



Neuraltalk+: neural image captioning with visual assistance capabilities

Himanshu Sharma¹  · Devanand Padha¹

Received: 4 December 2023 / Revised: 15 February 2024 / Accepted: 15 April 2024

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2024

Abstract

Image captioning is a technique that generates concise and meaningful descriptions of the visual contents present in an image. Image captioning frameworks generally employ an encoder-decoder-based pipeline to generate image descriptions. Multimodal meaning space, visual and semantic fusion, and influential recurrent decoding are some of the highlights of these frameworks. However, the lack of cutting-edge implementation schemes, such as ensemble feature extraction, context-aware fusion, and real-time captioning, limit their integration in the vision assistance domain. In this research work, we introduce Neuraltalk+, which comprises various structural and functional enhancements, and feature-based extensions, making it lightweight, robust, effective, and automated. Neuraltalk+ uses ensemble feature extraction to extract visual and spatial image features for efficient image comprehension. We then map these feature vectors with multimodal semantic knowledge using dual context-aware feature fusion followed by self-attention-assisted decoding. Lastly, we introduce two new features: real-time captioning and visual similarity comparison, which allow vision assistance and sight comprehension capabilities. Experimental analysis on the Flickr 8K and Flickr 30K datasets demonstrates that our model trains faster and generates improved quantitative (BLEU(72.08), METEOR(33.65), and CIDEr(143.5)) and qualitative results. Neuraltalk+ also demonstrates high performance in real-time captioning for both familiar and unfamiliar contexts. We also offer potential suggestions for extending our work in the future.

Keywords Deep learning · Encoder-Decoder · Image captioning · Image comprehension · Neuraltalk · Visual assistance

1 Introduction

The current era of digital information generates an exponential amount of multimedia data (images, audio, and videos) every second. Digital images are one of the most widely shared

✉ Himanshu Sharma
himanshusharma.csit@gmail.com

¹ Department of Computer Science and Information Technology, Central University of Jammu, Jammu, Kashmir 181124, India

contents over the web across multiple domains such as social media, news, e-learning, and medical among others. An image comprises a vast amount of visual entities with various attributes and actions [39]. A quick glance is enough for us to identify all the visual entities and thus infer a logical conclusion for any image [20]. This image-understanding ability even though seems simple is one of the complex and challenging problems for image-understanding models [33]. Image Captioning (IC) refers to the generation of a description for an image [16]. The generated description is comprehensive enough to include all the significant information about the image [7]. IC models not only recognize the visual context correctly but also describe them precisely using the appropriate phrases and sentences [22]. IC thus involves the joint application of two cutting-edge disciplines of Artificial Intelligence (AI) namely Computer Vision (CV) and Natural Language Processing (NLP).

IC has numerous direct and indirect applications across various domains. For instance, it assists visually impaired individuals in navigating their surroundings [17] and e-learning platforms in generating automated captions [21]. It is also employed in self-driving cars to aid drivers in traffic congestion [24]. In addition, IC aids robots and Unmanned Aerial Vehicles (UAV) to perceive and interact with their neighboring [3, 4]. Moreover, IC frameworks are essential components of multimodal search engines that retrieve images from web servers [42]. Some novel applications include medical IC and remote-sensing IC. Therefore, devising effective and reliable IC models is crucial, as they can directly affect the performance of many other application frameworks used every day.

The principal approaches of IC are template-based, retrieval-based, and encoder-decoder-based models [31]. Template-based models [22] leverage handcrafted visual features such as Global Image Descriptor (GIST), Scale Invariant Feature Transform (SIFT), and Histogram of Oriented Gradients (HOG) extracted from the image to produce captions. While these models produce grammatically correct descriptions, they are limited in their ability to convey visual facts. Retrieval-based architectures [9] produce descriptions by reusing existing phrases based on multimodal similarity. However, as the dataset grows, they suffer from performance issues. The state-of-the-art paradigm of IC is the end-to-end encoder-decoder framework [21] that leverages data-driven techniques to generate captions. The captioning is further improved by incorporating visual and semantic attention [38] into the decoder, which leads to descriptions having tight association with the query image. The encoder-decoder paradigm has brought a revolution in the area of IC frameworks by effectively correlating visual context and lexical information within the latent space. Furthermore, the end-to-end methodology enhances the robustness and ease of upgrading, making the integration of IC frameworks into vision aid domains possible.

NeuralTalk (NT), developed by Karpathy and Fei Fei [20], is an IC framework widely renowned for its encoder-decoder paradigm. It learns a multimodal latent space that translates the input visual context to image descriptions word by word. It is freely available on the *github*¹ for educational and research purposes. Despite being released in 2015, it is still a crucial benchmark for nearly all IC models. Current frameworks draw upon several of its characteristics, including architecture, feature extraction, multimodal fusion, correlational objective function, and decoding strategy. Its methods of vocabulary initializations and handling bad tokens are still a baseline for current captioning and language modeling architectures. In recent years, various domains, including vision aid, have actively reinforced the implementation of IC. However, the NT and other cutting-edge IC models are fundamentally not structured for this purpose. It comprises complex computational models, a semi-end-to-end pipeline, and serial processing. Additionally, it primarily focuses on image rankings,

¹ **Neuraltalk:** <https://github.com/karpathy/neuraltalk>

resulting in descriptions lacking vital details for image interpretation. Moreover, data-driven concepts such as ensemble feature extraction, visual and spatial attribute fusion, and adaptive attention are absent. Even if explicitly tailored for visual aids, NT lacks built-in visual aid functionalities precisely devised for this purpose. Consequently, it requires supplementary handling, rendering it unfeasible for visual aid systems such as self-driving cars, road assistance, or aiding visually impaired individuals.

Motivated by the popularity and success of this model, we undertook this research work to design and develop a lightweight neural IC framework, Neurtalk+ (NT+), which exhibits architectural optimizations, captioning enhancements, and feature extensions. To our knowledge, no previous works have attempted to extend NT's performance and features. Our work involves introducing state-of-the-art algorithms and data-driven paradigms into the NT, making it lightweight for swift captioning and effective and automated for its possible incorporation into the visual aid domain. This research aims to develop a novel IC framework with reliable semantic knowledge and a fast captioning pipeline through structural and functional modifications. To achieve this, we follow a two-phase research methodology: analyzing baseline models for research gaps and then resolving them during the development of our NT+ framework. The main contributions of our work are as follows:

- i. We propose a novel, lightweight, and autonomous NT+ framework for visual comprehension and image captioning.
- ii. To enhance the training efficiency and description quality, we introduce dual context-aware feature fusion and adaptive attention-assisted decoding.
- iii. We design and incorporate two new vision assistance functionalities (real-time captioning and visual similarity comparison) into our framework.
- iv. Quantitative and qualitative experiments on the benchmark datasets Flickr 8K and Flickr 30K reveal that NT+ outperforms its baselines. Additionally, NT+ also yields competitive real-time performance on out-of-distribution-testing.

The remainder of this article is as follows. Section 2 presents a brief literature overview of the encoder-decoder-based IC paradigm. Section 3 comprehensively examines our research methodology, covering framework characteristics such as ensemble feature extraction, dual-context aware multimodal fusion, adaptive attention, and vision assistance extensions. The dataset, evaluation metrics, implementation details, ablation study, and comparative analysis are explored in Section 4. Section 5 discuss the limitations and future scopes of our study, followed by a conclusion in Section 6.

2 Related work

Image captioning is currently undergoing intensive research, and as a consequence, several design principles, training methodologies, and alignment algorithms have been proposed [16]. The classical encoder-decoder pipeline, attention assisted techniques, and transformer-based IC models are all strongly related to our research. Kiros et al. [21] pioneered the integration of the encoder-decoder architecture into IC, wherein a Convolutional Neural Network (CNN) and Log BiLinear (LBL) space are employed to derive a fixed-length visual feature vector, facilitating the generation of captions for query images. Karpathy et al. [20] devised an IC alignment network aimed at correlating image segments with language phrases utilizing a multimodal structured alignment cost function. The model decodes semantic context using a Recurrent Neural Network (RNN), specifically Long Short Term Memory (LSTM). Vinyals et al. [33] developed an integrated correlational model rather than a two-component

architecture, which is dynamically trained to optimize the log-likelihood to yield a target sequence of words that best fits the query image. Xu et al. [38] integrated an attention-based module into the captioning pipeline. The model dynamically evaluates the visual segments and assigns them an attention score to determine the image fragment to focus on for the next token generation. Jin et al. [19] refined the existing attention algorithm by incorporating scene-specific visual ordering within image regions, evaluating and giving attention scores to visual details by leveraging earlier predicted words. Wang et al. [34] underline the decoder's limitation in capturing global dependencies between current and previous words. They propose an incremental adaptive global context-based attention method that enhances target word prediction by capturing global information between words. Zhao et al. [42], and Effendi et al. [6] replaced the feature extractor with ResNet [13] for capturing the grid features and preserving the spatial properties between the visual entities.

The attention-assisted captioning pipeline is a standard benchmark in IC. However, explicit attention modules struggle to capture complex spatial and multiscale information in prolonged captions [14]. Thus, an implicit attention module in the decoder is necessary, enabling training with a joint multimodal loss function. This finding paved the way for integrating self-attention-driven layouts into IC, facilitating the adoption of transformer-based models. Jiang et al. [18] introduced an early transformer-based model in the IC framework. The model employs a self-attention-aided multi-gate transformer block, which evaluates image fragments with additional weights. This enhances the capture of visual and semantic representations, thereby improving caption quality. Transformer-based decoders improve caption length, but the CNN-based feature encoders lose positional and geometric details. Haque et al. [12] developed a capsule network-based feature extraction with a transformer decoder, preserving geometric details for expressive correlations in captions. Effendi et al. [6] devised an image-to-speech model, eliminating the need for a textual training corpus. The model utilized a vector-quantized variational autoencoder and a transformer-based decoder to learn direct associations between images and speech. To overcome the constraints of sequential contextual encoding and uniform weight assignment to visual regions during decoding, Shao et al. [30] proposed an end-to-end transformer-based dense IC architecture. This system incorporates a region-object correlation score unit to evaluate the relevance of visual areas, with semantic objects aiding in score determination. Despite resolving the sequential encoding issue, IC pipelines overlook potentially significant phrases during captioning, relying solely on visual context cues. Additionally, limited vocabulary diversity during training hinders the model's ability to produce diverse descriptions. To address this issue, Shao et al. [29] offered an IC design trained on a diverse vocabulary, assigning equal preference to textual and visual cues. Zhang et al. [41] proposed a fusion-enhanced multi-feature transformer-based framework. The model achieves multi-feature fusion by aligning semantic features with the visual attributes. Li et al. [23] also implemented a similar transformer-based multimodal decoder backed by a hybrid attention module for effective training. Recent analysis indicates a preference for transformer-based decoders over LSTMs in IC. However, transformers are not solely limited to decoders; deeper object detection blocks, including transformer-based fusion models [5], are increasingly integrated into IC architectures, rendering both encoder and decoder components entirely self-attention-driven. Yu et al. [40] and Guo et al. [11] utilized a hybrid attention module with a multimodal transformer-based decoder for precise caption generation. Fang et al. [8] and Wang et al. [36] devised a transformer-transformer-based framework integrating attention mechanisms for visual and semantic features.

The emergence of IC frameworks leveraging reinforcement learning is also notable. However, challenges such as exposure bias and inconsistent evaluation emerge. Zhou et al. [43] tackle this problem by introducing a reinforcement learning-based IC framework incorpo-

rating spatial attention and Generative Adversarial Network (GAN) modules. The model employs a discriminator network to address prior evaluation limitations, resulting in a more efficient framework. Zhou et al. [43] combined generation-based and retrieval-based models to enhance feature extraction and fusion in IC. The framework incorporates three modules—cross-modal feature enhancement, gated feature fusion, and cross-model decoder—to facilitate effective caption decoding with visual and enhanced features. Lian et al. [25] observed that the majority of the existing two-pass IC models employ captions to aid in the refining process by using the conventional attention blocks. They, therefore, proposed a novel cross-modification attention model that utilizes the image and caption tokens to deliver reliable features for refinement. Wei et al. [37] devised an enhanced captioning model to capture and combine the complexities of image features with semantic phrases. The model makes use of a two-phase CNN-Transformer pipeline. Transformer models are computationally expensive due to their depth and multiple hidden layers, resulting in heavyweight models [14]. It makes them hardware-intensive and time-consuming to train and generate captions. The aforementioned related works indicate a limited usefulness of captioning in aiding vision, underscoring an ongoing research area. Therefore, we designed our NT+ framework specifically to address this constraint. To improve efficiency and reliability, we developed a lightweight captioning framework. To the best of our knowledge, this is the first attempt to reconstruct an existing IC pipeline, modifying it at three fundamental levels.

3 Methodology

In this section, we provide a comprehensive overview of our research methodology, aimed at enhancing object recognition and description generation. Our methodology centers around the pivotal research question: how do structural, functional, and feature-level enhancements in the IC pipeline augment visual assistance and image interpretation capabilities? To address this, we redesigned the pipeline for lightweight execution, enhancing its ability to decode complex image comprehension skills with additional features compared to existing methods. Additionally, we prioritize componentization of the pipeline to facilitate seamless integration in future expansions. Our research methodology consists of two fundamental phases:

- i. **Analysis phase:** Study and analyze the baseline models to determine potential upgradable, replaceable, and redesignable areas, and
- ii. **Development phase:** Design and develop NT+ to adopt novel solutions for the identified concerns and implement new data-driven strategies resulting in a lightweight, compact, and robust framework.

We begin by examining the sequential flow of baseline models in our methodology, covering dataset preprocessing, vocabulary initialization, correlational mapping, model training, and captioning. Then, we systematically identify areas for structural, functional, and feature upgrades. After a thorough examination, we propose solutions for each concern. These solutions form the basis for designing our novel NT+ framework, which combines classical and cutting-edge approaches. Figure 1 illustrates both phases of our methodology. Generally, our model comprises three fundamental sets of enhancements from the baselines, namely, structural, functional, and feature-based. Structural enhancements aim to optimize the architectural pipeline by removing and replacing outdated libraries or non-efficient responsibilities. Structured enhancements instantly contribute to the lightweight and autonomous execution of the pipeline. The functional upgrades enclose improvements to the training and alignment network of the framework, including transformations to feature extraction, feature fusion, and

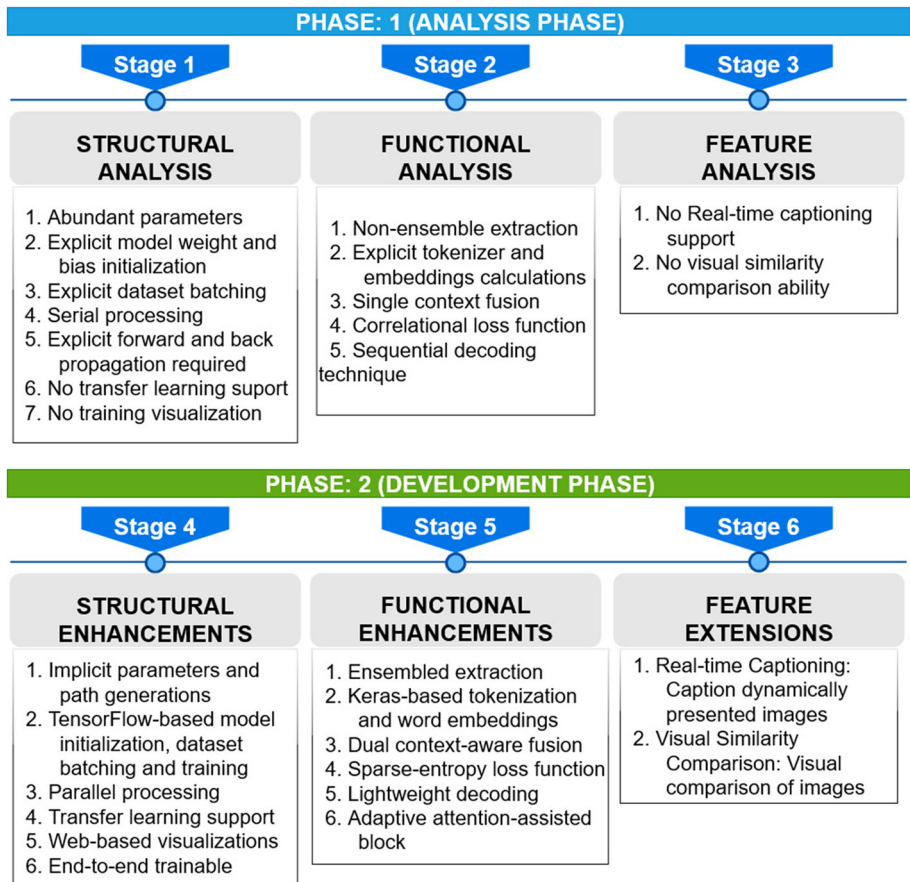


Fig. 1 The proposed methodology of our model: Analysis and Development phases

decoding. Functional enhancements impact the image comprehension capability. Feature-based advances are the addition of new components to the framework. We introduce two novel features, real-time captioning and visual similarity comparison, to our framework, which are missing in existing pipelines.

Table 1 offers an abstract comparison between NT+ and contemporary models in IC (attention-based [38], Bi-LSTM [35], and transformer model [18]). Our model exhibits substantial deviations from NT [20] and significant disparities from other baselines, as delineated in Table 1. The subsequent subsections will delve into the detailed discussion of each distinction, its significance, and its benefit.

3.1 Structural analysis and enhancements

Structural analysis and enhancements involve modifying the pipeline organization, data pre-processing, and flow techniques. While these modifications do not alter the training logic, they streamline the execution by replacing non-optimized libraries with state-of-the-art. These upgrades impact the execution and computational resource utilization, making the model

Table 1 Distinguishing features of Neurtalk+ compared to other cutting-edge image captioning frameworks

Differences		Neurtalk [20]	Show & Tell [33]	Bi-LSTM [35]	Show, Attend & Tell [38]	Multigate Transformer [18]	Neurtalk+
Structural	End-to-end	✗	✗	✓	✓	✓	✓
	Implicit parallelism	✗	✗	✗	✗	✓	✓
	Low training overhead	✗	✗	✗	✗	✗	✓
Functional	Ensemble feature extraction	✗	✗	✗	✗	✗	✓
	Context-aware fusion	✗	✗	✗	✗	✗	✓
	Attention-assisted decoding	✗	✗	✗	✓	✓	✓
Feature	Real-time captioning	✗	✗	✗	✗	✗	✓
	Visual similarity comparison	✗	✗	✗	✗	✗	✓
	Web-application	✗	✗	✗	✗	✗	✓

lightweight and swift. In the structural analysis step of our methodology (*Stage 1* of Fig. 1), we investigated several baseline implementations to identify the segments to be removed, altered, or replaced. Based on this analysis, we introduce three primary kinds of structural upgrades (pipeline-based, parallel processing, and library-based) to our NT+ framework.

3.1.1 Pipeline updates

The earlier captioning pipeline is partially end-to-end and comprises two disconnected components: the visual feature extractor API and the encoder-decoder component, as illustrated in Fig. 2. It results in a non-trainable and non-finetuneable feature extractor. In addition, the explicit saving, loading, and mapping of feature files hinder the baseline framework’s

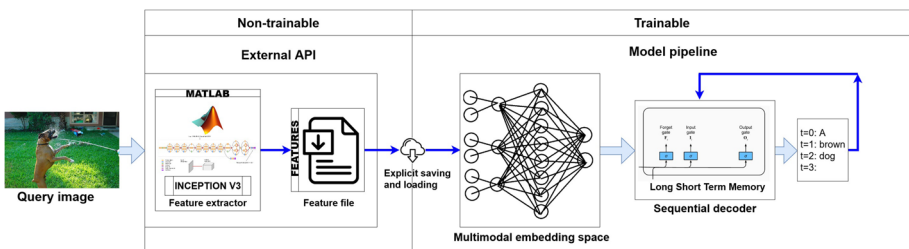


Fig. 2 Architectural pipeline of Neurtalk framework (Karpathy & Fei Fei [20])

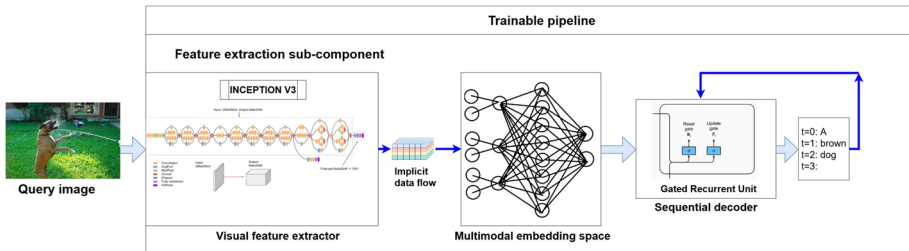


Fig. 3 The fully end-to-end and trainable pipeline of our framework Neuraltalk+

end-to-end automation capability, making it non-suitable for real-time application scenarios (vision assistance, traffic assistance, and image comparisons).

In contrast, our framework has an end-to-end trainable pipeline, where the feature extractor is trainable and finetuneable, as illustrated in Fig. 3. Furthermore, our framework operates autonomously, making it suitable for diverse real-time applications and image comprehension domains. Several additional segments, such as implicit resource initializations, reduced user-defined parameter context, log saves, and modular designs, have also been introduced to make the framework lightweight, flexible, and expandable in the future.

3.1.2 Parallel processing

The baseline frameworks have no inbuilt parallel processing implementations for training and evaluation. However, computationally high-speed Graphic Processing Units (GPUs) are effortlessly available and affordable nowadays. Lack of GPU support limits the model to efficiently use the underlying computational resources, resulting in longer training times. In response to this limitation, we integrate NT+ with implicit GPU support, allowing for swift computations and reducing training times. Moreover, our framework facilitates easy enablement and disablement of GPU support without necessitating any alterations to the pipeline configuration.

3.1.3 Library updates

The baseline IC frameworks mostly employ the native Python implementations in conjunction with several additional libraries (such as *argparse*, *numpy*, *scipy*, *JSON*, and *pickle*, among others). The data flow and training processes, including dataset batching, loss computation, gradient estimation, and backpropagation, are all explicitly executed. All these issues impose limitations on the extendability and restructuring of the frameworks. In contrast, our NT+ frameworks operate and exploit the optimized *TensorFlow* and *Keras* resources. Our ensemble feature extractor, tokenizer, word encoder, attention module, and multimodal decoder are all developed using these cutting-edge libraries. It enables the effective utilization of underlying hardware, which is particularly beneficial due to the limited availability of resources in real-time domain. A contrast between the architectural libraries used in the baseline and our model is illustrated in Table 2.

Table 2 Comparison of application libraries used in Neuraltalk and Neuraltalk+

Libraries	Usage description	Neuraltalk	Neuraltalk+
airium	Writes python context to HTML documents	✗	✓
argparse	Handles dynamically passed arguments	✓	✗
json	Loads and stores json data	✓	✓
matplotlib	Draws graphs and images	✗	✓
numpy	Performs numeric computations	✓	✓
os	Loads os and machine specific details	✓	✓
pickle	Loading and stores files	✓	✓
PIL	Image manipulation	✗	✓
random	Random numbers generator	✓	✓
re	Preprocesses text dataset	✗	✓
scipy	MATLAB file processing	✓	✗
time	Date and time details	✓	✓
TensorFlow	Deep learning framework	✗	✓
tqdm	Progress bar	✗	✓

3.2 Functional analysis and enhancements

The functional upgrades are one of the significant contributions of our work. In this analysis (*Stage 2* of Fig. 1), we empirically analyze the baseline frameworks to identify obsolete strategies and techniques that require disposal and alteration. Based on this, we introduce an ensemble feature extraction network incorporating the recurrent and non-recurrent visual features into the latent space. We also present a novel dual context-aware multimodal fusion that fuses the visual information with every semantic phrase, making the multimodal correlational space more intricate. Furthermore, we incorporate an adaptive attention module in the decoder to capture correlational dependencies within the fused sequences and previously generated words. In particular, we introduce three different kinds of functional upgrades in NT+, including ensembled feature extraction, context-aware feature fusion alignment, and adaptive attention-assisted decoding.

3.2.1 Ensemble feature extraction

The initial phase of the IC pipeline is feature extraction, which involves extracting visual information from image regions. It captures significant identities of the input image, enabling subsequent multi-modeling, feature fusion, and decoding. The feature extractor typically employs a deep CNN leveraging multiple convolutional and pooling layers. The convolutional filters of various sizes capture diverse features containing detailed information about the visual entities and their actions and attributes. The classical convolutional models, however, lack spatial information, which is essential for establishing connections between visual entities. Consequently, even though the generated descriptions contain rich object and semantic details, they lack spatial knowledge. However, in visual assistance, associations are crucial for comprehensive context. To address the lack of spatial information during the feature extraction, we introduce an ensemble feature extraction network, where the spatial and visual features are individually extracted.

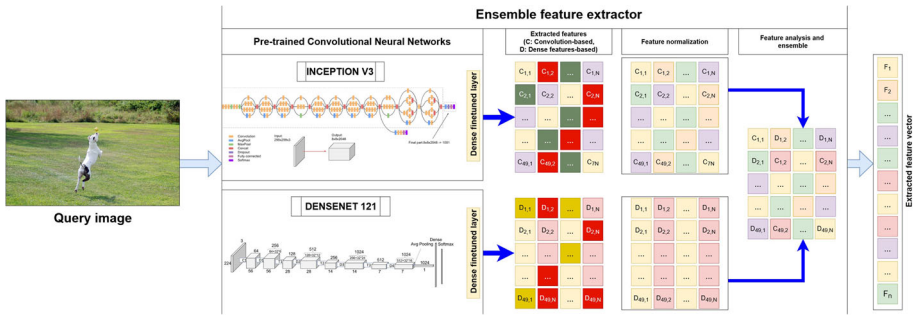


Fig. 4 An ensemble feature extraction model using Inception V3 and DenseNet-121 networks

We leverage a dual CNN (Inception V3 and DenseNet-121) ensemble network to preserve the spatial features in our pipeline. The DenseNet-121 has around 7 million trainable parameters and has dense feature associations, thereby saving spatial attributes. A finetuning layer is introduced at the top of both networks for realigning feature vectors into the equivalent feature space. The ensemble extractor extracts both the spatial and visual features from the images. We then employ a preprocessing layer to normalize the feature values removing any variations and noises from the extracted attributes. As depicted in Fig. 4, the analysis and fusion block fuses and reintroduces the spatial information into the feature vector wherever necessary. The extracted features are then preserved and used during the execution of the rest of the pipeline.

3.2.2 Self-attention block

Attention is a predominant technique in deep learning models that considerably affects their training and efficiency. It draws inspiration from human and animal visual systems, which selectively focus on vital contextual information while disregarding irrelevant details. A similar strategy is implemented in our IC framework to attend only to the relevant image segments [38], leading to prompter model training and elaborate captioning as depicted in Fig. 5.

The baseline NT model does not use any attention mechanism as outlined in Algorithm 1. Consequently, it cannot dynamically shift its gaze on influential visual regions while captioning, resulting in erroneous descriptions in numerous instances. In contrast, NT+ incorporates a self-attention block leveraging the Bahdanau attention mechanism [1] to dynamically

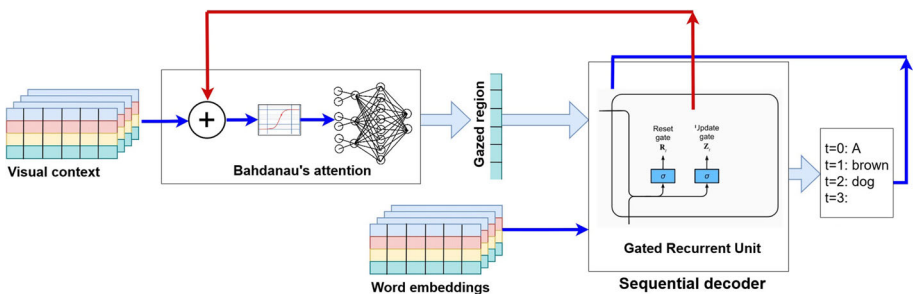


Fig. 5 Attention mechanism (Bahdanau et al. [1]) in Neuraltalk+

compute the significance of the image regions corresponding to the previously generated phrase. Based on the selected image region, context-aware fusion is applied to obtain the visual-semantic cross-domain embedding. This way, the caption prediction is guided by both modalities resulting in a dual context-aware decoding technique.

Algorithm 1 Training and backpropagation sequence of Neurtalk (Baseline)

Require: Embedding size (es), Caption size (S), Batch size (bs), Dataset (D), Pipeline weights (W)

```

1: while batch in  $D$  do
2:   while (image, caption) in batch do
3:      $E_i^{1*es} \leftarrow \text{feature\_encoder}(\text{image})$  ▷ Encode image features
4:      $E_c^{s*es} \leftarrow \text{caption\_encoder}(\text{caption})$  ▷ Encode captions
5:      $E_m^{(1+s)*es} = MF_{\hat{Z}}(E_i^{1*es}, E_c^{s*es})$  ▷ Multimodal fusion
6:     while  $E^{1*es}$  in  $E_m^{(1+s)*es}$  do ▷ Caption prediction
7:        $P_{caption} \leftarrow \text{decoder}(E^{1*es})$ 
8:        $L_i \leftarrow \text{loss\_function}(P_{caption}, \text{caption})$ 
9:     end while
10:     $L_{bs} + \leftarrow L_i$ 
11:  end while
12:   $G \leftarrow \theta(\text{feature\_encoder}, \text{caption\_encoder}, L_{bs})$  ▷ Gradient calculation
13:   $W \leftarrow \text{backpropagation}(W, G)$  ▷ Model update
14: end while

```

We implement a hard variant of region-based attention, where only a single image subregion is gazed at each decoding step. Our self-attention block leverages the extracted image features and the decoder’s hidden states. The visual attributes first undergo normalization using the hyperbolic tangent (\tanh) activation, following which the activation scores are computed using the Hadamard addition of visual and hidden attributes using an Artificial Neural Network (ANN). These scores determine the significance of the image region to be attended to and fused in the current decoding step. The feature vector of the selected subregion is known as dynamic context (I_f). Equations (1), (2), and (3) provide the mathematical representation of the steps performed for obtaining dynamic context.

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{256} \exp(e_{t,k})} \tag{1}$$

$$a(h_{t-1}, F_i) = v_a^T \tanh(W_a h_{t-1} + U_a I_e) \tag{2}$$

$$I_f = a(s_{t-1}, I_e) \tag{3}$$

where $\alpha_{t,i}$ represents the attention weight assigned to the i^{th} image region at time step t , $e_{t,i}$ is the attention energy or compatibility score between the previous decoder state h_{t-1} and the image regions I_e at time step t , a is the alignment function, v_a , W_a , and U_a are the learnable parameters, and \tanh is the hyperbolic tangent function. The output vector I_f is the dynamically selected visual region to be attended by the decoder at the current time step.

3.2.3 Dual context-aware fusion

The cross-domain fusion facilitates the interactions between the visual and the word embeddings to align the image regions with their respective semantic phrases. It enables the model to learn a multimodal latent space capable of translating the visual context into description sentences. In baseline models, the visual features (I_e) and the semantic embeddings

($C = \{w1_e, w2_e, \dots, wk_e\}$, where k is the maximum caption length and $wi_e \in \text{vocabulary}$) are organized sequentially in the semantic space (\hat{Z}) by positioning visual features above the semantic phrases as illustrated in (4).

$$S_{v,s} = MF_{\hat{Z}}(I_e, [w1_e, w2_e, \dots, wk_e]) \tag{4}$$

where $S_{v,s}$ is the fused encoding, and $MF_{\hat{Z}}$ is the cross-domain fusion function. The complete operational procedure of one of our baseline model (NT) is outlined in the Algorithm 1. As depicted in it, the decoder’s hidden states are initially set by the visual features. Subsequently, the decoding process relies entirely on previously generated semantic information. However, there are instances where the decoder makes incorrect word predictions, leading to inaccurately generated descriptions. The limitation arises from the decoder’s inability to reference the visual information during each prediction step. Consequently, the visual information holds minimal significance in the decoding process of baseline models. Furthermore, this drawback makes the training process complex, and time-consuming and restricts IC to be employed in more complicated and innovative domains

We present a novel multimodal decoding technique in NT+ where the visual information is gazed at for the prediction of each word. The joint direction of cross-modal attributes ensures improved token prediction. By considering both the semantic and visual details, the decoder accurately predicts the subsequent word, even when the previously predicted token is incorrect. As a result, the generated captions are closely aligned with the query image. Additionally, we replace the traditional multimodal fusion method with a novel cross-domain approach where visual features are merged with each word embeddings individually. Employing the novel fusion technique, a combined visual-semantic cross-domain embedding is generated that surpasses the effectiveness of earlier used features. Equation (5) describes our multimodal fusion technique.

$$S_{v,s} = \sum_{i=1}^k (CF_{\hat{Z}}(I_f, wi_e)) \tag{5}$$

where $S_{v,s}$ is the fused encoding, k is the maximum caption length, $CF_{\hat{Z}}$ is the cross-domain context-aware fusion function, and I_f and wi_e are the dynamic context and individual word embeddings respectively. Once the multimodal fusion completes, the context is fed to the decoder for predicting the subsequent word token, as depicted in Algorithm 2.

The respective equations of the update gate, reset gate, and hidden states of our GRU-based decoder model are illustrated in (6), (7), and (8).

$$z(t) = \sigma(W_z x(t) + U_z h(t - 1) + b_z) \tag{6}$$

$$r(t) = \sigma(W_r x(t) + U_r h(t - 1) + b_r) \tag{7}$$

$$h(t) = (1 - z(t)) \odot h(t - 1) + z(t) \odot h(\tilde{r}) \tag{8}$$

where wz , uz , wr , ur , bz , and br are the trainable weights and biases of the update and reset gates. $x(t)$ stands for the input at the current time frame, $h(t-1)$ depicts the previous hidden state, σ denotes the sigmoid activation and \odot signifies the elementwise multiplication. The LSTM network is a more advanced and complex architecture compared to the GRU. This complexity enables the decoder to exhibit improved performance in decoding tasks. The LSTM network incorporates an additional output gate to regulate the information flow.

Algorithm 2 Training and backpropagation sequence of Neurtalk+

```

Require: Embedding size ( $es$ ), Caption size ( $S$ ), Batch size ( $bs$ ), Dataset ( $D$ ), Pipeline weights ( $W$ )
1: while  $batch$  in  $D$  do
2:   while ( $image, caption$ ) in  $batch$  do
3:      $E_i^{n*es} \leftarrow ensembled\_feature\_encoder(image)$  ▷ Encode images
4:      $E_c^{s*es} \leftarrow caption\_encoder(caption)$  ▷ Encode captions
5:      $H \leftarrow E_c^{1*es}$  ▷ The start token embedding
6:     while  $E_c^{1*es}$  in  $E_c^{(s-1)*es}$  do ▷ Caption prediction
7:        $A^{1*es} \leftarrow attention\_block(E_i^{n*es}, H)$  ▷ Attention calculation
8:        $P_{caption}, \leftarrow decoder(CF_z(A^{1*es}, E_c^{1*es}))$  ▷ Context-aware fusion
9:        $L_i \leftarrow loss\_function(P_{caption}, caption)$ 
10:       $H \leftarrow E_c^{1*es}$ 
11:    end while
12:     $L_{bs}+ \leftarrow L_i$ 
13:  end while
14:   $G \leftarrow \theta(feature\_encoder, caption\_encoder, L_{bs})$  ▷ Gradient calculation
15:   $W \leftarrow backpropagation(W, G)$  ▷ Model updation
16: end while

```

Moreover, unlike the GRU, the LSTM decoder uses an explicit mechanism for updating the cell state, enabling the preservation and manipulation of both short-term and long-term memory. However, due to the incorporation of extra parameters, it takes more resources and training period. Equations (9), (10), (11), (12), and (13) depicts the forget gate, input gate, output gate and cell and hidden state update equations for our LSTM-based decoder respectively.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{9}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{10}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{11}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{12}$$

$$h_t = o_t \odot \tanh(\tilde{C}_t) \tag{13}$$

where $f_t, i_t, o_t, \tilde{C}_t$, and h_t are the outputs of forget gate, input gate output gate, cell state, and hidden state respectively. W and b denote the weight matrices and bias vectors, h_{t-1} is the previous hidden state. The σ and \odot functions denote the sigmoid activation and element-wise multiplication respectively.

3.3 Feature analysis and enhancements

IC finds extensive application across many domains like social media, e-commerce, visual aid, and traffic assistance. As a result, there is a growing need for real-time IC frameworks. However, the limitations of baseline IC models, such as reliance on external feature extraction APIs, distributed pipelines, and non-lightweight execution, hinder their use in real-time image comprehension domains. By contrast, NT+ employs a fully implicit execution pipeline, enabling the expansion of functionality. Accordingly, in the feature enhancement phase (*Stage 6* of Fig. 1), we have introduced two new components (real-time captioning and visual similarity comparison) to our framework to facilitate image comprehension and visual assistance capabilities.

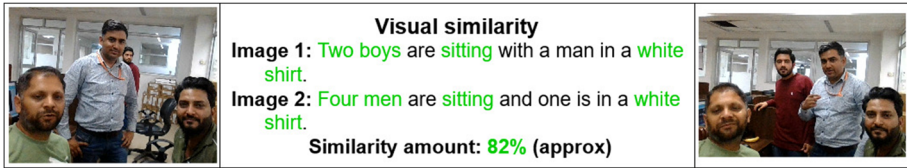


Fig. 6 Results of visual similarity between two sample images

3.3.1 Visual similarity

While the visual similarity segment of our framework is in its preliminary stages of development, we have demonstrated its potential value in real-time scenarios, particularly in image comparison. It is a crucial part of multimodal search engines that retrieve images based on multimodal queries. Presently, we leverage the generated semantic information for computing visual similarity. In the future, we can explore more innovative techniques, such as visual context or multimodal vector analysis. Figure 6 presents the visual similarity outcomes obtained by our model for a pair of input images.

3.3.2 Real-time captioning

Real-time captioning allows the interpretation of any image presented to the model dynamically. Our framework has a user-friendly and swift dynamic captioning functionality. It is valuable for real-time visual aid applications, including blind assistance and traffic aid systems, to comprehend the surrounding environments and take proper responses. NT+ executes autonomously to extract and save the feature of the query image, followed by the captioning stage to generate descriptions. The whole process is automated and does not demand any human intervention. Additionally, owing to our framework's modular and componentized architecture, the inclusion of supplementary visual support, such as image-to-voice conversion or visual question answering, can be seamlessly integrated without the need for future system alterations. Figure 7 illustrates results, showcasing dynamically captioned images by our model. Our model produces captions that align more closely with observed semantics, providing comprehensive inference of visual traits while maintaining syntactic correctness in description phrases. Consequently, the captions generated by NT+ are sufficiently reliable for integration into visual assistance systems.

4 Experiments

To assess the efficacy of our proposed NT+ framework, we perform a comprehensive set of experiments on the popularly adopted Flickr 8K and Flickr 30K datasets, which serves as a widely used public benchmark for IC. Subsequently, we present detailed information regarding the architecture, dataset preprocessing, search spaces, evaluation metrics, testing procedures, and evaluation strategy employed in our study. We further conduct ablation experiments and offer comparative results for thorough analysis. Finally, we visualize the performance of our model in comparison to the baseline frameworks to demonstrate the effectiveness of our proposed model.


	<p>Ground Truth</p> <ol style="list-style-type: none"> 1. A group of children run a footrace in the snow. 2. A group of young boys race on a snowy day. 3. School kids racing in the snow. 4. The children are running in the snow with fences in the background. 5. The young runners are racing through the snow. 	<p>Predictions (beam search=3)</p> <ol style="list-style-type: none"> 1. A group of children are playing in the snow. 2. Children are playing in the snow. 3. Many children are playing in the snow.
	<p>Ground Truth</p> <ol style="list-style-type: none"> 1. A dog is running along a beach on a sunny day. 2. A dog runs along the shore at a beach. 3. A lean dog runs along the beach. 4. A dog is running at the beach in front of the ocean and a bright blue color sky. 5. Slender dog running in the sand on a sunny day. 	<p>Predictions (beam search=3)</p> <ol style="list-style-type: none"> 1. A dog is running on the beach. 2. The dog is running on the beach. 3. Dog is running on the beach.
	<p>Ground Truth</p> <ol style="list-style-type: none"> 1. A motorcycle racer leans his bike. 2. A motorcyclist is driving down a road on their motorbike. 3. A motorcyclist is riding their sponsored car along a roadway that has recently turned. 4. A motorcyclist on the street. 	<p>Predictions (beam search=3)</p> <ol style="list-style-type: none"> 1. Motorcyclist is riding a red motorcycle. 2. A man on a red motorcycle. 3. An orange motorcycle and a racer.
	<p>Ground Truth</p> <ol style="list-style-type: none"> 1. A group of people stand in the sand looking out at the water. 2. Four people look out toward the ocean. 3. Five people standing in front of a body of water. 4. A lady and three men looking at the ocean. 	<p>Predictions (beam search=3)</p> <ol style="list-style-type: none"> 1. Three people are standing in front of the beach. 2. A group of people are standing near the beach. 3. Four people are standing near the beach.
	<p>Ground Truth</p> <ol style="list-style-type: none"> 1. A brown dog running in a yard. 2. A dog runs on the grass in a yard , a gazebo in the background. 3. A dog trotting through someone 's yard. 4. A large brown dog is running through a grassy backyard. 	<p>Predictions (beam search=3)</p> <ol style="list-style-type: none"> 1. A brown dog is running through the grass. 2. One brown dog is running through the grass. 3. Dog is running through the grass.
	<p>Ground Truth</p> <ol style="list-style-type: none"> 1. A crowd wearing red , cheers on the red football team. 2. Football players in red congratulate each other as crowds in red cheer behind. 3. Two Oklahoma Sooner football players talk on the sideline. 	<p>Predictions (beam search=3)</p> <ol style="list-style-type: none"> 1. Football players in a crowd 2. Oklahoma football player in a crowd. 3. American footballers in a crowd

Fig. 7 Results of real-time image captioning against the ground truth descriptions

4.1 Dataset

IC frameworks need a substantial amount of (*image, caption*) training pairs for generating reasonable captions. The widely utilized Flickr 8K dataset, introduced by Hodosh et al. [15] and Flickr 30K introduced by Gong et al. [10], serves as prominent benchmark for training and evaluating IC frameworks. Flickr 8K dataset consists of 8,091 images, each associated with five distinct captions, resulting in a total of 40,455 unique image-caption pairings. Flickr 30K is a much larger dataset containing more complex visual representations of people, animals, and indoor and outdoor scenes. We standardize and lowercase all the captions. To ensure consistency with the baselines, Karpathy’s splits has been used in both the datasets. Specifically, our model trains on 30,455 unique (*image, caption*) pairs, followed by fine-tuning on 5,000 new samples and validates on unseen 1,000 images on Flickr 8K. In the case of Flickr 30K, we train on 113,915 unique (*image, caption*) pairs followed by validation on 15,000 new samples and testing on 6,000 unseen images.

4.2 Evaluation metrics

We evaluate the performance of our framework against the baseline models using the widely-used Bilingual Evaluation Understudy (BLEU) score, Metric for Evaluation of Translation

with Explicit Ordering (METEOR), and Consensus-based Image Description Evaluation (Consensus-based Image Description Evaluation).

4.2.1 BLEU

BLEU [28] is a popular machine translation evaluation metric using a modified n-gram precision approach to compare candidate captions with reference sentences. The BLEU score for a candidate caption against a ground truth sentence is calculated as illustrated in (14).

$$BLEU = \frac{\sum_{C \in Cand} \left(\sum_{T_n \in C} Count_{clip}(T_n)_{C,C'} \right)}{\sum_{C' \in Ref} \left(\sum_{T_n \in C'} Count(N) \right)} \quad (14)$$

where T_n represents the n-gram tokens occurring in the candidate and reference captions, based on the user-defined n-gram size. $Count()$ computes the total number of unique tokens in the caption, while the $Count_{clip}$ determines the minimum frequency of shared tokens between the C_{and} and C_{ref} captions.

4.2.2 METEOR

METEOR [2] evaluates the input caption tokens against the WordNet vocabulary for semantics. The METEOR [2] score includes a porter module and a synonym module that help map word tokens to their semantic meanings. Precision (P) and Recall (R) are calculated based on all such mappings. The F_{mean} is then calculated using P and R, as shown in (15)

$$F_{mean} = \frac{10PR}{R + 9P} \quad (15)$$

The final METEOR score is calculated as illustrated in (16).

$$METEOR = F_{mean} \times (1 - Penalty) \quad (16)$$

where $Penalty$ is calculated for each unaligned n-gram in the candidate caption.

4.2.3 CIDEr

The CIDEr [32] score is a more comprehensive evaluation metric than BLEU [28] and METEOR [2], as it incorporates the underlying concepts of the reference captions to determine whether the candidate captions accurately depict the same idea. Unlike traditional n-gram matching, the CIDEr score provides a consensus-based accuracy score that is both normalized and reliable.

4.3 Implementation details

We use an ensemble of object recognition and feature extraction models, including Inception V3 and DenseNet-121 for extracting visual image regions. A feature encoder network is then used to turn the visual information into a multimodal semantic embedding. We eliminated all words that appeared fewer than 5 times in the training sets while creating our vocabulary, resulting in a 2768 and 11569-word vocabulary size for Flickr 8K and Flickr 30K respectively. Any term that is not in our vocabulary is represented by the 'UNK' token, which stands for

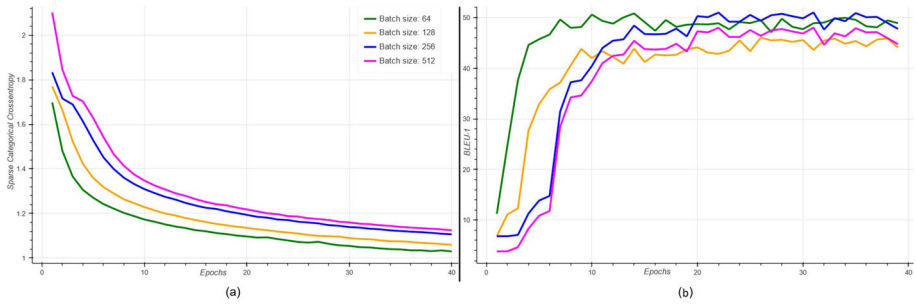


Fig. 8 Visualizations of our model training on Flickr 8K: a) Sparse categorical cross-entropy loss per epoch, b) Average BLEU-1 evaluation scores per epoch during training

unknown token. Our attention network examines visual areas to identify their significance for predicting present tokens. We use a single hidden layer ANN that aids the GRU or LSTM-based decoder. The decoder model is linked to an output-dense layer with hidden units equal to the vocabulary sizes. To deal with the pipeline’s complexity, we include a 0.2 dropout in each trainable component. For backpropagation and learning, NT+ employs a sparse categorical cross-entropy loss function and Adam’s optimizer with a learning rate of $1e-3$.

4.4 Training strategy

During training, the model employs a variety of batch sizes (32, 64, 128, 256, 512, 1024, and 2048). Because of the extra gradient calculation and backpropagation, smaller batches require more training time per epoch than the larger ones. As a result, as demonstrated in Fig. 8(a), models with small batch sizes converge faster than models with higher batch sizes.

We train each of our variants for a minimum of 40 epochs while analyzing their performance on the evaluation set. Since the smaller batch training receives more modeling adjustments, these variants attain satisfactory evaluation scores quicker than the larger batches, as shown in Fig. 8(b). Finally, the test split is processed, and the resulting captions are compared to the ground truth to determine quantitative findings. Flickr 30K comprises a substantial training samples including diverse indoor and outdoor scenes, landscapes, people, and animals. Consequently, the model trains and converges more efficiently as illustrated in

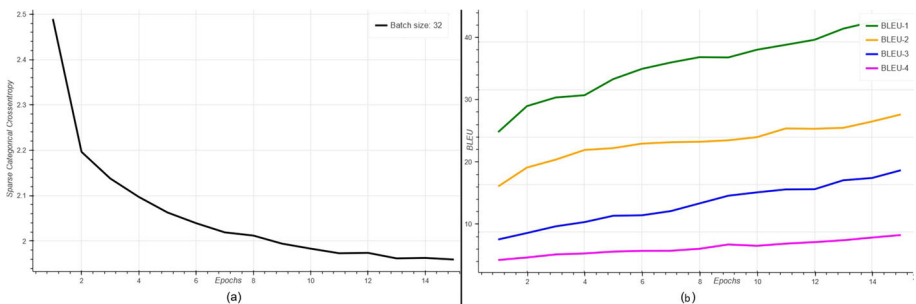


Fig. 9 Visualizations of our model training on Flickr 30K: a) Sparse categorical cross-entropy loss per epoch, b) Average BLEU-1, BLEU-2, BLEU-3 and BLEU-4 evaluation scores per epoch during training

Table 3 Ablative experimental results for feature extractor ensembles

Extractor	Feature Shape	Batch Size	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Inception V3	2048	32	40.65	21.63	8.31	2.78	20.23	129.21
		64	41.23	21.76	9.63	2.91	21.61	128.51
		128	42.49	22.80	10.15	3.00	20.96	129.84
		256	43.14	23.04	10.83	2.96	22.01	129.48
		512	41.36	22.35	9.82	2.95	21.78	127.06
DenseNet-121	1024	32	45.56	29.37	11.14	2.84	22.07	129.35
		64	45.41	30.16	10.46	3.01	23.79	129.50
		128	46.73	30.68	12.36	2.92	24.35	130.25
		256	47.27	31.03	12.52	4.03	26.69	131.47
		512	48.59	32.39	13.03	4.34	28.56	130.56
ResNet-101	4096	32	43.39	25.22	9.92	2.63	23.77	129.49
		64	43.52	26.31	9.74	2.49	23.83	128.81
		128	44.61	26.07	10.17	3.11	24.35	129.62
		256	45.90	27.14	11.01	3.20	23.92	129.68
		512	46.33	27.56	10.83	2.96	22.46	129.51

The best result in each column is highlighted in bold

Fig. 9(a). Additionally, it offers a large training vocabulary, enabling our model to generate more efficient and robust captions, as indicated by the BLEU evaluation graph shown in Fig. 9(b).

4.5 Ablation study

We conducted ablation experiments on the Flickr 8K dataset to analyze the NT+ framework's capabilities. The investigations focused on feature extraction, context-aware fusion, decoding techniques, and attention mechanisms. We examined how different combinations of feature extraction ensembles, context-aware techniques, and attention modeling influence outcomes. Table 3 shows results from our ablation study, focusing on the feature extractor ensemble. We evaluate our framework using three feature extractor blocks: Inception V3, ResNet-101, and DenseNet-121. These extractors generate feature shapes of 2048, 4096, and 1024 respectively. All three are prominent extractor blocks with varying abilities. Inception V3 focuses on parallel convolutional operations, while ResNet-121 is a residual network-based extractor. DenseNet-101 employs dense connections to encourage feature reuse and gradient flow. Table 3 illustrates DenseNet-101's precise capture of visual characteristics and retention of spatial information. Additionally, DenseNet-101 achieves a reduced feature size of 1024, aiding quick and efficient model training, aligning with our goal of a lightweight and fast pipeline. Consequently, we proceed with DenseNet-121 and Inception V3 for the remaining ablation studies, given their brisk operational speed.

Table 4 summarizes our ablation study on the feature fusion component. We tested our pipeline with diverse image features, including spatial and visual attributes, in the fusion module. The results in Table 4 show that context-aware fusion, merging spatial and visual attributes, outperforms other strategies. It underscores the importance of including details from both visual and spatial contexts in caption generation.

Table 4 Ablative experimental results for dual context-aware feature fusion (I: Inception-V3, D: DenseNet-121, G: GRU, and embedding sizes: 32, 64, 128, 256, 512, 1024 and 2048)

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Visual context-aware fusion	IG_32	41.48	21.6	8.65	2.89	21.93	129.15
	IG_64	41.43	22.58	10.31	3	20.36	127.84
	IG_128	42.51	23.35	9.15	2.94	20.76	129.21
	IG_256	37.68	23.6	8.8	2.84	22.32	129.47
	IG_512	40.93	24.04	8.82	2.42	21.46	129.21
Spatial and visual context-aware fusion	DG_32	45.64	32.38	13.73	4.36	28.56	130.5
	DG_64	45.46	30.97	12.63	3.15	24.77	131.35
	DG_128	45.71	32.72	12.14	3.63	24.94	129.69
	DG_256	47.23	29.98	11.46	2.18	23.97	128.45
	DG_512	48.59	31.06	10.32	2.08	22.87	129.49

The best result in each column is highlighted in bold

Table 5 Ablative experimental results for decoding techniques (nD: Non-dual context-aware fusion, G:GRU, L: LSTM, and embedding sizes: 32, 64, 128, 256, 512, 1024 and 2048)

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
GRU-assisted captioning	nD-G_128	41.48	21.6	8.65	2.89	21.93	129.15
	nD-G_256	41.43	22.58	10.31	3	20.36	127.84
	nD-G_512	42.51	23.35	9.15	2.94	20.76	129.21
	nD-G_1024	37.68	23.6	8.8	2.84	22.32	129.47
	nDG_2048	40.93	24.04	8.82	2.42	21.46	129.21
	DG_128	45.64	32.38	13.73	4.36	28.56	130.5
	DG_256	45.46	30.97	12.63	3.15	24.77	131.35
	DG_512	45.71	32.72	12.14	3.63	24.94	129.69
	DG_1024	47.23	29.98	11.46	2.18	23.97	128.45
	DG_2048	48.59	31.06	10.32	2.08	22.87	129.49
LSTM-assisted captioning	nD_L_128	47.55	24.46	12.71	3.27	27.01	130.19
	nD_L_256	46.29	26.54	12.61	3.12	25.78	130.07
	nD_L_512	45.36	27.27	11.72	3.14	26.9	129.82
	nD_L_1024	43.7	28.58	12.42	3.94	27.32	129.13
	nD_L_2048	44.34	29.63	12.27	4.17	27.25	128.32
	DL_128	49.71	30.5	15.31	6.47	29.06	142.45
	DL_256	49.36	29.92	15.57	6.93	33.49	142.35
	DL_512	49.6	29.53	16.12	6.96	33.44	141.69
	DL_1024	46.84	27.88	14.6	6.61	33.2	136.45
	DL_2048	45.79	26.25	15.74	5.68	32.66	139.49

The best result in each column is highlighted in bold

Furthermore, different techniques exist for decoding multimodal context into captions. The GRU-based decoding method employs a recurrent approach without an output control gate and fewer training parameters. However, GRU typically retains less historical context, resulting in more generic captions. In contrast, the LSTM-based decoder leverages a more potent recurrent mechanism with an output gate to predict the next token of the caption.

Consequently, the generated word is effectively visualized and reused within the decoder. The results in Table 5 reveal that the LSTM-based decoder yields captions more closely aligned with the query image. The LSTM decoder outperforms the GRU-based decoder in capturing and analyzing complex dual-context semantics during decoding. Additionally, LSTM focuses solely on relevant semantic information via the output gate. The resultant captions are descriptive, accentuating prominent visual subregions and their distinctive features and actions. Consequently, the captions better align with the ground truth, enhancing the overall effectiveness of the model.

We also conduct ablation experiments on our attention module, exploring attention and non-attention-assisted training. The attention block receives input from the 49 detected image sub-regions extracted during feature extraction before predicting each word. It scores the visual regions based on the previously predicted token and the visual embedding. We select and fuse the region with the highest attention score to create the dual-context multimodal embedding. Our attention-based decoding outperforms the baseline model, which lacks attention in both the encoding and decoding stages. The experimental results of our attention ablation study are presented in Table 6.

To illustrate the statistical distribution observed in our ablation investigations, we employ Kernel Density Estimation (KDE) plots to visualize our BLEU scores. Figure 10 depicts KDE plots for both BLEU-1 and BLEU-2 scores, respectively. The narrow bases of both BLEU-1 and BLEU-2 distributions indicate low variability among our data, implying no underfitting during training and testing. The shady regions in Fig. 10(a) & (b) highlight the most frequently occurring scores, with our model variants consistently achieving BLEU-1 scores within the range of (46-51) and BLEU-2 scores within the range of (30-33). These ranges notably exceed the baseline scores, indicating the effectiveness of our model variants.

Figures 11 depict KDE plots specifically for BLEU-3 and BLEU-4 scores. In Fig. 11(a), the distribution of BLEU-3 scores indicates the presence of two distinct kernel point ranges, namely (10-12) and (15-16). This observation highlights a discernible variability within the BLEU-3 data, underscoring the necessity for additional training to facilitate convergence towards a singular point. Likewise, the KDE plot for BLEU-4 scores, depicted in Fig. 11(b), manifests a similar pattern to BLEU-3, albeit not identical, with variability occurring within data point ranges of (2-4) and (7-8). The statistical visualization analysis reveals that our model effectively captures BLEU-1 and BLEU-2 metrics during training, exhibiting negligible data dispersion and no outliers. However, for BLEU-3 and BLEU-4, our model converges towards two distinct data ranges, suggesting some variability and the presence of an outlier within the ablation dataset.

4.6 Comparative experiments

We evaluate our work quantitatively and qualitatively against various standard state-of-the-art IC models. On both the Flickr 8K and Flickr 30K datasets, our model offers effective results and more reliable captions than the baselines.

Table 6 Ablative experimental results for attention and non-attention assisted captioning (I: Inception V3, G: GRU, L: LSTM, D: DenseNet-121)

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Non-attention assisted captioning	IG_128	41.48	21.6	8.65	2.89	21.93	129.15
	IG_256	41.43	22.58	10.31	3	20.36	127.84
	IG_512	42.51	23.35	9.15	2.94	20.76	129.21
	IG_1024	37.68	23.6	8.8	2.84	22.32	129.47
	IG_2048	40.93	24.04	8.82	2.42	21.46	129.21
	IL_128	47.55	24.46	12.71	3.27	27.01	130.19
	IL_256	46.29	26.54	12.61	3.12	25.78	130.07
	IL_512	45.36	27.27	11.72	3.14	26.9	129.82
	IL_1024	43.7	28.58	12.42	3.94	27.32	129.13
	IL_2048	44.34	29.63	12.27	4.17	27.25	128.32
	DG_128	45.64	32.38	13.73	4.36	28.56	130.5
	DG_256	45.46	30.97	12.63	3.15	24.77	131.35
	DG_512	45.71	32.72	12.14	3.63	24.94	129.69
	DG_1024	47.23	29.98	11.46	2.18	23.97	128.45
	DG_2048	48.59	31.06	10.32	2.08	22.87	129.49
	DL_128	49.71	30.5	15.31	6.47	29.06	142.45
	DL_256	49.36	29.92	15.57	6.93	33.49	142.35
	DL_512	49.6	29.53	16.12	6.96	33.44	141.69
	DL_1024	46.84	27.88	14.6	6.61	33.2	136.45
	DL_2048	45.79	26.25	15.74	5.68	32.66	139.49
Attention assisted captioning	IG_128	44.82	25.15	10.84	4.29	25.85	131.34
	IG_256	44.02	25.47	12.43	5.14	25.58	130.85
	IG_512	46.77	26.63	12.37	5.38	25.66	131.21
	IG_1024	44.84	26.41	10.88	4.5	27.23	132.04
	IG_2048	46.39	27.27	11.17	4.55	26.7	132.75
	IL_128	50.79	27.09	16.25	5.45	33.65	134.15
	IL_256	49.42	31.91	15.97	6.93	33.49	134.21
	IL_512	48.09	30.89	15.9	7.24	33.44	134.19
	IL_1024	47.73	31.72	15.3	6.64	33.2	134.32
	IL_2048	49.56	31.83	15.18	6.65	32.66	133.21
	DG_128	45.64	32.38	15.31	6.47	29.06	142.45
	DG_256	45.46	30.97	15.57	6.93	33.49	142.35
	DG_512	45.71	32.72	16.12	6.96	33.44	141.69
	DG_1024	47.23	29.98	14.6	6.61	33.2	136.45
	DG_2048	48.59	31.06	15.74	5.68	32.66	139.49
	DL_128	52.06	33.88	17	7.27	33.65	143.5
	DL_256	51.49	33.07	15.9	5.54	29.06	143.39
	DL_512	51.93	33.94	16.69	5.72	29.19	141.68
	DL_1024	49.3	31.08	14.7	4.94	28.9	137.82
	DL_2048	48.1	29.57	13.61	4.2	26.56	139.58

The best result in each column is highlighted in bold

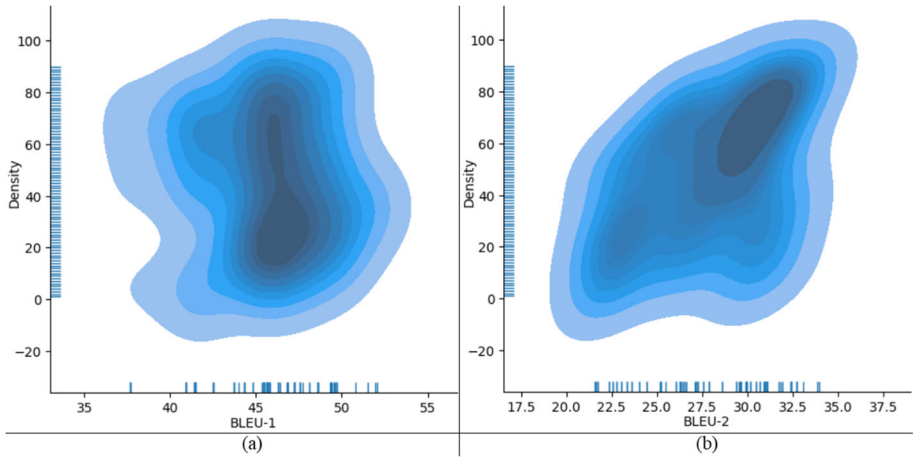


Fig. 10 Kernel density estimation plot of BLEU Scores: a) Distribution of BLEU-1 scores, b) Distribution of BLEU-2 scores

4.6.1 Quantitative evaluation

To demonstrate the effectiveness of our model, we use an offline evaluation method to compare it against various conventional and state-of-the-art baseline models. Specifically, our model is quantitatively compared to the ten most influential and popular IC models. These include the Neurtalk [20], the Show & Tell framework [33], the BiLSTM network [35], Show, Attend, & Tell model [38], and six different transformer-based cutting-edge approaches [6][18][12][23][40][36]. Our model outperforms Neurtalk [20] by implementing ensemble feature extraction, followed by dual-context aware multimodal fusion and attention-assisted caption decoding, resulting in superior performance across all evaluation metrics. Likewise, we quantitatively compare our model to Show and Tell IC developed by Vinyals et al. [33].

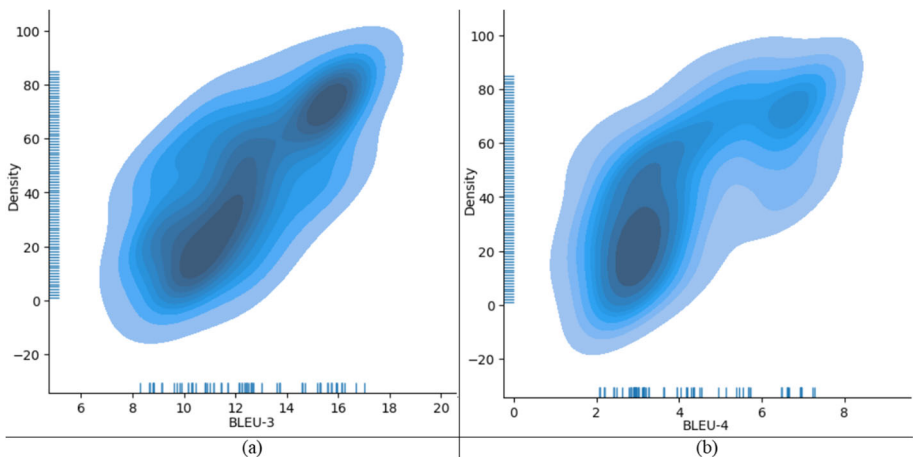


Fig. 11 Kernel density estimation plot of BLEU Scores: a) Distribution of BLEU-3 scores, b) Distribution of BLEU-4 scores

Table 7 Quantitative evaluation of our model with various state-of-the-art baselines on Flickr 8K [27] dataset (D: DenseNet-121, L: LSTM, I: Inception V3)

Authors(s)	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Karpathy & Fei Fei [20]	Neuraltalk	41.41	24.63	9.22	3.39	19.72	–
Vinyals et al. [33]	Show & Tell	46.6	–	–	4.6	21.06	–
Wang et al. [35]	Bi-LSTM	41.9	28.3	12.7	5.0	–	–
Xu et al. [38]	Show, Attend & Tell	45.7	26.4	10.34	2.36	23.33	133.51
Effendi et al. [6]	Grid Feature Captioner	–	–	–	4.62	14.1	123.2
Jiang et al. [18]	Multigate Attention	46.23	–	–	5.7	26.2	117.9
Haque et al. [12]	Positional & Geometric Semantics	42.5	23.0	13.6	3.2	24.5	–
Li et. al [23]	Hybrid Transformer	51.2	25.7	18.4	3.97	29.1	131.5
Yu et al. [40]	Multimodal Transformer	49.6	29.4	18.7	5.75	29.4	130.9
Wang et al. [36]	End-to-end Transformer	51.8	30.4	19.4	6.08	29.9	141.0
Neuraltalk+ (Our model)	DL_128	52.06	33.88	19.68	7.27	31.84	143.5
	DL_256	51.49	33.07	18.9	6.54	30.04	143.39
	DL_512	51.93	33.94	18.69	5.72	29.91	141.68
	IL_128	50.79	27.09	17.25	5.45	30.53	134.15

The best result in each column is highlighted in bold

Our model also outperforms theirs, leveraging spatial and visual features for context-aware multimodal embedding, resulting in more image-aligned captioning.

Furthermore, our model surpasses the results of the bidirectional decoding-based IC framework proposed by Wang et al. [35], even though we do not use bidirectional recurrent context while decoding. Similarly, we also consider the attention-based model, Show, Attend, and Tell, by Xu et al. [38], for quantitative comparison. While the Show, Attend, and Tell [38] model gazes at visual regions during each decoding step, its feature fusion lacks a dual-context approach, leading to the loss of significant contextual information. As a result, the generated captions lack particulars on object inter-associations and visual elements, resulting in generic captions. We also include top transformer-based IC paradigms for quantitative assessment in the IC pipeline. It aims to assess our model's performance with state-of-the-art IC frameworks. Despite not using a transformer-based language model, our model achieves comparable or slightly better results against these transformer-based models on both Flickr 8K and Flickr 30K datasets. It validates the effectiveness in image comprehension due to the structural and functional transformation in the captioning pipeline, thus rendering them more beneficial in the domain of vision aid, as stated in our research question. We execute and obtain the results of all these baseline models on Flickr 8K and Flickr 30K datasets using offline evaluation. Table 7 shows the quantitative comparison analysis of our model against these baselines on the Flickr 8K dataset.

Table 8 depicts the summary of our experiment validating the quantitative effectiveness of our model against the above-discussed baselines on a more robust and challenging Flickr 30K dataset. As can be seen, our model outperforms the baseline models on Flickr 30K dataset as well. Finally, we also note that we obtained these performances on a single predicted caption without using beam searches.

Table 8 Comparative evaluation of our model with various state-of-the-art baselines on Flickr 30K [10] dataset (D: DenseNet-121, L: LSTM, I: Inception V3)

Authors(s)	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Karpathy & Fei Fei [20]	Neuraltalk	57.3	36.9	24.0	15.7	–	–
Vinyals et al. [33]	Show & Tell	66	–	–	27.7	23.7	–
Wang et al. [35]	Bi-LSTM	58.9	39.3	25.9	17.1	–	–
Xu et al. [38]	Show, Attend & Tell	67.0	45.7	31.4	21.3	20.3	–
Effendi et al [6]	Grid Feature Captioner	–	–	–	37.78	17.40	129.34
Jiang et al. [18]	Multigate Attention	70.6	55.2	40.7	38.5	28.9	129.6
Haque et al. [12]	Positional & Geometric Semantics	65.4	51.6	38.29	34.21	18.7	–
Li et. al [23]	Hybrid Transformer	69.8	54.3	39.8	38.4	29.3	130.9
Yu et al. [40]	Multimodal Transformer	76.2	–	–	36.6	28.7	134.1
Wang et al. [36]	End-to-end Transformer	79.1	–	–	40.9	30.2	138.2
Neuraltalk+ (Our model)	DL_128	78.06	56.88	40.78	41.27	33.65	143.5
	DL_256	75.49	54.07	38.9	38.54	29.06	143.39
	DL_512	71.93	53.94	38.69	36.72	29.19	141.68
	IL_128	71.79	51.09	37.25	35.45	33.65	134.15

The best result in each column is highlighted in bold

4.6.2 Qualitative evaluation

Despite having found that quantitative results show conclusively that our model yields more efficient results than baseline models, we also performed a qualitative evaluation to gain an intuitive understanding of the experimental outcomes and to observe how the quality of information generated in their captions differs from ours. Figure 12 depicts the qualitative results of our experimentation. During the qualitative analysis, we compare our model with the predicted captions from four baseline models: Neuraltalk [20], Show and Tell [33], BiLSTM Network [35], and Show, Attend, and Tell [38].

We evaluate these models using a diverse set of images with varying visual complexity. In Fig. 12, we observe that our model produces captions that are more closely related to the visual content of the images. Specifically, our model demonstrates improved object detection compared to the baselines, which can be attributed to the incorporation of both spatial and visual features during the feature extraction process. Furthermore, the recognized characteristics are more reliable in our predictions than the current models, demonstrating the worth of our dual context-aware feature fusion strategy. In addition, whenever multiple elements are present in the image, our approach attempts to interpret interactions between them. This ability seems to be less efficient in the baselines, which might be due to a lack of attention and an effective cross-domain decoding method. Although the majority of our captions are correct and reliable, our model does produce some erroneous, as seen in some predicted samples such as "A man in a white shirt is standing in a stream", where our model wrongly recognizes the shirt. We believe that the quality of our captions may be improved further by leveraging a more robust dataset during model training, resulting in a more rich semantic embedding space.

	<p>Baselines</p> <p>B1: A man in a red shirt is standing on a bench with a man in a black shirt. B2: A woman is jumping close to the beach. B3: A man in black shirt standing. B4: A man working near a beach.</p>	<p>Predictions (beam search=3)</p> <ol style="list-style-type: none"> 1. A man standing near a tree. 2. A boy is standing near a tree. 3. Young boy in a tree.
	<p>Baselines</p> <p>B1: A man in a white shirt and a black dog is running through the snow. B2: A boy is walking through a beach. B3: Blue sky and a beach. B4: A black dog is running through water.</p>	<p>Predictions (beam search=3)</p> <ol style="list-style-type: none"> 1. Two dogs are playing with a beach. 2. A black dog is running on the beach. 3. The dogs are playing with a beach.
	<p>Baselines</p> <p>B1: A man in a blue shirt is standing on a rock overlooking a lake. B2: A man smiling next to a bus. B3: A mountain rock. B4: A man in a lake.</p>	<p>Predictions (beam search=3)</p> <ol style="list-style-type: none"> 1. A man in a white shirt is standing in a stream. 2. An adult in a white shirt is standing in a stream. 3. A man in a white shirt is standing in a stream.
	<p>Baselines</p> <p>B1: A man in a black shirt and a woman in a black shirt and a woman in a black shirt. B2: A boy and a small girl in white. B3: A boy and a woman in black and a woman. B4: A boy in black and white and white and white.</p>	<p>Predictions (beam search=3)</p> <ol style="list-style-type: none"> 1. An older man and a woman in a white shirt is sitting on a wooden bench 2. Two people are sitting on a wooden bench. 3. A women and a man are sitting on a wooden bench.
	<p>Baselines</p> <p>B1: A boy in a blue shirt is jumping into the air. B2: A man floats in the air. B3: A man is jumping in the mid air. B4: A boy is playing in green grass.</p>	<p>Predictions (beam search=3)</p> <ol style="list-style-type: none"> 1. A skateboarder is jumping on a ramp. 2. A boy is jumping on a ramp. 3. The skateboarder is jumping on a ramp.
	<p>Baselines</p> <p>B1: A group of people are standing on a beach. B2: A little girl and a boy lying in the water. B3: A boy is walking through the water. B4: A group of people in ocean.</p>	<p>Predictions (beam search=3)</p> <ol style="list-style-type: none"> 1. Two children are standing in the water. 2. A boy and a girl is standing in the water. 3. A girl in a body of water.

Fig. 12 Results of qualitative comparisons of our models’ predictions against baselines

5 Limitations and future scopes

Although our work surpasses numerous baseline models in terms of quantitative and qualitative benchmarks, certain areas have not been thoroughly investigated due to the intricate pipeline architecture, the complexity of training resources, and the wide range of possible extensions. The following are the areas that demand further exploration:

- i. **Lack of inherent personalized assistance:** The NT+ framework offers visual assistance through object detection, activity recognition, decoding inter-associations, real-time captioning, and visual comparisons. However, the model concentrates primarily on common contextual elements [26] prevalent in datasets like Flickr 8K and Flickr 30K. As a result, NT+ cannot offer tailored assistance like customized personal identification, 3D object recognition, and pedestrian detection inherently. Nonetheless, we’ve included support for training our framework on personalized domains, enabling it to feature personalized assistance.
- ii. **Investigating varied fusion methods:** Presently, our NT+ frameworks employ a dual-context feature fusion approach. While effective, evidenced by quantitative and qualitative analyses, alternative fusion methodologies merit consideration for future framework development. These include bidirectional feature fusion, transformer-boosted fusion, and graph neural network-based multimodal fusion. The investigation

- of these fusion techniques highlights a clear research gap in our work, necessitating additional attention for future exploration.
- iii. **Incorporating transformer-based decoders:** Our model engages an encoder-decoder architecture for IC, utilizing a CNN-based visual feature extractor alongside a word encoder. The decoder employs LSTM or GRU, yielding excellent results after extensive training. However, future research directions include integrating state-of-the-art deep learning techniques, specifically focusing on self-attention, cross-attention, and transformer decoders, to enhance our model further.
 - iv. **Multilingual captioning:** NT+, trained solely in English corpora, currently produces captions exclusively in English. However, enabling vision-assistive IC across diverse linguistic domains necessitates the crucial adoption of multilingual decoding. Hence, the augmentation of our captioning algorithms to cage training across diverse corpora emerges as a vital measure for future advancement.
 - v. **Knowledge-transfer approaches:** While we use transfer learning-based visual feature extraction approaches, further exploration is required to investigate the application of domain-based knowledge-transfer strategies. IC frameworks are generally very complex to train. Therefore, leveraging existing knowledge can mitigate the need for training novel models from scratch, leading to resource and time savings.
 - vi. **Dual-context attention:** Although we employ self-attention to calculate the significance of visual image regions during captioning, exploring the use of attention for semantic sequences, focusing on words with contextual information and ignoring the rest, presents another domain for further research in IC.
 - vii. **Pre-trained networks:** The increasing popularity of generative pre-trained embeddings offers a research opportunity in IC. These networks, trained on extensive image and text datasets, can be potentially employed in IC with fine-tuning. Given the limitations of training IC models from scratch to incorporate new visual or semantic information, the use of pre-trained networks becomes crucial for future advancements.
 - viii. **Reinforcement learning:** While supervised and unsupervised learning approaches are prevalent in multimodal translation problems, the introduction of reinforcement learning has shown promising results. Exploring the application of reinforcement learning in IC can potentially enhance the performance and capabilities of the models.
 - ix. **Domain-specific applications:** Our IC framework facilitates environmental comprehension for visually impaired individuals. However, in complex domains like traffic assistance, medical applications, and remote sensing, domain-specific training is necessary for broader utilization. The integration of IC frameworks into customized application domains is an emerging research area that needs further exploration.

6 Conclusions

This study introduces a novel, lightweight, autonomous, and efficient image captioning framework for vision assistance. Through a two-staged systematic methodology, we identify areas for improvement in existing pipelines by empirically analyzing their architectures, execution sequences, training algorithms, and semantic mappings. We address these concerns by incorporating ensemble feature extraction, dual-context aware feature fusion, and attention-assisted decoding into our framework. Additionally, we implement real-time captioning and visual similarity comparison capabilities. Our model is trained and evaluated on two benchmark datasets, demonstrating quantitative and qualitative improvements over vari-

ous baselines. Neurltalk+ produces semantically and syntactically efficient captions closely aligned with query images. It also supports captioning and similarity comparison for out-of-training samples. Furthermore, our model's scalability allows for potential expansions, including multilingual training, customized object identification, domain-specific training, and integration of cutting-edge data-driven strategies, which warrant future exploration.

Acknowledgements The authors extend sincere gratitude to the Editor and Reviewers for their insightful remarks and helpful opinions, which contributed to the enhancement of the work.

Research data policy and data availability The generated framework resources and training logs are available from the corresponding author upon reasonable request.

Declarations

Conflicts of interest All the authors declare that they do not have any conflict of interest.

References

1. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. [arXiv:1409.0473](https://arxiv.org/abs/1409.0473)
2. Banerjee S, Lavie A (2005) 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
3. Chen Z, Wang J, Ma A, Zhong Y (2022) Typeformer: Multiscale transformer with type controller for remote sensing image caption. *IEEE Geosci Remote Sens Lett* 19:1–5
4. Cheng Q, Zhou Y, Fu P, Xu Y, Zhang L (2021) A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. *IEEE J Sel Top Appl Earth Obs Remote Sens* 14:4284–4297
5. Chu F, Cao J, Shao Z, Pang Y (2022) Illumination-guided transformer-based network for multispectral pedestrian detection. In: CAAI international conference on artificial intelligence. Springer, pp 343–355
6. Effendi J, Sakti S, Nakamura S (2021) End-to-End Image-to-Speech Generation for Untranscribed Unknown Languages. *IEEE Access* 9:55144–55154
7. Fang H, Gupta S, Iandola F, Srivastava RK, Deng L, Dollár P, Gao J, He X, Mitchell M, Platt JC, et al. (2015) From captions to visual concepts and back. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 1473–1482
8. Fang Z, Wang J, Hu X, Liang L, Gan Z, Wang L, Yang Y, Liu Z (2022) Injecting semantic concepts into end-to-end image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 18009–18019
9. Farhadi A, Hejrati M, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth D (2010) Every picture tells a story: Generating sentences from images. In: Computer Vision—ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5–11, 2010, Proceedings, Part IV 11. Springer, pp 15–29
10. Gong Y, Wang L, Hodosh M, Hockenmaier J, Lazebnik S (2014) Improving image-sentence embeddings using large weakly annotated photo collections. In: European conference on computer vision. Springer, pp 529–545
11. Guo L, Liu J, Zhu X, Yao P, Lu S, Lu H (2020) Normalized and Geometry-Aware Self-Attention Network for Image Captioning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Seattle, WA, USA, pp 10324–10333
12. Haque AUI, Ghani S, Saeed M (2021) Image captioning with positional and geometrical semantics. *IEEE Access* 9:160917–160925
13. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 770–778
14. Herdade S, Kappeler A, Boakye K, Soares J (2019) Image captioning: Transforming objects into words. In: Conference and workshop on neural information processing systems
15. Hodosh M, Young P, Hockenmaier J (2013) Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *J Artif Intel Res* 47:853–899

16. MZakir Hossain, F Sohel, MF Shiratuddin, H Laga, (2019) A comprehensive survey of deep learning for image captioning. *ACM Comput Surv* 51(6)
17. Hossain MZ, Sohel F, Shiratuddin MF, Laga H, Bennamoun M (2021) Text to image synthesis for improved image captioning. *IEEE Access* 9:64918–64928
18. Jiang W, Li X, Hu H, Lu Q, Liu B (2021) Multi-Gate Attention Network for Image Captioning. *IEEE Access* 9:69700–69709
19. Jin J, Fu K, R Cui, F Sha, C Zhang (2015) Aligning where to see and what to tell: image caption with region-based attention and scene factorization. [arXiv:1506.06272](https://arxiv.org/abs/1506.06272)
20. Karpathy A, Fei-Fei L (2014) Deep visual-semantic alignments for generating image descriptions. *IEEE Trans Pattern Anal Mach Intell* 39:664–676
21. Kiros R, Salakhutdinov R, Zemel RS (2014) Multimodal neural language models. In: *International Conference on Machine Learning*
22. Kulkarni G, Premraj V, Ordonez V, Dhar S, Li S, Choi Y, Berg AC, Berg TL (2013) Babytalk: Understanding and generating simple image descriptions. *IEEE Trans Pattern Anal Mach Intell* 35(12):2891–2903
23. Li J, Yao P, Guo L, Zhang W (2019) Boosted Transformer for Image Captioning. *Appl Sci* 9(16):3260
24. Li W, Qu Z, Song H, Wang P, Xue B (2020) The traffic scene understanding and prediction based on image captioning. *IEEE Access* 9:1420–1427
25. Lian Z, Zhang Y, Li H, Wang R, Hu X (2023) Cross modification attention-based deliberation model for image captioning. *Appl Intell* 53(5):5910–5933
26. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: *European conference on computer vision*. Springer, pp 740–755
27. Ordonez V, Kulkarni G, Berg T (2011) Im2text: Describing images using 1 million captioned photographs. *Adv Neural Inf Process Sys* 24
28. Papineni K, Roukos S, Ward T, Zhu W-J (2002) 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
29. Shao Z, Han J, Debattista K, Pang Y (2023) Textual context-aware dense captioning with diverse words. *IEEE Trans Multimed*
30. Shao Z, Han J, Marnerides D, Debattista K (2022) Region-object relation-aware dense captioning via transformer. *IEEE Trans Neural Netw Learn Sys*
31. Sharma H, Padha D (2023) A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues. *Artif Intel Rev* 1–43
32. Vedantam R, Zitnick CL, Parikh D (2015) CIDEr: Consensus-based Image Description Evaluation. [arXiv:1411.5726](https://arxiv.org/abs/1411.5726)
33. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: A neural image caption generator. 2015 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Boston, MA, USA, pp 3156–3164
34. Wang C, Gu X (2022) Image captioning with adaptive incremental global context attention. *Appl Intel* 1–23
35. Wang C, Yang H, Bartz C, Meinel C (2016) Image captioning with deep bidirectional lstms. In: *Proceedings of the 24th ACM international conference on Multimedia*. pp 988–997
36. Wang Y, Xu J, Sun Y (2022) End-to-end transformer based model for image captioning. *Proc AAAI Conf Artif Intel* 36:2585–2594
37. Wei J, Li Z, Zhu J, Ma H (2023) Enhance understanding and reasoning ability for image captioning. *Appl Intell* 53(3):2706–2722
38. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: *International conference on machine learning*. PMLR, pp 2048–2057
39. Yang Y, Teo C, Daumé III H, Aloimonos Y (2011) Corpus-guided sentence generation of natural images. In: *Proceedings of the 2011 conference on empirical methods in natural language processing*. pp 444–454
40. Yu J, Li J, Yu Z, Huang Q (2019) Multimodal transformer with multi-view visual representation for image captioning. *IEEE Trans Circuits Syst Video Technol* 30(12):4467–4480
41. Zhang J, Fang Z, Wang Z (2022) Multi-feature fusion enhanced transformer with multi-layer fused decoding for image captioning. *Appl Intel* 1–17
42. Zhao W, Wu X, Luo J (2020) Cross-domain image captioning via cross-modal retrieval and model adaptation. *IEEE Trans Image Process* 30:1180–1192
43. Zhou D, Yang J, Bao R (2021) Collaborative strategy network for spatial attention image captioning. *Appl Intel* 1–16:100

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.