

Feature extraction semilearning and augmented representation for image captioning in crowd scenes

Khang Tan Tran Minh Nguyen*

ABSTRACT

Image captioning has been an interesting task since 2015. The topic lies in the gap between Computer Vision and Natural Language Processing research directions. The problem can be described as follows: Given the input as a three-channel RGB image, a language model is trained to generate the hypothesis caption that describes the images' contexts. In this study, we focus on solving image captioning in images captured in a crowd scene, which is more complicated and challenging. In general, a semilearning feature extraction mechanism is proposed to obtain more valuable high-level feature maps of images. Moreover, an augmented approach in the Transformer Encoder is explored to enhance the representation ability. The obtained results are promising and outperform those of other state-of-the-art captioning models on the CrowdCaption dataset.

Key words: crowd scene, image captioning, transformer

INTRODUCTION

With the development of technology, automatic image processing now plays an important role in many fields in life to provide information to humans as quickly as possible without increasing the number of employees but guaranteeing accuracy. In parallel, step-by-step deep learning has become the key to solving many real-world problems to satisfy society's requirements. One of the most interesting tasks worldwide is image captioning¹. With the given image, an automatic system can supply humans with a description of its context, emphasizing some highlighted visual events. In addition, day-to-day crowd scenes include much information for analysis in specific fields, such as the management and service of smart cities, intelligent transportation, and public security risk². We suppose that the most important application of automatic description in crowd scenes is security because of several highly possibly abnormal occurrences, such as violence, stealing, and sexual assault. However, crowd images are much more complicated because of the large number of people and activities involved in recognizing and analyzing them. Therefore, there is a high demand for studying the problem of automatic description generation for crowd scenes. The large differences between usual image captioning and image captioning in crowd scenes include objects, contexts, and interactions. Images capturing crowd scenes include lots of people in complicated and various contexts. More-

over, words describing the interactions between objects appear much more frequently. Therefore, there is a need to explore approaches that effectively represent objects in images and contexts simultaneously to help caption models more easily learn and generate descriptions. Even with this, several techniques have also been researched to increase the quality of these representations to avoid underfitting because of the complex information in crowd scenes.

To solve the image captioning problem, previous studies model the problem as a machine translation problem. Given the input X as the image, the trained model predicts the sequence of tokens $Y = \{y_1, y_2, \dots, y_n\}$, in which y_t is the predicted word at time step t and n is the maximum length of the predicted sentence. From y_1 to y_n , there are two specific tokens $\langle \text{bos} \rangle$ and $\langle \text{eos} \rangle$ that indicate where the generated caption starts and ends, respectively. Most of the studies considered an autoregressive approach to solving this problem, which means that the next predicted word conditioned on the previous word can be formulated as $y_t = \text{model}(X|y_{1:t-1})$. In this study, we also follow the autoregressive approach. Figure 1 illustrates the input-output definition.

There are two common types of images in the model space: region-based features and grid-based features. Region-based features of images are extracted using the Faster R-CNN model, which is a famous object detection model. Once extracted, region-based features are embedding vectors that represent objects detected in images. In contrast, grid-based features in-

University of Information Technology,
Viet Nam National University Ho Chi
Minh City, Ho Chi Minh City, Viet Nam

Correspondence

Khang Tan Tran Minh Nguyen, University of Information Technology, Viet Nam National University Ho Chi Minh City, Ho Chi Minh City, Viet Nam

Email: khangnttm@uit.edu.vn

History

- Received: 2022-12-28
- Accepted: 2023-11-28
- Published Online: 2023-12-31

DOI :

<https://doi.org/10.32508/stdj.v26i4.4028>

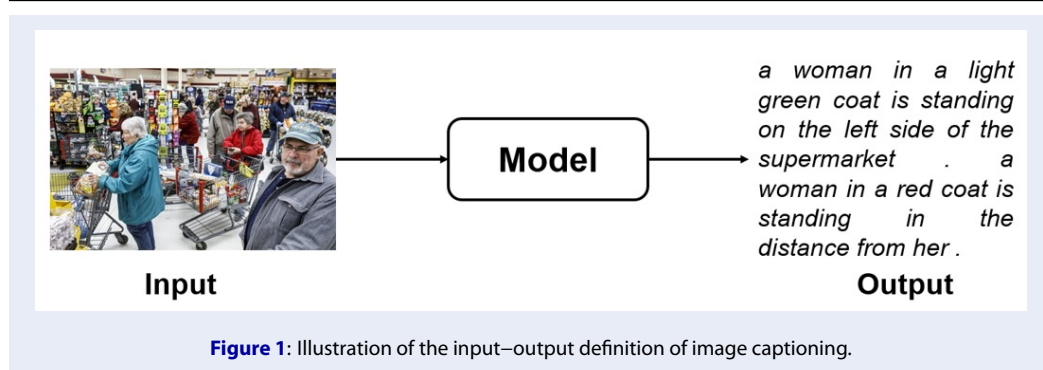


Copyright

© VNUHCM Press. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.



Cite this article : Nguyen K T T M. **Feature extraction semilearning and augmented representation for image captioning in crowd scenes.** *Sci. Tech. Dev. J.* 2023; 26(4):3128-3138.



volve high-level information extracted via convolutional neural networks. Therefore, the embedding vectors represented for grids on images, including global information, are considered.

With respect to architecture, the earliest studies from the latest studies used RNN-based models to solve image captioning problems: long short-term memory (LSTM), the transformer³, and the GPT. In the 2020s, transformer-based models were most utilized to solve image captioning problems because of their effective parallel computing mechanism.

After surveying some previous studies^{2,4-9}, we recognized three points: 1) Almost all current captioning models use the visual features extracted from the nonlearned pretrained feature extractor, which means that the feature extraction task is not updated during the training process. 2) Neither grid-based features nor region-based features are used; few studies have evaluated two of these features. 3) Recent transformer-based models only use the feature maps produced from the last encoder layer; however, there are N encoder layers. Therefore, exploring all high-level features from all encoder layers may yield better results.

- Considering the three points mentioned above, in this study, we adopt the Transformer-based model as the primary captioning model and address the main points listed below:
- We propose a feature extraction semilearning approach (FES), which simultaneously uses the feature maps extracted from a frozen pretrained model and its counterpart extracted from a learned model.
- We propose an approach that fuses two types of representations, region-based features and grid-based features; this proposed module is called the grid-based and region-based feature fusion mechanism (FGR).

- We propose a mechanism that aggregates high-level features from all stacked encoder layers called augmented encoder representations (AERs).
- The experimental results show that my approach achieves promising results and achieves the highest BLEU, METEOR, and ROUGE scores on the CrowdCaption dataset.

We structure the rest of the paper as follows: Section 2 describes my approach clearly; Section 3 reports the experimental results on the CrowdCaption dataset with careful discussion and analysis; and Section 4 summarizes the contributions and raises some problems for further research.

METHODOLOGY

Transformer-based architecture

In this subsection, we describe the Transformer-based model as the base knowledge. First, the Transformer architecture was proposed by Vaswani et al.³ in 2017 and is still commonly used. We can formulate the Transformer architecture as shown in Equations 1 and 2:

$$Z = \text{TransformerEncoder}(X) \tag{1}$$

$$Y = \text{TransformerDecoder}(X) \tag{2}$$

where X denotes the visual features of the input images and Z denotes the encoded information using the transformer encoder. Then, Z is fitted into the Transformer Decoder to learn the decoding from the encoded features to the generated caption.

Both the Transformer Encoder and Transformer Decoder include stacked layers. The outputs of the previous layers will be the inputs of the next layers. Through stacked layers, deeper and deeper information is encoded to explore the high-level information. The key idea of the stacked transformer layers is the self-attention mechanism, which is used to learn the

relation between tokens in visual features and annotated captions. Given the sequence of tokens (that is, maybe visual features or annotated captions) $S = \{s_1, s_2, \dots, s_k\}$, self-attention is formulated as Equations 3 and 4:

$$Q = SW^Q; K = SW^K; V = SW^V \quad (3)$$

$$Z = softmax\left(\frac{QK}{\sqrt{d_k}}\right) \times V \quad (4)$$

where W^Q , W^K and W^V represent the learned weight matrices used to transform the dimensions of S to fixed dimensions (called d_{model}). Q and K are used to calculate the attention weights, which indicate which tokens are worth paying more attention to, and perform matrix multiplication with V . Z is now the high-level encoded latent space. However, there are h self-attention heads used to calculate different attention aspects; then, they are concatenated together. This is called the multihead self-attention mechanism, which can be formulated as Equations 5 and 6:

$$head_i = self - attention(Q, K, V) \quad (5)$$

$$Z = Concatenate(head_1, \dots, head_h) \quad (6)$$

The full architecture of the i -th encoder layer in a Transformer Encoder can be formulated as follows (Equations 7, 8 and 9):

$$Z_i = MultiheadSelfAttention(H_i, H_i, H_i) \quad (7)$$

$$Z_i = LayerNorm(Z_i + H_i) \quad (8)$$

$$Z_i = LayerNorm(FeedForward(Z_i)) \quad (9)$$

where H_i is the encoded feature from the previous encoder layer. If $i=0$, then $H=X$. $LayerNorm$ is the normalization layer that scales the values included in the encoded features. $FeedForward$ is simply the fully connected layer.

In the Transformer Decoder, the model learns the relation features between the visual encoded features and annotated captions. This can be formulated as follows (Equations 9, 10, 11, and 12):

$$Z_i = MaskMultiheadSelfAttention(H_i, H_i, H_i, Mask) \quad (9)$$

$$Z_i = LayerNorm(Z_i + H_i) \quad (10)$$

$$Z_i = MultiheadSelfAttention(Z, Z_i, Z_i) \quad (11)$$

$$Z_i = LayerNorm(FeedForward(Z_i)) \quad (12)$$

where H_i is the encoded feature from the previous decoder layer; if $i=0$, then $H = W = \{w_1, w_2, \dots, w_n\}$, and W is the sequence of words in the annotated caption. $MaskMultiheadSelfAttention$ is simply multi-head self-attention. However, this approach masks the positions of words in future time steps. Figure 2 fully illustrates the whole architecture of the Transformer.

In the Transformer Decoder, N stacked decoder layers produce high-level relation features between visual encoded features and embedding caption features $Z^0 \in \mathbb{R}^{d_{model}}$. It is then fit into a fully connected layer followed by softmax activation to obtain the probability vector of the predicted words at each timestep:

$$\hat{Y} = \log softmax(FC(Z^o)) \quad (13)$$

Transformer architectures are widely used to solve machine translation problems; the most common example is language translation. In language translation, the Transformer Encoder is used to learn the relationships between words in sentences in the source language at the same time; these relations are subsequently formed as high-level encoded features. The Transformer Decoder takes those encoded features and sentences in the target language to learn the relationships between them, finds the patterns to learn, and translates them. However, in the image captioning problem, the Transformer encoder takes image features as input whose shape is $H \times W \times C$. Then, the approach learns the relation features between $H \times W$ feature patches instead of words in language translation problems. After that, the encoded features are processed in the same way as machine translation, which is fitted into the Transformer Decoder along with ground-truth captions to learn and generate hypothesis captions for the new samples.

Existing common captioning models, such as the Meshed-Memory Transformer⁵, AoANet⁸, and X-Lan⁹, are based on the Transformer architecture. One of the most prominent methods is the mesh-memory transformer. In detail, the mesh-memory transformer introduces the memory self-attention mechanism, which is used to enhance the quality of attention weights via prior knowledge. Moreover, a meshed-like computation was also presented to use all the features computed by the Transformer encoder layers in the Transformer Decoder. With N features from the Transformer Encoder, Cornia et al.⁵ computed N times to produce high-level features and applied a weighted sum to obtain consistent features in the Transformer Decoder. Our study proposes a new approach to feature fusion (named Augmented Encoded

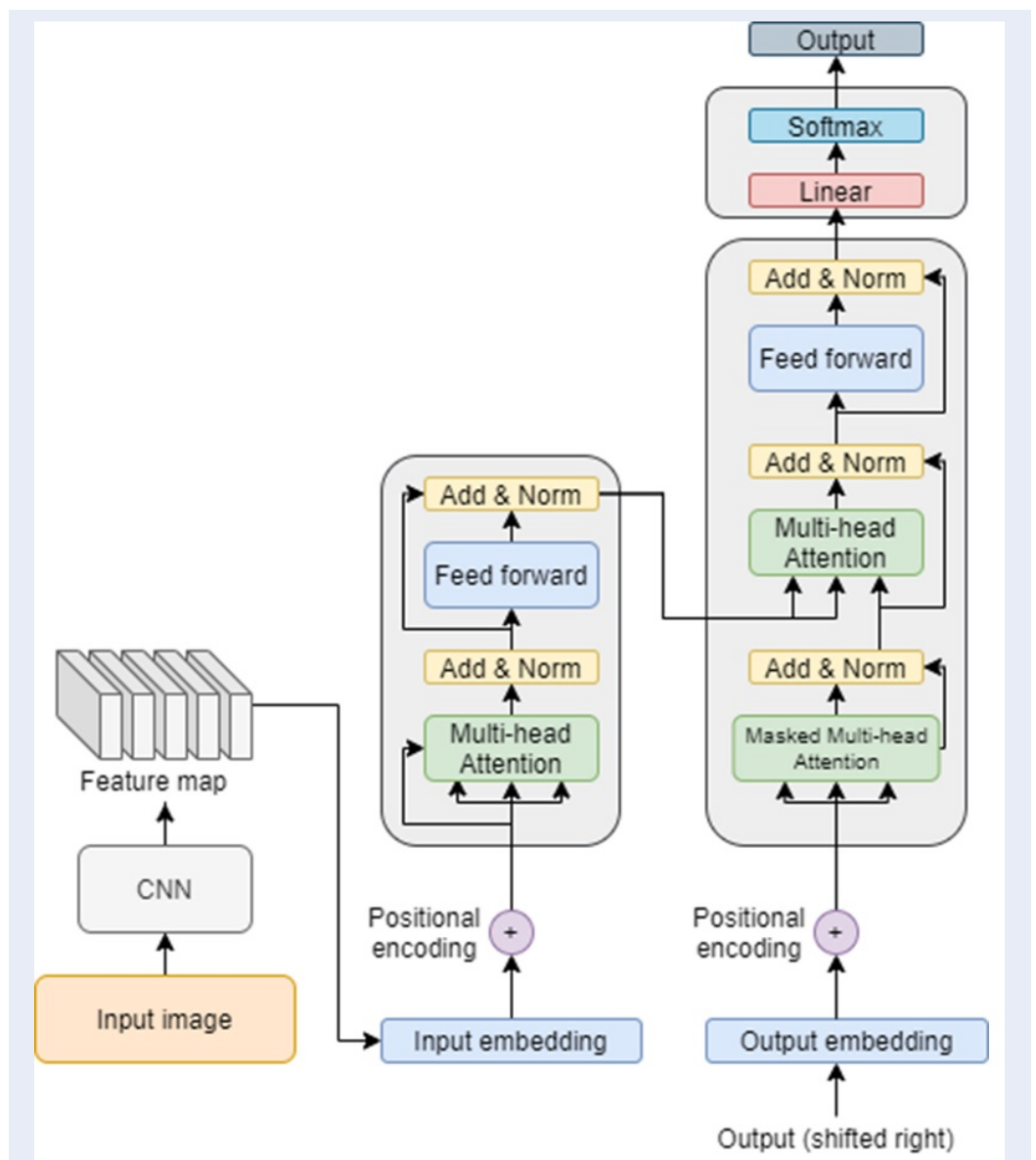


Figure 2: An illustration of the transformer-based model receiving images as inputs.

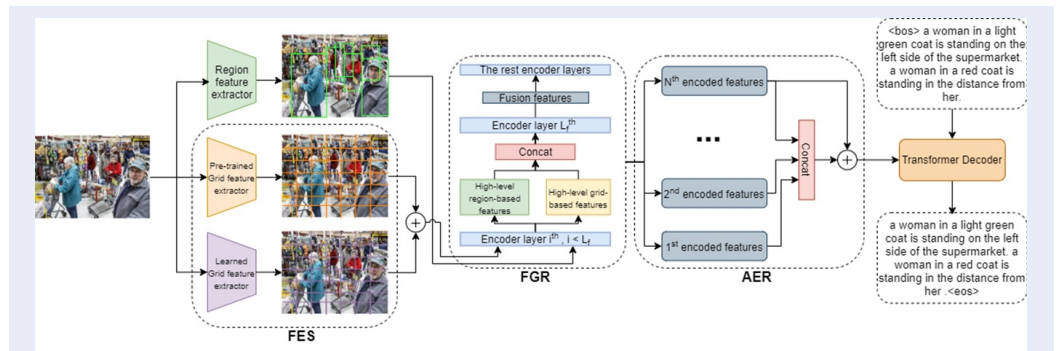


Figure 3: Illustration of our whole framework

Representation - AER), which combines features into a unified form via the Transformer Encoder and uses it to produce high-level features one time in the Transformer Decoder. Moreover, a feature fusion scheme (FGR) is presented to combine two types of features (region-based and grid-based) for better representation in the model space. With grid-based features, we also introduce a semilearning scheme (FES), which is used to effectively update a part of the feature extraction network to better fit the contexts in the experimental dataset. Our whole architecture is illustrated in Figure 3.

Fusion mechanism for region-based and grid-based features (FGR)

Region-based and Grid-based Feature Extraction

As mentioned in Section I. Introduction, there are two types of features: region-based features and grid-based features. Region-based features locally represent objects that appear in the image, while grid-based features globally represent the whole context. To extract region-based features, we use a Faster R-CNN model¹⁰ to detect all instances appearing in the input image and subsequently obtain the embedding vectors $R \in \mathbb{R}^{N \times 2048}$. Each 2048-d vector represents one detected instance, and N is the number of detected instances.

To extract the grid-based features, we use a ResNeXt-based convolutional neural network backbone¹¹ to extract high-level features. The feature maps obtained from the last convolutional layer are $G \in \mathbb{R}^{H \times W \times 2048}$. Then, we use an adaptive average pooling mechanism to compress (H×W) grids into (7×7 grids). Finally, we obtain grid-based features $G \in \mathbb{R}^{N \times 2048}$, where N is the number of grids (N=49).

Fusion Mechanism

In this section, we describe an approach for fusing these two types of features in the Transformer Encoder.

There are N stacked layers in the Transformer encoder $\{EncoderLayer_i\}_{i=0}^{N-1}$. Suppose R are region-based features and G are grid-based features; the proposed calculation process is formulated as follows:

$$\begin{cases} Z_i^r = EncoderLayer_i(R); i < l_f \\ Z_i^g = EncoderLayer_i(G); i < l_f \\ Z_i = Concatenate(Z_i^r, Z_i^g); i = l_f \\ EncoderLayer_i(Z_i); i = l_f \\ Z_i = EncoderLayer_i(Z_i), otherwise \end{cases} \quad (14)$$

In the first l_f layers, we calculate the high-level latent space of region-based features and grid-based features

separately. In the l_f^{th} encoder layer, we concatenate these two high-level representations achieved from previous encoder layers and continue to perform operations in the encoder layer. The remaining encoder layers process the fused representations of region-based and grid-based features.

However, R and G are both 2D representations (include K embedding vectors for representing K objects/grids). Therefore, we propose an approach to provide information on the relationships between grids or regions' positions in an image (bounding box coordinates).

Given K bounding box coordinates of detected objects or grids, we first calculate their center coordinates, width, and height:

$$(cx_i, cy_i) = \frac{x_i^{min} + x_i^{max}}{2}, \frac{y_i^{min} + y_i^{max}}{2} \quad (15)$$

$$w_i = (x_i^{max} - x_i^{min}) + 1 \quad (16)$$

$$h_i = (y_i^{max} - y_i^{min}) + 1 \quad (17)$$

Following⁴, the relation between bounding boxes i and j or between grids i and j is represented by r_{ij} , which is calculated as follows:

$$r_{ij} = \left(\log \frac{|cx_i - cx_j|}{w_i}, \log \frac{|cy_i - cy_j|}{h_i}, \log \frac{w_i}{w_j}, \log \frac{h_i}{h_j} \right) \quad (18)$$

$$G_{ij} = FullyConnectedLayer(r_{ij}) \quad (19)$$

$$\lambda_{ij} = ReLU(w_g^T G_{ij}) \quad (20)$$

where λ_{ij} is the final representation of the relation r_{ij} between bounding boxes i and j. With region-based features, λ_{ij} is built based on detected bounding boxes i and j, called λ_{ij}^r . With grid-based features, λ_{ij} is constructed based on grids i and j, called λ_{ij}^g .

Then, in the self-attention operation in stacked layers in the Transformer Encoder, we follow the study⁴ that adds the relation weights to the attention weights:

$$Z = softmax \left(\frac{QK}{\sqrt{d_k}} + \lambda \right) \times V \quad (21)$$

This new self-attention mechanism helps the model become aware of the relation information between bounding boxes and between grids.

Augmented encoded representation (AER)

In the Transformer Encoder, previous studies only considered the final encoded features from the last encoder layer. To enhance the final representation in the encoder, we use the approach to enhance the features from the previous layer with previously encoded features. The enhanced encoded representation is calculated by the following equations:

$$Z = \text{Concatenate}(Z_1, Z_2, \dots, Z_N) \quad (22)$$

$$Z_{final} = \beta \times \text{MLP}(Z) + Z_N \quad (23)$$

, $Z \in \mathbb{R}^{3 \times d_{model}}$ is the concatenated representation of $\{Z_i\}_{i=0}^{N-1}$. The multilayer perceptron is used to transform $3 \times d_{model}$ to d_{model} dimensions. β is the trade-off hyperparameter used to control the contribution of Z to the final representation. Because Z_N includes more valuable information from previous stacked layers, β should be quite small to ensure a sufficient contribution. In this study, we set $\beta = 0.2$.

Feature Extraction Semilearning (FES)

In crowd scenes, contexts are complicated. However, previous studies commonly use a single pretrained feature extractor to obtain high-level representations of images. This pretrained model is frozen, which means that it is only used for feature extraction, and the weights are not updated during the training process. Recognizing that learning more valuable information can be disadvantageous for the model, we propose the feature extraction semilearning scheme. In detail, given two networks: Net₁ and Net₂. Net₁ is a frozen pretrained model, which is the Faster R-CNN model mentioned in Section 2.1. Net₂ is a learned convolutional neural network model that is based on the ResNet¹² architecture. We do not update the weights in Net₁ because Net₁ is well trained on a large-scale dataset. Instead, we update the weights, including those in the last ResNet-based convolution block in Net₁, and train them end-to-end with the language model (transformer-based model). Finally, the feature maps extracted from Net₂ are added to their counterparts extracted from Net₁, but there is a trade-off hyperparameter α to control the contribution of the new high-level feature maps. The process can be formulated as follows:

$$F_1 = \text{Net}_1^{frozen}(\text{images}) \quad (24)$$

$$F_2 = \text{Net}_2^{unfrozen}(\text{images}) \quad (25)$$

$$F_{1+2} = F_1 + \alpha F_2 \quad (26)$$

where F_i is the grid-based feature, which means that we perform only feature extraction semilearning with a grid-based representation. α is set to 0.3 in this study.

RESULTS AND DISCUSSION

CrowdCaption dataset

Wang et al.² suggested that although the image captioning task has attracted the most interest from researchers worldwide, captioning for crowd scenes has rarely been explored due to the shortage of relevant datasets. To motivate the research community to bridge the gap in crowd captioning in an image captioning task, Wang et al.² constructed a novel and high-quality dataset, the CrowdCaption dataset. The CrowdCaption dataset includes 11,161 images in total, with an average of 4.4 captions per image. The average length of the captions in this dataset is 20. In this study, we use the official train-valid-test split to train the captioning model and evaluate its performance. Figure 4 shows some samples from the CrowdCaption dataset. Figure 5 shows the statistics about the numbers of images and captions in the training, validation and testing sets.

Evaluation metrics

We use four standard metrics to evaluate the performance of my proposed approach, namely, BLEU (B)¹³, METEOR (M)¹⁴, ROUGE (R)¹⁵, and CIDEr (C)¹⁶. The BLEU is a common metric used to evaluate machine translation tasks. The METEOR score uses a fragmentation measure to assess the order of unigrams between hypothesis and reference captions, while the BLEU score considers only matching. The ROUGE-L score is the ROUGE metric that considers L-gram matching; this metric is designed as a recall-related metric because the denominator used to calculate the correct percentage is the total sum of the number of L-grams appearing in the reference captions. In contrast, this number in the BLEU score is the total number of n-grams that occur in hypothesis captions. The CIDEr score considers distinguishing the n-grams that are rare or common in the vocabulary by vectorizing them using TF-IDF¹⁷ to obtain their frequency weights. Since the CIDEr score evaluates the diversity of words in generated captions, it was considered the most crucial metric in previous studies.

Implemental Details

For training my proposed model, we follow the training scheme proposed by⁷, which includes two training stages: we first train the model using cross-entropy (XE) loss:

$$l_{XE}(\theta) = -\sum_{t=1}^T \log(p_{\theta}(w_t^* | w_{1:t-1}^*)) \quad (22)$$

where θ denotes the parameters of the proposed model and $w_{1:t-1}^*$ is the target ground-truth sentence.

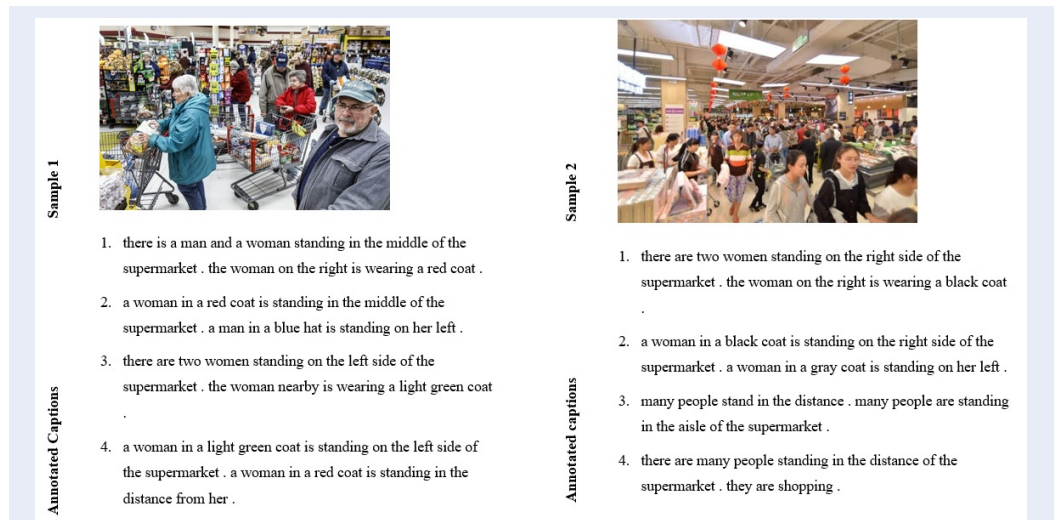


Figure 4: Some examples in the CrowdCaption dataset.

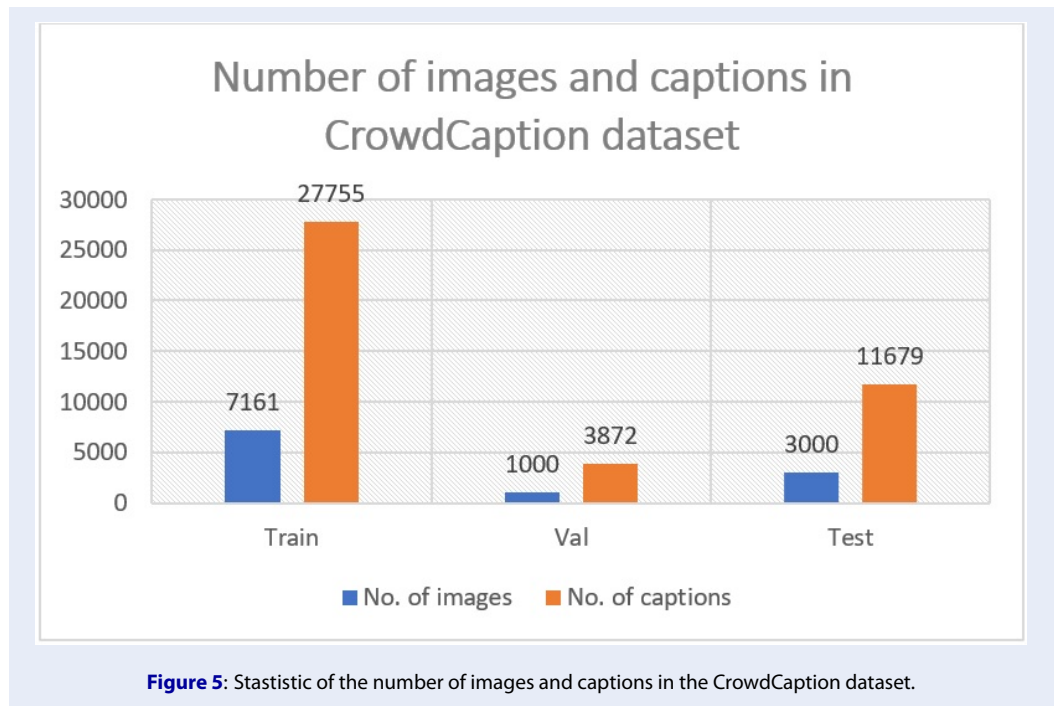


Figure 5: Statistic of the number of images and captions in the CrowdCaption dataset.

Following⁷, we apply self-critical sequence training (SCST) after training cross-entropy loss (XE) to refine the generated captions:

$$L_{RL}(\theta) = -E_{w_{1:T} \sim p_{\theta}} [r(w_{1:T})] \quad (23)$$

where the reward $r(\cdot)$ denotes the CIDEr score. For more detail about the training configuration, we set the maximum length of the generated captions to 50. There are 03 stacked encoder and decoder layers in

the Transformer model. d_{model} is set to 512. The number of heads in multihead self-attention is 8. We used Adam as the optimizer for updating the weights of the models. In the first stage (training the model using the XE loss), we used a learning rate = 0.0001. In the second stage (training the model using the SCST), we set the learning rate = 5e-6. The beam search is used to obtain the most suitable captions, and the beam size is set to 3.

Experimental Results

Ablation studies

To evaluate the performance of the proposed modules, we conduct an ablation study and report the results in Table 1. With the default setting of the transformer architecture using grid-based features as a visual representation, we achieved 29.58 B@4, 22.19 METEOR, 48.42 ROUGE-L, and 62.30 CIDEr, which are the lowest among the experiments. The region-based and grid-based feature fusion mechanism (FGR) helps to improve the results for all B@4, METEOR (+0.06), ROUGE-L (+0.47), and CIDEr (+0.84) models. This proves the effectiveness of the proposed fusion mechanism. The Transformer-based model is now aware of specific regions, including objects in images, leading to more knowledge and generating better-quality captions. The augmented encoded representation (AER) in the Transformer Encoder also obtains better results when combined with FGR. AER helps explore some hidden valuable information stored in previous encoder layers and add this information to the encoded features from the last encoder layer. This helps to enhance the representation ability of encoded features before fitting them to the Transformer Decoder. Using AER boosts the results on the CIDEr metric (+1.73). Finally, the experiments using the semilearning feature extraction mechanism along with FGR and AER improve the performance on all the metrics except for CIDEr (B@4 + 0.14, METEOR + 0.12, ROUGE-L +0.16). These results are promising because they prove that the semilearning mechanism has the ability to provide additional knowledge to the captioning model. Updating learned parameters in the last ResNet convolution block means adjusting awareness of feature extraction along with generating captions, which helps to recognize more patterns related to texts. Then, the results are improved.

Comparison to state-of-the-art models.

In this section, we report my achieved results compared to those of other state-of-the-art models in Table 2, including the following:

- Show, attend and tell¹⁸: The fundamental baseline used a convolutional neural network to extract high-level features of the image. The main captioning model was based on long short-term memory (LSTM). N stacked LSTM layers are used to decode the high-level features to generate captions. In addition, an attention mechanism was proposed to assign high weights to crucial pixels.

- ConceptualCaptions⁶: The baseline uses ResNetv2 as the feature extractor and the Transformer model as the main captioning model.
- Up-Down¹⁰: The updown method uses a Faster R-CNN model to extract region-based features. Then, LSTM layers are used to train the captioning task.
- Meshed-Memory Transformer⁵: A Transformer-based model that uses the proposed Memory Self-attention mechanism in encoder layers. The memory self-attention mechanism expands the keys and values in the self-attention mechanism with learned slots. These learned slots are used to learn prior knowledge via an attention operation. A connectivity mechanism between multilevel encoder outputs and multilevel decoder layers was also proposed.
- AoANet⁸: AoANet is a transformer-based method that uses a new self-attention mechanism called attention on attention. After performing a self-attention operation, the results are concatenated with the value matrices and then divided into two fully connected layers to learn the transformation representation. One of these two layers produces the representation called the "information vector." The other layer is followed by the sigmoid function and generates the representation called the "attention gate." These layers are subsequently combined via elementwise multiplication with the aim of obtaining more useful knowledge.
- X-LAN⁹: X-LAN is a transformer-based method that leverages the X-linear attention block as a self-attention mechanism. This new self-attention mechanism helps to build second-order interactions and jointly fuses the representations of visual features and hidden states. An X-linear attention block is used in both the Transformer Encoder and Transformer Decoder to exploit higher-order intra- and intermodal interactions between multimodal representations.

The experimental results show that our proposed module outperforms other state-of-the-art methods on the BLEU@4, METEOR, and ROUGE-L metrics. Lee et al.¹⁸ and Up-Down¹⁰ used only LSTM layers to learn to decode from high-level features to captions. Like RNN-based architectures, LSTM cannot process sentences in parallel; instead, the next predicted word should be computed by the previous hidden

Table 1: Ablation studies using FGR, AER and FES

Methods	Grid-based network	B@1	B@2	B@3	B@4	METEOF	ROUGE L	CIDEr
Transformer	X101	66.42	50.16	38.43	29.58	22.19	48.42	62.30
Transformer + FGR (proposed)	X101	66.17	50.49	39.19	30.51	22.27	48.89	63.14
Transformer + FGR + AER (proposed)	X101	66.31	50.54	39.16	30.47	22.33	48.89	64.87
Transformer + FGR + AER + FES (proposed)	X101 (frozen) + R101 (unfrozen)	66.86	50.92	39.46	30.61	22.45	49.05	64.26
Transformer + FGR + AER + FES (proposed)	X152 (frozen) + R152 (unfrozen)	66.03	50.24	38.98	30.41	22.22	48.69	65.0

Table 2: Comparison with other state-of-the-art methods

Methods	BLEU@4	METEOR	ROUGE-L	CIDEr
Show, attend and tell ¹⁸	28.29	21.47	44.71	61.78
ConceptualCaptions ⁶	28.36	21.36	43.58	63.22
Up-Down ¹⁰	29.70	22.01	45.04	64.79
Meshed-Memory ⁵	29.33	21.55	43.74	63.89
AoAnet ⁸	29.11	21.73	44.44	65.01
X-LAN ⁹	29.59	22.06	44.69	66.07
The proposed approach (X101)	30.61	22.45	49.05	64.26

units. Although gates are designed to avoid the lack of memory through the decoding step, information has still not been fully explored to generate high-quality captions. ConceptualCaptions⁶, Meshed-memory⁵, AoAnet⁸, and X-LAN⁹ are all transformer-based architectures. However, they have the same disadvantages: only region-based features are used to represent images; these features provide only foreground information and lack background signals. For example, to clearly describe the interactions between people, people should use not only the features of people but also the general high-level features of the scenes around them. Despite the significant improvement in the self-attention mechanism, these models cannot perform well on the CrowdCaption dataset.

The BLEU@4, METEOR, and ROUGE-L metrics are used to evaluate the similarity between two sentences. These higher values of these metrics prove that the captions generated from my proposed modules are closer to the ground-truth sentences. However, the diversity of words may still be limited, which is the

reason why the CIDEr score is lower than that of other methods.

Qualitative Results

To provide a better illustration of how much of my proposed modules help the Transformer-based model more effectively, we show some generated captions compared to ground-truth sentences in Figure 6. Words that describe the context nearly accurately are highlighted in green. In contrast, red text is unsuitable for expressing events or characteristics appearing in images. By observation, we explore whether our model is better at increasing awareness of objects' positions. In the third example picture in Figure 6, our proposed approaches describe exactly the man's positions and the shirt's color. For example, in the second picture, our model recognized that only a group of women standing in front are holding the tennis. In contrast, the Transformer model wrongly supposes that the men on the back are holding it. With an unclear context similar to that of the first example pic-

ture, our model seems to be more sensitive when the group of people is taking photos, which is the most suitable description for this picture.

CONCLUSION

In this paper, we propose a region-based and grid-based feature fusion mechanism for taking advantage of these representations in the Transformer-based model (FGR). In addition, augmented encoded representation (AER) is investigated to provide valuable transformer information produced from all the encoder layers. Finally, the proposed semilearning approach (FES) helps the captioning model learn to adjust the feature extraction task, aiming to produce better-quality high-level features that improve performance. All of these proposed modules are used to enhance the results for image captioning in crowd scenes. The experimental results show that my proposed approach is promising and comparable with other existing state-of-the-art methods. In the future, we aim to address some current drawbacks by proposing graph-based representations for images or embedding semantic features in captioning models to capture more valuable information between humans in various contexts.

ABBREVIATIONS

None.

FUNDING

None.

CONFLICT OF INTEREST

The author declares that they have no competing interests.

REFERENCES

- Chen X, Fang H, Lin TY, Vedantam R, Gupta S, Dollár P et al. 2015. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325;.
- Wang L, Li H, Hu W, Zhang X, Qiu H, Meng F et al. What happens in crowd scenes: A new dataset about crowd scenes for image captioning. IEEE Trans Multimedia. 2022;25:5400-12; Available from: <https://doi.org/10.1109/TMM.2022.3192729>.
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30;.
- Herdade S, Kappeler A, Boakye K, Soares J. Image captioning: transforming objects into words. Adv Neural Inf Process Syst. 2019;32;.
- Cornia M, Stefanini M, Baraldi L, Cucchiara R. Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 10575-84; Available from: <https://doi.org/10.1109/CVPR42600.2020.01059>.
- Sharma P, Ding N, Goodman S, Soricic R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th annual meeting of the Association for Computational Linguistics (volume 1: long Papers); 2018, July. p. 2556-65; PMID: 30441875. Available from: <https://doi.org/10.18653/v1/P18-1238>.
- Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1179-95; Available from: <https://doi.org/10.1109/CVPR.2017.131>.
- Huang L, Wang W, Chen J, Wei XY. Attention on attention for image captioning. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 4633-42; Available from: <https://doi.org/10.1109/ICCV.2019.00473>.
- Pan Y, Yao T, Li Y, Mei T. X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 10968-77; Available from: <https://doi.org/10.1109/CVPR42600.2020.01098>.
- Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S et al. Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 6077-86; Available from: <https://doi.org/10.1109/CVPR.2018.00636>.
- Jiang H, Misra I, Rohrbach M, Learned-Miller E, Chen X. In defense of grid features for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 10264-73; Available from: <https://doi.org/10.1109/CVPR42600.2020.01028>.
- He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-8; PMID: 26180094. Available from: <https://doi.org/10.1109/CVPR.2016.90>.
- Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics; 2001. p. 311-8; Available from: <https://doi.org/10.3115/1073083.1073135>.
- Denkowski M, Lavie A. Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation; 2014, June. p. 376-80; Available from: <https://doi.org/10.3115/v1/W14-3348>.
- ROUGE LC. A package for automatic evaluation of summaries. In: Proceedings of the workshop on text summarization of ACL, Spain; 2004, July;.
- Vedantam R, Zitnick CL, Parikh D. Cider: consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 4566-75; Available from: <https://doi.org/10.1109/CVPR.2015.7299087>.
- Robertson S. Understanding inverse document frequency: on theoretical arguments for IDF. J Doc. 2004;60(5):503-20; Available from: <https://doi.org/10.1108/00220410410560582>.
- Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R et al. Show, attend and tell: neural image caption generation with visual attention. In: International conference on machine learning; 2015, June. p. 2048-57. PMLR;.




	<p>Ground-truth: on the left side of the tree stand five people . on the far left is a man wearing sunglasses .</p> <p>Transformer: there are many people standing in the square . they are watching a performance .</p> <p>Transformer + FGR + AER: a group of people are standing on the left side of the crowd . they are looking at something .</p> <p>Transformer + FGR + AER + FES (X101): a few men and women are standing in a row . they are all wearing sunglasses .</p> <p>Transformer + FGR + AER + FES (X152): some people are standing in front of the crowd . they are taking photos .</p>		<p>Transformer + FGR + AER + FES (X152): there is a group of people behind . they stand behind .</p> <p>Ground-truth: there are two men on the left . the man on the right among them wears a white shirt .</p> <p>Transformer: there are three men in the distance . the man on the far left of them is wearing glasses .</p> <p>Transformer + FGR + AER: there is a man in a white shirt in the distance . there are two men sitting in front of him .</p> <p>Transformer + FGR + AER + FES (X101): there are two men on the left . the man on the right is wearing a white shirt .</p> <p>Transformer + FGR + AER + FES (X152): there are two men in the back . the man on the right of them is wearing a white shirt .</p>
	<p>Ground-truth: there is a group of women holding tennis rackets in front . they are all wearing white vests .</p> <p>Transformer: there is a group of men and women in front . the man on the far right among them is holding a tennis racket .</p> <p>Transformer + FGR + AER: there is a group of people in the back . they stand .</p> <p>Transformer + FGR + AER + FES (X101): there is a group of women in front . they are all holding tennis rackets .</p>		

Figure 6: Illustration of generated captions from the proposed approach.