IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 May 2025, accepted 14 June 2025, date of publication 24 June 2025, date of current version 17 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3582655

RESEARCH ARTICLE

News Image Captioning via Separate Attention on Entity Categories

SONALI AJANKAR[1,2], (Member, IEEE), AND TANIMA DUTTA[1], (Senior Member, IEEE)
[1]Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, Uttar Pradesh 221005, India
[2]Department of Master of Computer Application, Veermata Jijabai Technological Institute (VJTI), Mumbai, Maharashtra 400019, India

Corresponding author: Sonali Ajankar (sonaliajankar.rs.cse19@itbhu.ac.in)

This work was supported by the Science and Engineering Research Board (SERB) Mathematical Research Impact Centric Support (MATRICS) under Grant MTR/2021/000604.

ABSTRACT News image captioning involves generating descriptive and informative captions for news images by utilizing news article context. This task aims to capture detailed information, including multiple types of named entities like person, organization, location, events etc. However, identifying named entities from an image is a challenging task; to address this, we propose a novel approach of categorizing the key named entities, into person entities and geoOrg entities, and providing distinct attention to these categories, which ensure the focused extraction of relevant information from the image. Despite this approach, a single news image falls short of providing comprehensive background details, leading to a lack of useful context. Nevertheless, it is possible that the missing context in one image is present in another image of the same article. To address this, we propose to incorporate the nearby image as an addon input, as structural proximity implies contextual relevance. This multimodal cue facilitates the accumulation of contextualized features that effectively capture contextually rich information from the article. Experimental results demonstrate the effectiveness of our proposed approach in the GoodNews, NYTimes800k and DM800K datasets for news image captioning, achieving an improvement of 0.4 BLEU-4 score over the state-of-the-art on the DM800K dataset.

INDEX TERMS Deep learning, news image captioning, named entities, integrated vision-language, social network.

I. INTRODUCTION

Image captioning [1], [2], [3], [4], [6], [7], [8], [9], [10], is the task of narrating the visual content of an image in natural language. Further, the News Image Captioning (NIC) [11], [12], [13], [14] is a sub-task of image captioning, which generates natural language sentences for the news image with the help of the associated article. NIC is considered to be challenging due to image understanding and the generation of the entity-rich captions. NIC aims to provide more contextually relevant captions, offering a deeper understanding of the image's content and its relation to the news story. In our modern world, where communication of news articles relies on smoothly combining written stories with visual hints, news image captioning takes center stage, allowing us to provide insightful descriptions of news-related pictures. This is particularly crucial within our social systems, where news images convey information and emotions, fostering connections. As we delve into news image captioning, we discover its role in enhancing comprehension and communication within our news-savvy society. The explosive growth of online news content and its sharing on online platforms has substantially increased the importance of the automatic generation of NIC. The news media sources are essential platforms for disseminating news. News articles containing multiple named entities, such as celebrity names or political figures, tend to generate more attention and discussion on social networks.

In news articles, entities play a pivotal role in generating entity-aware captions for news images. Each category provides distinct semantic information that contributes to a deeper understanding of visual and textual content. Towards

The associate editor coordinating the review of this manuscript and approving it for publication was Szidonia Lefkovits.

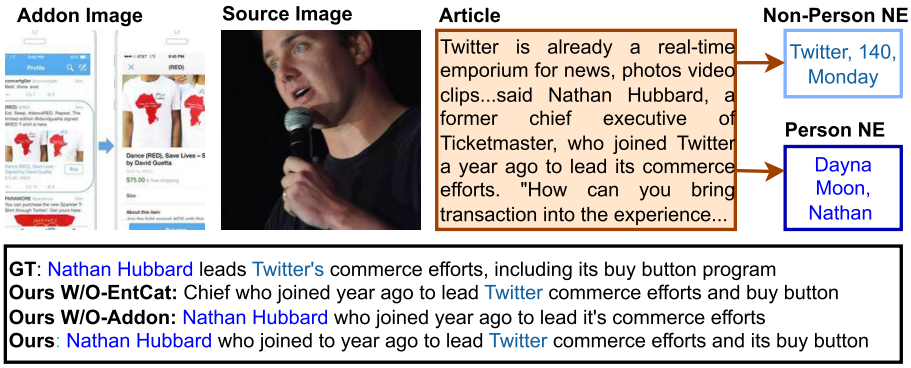

FIGURE 1. A news article, along with its entity categories, addon and source images, is depicted in the figure. The dark blue and light blue colors are used to depict person entities and other entities, respectively. Where, 'GT', 'W/O', and 'EntCat' represents, ground truth, without and entity categorization, respectively.

this, we proposed categorizing entities and providing distinct focus on two categories of entities *i.e.* person and geoOrg entities. our proposed method seamlessly integrates facial features extracted from the source image. This integration enables focused attention on person entities, Figure 1 illustrated that without entity categorization (W/O-EntCat) the proposed architecture fails to identify the correct person entity ''Nathan Hubbard'' in the generated caption. This highlights the importance of providing explicit attention on the each entity category. By mimicking the human cognitive process of prioritizing and categorizing relevant information, this strategy captures fine-grained crucial features for producing captions.

In general, a news article consists of a piece of writing about a specific news topic, accompanied by a few images representing key events related to the story; such two nearby images of an article are shown in Figure 1. A challenge arises when trying to grasp the complete background of an image using a single photograph, as it often captures a specific viewpoint without conveying the full story. For instance, Figure 1 illustrates a pair of images associated with an article. Our objective is to generate an appropriate caption for the source image of Figure 1. However, this source image alone, featuring an individual holding a microphone, lacks the crucial background context essential for predicting the caption accurately. To address the aforementioned limitation of insufficient visual context, we propose incorporating an additional nearby image from the same article. This nearby image is selected as the preceding or succeeding image of the source image. As structural proximity typically correlates with contextual relevance (e.g., images from the same event). In this case, addon image of Figure 1 serves as supplementary visual information. This addition aids in identifying and narrowing down multiple relevant concepts from the article, emphasizing coarse-grained features. This approach mirrors the human cognition process, leveraging prior knowledge to comprehend and interpret scenes.

The prior news image captioning literatures [11], [15], and [16] prefers to choose two-stage template-based approaches for generating the captions for news images. The generated caption in the first stage has an entity-type token and omits the inclusion of named entities, constituting a core elements of news caption. Subsequently, in the second stage, various methods were employed by different literature to replace these token types with the entities extracted from the article. Like, techniques in [11] and [15] aligned the produced template caption with the article data, while [16] incorporated external web-based information. To encode the article's named entities [17] introduced a Named Entity Embedder utilizing Wikipedia2Vec [18]. Moreover, the prioritization of person and geoOrg entities (i.e., GPE, ORG, LOC, EVENT) in our work is motivated by visual grounding through facial, landmark, and scene cues and linguistic relevance. These entities exhibit a stronger visual grounding facilitating effective cross-modal alignment via our entity-specific attention mechanisms. From a linguistic perspective, news captions typically emphasize ''who'', and ''where'' information, aligning well with person and geoOrg categories. Therefore, making sure they are accurate is important in writing quality captions. Finding the named entities and important context from a long article is a crucial task. To handle this, our proposed architecture works by attending to the person and geoOrg entities separately, and by integrating the addon image to gather additional relevant information from articles. This approach ultimately improves the caption generation process.

Our framework's key contributions are outlined as follows:

- The presence of named entities makes the news image captioning task more challenging. To tackle this, we select the dominant entity types and categorize them into two groups: person (e.g., individuals or public figures) and geoOrg (e.g., organizations, geopolitical entities, or locations, events) entities. Further, we introduce an innovative method to distinctly focus on these two categories. Specifically we introduce Cross-attentional Entities' Features (CEF) module which applies specialized attention to facial features for person entities and general image features for geoOrg entities.

This approach leads to the identification of emphasized entities' features.

- Typically, a news article covers a single narrative or event supported by set of images. Sometimes, source images only cover limited details like unclear background information such as place, event, etc., that can be present in adjacent images of the same article. To enhance the capture of contextual information from the article, we propose to incorporate a surrounding image as the addon input. This addon image serves to supplement missing visual context and provides complementary information that may not be present in the primary source image.
- Furthermore, we conduct substantial experiments on two widely used benchmark datasets i.e. GoodNews [11] and NYTimes800K [12]. Further, to test the generalizability of our framework we perform preliminary experiments on more recent DM800K [19] news image captioning datasets. The outcomes demonstrate that our suggested framework showcases performance advancements, thereby validating the robustness and adaptability of the proposed method across diverse news image captioning scenarios.

The subsequent sections of the paper are structured as follows. A brief review of image captioning, integrated vision-language framework and news image captioning literature is discussed in the next section II. Section III details the proposed architecture, and the subsequent exploration of experimental results on this news image captioning task is presented in Section IV. Finally, we conclude the paper and provide future work in Section V.

II. RELATED WORK

A. IMAGE CAPTIONING

In recent years, multi-modal deep learning applications [20], [21], [22], [23], [24] have accumulated significant attention, with notable advancements in various fields. One prominent area is image captioning, which aims to generate descriptive text for images automatically. This task has seen considerable progress, leading to the development of various methodologies and approaches. Initially, methods combining Convolutional Neural Networks and Recurrent Neural Networks [25], [26], [27] laid the groundwork for subsequent advancements. Further studies have explored different approaches [3], [10], [28], [29], [30], [31], [32] incorporating variations in attention mechanisms, visual and textual feature extraction techniques, language models, standard / pre-training / domain-specific datasets, etc.

The emergence of transformer [33] based models, known for their success in Natural Language Processing (NLP) tasks, has brought about a significant shift in image captioning methodologies. In a recent transformer-based approach, Luo et al. [34] introduced the dual-level cooperative transformer for addressing the supplementary nature of grid and region features. This model aimed to enhance semantic comprehension by leveraging both levels of features. Additionally, Ma and colleagues [35] proposed a transformer-based captioning framework incorporating channel-wise and spatial attention mechanisms to enhance semantic understanding. Ji et al. [36] proposes an approach called metric-oriented focal mechanism to address the challenge of varying difficulty levels in image captioning. BLIP-2 [37] integrates ViT-G [38] a frozen image encoder with a Flan-T5 [39] (or OPT [40]) large language model through a lightweight querying transformer resulting in efficient zero-shot image captioning and VQA. It excels at transferring visual understanding to language tasks with minimal supervision. Moreover, semantic-enhanced dual attention transformer model of [41] introduces a locality-sensitive transformer framework for generating captions for images and enhancing local visual modeling through unique locality-sensitive fusion and attention techniques. Recent study [42] have demonstrated enhancements in image captioning task through the integration of text-level chain-of-thought reasoning and exploring in-context learning. Furthermore, our proposed framework leverages transformers, harnessing their multi-head self-attention mechanism to produce contextual captions for news images.

B. INTEGRATED VISION-LANGUAGE FRAMEWORK

The transformer architecture has emerged as the cornerstone of vision-language integration due to its scalability and flexibility in processing both sequential and non-sequential data. Transformer has Initially popularized in natural language processing [33], [43] and computer vision [44], [45]. The Vision Transformer (ViT) [44] and its subsequent improvements have optimized computational efficiency [45], [46] and architectural design [47], [48] making them suitable for integrating visual and textual data. These advancements has become a major area of research, with the goal of bridging the gap between these modalities.

One of the influential techniques in the vision-language alignment work has been contrastive learning. Models like CLIP [49] and its successors [50], [51], [52], [53] have demonstrated impressive performa Contrastive learning frameworks have proven pivotal by linking images with their corresponding textual descriptions through dual encoders, which project these modalities into a shared embedding space, as demonstrated in various studies [49], [54]. Although highly effective in capturing high-level associations and enabling zero-shot learning, these methods often struggle with tasks that require fine-grained compositional reasoning [55]. For bridging the gaps in vision-language alignment, our architecture utilizes CLIP for visual feature extraction. OFA (One For All) [56] is a unified vision-language model that reformulates modalities like vision, language, cross-modal into a sequence-to-sequence task using a single transformer framework. It supports diverse tasks like captioning, VQA, and grounding via instruction tuning and shared token space.

C. NEWS IMAGE CAPTIONING

News image captioning is a tricky task since it handles the extensive coverage of entities in the generated caption. For

the news image captioning task, Biten et al. [11] proposed a two-stage template-based LSTM framework, which produces news image caption with a placeholder for the entities and then used the context or sentence attention to replace the entity tags from the generated sentence. They provide hidden state attention to attend to the image and article features. Yang et al. [57] incorporates transformer into the captioning process with the object recognition ImageNet model [58] and scene recognition model [59]. Further, Tran et al. [12] suggested an end-to-end transformer model that encoded the article sentences by using weighted RoBERTa. Then, a decoder takes various input embeddings to generate a caption. Jing et al. [60] proposed a sentence correlation analysis method to locate the semantically similar context for the accompanying image. They used real-world knowledge to leverage the semantic relevance of the named entities. Yumeng et al. [61] proposed template-based image-text matching using image captions, which first generate the template caption and then fill the placeholder by choosing named entities related to the knowledge graphs obtained from images from the news. JoGANIC [17] first broke the captions into templates of who, when, where, context, and misc. Further, using the Wikipedia knowledge base, train the neural entity predictor approach to predict named entities from the article. Then, the caption is generated by the transformer decoder's distinct prediction heads using template guidance. In news image captioning models [11], [12], [17], the encoded images and articles work separately, neglecting their association. Moreover, visual news captioner [13] introduced combined multi-modal learning and incorporated a visual selective layer and multi-head attention on attention. The author also presented a multi-head pointer-generating approach for selecting words from the source input and the name entity list. Zhao et al. [62] proposed the cross-modal entity matching method using a multi-modal knowledge base from Wikipedia to minimize the image and text embedding representation. NewsMEP [14] proposed to use CLIP for visual features and incorporated a prompting mechanism for generating captions for news images. Zhang et al. [63] proposed to improve caption generation by incorporating discourse information and enhancing style coherence in the generated captions. Zhou et al. [64] proposed effectively choosing context information related to the image to generate more accurate and contextually appropriate captions. Motivated by the aforementioned discussion, this paper introduces unique news image captioning architecture aimed at effectively leveraging the additional visual features to capture more relevant details from the article.

Kalarani et al. [65] proposed to used a unified vision-language model based on the OFA architecture on three tasks i.e. keyword extraction, contextual visual entailment, and news image captioning. Zhang and Wan [66] study explores that recent models achieve competitive performance even without vision features. Notably, visual features assist in generating specific entities when textual context provides sufficient information. Qu et al. [67] introduce a face-naming module to refine name embeddings, and a sentence retrieval strategy using CLIP is proposed to link image content to relevant segments in the article. Chen et al. [19] proposed a novel task of multi-image captioning and introduced a new dataset DM800K. Xu et al. [68] develops a rule-driven approach that incorporates news-aware semantic rules to predict captions. To tackle named entity challenges in news image captioning, we categorize them into person and geoOrg entities and leverage facial and image features through the Cross-attentional Entities' Features (CEF) module. Additionally, our framework integrates nearby images to enrich contextual information from the article and generate entity-rich news image captions.

III. PROPOSED WORK

This section covers a proposed model, depicted in Figure 2. In this architecture, we propose to augment the addon image and provide separate attention to the two categories of named entities. We utilize a prevalent encoder-decoder architecture for the task of news image caption generation. In the captioning encoder, we initially encode both visual and textual components. Subsequently, we enhance article features by incorporating information from the source image and addon image. Additionally, we improve entities' features by leveraging facial features and source image features. The captioning decoder produces the caption sequence word by word with the help of gated units to filter the irrelevant features.

A. CAPTIONING ENCODER

The captioning encoder examines the inputs and transforms the data into context vectors.

1) VISUAL AND TEXTUAL FEATURE REPRESENTATIONS

This module describes the feature extraction from multi-modal entities such as images, faces, articles and named entities. To produce visual features, we have utilized the CLIP encoding module proposed by Radford et al. [49]). We obtain the visual features from CLIP and project them using a two-layer MLP, resulting in a 1024-dimensional feature vector for the source image, denoted as S^I. Similarly, we obtain the additional sequence image (addon image) features, denoted as S^O. The image features generated by CLIP improve its ability to associate with text representations and help to generalize to new classes or concepts.

For detecting the bounding boxes of the faces from the image, we have utilized the pre-trained MTCNN [69] model. Subsequently, we select up to four faces of high confidence. Additionally, we forwarded these detected faces bounding boxes to the pre-trained FaceNet [70][1] framework and to obtain the feature set for the image's faces as $F = \{f_i \mid 0 \leq i \leq N^F\}$, the embedding dimension of each face is 512. N^F denotes the number of faces.

[1] https://github.com/timesler/facenet-pytorch

To obtain a richer representation of the article that provides contextual embeddings for text, we used the RoBERTa [71] language representation model. RoBERTa encoder generates the features set denoted as $A = \{a_i \mid 1 \leq i \leq N^A\}$, where N^A denotes the input length. The embedding dimension of each token is 1024. Byte-pair encoding, as used in RoBERTa, efficiently encodes linguistically and named entity-rich articles, overcoming the limitations of pre-trained word embeddings like word2vec [72] and GloVe [73], which struggle to encode out-of-vocabulary words in such contexts.

We used a spaCy [74] named entity recognizer for retrieving named entities from the article.[2] The article's named entities' features set is computed using RoBERTa, as shown above. We separate the person entities having entity type PER, which are denoted as set $E^P = \{e_i^P \mid 0 \leq i \leq N^{E^P}\}$, where N^{E^P} represents the number of the person entities in an article. Further, we consolidated dominant entities of types like Geopolitical entities (GPE), organizations (ORG), locations (LOC), and event (EVNT) in the news article, leading to the formation of the geoOrg entity set, denoted as $E^{GO} = \{e_i^{GO} \mid 0 \leq i \leq N^{E^{GO}}\}$, where $N^{E^{GO}}$ represents the number of geoOrg entities in an article.

2) MULTI-HEAD SELF-ATTENTION (MSA)

The standard transformer multi-head attention, as described in [33], is defined as the scale dot-product attention, supplied queries Q, keys K and values V, which is shown below:

$$\text{Attention}(Q, K, V) = \delta(\frac{QK^\top}{\sqrt{d_k}})V, \tag{1}$$

where d_k is the scaling factor. $\delta(\cdot)$ is an softmax activation fucntion. Extending the attention mechanism of equation (1) into several heads employed for diverse attention distributions. Multi-head attention is defined as follows:

$$f_{\text{MSA}}(Q, K, V) = \text{Concat}(H_1, \ldots, H_{N^h})M_h, \tag{2}$$

$$H_i = \text{Attention}(QM_{qi}, KM_{ki}, VM_{vi}), \tag{3}$$

where $\{M_h, M_{qi}, M_{ki}, M_{vi}\}$ are learnable parameters. H_i represents $i-th$ attention head. N^h are the total number of attention heads and $N^h = 8$.

3) SELF-ATTENTIONAL ARTICLE FEATURES (SAF)

The textual information comprises high-level semantic data, making it more discriminative compared to images framed with pixels. This module describes the generation of enhanced textual features. To that end, we utilized the f_{MSA} of equation (2) to extract long-range semantic relationships from the article context as follows:

$$\hat{a}_i = f_{\text{MSA}}(a_i, A, A). \tag{4}$$

Likewise, we obtained geoOrg entities' features denoted as $\hat{e}_i^{GO}$ by passing e_i^{GO} and E^{GO} in the above equation. The multi-head configuration of MSA is helpful in understanding complex relationships.

[2] github.com/pemagrg1/Natural-Language-Processing-NLP-using-Spacy

4) CROSS-ATTENTIONAL ARTICLE FEATURES (CAF)

In real-world situations, news articles frequently cover various intricate sequences of events on a particular topic. While only a small fraction aligns with the image. Thus, the proposed incorporation of addon image features helps to gather the more relevant background context of that topic from the article. First, we utilized the source image features as a query to enhance article features of equation (4) as depicted below:

$$\check{a}_i^I = f_{\text{MSA}}(s_i^I, \hat{A}, \hat{A}), \tag{5}$$

where $\check{a}_i^I$ denotes the enhanced article features obtained by passing source image features as query and article features obtained in SAF module as key and value to MSA module, $\hat{A} = \{\hat{a}_i\}$ and $s_i^I \in S^I$ Then, we utilize the features of an addon image as a query in the following manner:

$$\check{a}_i^O = f_{\text{MSA}}(s_i^O, \hat{A}, \hat{A}), \tag{6}$$

where $s_i^O \in S^O$. Subsequently, for projecting the spatial article features, we used linear layer followed by Gaussian Error Linear Unit (GELU) [75] activation function and layer normalization, which is illustrated below:

$$\check{\check{a}} = \text{LN}(\text{GELU}([\check{a}^I; \check{a}^O]M_a + b_a), \tag{7}$$

where $\{M_a, b_a\}$ are learnable parameters. LN$(\cdot)$ represents layer normalization. [ ] represents concatenation. $\check{\check{a}}$ represents an enhanced article features by the encoder, which will be passed to the decoder further. GELU is effective in capturing complex patterns.

5) CROSS-ATTENTIONAL ENTITIES' FEATURES (CEF)

The importance of named entities in generating informative captions is significant, and its predictability is often challenging. Notably, person entities are the most prevalent type of entities present in captions. Inspired by this observation, we introduced the CEF module. In this module, we begin by concatenating the facial features f_1 to f_{N^F} of an image followed by a linear layer as shown below:

$$F = [f_1; \ldots; f_{N^F}]M_f + b_f, \tag{8}$$

where $\{M_f, b_f\}$ are learnable parameters. The weights assigned to facial features have been proven to be effective in learning the importance of faces.

Further, we utilized the facial features of the source image to align the transformed person entities' features as given below:

$$\check{e}_i^{FP} = FC(f_{\text{MSA}}(f_i, \acute{E}^P, \acute{E}^P)), \tag{9}$$

where $f_i \in F$ and $\acute{E}^P$ is obtained by applying a linear layer followed by GELU to person entities' features (E^P) for matching dimensions. Moreover, $\hat{E}^{GO} = \{\hat{e}_i^{GO}\}$ undergoes transformation via a linear layer and subsequent GELU activation to yield $\acute{E}^{GO}$. Then we utilized the source image

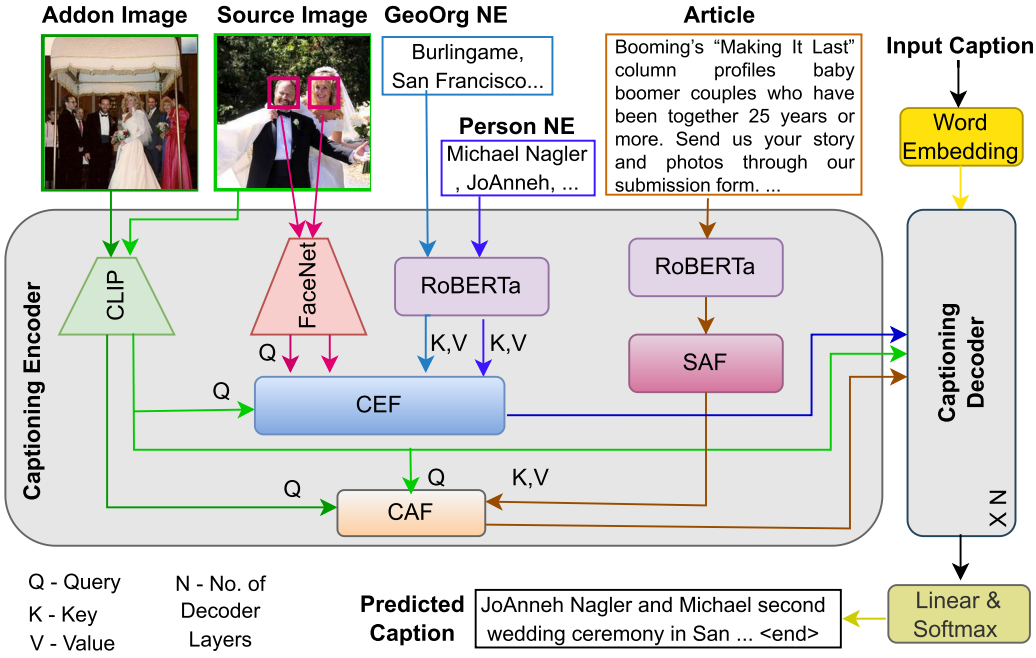

FIGURE 2. Outline of the proposed architecture. In the encoder layer, we employed FaceNet to encode faces, CLIP to encode images, and RoBERTa to encode text. The self-attentional Article Features (SAF) module is used to enhance textual features. The Cross-attentional Article Features (CAF) and Cross-attentional Entities' Features (CEF) components play a crucial role in determining the related context from the article and relevant named entities' features. The architecture consists of N captioning Decoder Layers.

features as a query to the enhanced features of geoOrg entities as shown below:

$$\check{e}_i^{GO} = FC(f_{\text{MSA}}(s_i^I, \acute{E}^{GO}, \acute{E}^{GO})). \tag{10}$$

This attention helps to align other named entities, such as GPE, ORG, LOC and EVNT, with visual details of the source image. After providing the $\check{e}_i^{FP}$ and $\check{e}_i^{GO}$ features to equation (7), we obtained attended entities' features relevant to the source image denoted as $\check{\check{e}}_i$. We are considering only source image features to attend to entities' features. The proposed distinct attention mechanisms for person and geoOrg entities are designed not only for fine-grained entity alignment but also to support contextually grounded caption generation.

B. CAPTIONING DECODER

The transformer decoder generates the target representation $Y = \{y_t\}_{t=1}^{T}$, where y_t represents the token produced at $t-th$ time step, T is the length of the caption. The input matrix representation to the captioning decoder at $t-th$ time step is shown below:

$$W_{<t} = [w_0, \ldots, w_{t-1}], \tag{11}$$

where $w_t \in R^{d\times 1}$ is the vector representation of the $t-th$ word, obtained by combining word embedding and positional encoding. w_0 is the feature vector of the token that represents the beginning of a sentence.

For each layer in the decoder, the first sub-layer of each layer is Masked Multi-head Self-Attention (MaskMSA), which aims to capture target-side relationships with previously predicted words as follows:

$$c_t^l = \text{LN}(w_t^{l-1} + f_{\text{MaskMA}}(w_t^{l-1}, W_{<t}^{l-1}, W_{<t}^{l-1})), \tag{12}$$

where $W_{<t}^{l-1}$ is the output of the $(l-1)^{th}$ layer, $0 \le l < N$. N denotes the number of decoder layers.

Further, the decoder has two more sub-layers to look for translation-relevant source semantics in order to fill the gap between the source and target. The second sub-layer strengthens the decoder by using the encoder's output as a key and value. The output c_t^l of MaskMSA is supplied to the MHEDMA (explained in the subsequent point within this subsection) as follows:

$$\check{\check{z}}_t^l = f_{\text{MHEDMA}}(c_t^l, S^I, \check{\check{A}}, \check{\check{E}}), \tag{13}$$

where $\check{\check{z}}_t^l$ possesses information based on the importance of modalities and the contextual encoder output $\check{\check{A}} = \{\check{\check{a}}_i\}$, $\check{\check{E}} = \{\check{\check{e}}_i\}$ and source image embedding S^I is passed to all of the decoder layer's MHEDMA sub-layer. After applying the third sub-layer, *i.e.*, Feedforward Neural Network (FNN), the $l-th$ layer's output is calculated as follows:

$$x_t^l = \text{FFN}(\check{\check{z}}_t^l), \tag{14}$$

where FFN$(\cdot)$ is composed of two feed-forward layer with GELU activation function. The output x_t^l is forwarded to the

next captioning decoder layer. Finally, $N - th$ captioning decoder layer output is passed into the linear layer followed by a softmax activation to deduce the word y_t.

1) MULTI-HEAD ENCODER-DECODER MULTI-MODAL ATTENTION (MHEDMA)

This layer associates its previously generated word vectors over all contextualized encoder outputs in order to condition the probability distribution of the following target vectors:

$$\bar{s}_t^I = f_{\text{MSA}}(c_t, S^I, S^I). \tag{15}$$

Likewise, we obtained $\bar{a}_t$ and $\bar{e}_t$ by passing $\check{A}$ and $\check{E}$ to the aforementioned equation, respectively. Further, the Gated Unit (GU) filters the features that have a low correlation for predicting the next word, as follows:

$$\tilde{s}_t^I = f_{\text{GU}}^I(c_t, \bar{s}_t^I), \tag{16}$$

$$= ([c_t + \bar{s}_t^I]M_{s1} + b_{s1}) \odot tanh([c_t + \bar{s}_t^I]M_{s2} + b_{s2}), \tag{17}$$

where $\{M_{s1}, M_{s2}, b_{s1}, b_{s2}\}$ are learnable parameters. $\odot$ denotes hadamard product. *tanh* is activation function. In a similar manner, we obtained $\tilde{a}_t$ and $\tilde{e}_t$ by passing $\bar{a}_t$ and $\bar{e}_t$ to the aforementioned equation, respectively. Subsequently, we aggregated these filtered features. This fusion was strengthened by integrating a residual connection mechanism, accompanied by subsequent layer normalization. Following this, an affine transformation was applied, followed by GELU activation for non-linearity. Continuing to enhance the process, we introduced another residual connection paired with layer normalization. This progressive series of steps ultimately led to the enriched output, denoted as $\check{\check{z}}_t$:

$$u_t = \tilde{s}_t^I + \tilde{a}_t + \tilde{e}_t, \tag{18}$$

$$z_t = \text{LN}(u_t M_u + c_{<t} M_c), \tag{19}$$

$$\check{z}_t = \text{GELU}(M_{\check{z}}(z_t M_z + b_z) + b_{\check{z}}), \tag{20}$$

$$\check{\check{z}}_t = \text{LN}(\check{z}_t + z_t), \tag{21}$$

where $\{M_u, M_c, M_{\check{z}}, M_z\, b_z, b_{\check{z}}\}$ are learnable parameters.

C. MODEL TRAINING

Following previous work, the news image captioning is defined as the task of producing caption $Y = \{y_1, y_2, \ldots, y_T\}$, given an article A and a source image S^I from that article. First, the proposed multi-modal neural news image captioning architecture is trained by minimizing the negative log-likelihood of a captioning model that is additionally conditioned on the addon image. The loss function for the proposed news image captioning architecture is shown as follows:

$$\pounds(\theta) = -\sum_t log(P(y_t|y_{<t}, S^I, A, S^O)), \tag{22}$$

where $P(y_t)$ is the predicted probability for each target word, which is dependent upon S^I, A, S^O, and ground-truth tokens up to t. Then we fine-tune our model with the CIDEr-D [10] score as a reward. The baseline reward is calculated from a mean of top-k words of beam search rather than a greedy approach.

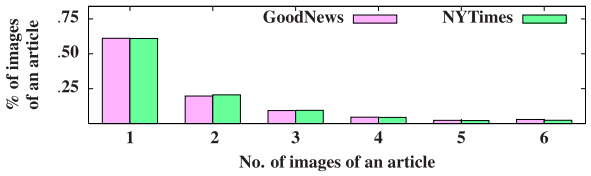

FIGURE 3. Training data statistics shows the percentage of the number of images present in an article. The pink and green bar represents the statistics for the GoodNews and NYTimes800K respectively.

IV. EXPERIMENTAL RESULTS

A. DATASETS AND EVALUATION METRICS

In our experimental setup, we utilized three datasets available for this task: GoodNews [11], NYTimes800k [12] and DM800K [19]. These datasets contain images, captions, and news DM800K is the largest dataset among the three datasets. The ground-truth captions of news images are written by skilled journalists, which results in vocabulary-rich articles. These datasets annotate each image with one ground-truth caption that is an average of eighteen words long in GoodNews and NYTimes800k datasets and twenty two words long in DM800k dataset. On the other hand, the image caption benchmark MSCOCO [76] dataset possesses five ground-truth descriptions for each image, with an average caption length of 12. For the experiment with the GoodNews dataset, the train split, the val split, and the test split comprises 410K, $21K$, and $21K$ examples, respectively. Further, the NYTimes800k dataset comprises the train split, the val split, and the test split of $720K$, $21K$, and $21K$, respectively. The DM800K dataset comprises the train split, the val split, and the test split of $880K$, $60K$, and $40K$, respectively. The percentage of the number of images present in an article of the GoodNews and NYTimes datasets is shown in Figure 3. The plot illustrates that around 40% of articles have more than one image for GoodNews and NYTimes800K datasets. On an average GoodNEws, NYTimes800k and DM800k datasets have 1.8, 1.78 and 4.45 number of images per articles, respectively, and 96%, 97% and 70% percentage of captions with named entities, respectively. Therefore, our proposed architecture with the combination of an addon image and separate entity categorization has shown good results.

For evaluating the produced news picture caption, we utilized the word-based metrics BLEU-4 [77], METEOR [78], ROUGE-L [79], CIDEr [80]. After lower-casing and removing punctuation from captions, we employed the COCO caption evaluation toolkit3.[3] These metrics assess the degree of the word overlap between created and actual captions. As the task targets the generation of entity-rich captions, our evaluation focuses on named entities within the generated

[3]https://github.com/tylin/coco-caption/

captions. Thus, the precision and recall metrics of the named entities are then calculated using matching counts between named entities in the generated captions and in the ground truth captions.

B. IMPLEMENTATION DETAILS

In our work, we configure the embedding dimension to 1024. We used 16 heads (H) and ten decoder layers (N). Furthermore, we used Adam [81] optimizer with the following parameters: $epsilon = 10^{-6}$, $beta_1 = 0.9$ and $beta_2 = 0.99$. During the initial 5% of training iterations, we warm up the learning rate to 10^{-4}, followed by a gradual linear decrease. After each attention and linear layer, dropout is applied with a keep probability of 0.9. Our approach consistently generates captions with an average length of 15 words. Subsequently, during the optimization phase focusing on CIDEr-D improvement, a consistent learning rate 5×10^{-6} is maintained. We set the batch size to 16. Moreover, we adopt a beam size of five.

To obtain the visual features from news images we use the ViT-B/32[4] backbone of CLIP. A two-layer MLP with GELU [75] activation projects image embeddings into the 1024-dimensional space. The pre-trained RoBERTa model is used to generate the embedding for each word. Token embeddings from its final transformer layer are passed through a learnable linear projection to align with the decoder's 1024-dimensional space. The context for a token embedding is selected from the news headline, first paragraph, and sentences surrounding the source image of the article. Further, in the context, we append the named entities of the captions that are not present in the article. For three datasets, we input the context of a maximum of 512 tokens. For the addon image, we use the preceding image of the article, and in the absence of preceding image, we select the succeeding image as a nearby image. When no addon image exists, the presence_flag will be set to 0, which encourages the model to focus on visual details from the source image and context.

For improving face detection success rates on low-resolution images, we have applied the contrast-limited adaptive histogram equalization technique [82], to enhance image clarity and contrast. When facial regions are unavailable, a zero tensor is passed to the CEF module to indicate the absence of face-related visual information. This allows the model to focus its attention on other visual and textual cues available in the source image.

Our training pipeline is implemented using PyTorch [83] in conjunction with the AllenNLP [84] framework. The code for the RoBERTa model is adapted from Fairseq.[5] We utilized three NVIDIA GeForce GTX 1080 Ti GPUs to perform experiments on three news image captioning datasets.

[4]https://github.com/openai/CLIP
[5]https://github.com/pytorch/fairseq

C. ABLATION RESULTS

Tables 1 and 2 summarize the results of different setups of the proposed architecture. Table 3 summarizes the results of the different ablation model setups on GPE, ORG, LOC and EVNT entities, which are part of our GeoOrg NE category. The performance of the model has been analyzed through ablation results on generic metrics, named entities, and person entities. To validate the effect of our proposed framework, we selected the GoodNews and NYTimes800k datasets and performed ablation results with two variants of the proposed model.

1) IMPACT OF ENTITY CATEGORIZATION

Table 1 and 3, showcases the impact of entity categorization on news image captioning. The model configuration without the categorization of entities is denoted as W/O-EntCat. In this model variant, we incorporated all entities as input in place of both the categories *i.e.* Person NE and GeoOrg NE. Therefore, no separate attention will be provided to entity categories. We observed significant performance drops in the results of the W/O-EntCat configuration. The results depict that using separate attention on person entities' and geoOrg entities' features have notably elevated the precision and recall performance of person entities and geoOrg entities thereby enhancing the results of all metrics. In particularly, leveraging facial details from the source image to capture information about person entities boosted person entity results. These outcomes indicated the success of our proposed combination, which resulted in a performance boost.

2) IMPACT OF ADDON IMAGE

The outcomes of the impact of the addon image on both datasets are presented in Tables 2 and 3. The model configuration without the addon image input is denoted as W/O-Addon. In this model variant's CAF module, we are using only the source image features to emphasize the article features. We observed a performance drop in results with this setup, particularly in metrics other than precision and recall of person entities. The results suggest that the addon image plays a crucial role in capturing relevant context from the article. Effective integration of the addon image has benefited the model well, showing performance boosts in results.

These ablation findings show that the incorporation of entity categorization demonstrating the effectiveness of structured entity-aware attention and additional visual features leads to context improvement resulting in overall performance improvement.

D. ANALYSIS OF COMPETITIVE METHODS

Table 4 summarizes the evaluation metrics result of several news images captioning works on three datasets: GoodNews, NYTimes800k and DM800K. In addition, we compare our results with two influential image captioning frameworks, such as BLIP-2 [37] (with Flan-T5.xl version text decoder) and OFA [56] in zero-shot setting with the prompt provided

TABLE 1. Impact of presence of entity categorization in news image captioning.

		GoodNews		NYTimes800k	
		Ours	W/O-EntCat	Ours	W/O-EntCat
BLEU4		**8.02**	7.75	**9.14**	8.44
ROUGE		**24.93**	22.52	**26.72**	22.59
METEOR		**14.02**	12.66	**13.48**	12.54
CIDEr		**77.52**	72.23	**78.31**	72.81
Named Entities	Precision	**27.92**	26.14	**30.07**	28.22
	Recall	**27.51**	25.84	**29.25**	27.76

TABLE 2. Impact of addon image in news image captioning.

		GoodNews		NYTimes800k	
		Ours	W/O-Addon	Ours	W/O-Addon
BLEU4		**8.02**	7.96	**9.14**	8.89
ROUGE		**24.93**	22.14	**26.72**	24.89
METEOR		**14.02**	13.62	**13.48**	12.92
CIDEr		**77.52**	74.96	**78.31**	75.57
Named Entities	Precision	**27.92**	26.81	**30.07**	28.66
	Recall	**27.51**	26.47	**29.25**	28.56

TABLE 3. Impact of entity categorization and addon image on the precision and recall results of PER, GPE, ORG, LOC and EVNT entities across two datasets of News Image Captioning.

		GoodNews			NYTimes800k		
		Ours	W/O-EntCat	W/O-Addon	Ours	W/O-EntCat	W/O-Addon
PER	Precision	**34.14**	32.32	33.72	**41.81**	39.63	40.59
	Recall	**30.62**	27.14	28.74	**39.92**	38.08	37.42
GPE	Precision	**27.57**	25.19	25.29	**34.67**	33.45	33.32
	Recall	**30.42**	29.45	29.53	**38.78**	37.27	37.25
ORG	Precision	**22.45**	21.57	21.48	**23.54**	22.46	22.37
	Recall	**22.51**	20.81	20.67	**23.78**	22.62	22.70
LOC	Precision	**15.45**	14.78	14.85	**14.56**	13.75	13.81
	Recall	**14.88**	12.94	13.18	**13.41**	12.57	12.82
EVNT	Precision	**14.67**	13.79	13.72	**14.87**	13.95	13.87
	Recall	**13.16**	12.51	12.43	**14.12**	13.42	13.37

as 'Generate the caption for news image?' without any task-specific fine-tuning. The results show that OFA outperforms BLIP-2, demonstrating its stronger capability in generating contextually rich and semantically aligned captions, especially prominent score of the METEOR and ROUGE metrics, indicating that OFA produces more fluent and semantically relevant captions. BLIP-2 and OFA perform better on DM800K as only 70% of its captions contain named entities, making it less context dependent than GoodNews and NYTimes800K, which have 96% captions having named entity. As, these models primarily focus on general image captioning and do not incorporate contextual information from associated news articles, limiting their effectiveness in the news domain.

We compare our proposed model against a range of baselines and competitive approaches designed for news image captioning, such as the framework proposed by Biten [11] proposed a framework that utilizes GloVe embedding or tough-to-beat [86] to derive features from the article, and in the following stage, the framework employs the named entity insertion method. ICECAP [15] focuses on retrieving the related sentence from the article and generating the captions from the selected sentences. Yang et al. [57] incorporates transformer into the captioning process with the object recognition ImageNet model and scene recognition Places365 [59] model. Trans & Tell [12] model, which utilizes separate attention over faces, images, objects, and text. Models [11], [12], and [57] exhibit mediocre performance due to a lack of crucial entity information necessary for caption generation. Zhao et al. [62] propose to enhance entity-aware image captioning through the integration of a multi-modal knowledge graph. Visual News [13] incorporated entities as input and addressed the out-of-dictionary challenge by utilizing the multi-head pointer generation method. Moreover, to enhance the quality of the generated caption, JoGANIC [17] followed journalists' guidelines by considering the structure of the

TABLE 4. Outcomes of proposed news image captioning model on GoodNews, NYTimes800k and DM800K datasets using BLEU, ROUGE, METEOR, and CIDEr evaluation metrics. It also illustrates the precision and recall of named entities.

	Method	BLEU-4	ROUGE	METEOR	CIDEr	Named Entities	
						Precision	Recall
GoodNews	BLIP-2 [37]	1.38	8.62	4.83	6.65	-	-
	OFA [56]	2.36	10.5	7.1	8.40	-	-
	GoodNews [11]	0.89	12.20	4.37	13.1	8.23	6.06
	ICECAP [15]	1.96	15.70	6.01	26.08	-	-
	Yang *et al.* (ImageNet) [57]	4.46	20.56	8.38	44.16	-	-
	Yang *et al.* (Places365) [57]	4.52	20.41	8.42	44.04	-	-
	Trans & Tell [12]	6.05	21.4	-	53.8	22.2	18.7
	Zhao *et al.* [62]	6.14	21.46	6.32	54.02	-	-
	Visual News [13]	6.1	21.6	8.3	55.4	22.9	19.3
	JoGANIC [17]	6.83	23.05	11.25	61.22	26.87	22.05
	NewsMEP [14]	8.30	23.17	12.23	63.99	23.43	23.24
	Zhang *et al.* [63]	7.94	28.68	13.97	64.51	29.69	27.37
	Focus with Context [64]	6.3	22.4	-	60.3	24.2	20.9
	Kalarani *et al.* [65]	7.14	24.3	11.21	72.33	24.37	20.09
	Zhao *et al.* [62]	6.14	21.46	6.32	54.02	—	—
	JoGANIC* [66]	5.69	20.69	-	50.42	20.98	18.05
	Tell(EG+SA) [66]	6.05	21.18	-	53.10	21.48	18.70
	VisContext w/ BART$_{large}$ [67]	7.91	22.61	11.67	69.11	24.38	23.56
	ConCaps [19]	6.2	21.7	10.3	54.3	22.2	18.7
	IEK-NIC [85]	9.04	23.57	13.37	65.45	26.80	22.08
	Rule [68]	8.18	23.56	12.50	71.58	25.51	23.68
	Ours	8.02	24.93	**14.02**	**77.52**	27.92	**27.51**
NYTimes800k	BLIP-2 [37]	1.57	8.73	5.18	6.57	-	-
	OFA [56]	2.56	10.72	7.32	8.44	-	-
	Trans & Tell [12]	6.30	21.7	-	54.4	24.6	22.2
	Zhao *et al.* [62]	6.32	21.62	6.25	54.47	-	-
	Visual News [13]	6.1	21.9	8.1	56.1	24.8	22.3
	JoGANIC [17]	6.79	22.80	10.93	59.42	28.63	24.49
	NewsMEP [14]	9.57	23.62	13.02	65.85	26.61	28.57
	Zhang *et al.* [63]	7.57	25.67	12.64	62.31	30.04	25.53
	Focus with Context [64]	7.0	22.9	-	63.6	29.8	25.9
	kalarani *et al.* [65]	7.54	23.28	11.27	66.41	28.11	23.25
	Zhao *et al.* [62]	6.32	21.62	6.25	54.47	-	-
	JoGANIC* [66]	6.30	21.39	-	53.91	24.24	22.20
	Tell(EG+SA) [66]	6.20	21.51	-	53.59	23.81	21.87
	VisContext w/ BART$_{large}$ [67]	8.81	23.02	12.00	70.83	27.57	27.16
	ConCaps [19]	6.6	21.8	10.6	55.2	24.6	22.2
	IEK-NIC [85]	10.18	23.81	13.33	67.34	28.59	28.44
	Rule [68]	9.41	24.42	13.10	72.29	28.15	28.80
	Ours	9.14	**26.72**	**13.48**	**78.31**	**30.07**	**29.25**
DM800K	BLIP-2 [37]	2.73	9.24	5.74	7.62	-	-
	OFA [56]	2.96	11.81	8.11	8.86	-	-
	ConCaps +VertCoh+HoriCoh1 [19]	6.8	17.8	10.2	43.3	18.8	17.8
	ConCaps +VertCoh+HoriCoh1+HoriCoh2 [19]	6.5	17.6	10.1	43.5	18.5	17.6
	Ours	**7.2**	**18.5**	**11.25**	**49.4**	**19.14**	**18.3**

captions. Generating such a precise and straightforward template for news image caption is improper. NewsMEP [14] leverages prompting mechanisms to produce more precise and informative captions for news images. Entities are selected for prompting by calculating the correlation score using images and entities' features. Zhang et al. [63] proposed the semantic discourse extraction (DiscExt), which helps to capture the relevant context related to the image from the article. At the same time, contrastive style-coherent learning improves the coherence and consistency of the captions. Their approach uses a heuristic method for selecting oracle elementary discourse units during pretraining in the DiscExt module, which relies on maximizing the ROUGE-L score.

Focus with Context [64] model ensures that the generated captions are well-aligned with the image content and the associated news article by efficiently selecting a context. Kalarani et al. [65] constructed the model based on the OFA_{Large} architecture, where pre-training involves three tasks: generating captions based on images and context, predicting keywords from news articles, and classifying whether a given caption entails the image and context. Qu et al. [67] proposed model that is built on BART and integrates visual and name features using cross-attention modules and a face naming module focused only on person entities. Xu et al. [68] introduce the news-aware semantic rule consisting of entities, actions and semantic roles which improves caption quality significantly. The authors of [19] propose a

TABLE 5. Comparison of precision (P) and recall (R) of PER, GPE, ORG, LOC and EVNT entities.

	Methods	PER		GPE		ORG		LOC		EVNT	
		P	R	P	R	P	R	P	R	P	R
GoodNews	Trans & Tell [12]	29.2	23.1	-	-	-	-	-	-	-	-
	NewsMEP [14]	32.50	29.06	26.18	30.37	22.10	22.47	-	-	-	-
	Zhang *et al.* [63]	28.48	29.45	23.66	26.24	-	-	-	-	-	-
	JoGANIC* [66]	26.22	21.66	-	-	-	-	-	-	-	-
	Tell(EG+SA) [66]	27.97	22.92	-	-	-	-	-	-	-	-
	VisContext [67] w/ BART$_{large}$	30.33	26.32	-	-	-	-	-	-	-	-
	IEK-NIC [85]	34.4	30.15	-	-	-	-	-	-	-	-
	Rule [68]	33.89	30.09	-	-	-	-	-	-	-	-
	Ours	**34.14**	**30.62**	**27.57**	**30.42**	**22.45**	**22.51**	**15.45**	**14.88**	**14.67**	**13.16**
NYTimes800k	Trans & Tell [12]	37.3	31.1	-	-	-	-	-	-	-	-
	NewsMEP [14]	41.32	38.51	28.24	38.76	23.10	23.51	-	-	-	-
	Zhang *et al.* [63]	27.60	38.41	33.05	35.53	-	-	-	-	-	-
	JoGANIC* [66]	35.57	30.23	-	-	-	-	-	-	-	-
	Tell(EG+SA) [66]	35.50	30.47	-	-	-	-	-	-	-	-
	VisContext [67] w/ BART$_{large}$	38.53	34.22	-	-	-	-	-	-	-	-
	IEK-NIC [85]	41.2	34.1	-	-	-	-	-	-	-	-
	Rule [68]	41.46	39.88	-	-	-	-	-	-	-	-
	Ours	**41.81**	**39.92**	**34.67**	**38.78**	**23.54**	**23.78**	**14.56**	**13.41**	**14.87**	**14.12**

contrastive entity-aware multi-image captioning framework, abbreviated as ConCaps. This approach focuses on jointly generating a caption by integrating multiple images. The model architecture comprises a transformer-based caption generation framework coupled with a coherence mechanism based on contrastive learning. The table IV presents two variants of the ConCaps model, each employing a different set of coherence mechanisms. The first variant, ConCaps+VertCoh+HoriCoh1+HoriCoh2, integrates all coherence components, whereas the second variant, ConCaps+VertCoh+HoriCoh1, excludes HoriCoh2. In contrast, our approach does not perform joint captioning over multiple images. Instead, we treat one image as the target for caption generation and utilize additional images only as auxiliary sources of contextual information. This spatially or temporally related nearby image is not directly captioned but is selectively processed to extract entity-level and contextual cues that inform the caption generation for the source image. This distinction allows our method to maintain focus on the target image while still benefiting from external context, without merging multiple visual scenes into a single narrative as ConCaps does. Furthermore, our utilization of entity categorization followed by diverse attention mechanisms, focusing on both faces and entities of the source image, has led to enhanced outcomes.

Zhang and Wan [66] aimed to address the influence of visual features in news image captioning. They conducted in-depth analysis on different models built upon Tran et al. [12], JoGANIC [17] and visual news captioner [13]. One model variant is JoGANIC*, which adopts a transformer-based framework that incorporates component template guidance in accordance with journalistic principles. Another model variant is Tell (EG+SA), which implemented Tell (full) with the Entity Guidance (EG) technique and replaced dynamic convolution with the self-attention (SA) mechanism. They observe that incorporating vision features enhances models' ability to generate entities that are present in both textual context and captions. Towards this, our proposed model seamlessly integrates the abundance of visual features that enhance the model's generalization ability and help the model generate entity-rich captions. Ajankar et al. [85] introduce an IEK-NIC model to generate news image captions that incorporates the named entities generated by image-relevant entities generation using the Time-step-bounded Entity Constrained (TEC) beam search algorithm. TEC imposes constraints on named entities composed of multiple tokens during the decoding process, ensuring they are generated sequentially without splitting or truncation. In contrast to prior models, our proposed model integrates additional visual information to leverage the additional contextual cues from the associated article. In addition to that, we proposed to categorize entities and uses separate attention over each entity category to emphasize the features of person entities and geoOrg entities. This enhancement facilitates the incorporation of image-relevant entities, consequently yielding enhanced outcomes of key entities like PER, GPE, ORG, LOC and EVNT, as demonstrated in Table 5 and other NLP metrics as shown in Table 4. On both the GoodNews and NYTimes800K datasets, the model achieves higher METEOR and CIDEr scores, indicating strong alignment with reference captions at both lexical and semantic levels. The high named entity recall further demonstrates the model's ability to preserve factual correctness and relevant details. These improvements highlight the robustness and generalizability of the proposed framework, particularly its effectiveness in generating context-aware and entity-grounded captions. This improved results shows that the proposed Cross-attentional Entity Features module strengthens the entity grounding by focusing on face and image regions associated with these named

FIGURE 4. Qualitative results for the proposed architecture show the predicted caption for the source image while the addon image is passed as additional visual context. The dark blue and light blue colors are used to depict person entities and other entities, respectively.

entities, enabling more accurate alignment between visual cues and textual mentions. Moreover, MHEDMA in the decoder provides dynamic attention to the various contextualized encoder inputs. As a result, our model demonstrates the effectiveness of the proposed architecture. In particular, the model achieves higher CIDEr scores by using rich contextual information. Despite DM800K dataset have less percentage named entities, our model shows improved results over all metrics. We carefully reduce potential errors from irrelevant addon image integration, as such images are only used to provide additional story-related context.

In particular, while ConCaps with all coherence modules performs reasonably well, it lags behind on all major metrics, pointing to limitations in coherence modeling when used without deep contextual fusion strategies. DM800K dataset characterized by a lower proportion of named entities and high diversity in articles, in spite of that our model continues to outperform baselines, achieving the best scores of BLEU-4 (7.2), ROUGE (18.5), METEOR (11.25), and CIDEr (49.4) and precision (19.14) and recall (18.3) of all named entities. This shows that our proposed model design choices, especially in handling category-specific attention and leveraging temporal image sequences, are capable of producing rich contextual captions.

E. QUALITATIVE STUDIES

In this section, we outline qualitative samples that demonstrate the capabilities of the proposed model. Some example images are depicted in Figure 4. Each example in the figure shows two input images; one is an addon image for providing additional visual context to the model, and the other is a source image for which the model predicts the informative caption. 'GT' shows the ground truth caption of the source image, and 'Ours' showcases predicted output. The 'Trans & Tell' represents the predicted captions by using the architecture proposed by Tran et al. [12]. The predicted caption for the source image of Figure 4(a) depicts the correct identification of person entities "Roger Federer" and "Rafael Nadal") and the event entity ("The French Open") and context, while Trans & Tell model is not able to predict the correct context. This outcome was achieved by attending to facial features and geoOrg entities separately with the help of proposed CEF module, which enable the model to associate visual cues with relevant named entities. Similarly, in the predicted caption of Figure 4 (b), the model generated the correct person entity *i.e.* "Leandra Medine" demonstrate the proposed model capability of focusing on person entities using faces. Also, the addon image proves invaluable as the model successfully predicts

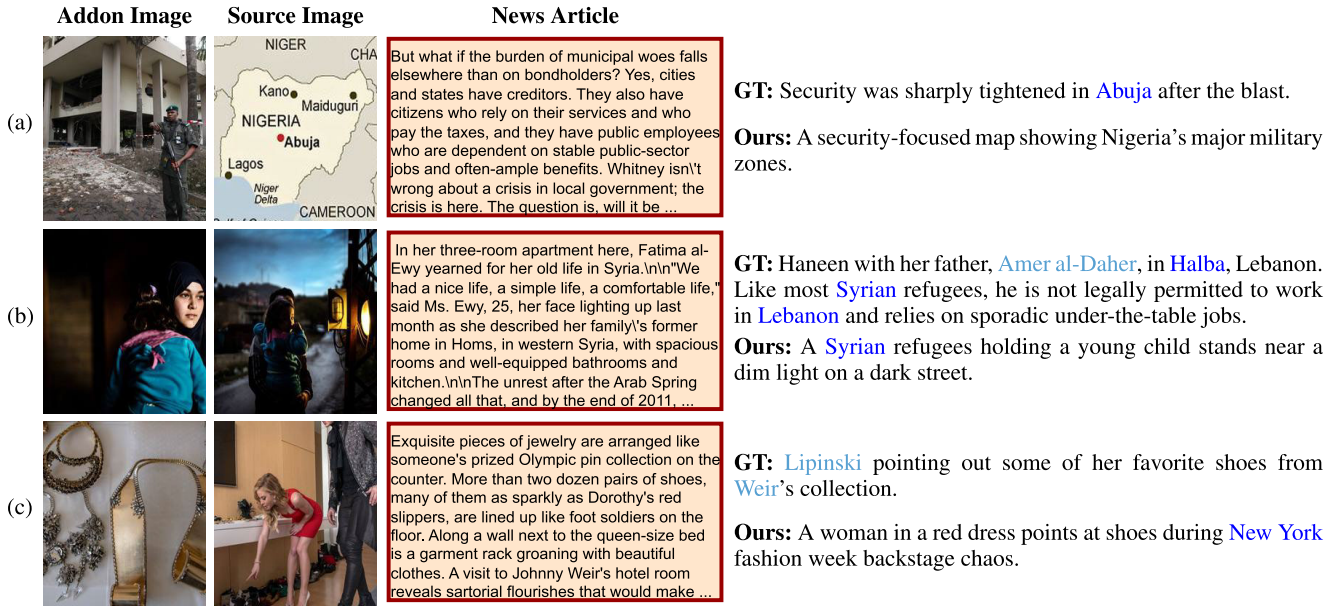

FIGURE 5. Figure illustrates some error cases of the proposed news image captioning model, showing the predicted captions for the source images while the addon images are provided as additional visual context. Light blue and dark blue colors are used to represent person entities and other entities, respectively.

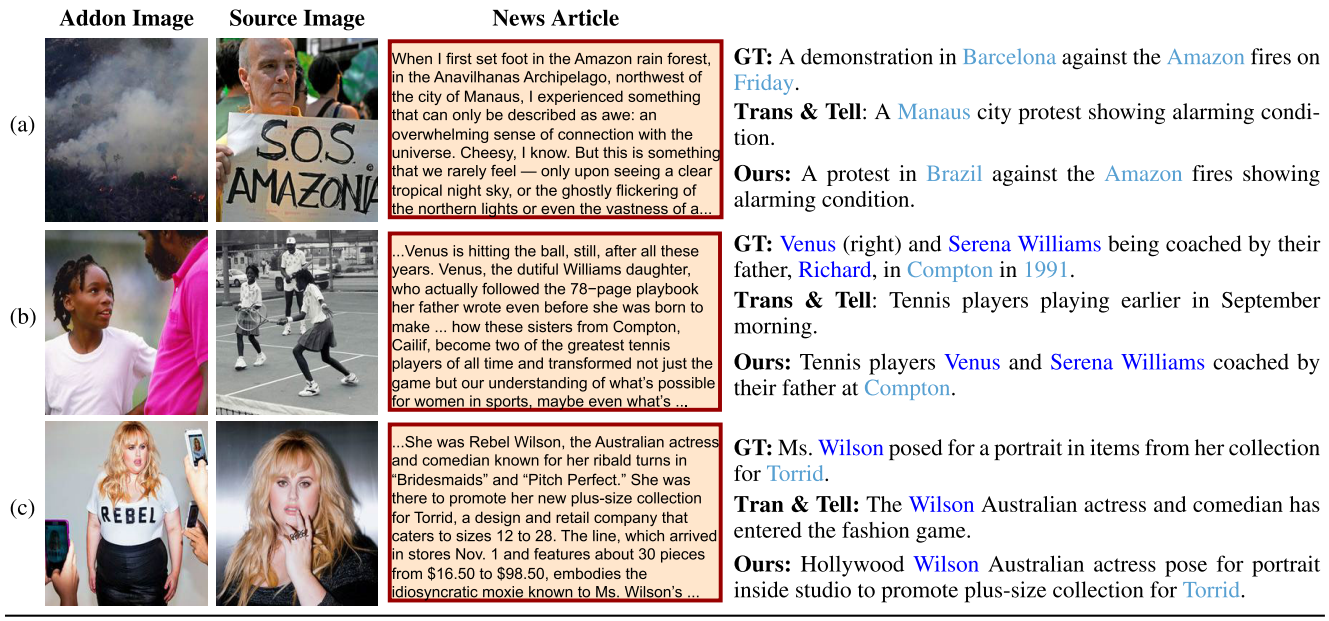

FIGURE 6. Additional examples of Qualitative results for the proposed architecture show the predicted caption for the source image while the addon image is passed as additional visual context.

related background information related to fashion, which is absent in the source image and in the prediction of the Trans & Tell model. Likewise, examining the result of Figure 4 (c), the correct identification of the person's name, "Afonso Dhlakama," and location "Mozambique," in the predicted caption emphasizes the efficacy of leveraging separate attention on the person and geoOrg entity categories. In addition, the proposed model demonstrates its capability by accurately predicting election-related information while generating a caption, which is missing in prediction of the Trans & Tell model. In a similar manner, examining the result for Figure 4 (d), the caption generated by our model demonstrates improved grounding by correctly capturing both the event type"bombarding" and the location "Eastern

Ghouta," while also emphasizing the severity of the medical condition. This improvement is aided by the nearby (addon) image, which provides additional visual context about the aftermath of the bombing not evident in the source image alone. Furthermore, the use of separate entity attention allows for more accurate identification and grounding of geo-political entities like "Eastern Ghouta," enabling the model to generate a more contextually rich and event-specific caption compared to the generic output of the Trans & Tell model.

The above observations demonstrate its ability to focus on the person and geoOrg entities separately and the usefulness of utilizing the addon image to capture missing information. The correct identification of entities demonstrates the positive impact of entity classification and multimodal cues in improving caption quality and contextual coherence.

V. CONCLUSION AND FUTURE WORK

This research provides a novel architecture for generating captions for news images. Our proposed framework includes attending to person entities using facial features and geoOrg entities using image features that resulted in the generation of entity-rich captioning. Additionally, we proposed to utilize additional visual information in the present research that assists in identifying the more contextual information by emphasizing the coarse-grained, ambiguous features of the article. This advancement sets the proposed framework apart from existing methodologies. The favorable advancement in outcomes over current frameworks demonstrates the efficiency of the proposed framework. There is potential for further exploration of robustness of the proposed model by conducting experiments on low-quality or noisy images to evaluate its generalization capabilities. In addition, in the future, incorporating advanced model optimization techniques such as structured pruning and quantization-aware training can help reduce computational complexity, decrease inference latency, and improve throughput without significantly compromising performance. These enhancements support readiness for scalable, low-latency applications in live news-streaming and mobile-device scenarios. Furthermore, although the framework performs well on English-language datasets such as GoodNews, NYTimes800k, and DM800K to explore its generalizability to non-English news content by developing a multilingual dataset for this task.

APPENDIX A
LINGUISTIC EVALUATION BASED ON VERB AND ADJECTIVE ALIGNMENT

In addition, to verify that the model produces syntactically and semantically coherent captions while enhancing entity-specific content. We conducted a light part-of-speech analysis on a representative sample of 500 generated captions. We compared verb and adjective usage in generated captions with ground-truth references and observed. We observed that the verb match rate is 82.9% on GoodNews, 84.2% on NYTimes800K and 81.2% on DM800K, whereas adjective match rate is 75.2% on GoodNews, 72.8% on NYTimes800K and 70.8% on DM800K datasets. These results reinforce that the model produces syntactically and semantically coherent captions while enhancing entity-specific content.

APPENDIX B
CLIP-BASED SIMILARITY ANALYSIS

We conducted an extra analysis on a sample of 500 examples, to check the effect of whether visually similar addon image and source image. For this analysis we use CLIP-based feature similarity score. We found that addon images with a similarity score below 0.6 also brought in useful new context. This suggests that our method for choosing proximity images keeps a good mix of relevance and variety, without repeating the same visual content. As illustrated in Fig. 4, addon images contribute missing cues across various scenarios. Among these examples Fig. 4 (b) and Fig. 6 (a) have CLIP similarity score less than 0.55 still provides the additional related contextual cues like fashion and fire, respectively, which are missing in the source image.

APPENDIX C
ERROR ANALYSIS

To evaluate the limitations of our proposed framework, we conducted a detailed error analysis using the validation sets of the GoodNews and NYTimes800K datasets. The observed errors fall into several categories, including entity misclassification, failures due to unclear visual details, hallucinated content and contextual omissions. We illustrate these with representative examples shown in Figure 5.

One major issue was entity misclassification. In Figure 5 (a), the ground truth discusses tightened security in Abuja after a blast, reflecting a specific geopolitical event. However, the image only shows a map with no visual cues about the incident, and the model generates a caption referring to "Nigeria's major military zones." This error arises from a misinterpretation of the scene, where the absence of event-related content leads the model to default to a broader, static geopolitical description. The lack of visual grounding makes it difficult to resolve the intended meaning, leading to incorrect entity-type attribution.

We also encountered challenges due to unclear or poor-quality visual input, which hindered accurate entity recognition. for example, in Figure 5 (b), the faces of Amer al-Daher and his daughter Haneen are not visible due to dim lighting. As a result, the model fails to associate the visual content with the work context provided in the article, generating a generic and under-informative caption. This demonstrates the model's sensitivity to low-quality or visually ambiguous scenes.

Another frequent limitation was hallucination, where the model generated content not grounded in the image or article. For instance, in Figure 5 (c), the ground truth caption refers to Lipinski pointing out some of her favorite shoes from

Weir's collection, yet the model-generated output includes ''New York'' location which is not mentioned in the article or not visible in the image. This hallucinated detail likely results from training bias.

The model also exhibited contextual omission, where important article-specific details such as named entities or background narrative were excluded. These omissions were especially frequent in images lacking strong visual cues to reinforce such details. The generated caption of Figure 5 (c) omits the important contextual details like Weir's Collection. In addition, incorporating adjacent images introduces the risk of context drift. For instance, as illustrated in Figure 5(c), the inclusion of an addon image led to the generation of irrelevant fashion-related information, deviating from the main content of the target image.

These findings collectively highlight the need for stronger entity-aware attention mechanisms, multimodal grounding, and robust disambiguation strategies, particularly when images lack sufficient context or specificity or when there is a risk of context drift from adjacent images.

APPENDIX D
ADDITIONAL EXAMPLES

We have shown the additional examples in this appendix to further illustrate the effectiveness of our proposed approach across diverse scenarios.

The example in Figure 4 (a) shows that our model correctly identifies the protest's location and its connection to the Amazon fires, while the Trans & Tell model misclassifies the geographical setting. This improvement is supported by addon visual cues, which provide evidence of the environmental crisis that is not visible in the source image alone. Combined with separate attention over geo-political entities, the model is better able to associate visual and textual signals with real-world events. Similarly, results in Figure 4 (b) shows that the model successfully recognizes the key person entities and their spatial and familial relationship, whereas the Trans & Tell model fails to ground any meaningful details. The correct identification of Venus and Serena Williams, along with the location Compton, is enabled by distinct attention mechanisms applied to person and location categories, enhancing the accuracy of entity grounding. Likewise, example in Figure 4 (c), our model accurately grounds both the subject Wilson and the brand Torrid, maintaining the promotional context reflected in the article. In contrast, the Trans & Tell model produces a vague description lacking key named entities. The benefit of separate entity attention is evident here, as the model is able to distinguish between the person and organization categories and incorporate both meaningfully into the caption.

Together, these examples emphasize that incorporating separate entity attention and additional visual cues enables our model to generate precise, contextually rich captions that more effectively reflect the underlying article content.

REFERENCES

[1] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, ''Show and tell: A neural image caption generator,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, ''Show, attend and tell: Neural image caption generation with visual attention,'' in *Proc. ICML*, vol. 3, Jul. 2015, pp. 2048–2057.

[3] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, ''Image captioning with semantic attention,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.

[4] J. Lu, C. Xiong, D. Parikh, and R. Socher, ''Knowing when to look: Adaptive attention via a visual sentinel for image captioning,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 375–383.

[5] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, ''Self-critical sequence training for image captioning,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1179–1195.

[6] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, ''Boosting image captioning with attributes,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4894–4902.

[7] A. Ueda, W. Yang, and K. Sugiura, ''Switching text-based image encoders for captioning images with text,'' *IEEE Access*, vol. 11, pp. 55706–55715, 2023.

[8] D.-J. Kim, T.-H. Oh, J. Choi, and I.-S. Kweon, ''Semi-supervised image captioning by adversarially propagating labeled data,'' *IEEE Access*, vol. 12, pp. 93580–93592, 2024.

[9] L. Cheng, W. Wei, X. Mao, Y. Liu, and C. Miao, ''Stack-VS: Stacked visual-semantic attention for image caption generation,'' *IEEE Access*, vol. 8, pp. 154953–154965, 2020.

[10] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, ''Meshed-memory transformer for image captioning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10575–10584.

[11] A. F. Biten, L. Gomez, M. Rusiñol, and D. Karatzas, ''Good news, everyone! Context driven entity-aware captioning for news images,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12458–12467.

[12] A. Tran, A. Mathews, and L. Xie, ''Transform and tell: Entity-aware news image captioning,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13032–13042.

[13] F. Liu, Y. Wang, T. Wang, and V. Ordonez, ''Visual news: Benchmark and challenges in news image captioning,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 6761–6771.

[14] J. Zhang, S. Fang, Z. Mao, Z. Zhang, and Y. Zhang, ''Fine-tuning with multi-modal entity prompts for news image captioning,'' in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 4365–4373.

[15] A. Hu, S. Chen, and Q. Jin, ''ICECAP: Information concentrated entity-aware image captioning,'' in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4217–4225.

[16] A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk, ''BreakingNews: Article annotation by image and text processing,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1072–1085, May 2018.

[17] X. Yang, S. Karaman, J. Tetreault, and A. Jaimes, ''Journalistic guidelines aware news image captioning,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 5162–5175.

[18] I. Yamada, A. Asai, J. Sakuma, H. Shindo, H. Takeda, Y. Takefuji, and Y. Matsumoto, ''Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia,'' 2018, *arXiv:1812.06280*.

[19] J. Chen, ''Transform, contrast and tell: Coherent entity-aware multi-image captioning,'' *Comput. Vis. Image Understand.*, vol. 238, pp. 1–28, Jan. 2024.

[20] I. Phueaksri, M. A. Kastner, Y. Kawanishi, T. Komamizu, and I. Ide, ''An approach to generate a caption for an image collection using scene graph generation,'' *IEEE Access*, vol. 11, pp. 128245–128260, 2023.

[21] Q. H. Ngo, T. Kechadi, and N.-A. Le-Khac, ''Domain specific entity recognition with semantic-based deep learning approach,'' *IEEE Access*, vol. 9, pp. 152892–152902, 2021.

[22] Y. Ma, X. Sun, J. Ji, G. Jiang, W. Zhuang, and R. Ji, ''Beat: Bi-directional one-to-many embedding alignment for text-based person retrieval,'' in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 4157–4168.

[23] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji, ''X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval,'' in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 638–647.

[24] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Proc. 16th Eur. conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2020, pp. 121–137.

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[26] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2014, pp. 3104–3112.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.

[28] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4634–4643.

[29] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2019, pp. 1–11.

[30] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10971–10980.

[31] B. Wang, C. Wang, Q. Zhang, Y. Su, Y. Wang, and Y. Xu, "Cross-lingual image caption generation based on visual attention model," *IEEE Access*, vol. 8, pp. 104543–104554, 2020.

[32] W. Jiang, X. Li, H. Hu, Q. Lu, and B. Liu, "Multi-gate attention network for image captioning," *IEEE Access*, vol. 9, pp. 69700–69709, 2021.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. NeurIPS*, vol. 30, Jun. 2017, pp. 5998–6008.

[34] Y. Luo, J. Ji, X. Sun, L. Cao, Y. Wu, F. Huang, C.-W. Lin, and R. Ji, "Dual-level collaborative transformer for image captioning," in *Proc. AAAI conf. Artif. Intell.*, vol. 35, May 2021, pp. 2286–2293.

[35] Y. Ma, J. Ji, X. Sun, Y. Zhou, Y. Wu, F. Huang, and R. Ji, "Knowing what it is: Semantic-enhanced dual attention transformer," *IEEE Trans. Multimedia*, vol. 24, pp. 1–14, 2022.

[36] J. Ji, Y. Ma, X. Sun, Y. Zhou, Y. Wu, and R. Ji, "Knowing what to learn: A metric-oriented focal mechanism for image captioning," *IEEE Trans. Image Process.*, vol. 31, pp. 4321–4335, 2022.

[37] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023, *arXiv:2301.12597*.

[38] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "EVA: Exploring the limits of masked visual representation learning at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19358–19369.

[39] H. Won Chung et al., "Scaling instruction-finetuned language models," 2022, *arXiv:2210.11416*.

[40] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.

[41] Y. Ma, J. Ji, X. Sun, Y. Zhou, and R. Ji, "Towards local visual modeling for image captioning," *Pattern Recognit.*, vol. 138, Jun. 2023, Art. no. 109420.

[42] Y. Xu, Y. Wu, M. Yang, and H. Chen, "Exploring diverse in-context configurations for image captioning," in *Proc. Adv. NIPS*, Jan. 2023, pp. 1–20.

[43] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, Jan. 2018, pp. 1–16.

[44] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[45] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2020, pp. 213–229.

[46] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. conf. Mach. Learn.*, Jul. 2021, pp. 10347–10357.

[47] C.-F. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 347–356.

[48] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 568–578.

[49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. conf. Mach. Learn.*, Jan. 2021, pp. 8748–8763.

[50] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. conf. Mach. Learn.*, Jan. 2022, pp. 12888–12900.

[51] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao, "Grounded language-image pre-training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10965–10975.

[52] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proc. Eur. conf. Comput. Vis.* Cham, Switzerland: Springer, Nov. 2024, pp. 38–55.

[53] H. Zhang, P. Zhang, X. Hu, Y.-C. Chen, L. Li, X. Dai, L. Wang, L. Yuan, J.-N. Hwang, and J. Gao, "GLIPv2: Unifying localization and vision-language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 36067–36080.

[54] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. conf. Mach. Learn.*, Jan. 2021, pp. 4904–4916.

[55] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "When and why vision-language models behave like bags-of-words, and what to do about it?" 2022, *arXiv:2210.01936*.

[56] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, vol. 162, Jul. 2022, pp. 23318–23340.

[57] Z. Yang and N. Okazaki, "Image caption generation for news articles," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 1941–1951.

[58] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. NeurIPS*, vol. 60, May 2017, pp. 84–90.

[59] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[60] Y. Jing, X. Zhiwei, and G. Guanglai, "Context-driven image caption with global semantic relations of the named entities," *IEEE Access*, vol. 8, pp. 143584–143594, 2020.

[61] Z. Yumeng, Y. Jing, G. Shuo, and L. Limin, "News image-text matching with news knowledge graph," *IEEE Access*, vol. 9, pp. 108017–108027, 2021.

[62] W. Zhao and X. Wu, "Boosting entity-aware image captioning with multi-modal knowledge graph," *IEEE Trans. Multimedia*, vol. 26, pp. 2659–2670, 2024.

[63] Z. Zhang, H. Zhang, J. Wang, Z. Sun, and Z. Yang, "Generating news image captions with semantic discourse extraction and contrastive style-coherent learning," *Comput. Electr. Eng.*, vol. 104, pp. 1–12, Dec. 2022.

[64] M. Zhou, G. Luo, A. Rohrbach, and Z. Yu, "Focus! Relevant and sufficient context selection for news image captioning," 2022, *arXiv:2212.00843*.

[65] A. Rajakumar Kalarani, P. Bhattacharyya, N. Chhaya, and S. Shekhar, "'Let's not quote out of context': Unified vision-language pretraining for context assisted image captioning," 2023, *arXiv:2306.00931*.

[66] J. Zhang and X. Wan, "Exploring the impact of vision features in news image captioning," in *Proc. Findings Assoc. Comput. Linguistics (ACL)*, 2023, pp. 12923–12936.

[67] T. Qu, T. Tuytelaars, and M.-F. Moens, "Visually-aware context modeling for news image captioning," 2023, *arXiv:2308.08325*.

[68] N. Xu, T. Zhang, H. Tian, and A.-A. Liu, "Rule-driven news captioning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 11, pp. 1–12, Nov. 2024.

[69] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

[70] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[71] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[72] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. NeurIPS*, vol. 26, Dec. 2013, pp. 3111–3119.
[73] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1–12.
[74] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "SpaCy: Industrial-strength natural language processing in Python," Zenodo, Tech. Rep., 2020, pp. 1–16.
[75] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[76] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, Jan. 2014, pp. 740–755.
[77] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2002, pp. 311–318.
[78] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Stat. Mach. Transl.*, 2014, pp. 376–380.
[79] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text summarization Branches Out*, Jul. 2004, pp. 74–81.
[80] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
[81] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[82] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. New York, NY, USA: Academic, 1994, pp. 474–485.
[83] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NIPS Workshop Autodiff*, Oct. 2017, pp. 1–4.
[84] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, "AllenNLP: A deep semantic natural language processing platform," in *Proc. Workshop NLP Open Source Softw. (NLP-OSS)*, 2018, pp. 1–6.
[85] S. Ajankar and T. Dutta, "Image-relevant entities knowledge aware news image captioning," *IEEE Multimedia Mag.*, vol. 29, no. 3, pp. 82–90, May 2022.
[86] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *Proc. Conf. Learn. Represent.*, Apr. 2017, pp. 1–16.

SONALI AJANKAR (Member, IEEE) received the M.Tech. degree from the Department of Computer Science and Engineering, Veermata Jijabai Technological Institute (VJTI), Mumbai, India, in 2010. She is currently pursuing the Ph.D. degree in computer science and engineering with Indian Institute of Technology (BHU) Varanasi, India. She has been an Assistant Professor with the Department of Master of Computer Application, VJTI. Her research interests include computer vision and deep learning.

TANIMA DUTTA (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from IIT Guwahati, Guwahati, India, in 2014. She was with the TCS Innovation Laboratories, Bengaluru, India. She is currently an Associate Professor with the Department of Computer Science and Engineering, IIT (BHU) Varanasi, Varanasi, India. Her research interests include multimedia, digital forensics, wireless sensor networks, and algorithms. She has received a TCS Research Fellowship and a SAIL Fellowship for pursuing her Ph.D. and B.Tech. degrees, respectively.