبنى ابيض). At the same time, this information was not present in the reference captions. Fig. 10 (d) shows a correct description provided by the model, nearly identical to the reference caption, describing "A man with a red shirt riding a surfing board," (رجل يرتدي قميصا حمراء يتزلج على لوح تزلج). Furthermore, gender distinctions are present in the generated captions. For instance, terms like "woman" (امرأة) and "man" (رجل) accurately incorporated. Additionally, the ARTIC model demonstrates the ability to produce verbs, a crucial aspect in Arabic, where verb conjugation varies based on gender and plurality. For instance, in Fig 10 (c), the word "woman" (امرأة) is appropriately conjugated with the prefix (ت) in the verb "wearing" (ترتدي). The resulting caption reads as (امرأه ترتدي زي ازرق مع صبي صغير يلعب في الشارع), "A woman wearing a blue dress is with a small boy playing in the street.". Similarly, in Fig. 10 (f), the word "man" (رجل) is paired with the prefix (ي) in the verb "wearing" (يرتدي). The corresponding caption "A man with a red shirt standing in front of a white building," (رجل يرتدي قميصا احمر اللون يقف امام مبنى ابيض). These instances demonstrate the model's capability for capturing Arabic grammar and gender-specific verb conjugations.

The model was able to handle plural forms of verbs, incorporating masculine plurals (ـــون), also known as (واو الجماعة), added to the end of plurals. This is illustrated in Fig. 10 (e) with the word "they are sitting down" (يجلسون) in the caption (مجموعة من ناس يجلسون على شاطئ يوم مشمس), "A group of people sitting on a beach on a sunny day.". Also, the model demonstrates proficiency in distinguishing singular instances, as seen in Fig. 10 (a) with the caption (رجل يقوم من منحدر على لوح تزلج), "A man rises from a cliff on a skateboard.". Furthermore, the model successfully distinguishes the presence of young individuals in the scene. For instance, Fig. 10 (i) identifies the term "A little boy" (طفل صغير) in the caption (طفل صغير يرتدي ستره حمراء يركض على طول مسار رمال), "A little boy wearing a red jacket running along a sand path."

Fig. 10 (a) generated a description that is "almost" identical to the reference captions but missing the detection of the object "buildings" (مباني). However, this object is correctly generated in Fig. 10 (h) with the word "building" (مبنى). However, Figure 10 (b) effectively describes a scene detecting elements like "soil road" (طريق ترابي) and objects such as "bicycle" (دراجه). Yet, it encounters difficulty generating the object "helmet" (خوذه) and mistakenly confusing it with the term "jacket"

(ستره). It is worth mentioning that the ARTIC model accurately captures specific scenes. For instance, Fig. 10 (e) identifies the term "sunny day" (يوم مشمس) as in the caption (مجموعه من ناس يجلسون على شاطئ يوم مشمس), "A group of people sitting on the beach on a sunny day." Also, Fig. 10 (j) recognizes the word "wave" (موجه) as in the caption (رجل يرتدي قميص ازرق يتزلج على موجه), "A man wearing a blue shirt surfing on a wave," and Fig. 10 (i) identifies the term "sand path" (مسار رمال) as in the caption (طفل صغير يرتدي ستره حمراء يركض على طول مسار رمال), "A little boy wearing a red jacket running along a sand path."

Fig. 12 shows instances where the ARTIC model generated inaccurate results, highlighted in red. For example, Fig. 12 (a) shows a caption describing a boy playing with a basketball, (صبي صغير يقفز من أجل كرة سلة), "A little boy jumping for a basketball," despite the image depicting "two men engaged in a hockey game.". Similarly, Fig. 12 (b) shows another mismatch. While the image features "A red airplane flying above a mountain and releasing a red substance in flames," the generated caption pertains to (رجل يرتدي ستة حمراء يقف على قمة جبل), "A man wearing a red jacket standing atop a mountain," accurately detecting the color "red" (حمراء) and the location "top of the mountain" (قمة جبل), but misinterpreting the context.

Unlike conventional approaches, ARTIC stands out as an encouragement of innovation in addressing the challenge of Arabic image captions. This improvement indicates a significant alignment with the formulations in ground truth captions, emphasizing the proficiency of the proposed architecture in crafting meaningful image descriptions. Our method's ability in feature-text extraction, using attention mechanisms to focus on salient image regions and features, sets it apart from existing approaches. The critical differences between ARTIC and other solutions, outlined in Table 2, emphasize the contributions of our study across the following dimensions:

- Enhanced robustness of predictions
- Comprehensive evaluation metrics
- Accurate model evaluation

6.2 Enhanced robustness of predictions

Unlike the other approaches, ARTIC employs an ensemble learning approach, a strategic combination of eight CNN models through a voting method, to refine the optimal caption for each image. This significantly enhances the architecture's performance and strengthens its generalizability and robustness. By combining predictions from diverse base models, ARTIC effectively reduces overfitting, ensuring a more reliable and adaptable solution. Modern techniques, such as the majority voting approach, as exemplified by [78], highlight the effectiveness of this ensemble strategy in producing accurate and robust results.

6.3 Comprehensive evaluation metrics

In this study, we adopted an approach that considered a range of metrics to understand the model's capabilities better. It has been observed that while ARTIC demonstrates superior performance in metrics such as ROUGE, METEOR, CIDEr, and SPICE,

there was a noticeable difference in BLUE-2,3 and -4 scores compared to the work of [61]. Although BLUE is a widely recognized metric for machine translation and image captioning, it has been criticized for its simplicity and potential lack of correlation with human judgment, especially in diverse and creative captions. Therefore, a balanced assessment that considers multiple metrics will provide a more accurate representation of the overall performance of a newly proposed model in this growing field of research.

The performance results of ARTIC using metrics such as ROUGE, METEOR, CIDEr, and SPICE signify its ability to generate accurate and diverse captions and its semantic richness and relevance to reference captions. These metrics offer a more complete picture of the model's proficiency in generating high-quality image captions in Arabic. The improved ROUGE scores, for instance, indicate a higher level of content overlap, potentially leading to more coherent and contextually relevant captions. Similarly, the enhanced SPICE scores suggest increased semantic similarity between generated and reference captions, highlighting the practical utility of the ARTIC approach. Our comprehensive analysis, consistent with the findings of [50], reveals that SPICE outperforms other artificial metrics in aligning with human judgments in model-generated captions. By incorporating SPICE into our analysis, we comprehensively understand its significance and impact, offering a unique perspective beyond the scope of conventional comparisons. At the time of our study, SPICE was not used in the literature.

6.4 Accurate model evaluation

The scarcity of data for Arabic image captioning study is an important factor in lower scores. Arabic Flickr8k dataset, which has 8,092 images and 24,276 captions as 3 captions per image, is currently the finest resource for Arabic image captioning. However, researchers in [29] have also attributed the lower scores to the morphological complexity of the Arabic language, resulting in sentences having significantly fewer words than most other languages and consequently much higher error penalties in metrics based on n-gram similarity such as BLEU. In comparison to English captioning, Arabic captioning exhibits fewer satisfactory outcomes. This discrepancy can be related to the richness and complexity of the Arabic language. Additionally, the ARTIC model relies on captions initially created in English and then translated into Arabic, a process that may only sometimes correctly capture the true essence of the image, events, or characters. Furthermore, the dataset for Arabic captioning is smaller than that for English captioning, and expanding the dataset size in the future will likely enhance the performance of the Arabic captioning task.

In particular, we encountered a publicly available dataset with a distinct splitting approach, a factor often disregarded in comparative studies, leading to a potential discrepancy in result interpretation. This difficulty emphasizes the need for standardized dataset practices to ensure fair and meaningful comparisons in image caption research. Therefore, unlike the other studies using the Flickr8k data set, we followed the recommendations of [88] regarding the significance of appropriate dataset partitioning, including training, validation, and testing. Our dataset-splitting methodology contributes to the robustness and applicability of deep learning methodologies in image description.

6.5 English image captioning analysis

In general, our model demonstrates proficiency in generating captions that are not only relevant but also accurate in describing the image content. As demonstrated in Fig. 11, we have provided several sentence examples produced by our caption method from the Flicker30K dataset to validate our model's effectiveness further. The highlighted text is used to identify the generated captions. As it was clear from Fig. 11(a) how the model can produce the word "racer" rather than just "a man". we can observe how the model determines gender (man) in figures Fig.11(b), Fig.11(c), Fig.11(d), Fig.11(i), and (woman) in Figure Fig.11(f). Objects as "saxophone" Fig.11(d), "helmet" Fig.11(b) , and "bag" Fig.11(i) was correctly recognized. An illustration of how the model can represent an object's location as "in front of" is provided in Fig.11(g). The description of the sites was available through the picture Fig. 11(e) in the sentence "on a railroad track," through the second picture Fig. 11(h) in a sentence "on a snowy mountain", and Fig. 11(i) "the street". Additionally, the example in Fig.11(b) "is riding" and Fig.11(c) "is throwing" effectively convey the setting. The "suite" was generated for Fig. 11(i) to clarify that it belongs to a certain outfit.

However, inaccurate captions are shown in Fig.13 from the Flicker30K dataset, highlighted in red. Fig.13(a), mistakenly add "and a" without needs, the caption describes the person who is sitting on the horse as a "man" however, the references contain options as: "man," "woman," or "person". On the other hand, the generated caption produces "a group of people sitting," but the caption includes "a car" incorrectly instead of the "screen"; the model has trouble recognizing these complicated settings.

7 Limitations and Research Opportunities

The proposed Model for Arabic Image Captioning faces some challenges: 1) It is noticeable that the model interprets some areas' color as the color of another area or clothes; a single factor might have multiple attributes, but learning to recognize attributes is still a challenging task in computer vision. 2) In some cases, the model fails to produce the correct number of elements in the target image; however, counting the number of items is a higher level of artificial intelligence than object recognition. 3) It is possible that the model fails to recognize complicated settings, which causes erroneous interpretations.

Future advancements in Arabic image captioning can address the limitations through targeted research. Firstly, effectively detecting and distinguishing items within images may require improving object recognition algorithms. Secondly, improving the preparation and collection of the Arabic image-captioning dataset could provide a greater foundation for training and assessment. In addition, creating and implementing more thorough evaluation metrics specific to the Arabic domain can offer a more comprehensive evaluation of the quality of captions. Finally, enhancing the visualization of image captioning models can be informative to see how the model pays attention to particular image sections when creating captions. These research avenues promise to enhance models' effectiveness and adaptability impact in Arabic image captioning applications.

8 Conclusion

This study investigated the complex task of Arabic image captioning, exploring the boundaries of its current capabilities. We have demonstrated the potential for significant advancements in this domain by deploying a transformer-based architecture with an attention mechanism. The thorough evaluation of the Arabic Flickr8K dataset, accompanied by standard metrics and measures, has revealed the robustness and effectiveness of the proposed model. In our experiments, we evaluated our model using Flickr30k English datasets. This approach ensures that the performance of our model is generalizable beyond the original training data. The results were promising, especially when compared to state-of-the-art models.

Significant achievements include surpassing the performance of previous work using the same dataset and establishing our model to generate high-quality and contextually relevant captions. The model's ability to navigate the complexities of Arabic grammar, including gender-specific conjugations and plural forms of verbs, signifies a substantial stride in multilingual image captioning. The comprehensive analysis of correct and incorrect captions has provided valuable insights into the model's strengths and areas for potential refinement. These findings contribute to the advancement of Arabic image captioning and a deeper understanding of the challenges of the rich Arabic language. We carefully considered the shortcomings of our methodology and suggested possible directions for further investigation to assist other researchers in improving the field of image captioning.

Beyond the immediate scope of Arabic image captioning, the research carries broader implications for the intersection of computer vision and natural language processing. The success of the transformer-based architecture prompts considerations for its application in other computer vision tasks requiring relational understanding. The careful handling of grammar, gender distinctions, and contextual complexities sets a roadmap for exploring multilingual models capable of accommodating diverse linguistic intricacies.

9 Declaration

Funding: No funding is available.

Conflict of Interest: The authors declare that they have no known competing financial interests or personal relations that could have appeared to influence the work reported in this paper.

Data Availability: The data used in this study, Arabic Flickr8k, is publicly available and can be downloaded from:

Text: https://github.com/ObeidaElJundi/Arabic-Image-Captioning/ tree/master/-data/Flickr8k_text

Images: http://nlp.cs.illinois.edu/HockenmaierGroup/Framing_Image_Description /Flickr8k_Dataset.zip

Code availability: The code used to generate the data results can be downloaded from GitHub: https://github.com/IsraaAbdullah/ARTIC.

Authors' contributions: All authors contributed equally and read and approved the final manuscript.

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] 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)

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

[4] Ghandi, T., Pourreza, H., Mahyar, H.: Deep learning approaches on image captioning: A review. arXiv preprint arXiv:2201.12944 (2022)

[5] Tien, H.N., Do, T.-H., Nguyen, V.-A.: Image captioning in vietnamese language based on deep learning network. In: International Conference on Computational Collective Intelligence, pp. 789–800 (2020). Springer

[6] 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

[7] 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)

[8] Pa, W.P., Nwe, T.L., *et al.*: Automatic myanmar image captioning using cnn and lstm-based language model. In: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), pp. 139–143 (2020)

[9] Alsabbagh, A.R., Mansour, T., Al-Kharabsheh, M., Ebdah, A.S., Al-Nahhas, S., Mahafza, W., Al-Kadi, O.: Minimedgpt: Efficient large vision–language model for medical visual question answering. Pattern Recognition Letters **189**, 8–16 (2025)

[10] Ayyoub, H.Y., Al-Kadi, O.S.: Learning style identification using semisupervised self-taught labeling. IEEE Transactions on Learning Technologies **17**, 1093–1106 (2024)

[11] Ahsan, H., Bhatt, D., Shah, K., Bhalla, N.: Multi-modal image captioning for the visually impaired. In: 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] Fudholi, D.H., Windiatmoko, Y., Afrianto, N., Susanto, P.E., Suyuti, M., Hidayatullah, A.F., Rahmadi, R.: Image captioning with attention for smart local tourism using efficientnet. In: IOP Conference Series: Materials Science and Engineering, vol. 1077, p. 012038 (2021). IOP Publishing

[13] 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)

[14] Hoxha, G., Melgani, F., Demir, B.: Toward remote sensing image retrieval under a deep image captioning perspective. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **13**, 4462–4475 (2020)

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

[16] 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)

[17] Fujiyoshi, H., Hirakawa, T., Yamashita, T.: Deep learning-based image recognition for autonomous driving. IATSS Research **43** (2019) https://doi.org/10.1016/j.iatssr.2019.11.008

[18] Shah, A.P., Lamare, J.-B., Nguyen-Anh, T., Hauptmann, A.: Cadp: A novel dataset for cctv traffic camera based accident analysis. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–9 (2018). IEEE

[19] 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)

[20] Huang, Q., Yang, L., Huang, H., Wu, T., Lin, D.: Caption-supervised face recognition: Training a state-of-the-art face model without manual annotation. In: European Conference on Computer Vision, pp. 139–155 (2020). Springer

[21] Elhagry, A., Kadaoui, K.: A Thorough Review on Recent Deep Learning Methodologies for Image Captioning. arXiv e-prints, 2107–13114 (2021) https://doi.org/10.48550/arXiv.2107.13114 arXiv:2107.13114 [cs.CV]

[22] Alyafeai, Z., Al-Shaibani, M.: Arbml: Democritizing arabic natural language processing tools. In: Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS), pp. 8–13 (2020)

[23] 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)

[24] Zhang, W., Nie, W., Li, X., Yu, Y.: Image caption generation with adaptive transformer. In: 2019 34rd Youth Academic Annual Conference of Chinese Association of Automation (YAC), pp. 521–526 (2019). IEEE

[25] 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)

[26] Al Badarneh, I., Hammo, B.H., Al-Kadi, O.: An ensemble model with attention based mechanism for image captioning. Computers and Electrical Engineering **123**, 110077 (2025)

[27] Madhfar, M.A.H., Qamar, A.M.: Effective deep learning models for automatic diacritization of arabic text. IEEE Access **9**, 273–288 (2020)

[28] Zakraoui, J., Elloumi, S., Alja'am, J.M., Yahia, S.B.: Improving arabic text to image mapping using a robust machine learning technique. IEEE Access **7**, 18772–18782 (2019)

[29] ElJundi, O., Dhaybi, M., Mokadam, K., Hajj, H.M., Asmar, D.C.: Resources and end-to-end neural network models for arabic image captioning. In: VISIGRAPP (5: VISAPP), pp. 233–241 (2020)

[30] Attai, A., Elnagar, A.: A survey on arabic image captioning systems using deep learning models. In: 2020 14th International Conference on Innovations in Information Technology (IIT), pp. 114–119 (2020). IEEE

[31] 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)

[32] 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, pp. 1–17 (2020)

[33] 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)

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

[35] 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)

[36] 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

[37] 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)

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

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

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

[41] 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)

[42] 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)

[43] Wang, D., Hu, H., Chen, D.: Transformer with sparse self-attention mechanism for image captioning. Electronics Letters **56**(15), 764–766 (2020)

[44] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics (2019). https://api.semanticscholar.org/CorpusID:52967399

[45] Cornia, M., Baraldi, L., Cucchiara, R.: Explaining transformer-based image

captioning models: An empirical analysis. AI Communications **35**(2), 111–129 (2022)

[46] 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

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

[48] 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)

[49] 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)

[50] 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

[51] Ghandi, T., Pourreza, H., Mahyar, H.: Deep Learning Approaches on Image Captioning: A Review. arXiv e-prints, 2201–12944 (2022) https://doi.org/10.48550/arXiv.2201.12944 arXiv:2201.12944 [cs.CV]

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

[53] Jindal, V.: A deep learning approach for arabic caption generation using roots-words. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31, pp. 4941–4942 (2017)

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

[55] Jindal, V.: Generating image captions in arabic using root-word based recurrent neural networks and deep neural networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, pp. 8093–8094 (2018)

[56] Al-muzaini, H.A., Al-yahya, T.N., Benhidour, H.: Automatic arabic image captioning using rnn-lstm-based language model and cnn. International Journal of Advanced Computer Science and Applications **9**(6) (2018) https://doi.org/10.14569/IJACSA.2018.090610

[57] Mualla, R., Alkheir, J.: Development of an arabic image description system. International Journal of Computer Science Trends and Technology (IJCST)–6 (3), 205–213 (2018)

[58] Hejazi, H., Shaalan, K.: Deep learning for arabic image captioning: A comparative study of main factors and preprocessing recommendations. International Journal of Advanced Computer Science and Applications **12**(11) (2021)

[59] Lasheen, M.T., Barakat, N.H.: Arabic image captioning: the effect of text preprocessing on the attention weights and the bleu-n scores. Int J Adv Comput Sci Appl **13**(7), 11 (2022)

[60] Emami, J., Nugues, P., Elnagar, A., Afyouni, I.: Arabic image captioning using pre-training of deep bidirectional transformers. In: Proceedings of the 15th International Conference on Natural Language Generation, pp. 40–51 (2022)

[61] Elbedwehy, S., Medhat, T.: Improved arabic image captioning model using feature concatenation with pre-trained word embedding. Neural Computing and Applications **35**, 1–17 (2023) https://doi.org/10.1007/s00521-023-08744-1

[62] Sabri, S.M.: Arabic image captioning using deep learning with attention. PhD thesis, University of Georgia (2021)

[63] Cho, S., Oh, H.: Generalized image captioning for multilingual support. Applied Sciences **13**(4) (2023) https://doi.org/10.3390/app13042446

[64] Colombo, F.: Transfer learning analysis of fashion image captioning systems (2020)

[65] Ray, P.P.: Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems **3**, 121–154 (2023) https://doi.org/10.1016/j.iotcps.2023.04.003

[66] Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL **2**, 67–78 (2014)

[67] 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)

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

[69] AlMahmoud, R.H., Hammo, B., Faris, H.: A modified bond energy algorithm with fuzzy merging and its application to arabic text document clustering. Expert

Systems with Applications **159**, 113598 (2020) https://doi.org/10.1016/j.eswa.2020.113598

[70] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015). PMLR

[71] 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)

[72] 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)

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

[74] 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

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

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

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

[78] 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)

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

[80] Alsayed, A., Qadah, T.M., Arif, M.: A performance analysis of transformer-based deep learning models for arabic image captioning. Journal of King Saud University-Computer and Information Sciences **35**(9), 101750 (2023)

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

[82] 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)

[83] 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

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

[85] 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)

[86] Shrimal, A., Chakraborty, T.: Attention beam: An image captioning approach (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 15887–15888 (2021)

[87] Zhao, W., Wu, X., Luo, J.: Cross-domain image captioning via cross-modal retrieval and model adaptation. IEEE Transactions on Image Processing **30**, 1180–1192 (2020)

[88] Eelbode, T., Sinonquel, P., Maes, F., Bisschops, R.: Pitfalls in training and validation of deep learning systems. Best Practice & Research Clinical Gastroenterology **52**, 101712 (2021)

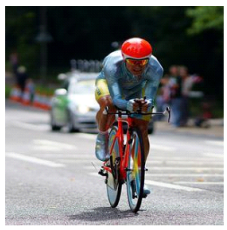
Figure 10: Samples of correct captions (highlighted in yellow) generated by the proposed model based on Arabic Flickr8k dataset

7710096952.jpg

['<start> the race car is being driven down the racetrack <end>', '<start> an orange dragster race car sits on the track <end>', '<start> a person is driving a red and black race car <end>', '<start> there is a red and black race car on a track <end>', '<start> a red race car driving on a racetrack <end>']

<start> a race car is driving on a track <end>

(a)



7700027818.jpg

['<start> a bicyclist wearing a red helmet sunglass and blue spandex clothing in a race on the street <end>', '<start> a bicycle racer is riding a bike on the street with people watching from the sideline <end>', '<start> a man who is wearing the same color a the bike is riding down the street <end>', '<start> a cyclist dressed in blue is riding up a road <end>', '<start> a man cycling on road and enjoying <end>']

<start> a man in a red helmet is riding a red bike <end>

(b)

7721977472.jpg

['<start> a man wearing a red and white uniform throw a pitch from the pitcher mound in a game of baseball <end>', '<start> a pitcher in red uniform is shown after throwing the baseball <end>', '<start> a baseball player in a red suit ha just thrown a ball <end>', '<start> a man in a red baseball uniform is throwing a pitch <end>', '<start> the baseball pitcher fire a ball towards the plate <end>']

<start> a young man in a red shirt is throwing a baseball <end>

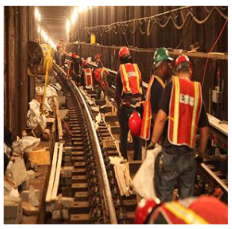
(c)

8088028369.jpg

['<start> a bald man with a beard is playing a saxophone in a dimlylit room with picture on the wall behind him <end>', '<start> a male saxophone player belt out a song in the dim light of a local establishment <end>', '<start> a man play the saxophone in a dark room lit with a red light <end>', '<start> a man is playing the saxophone <end>', '<start> a man is playing the saxophone <end>']

<start> a man in a black shirt play a saxophone <end>

(d)

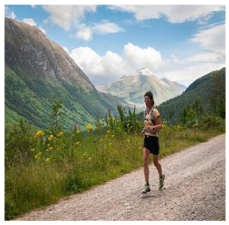


7773468788.jpg

['<start> worker in hard hat at a railroad track construction site <end>', '<start> construction worker working inside of a railway tunnel <end>', '<start> several worker trudging down a dimly lit subway tunnel <end>', '<start> several employee working on an underground rail track <end>', '<start> construction worker work laying subway track <end>']

<start> a group of construction worker are working on a railroad track <end>

(e)

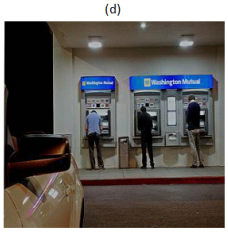


7797831448.jpg

['<start> a woman is walking down a trail with tall grass and a mountain in the background <end>', '<start> a hiker jog down a rocky road while blue sky and mountain cross the horizon <end>', '<start> woman jogging down a mountain trail with mountain in the background <end>', '<start> a woman in yellow shir and black short is running along <end>', '<start> a female runner running alone near a mountain <end>']

<start> a woman in a blue shirt and black short is walking down a trail mountain <end>

(f)



78879869.jpg

['<start> three men are using washington mutual atm outside near a parking lot <end>', '<start> three people are making transaction at an automated teller machine <end>', '<start> three men at a washington mutual atm machine in the evening hour <end>', '<start> three people are using washington mutual atm machine <end>', '<start> three people stand at washington mutual atm <end>']

<start> a group of people are standing in front of a building <end>

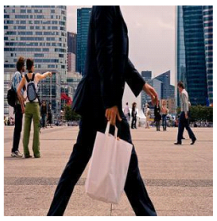
(g)

99171998.jpg

['<start> a group of people sit in the snow overlooking a mountain scene <end>', '<start> five people are sitting together in the snow <end>', '<start> a group of people sit atop a snowy mountain <end>', '<start> a group is sitting around a snowy crevasse <end>', '<start> five child getting ready to sled <end>']

<start> a group of people are in a snowy mountain <end>

(h)



986127455.jpg

['<start> man walking down the street is wearing a black suit and carrying a small white bag <end>', '<start> a person with a bag walking in a big city <end>', '<start> people walking a standing in a city park <end>', '<start> a woman in a black suit is walking by <end>', '<start> a person is walking with a white bag <end>']

<start> a man in a black suit is walking down the street with a white bag <end>

(i)

Figure 11: Samples of correct captions (highlighted in yellow) generated by the proposed model based on English Flickr30k dataset

3649307685_60c1294d2a.jpg

<start>['يلعب رجلان في لعبه رياضيه معا ', <end>
<start>'رجلان يلعبان رياضه يقاتلان من اجل كره ', <end>
<start>'لاعبان يتصارعان من اجل كره في لعبه هوكي'<end>]

<start>صبي صغير يقفز من اجل كره سله<end>

(a)

3692593096_fbaea67476.jpg

<start>['طائره تحلق فوق جبل في محاوله لاطفاء حريق
<start>', 'طائره صغيره تسقط ماده كيميائيه حمراء فوق<end>
<start> ', 'طائره صغيره حمراء تطير فوق قمه<end>
قمم جبال
جبل تسقط ماده حمراء على نار ['<end>

<start>رجل يرتدي ستره حمراء يقف على قمه جبل <end>

(b)

8089664348.jpg
['<start> a woman attired in a formal blue police or military uniform holding a flagpole is sitting astride a horse <end>', '<start> the photo is of a woman in a police uniform riding a horse <end>', '<start> a man in uniform hold a flag while sitting atop a horse <end>', '<start> a woman ha mounted a horse carrying a spear <end>', '<start> person on horseback dressed for a ceremony <end>']

<start> a man in a blue shirt and a is sitting on a horse <end>

(a)

7754645524.jpg
['<start> a crowd of people and their family are gathered in a field near an arena watching an event on an outdoor screen <end>', '<start> a crowd of people outside the olympic arena in great britain watching the olympics on a portable screen <end>', '<start> all the people in the park are trying to make themselves comfortable to enjoy the program set for them <end>', '<start> a crowd of people set on a lawn and watch the olympics on a jumbo screen <end>', '<start> a group of people have gathered in a grassy field on a sunny day <end>']

<start> a group of people are sitting in a car around a large crowd car on the sunny day <end>

(b)

Figure 12: Samples of incorrect captions (in red color) generated by the proposed model based on Arabic Flickr8k dataset

Figure 13: Samples of incorrect captions (in red color) generated by the proposed model based on English Flickr30k dataset