International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/ https://doi.org/10.22362/ijcert.v10i12.911
Volume 10, Issue 12, December 2023
© 2023, IJCERT All Rights Reserved

*Research Paper*

# Enhancing Image Descriptions with Image Transformers: A Journey into Advanced Image Captioning

Sushma Jaiswal[1*], Harikumar Pallthadka[2], Rajesh P. Chinchewadi[3], Tarun Jaiswal[4]

[1] *Guru Ghasidas Central University, Bilaspur (C.G.) and Post-Doctoral Research Fellow, Manipur International University, Imphal, Manipur, jaiswal1302@gmail.com , https://orcid.org/0000-0002-6253-7327*
[2] *Manipur International University, Imphal, Manipur, vc@miu.edu.in, https://orcid.org/0000-0002-0705-9035.*
[3] *Manipur International University, Imphal, Manipur, rajesh.cto@miu.edu.in .*
[4] *National Institute of Technology, Raipur, tjaiswal_1207@yahoo.com, https://orcid.org/0000-0003-3963-4548*

**Sushma Jaiswal: jaiswal1302@gmail.com*

Received: 17/10/2023, Revised: 23 /11/2023, Accepted: 09/12/2023 Published: 30/12/2023

**Abstract:** - *This research investigates using Image Transformers to improve image captioning. Deep learning algorithms have improved image captioning. We use Image Transformers, a sophisticated neural network design that captures intricate visual aspects and interactions in images, to improve image captioning. This study describes the creation and implementation of our innovative image captioning system, which seamlessly integrates Image Transformers with classic captioning structures. We show significant advances in visual description accuracy and contextually through trials and evaluations. This method improves image-caption correspondence by providing more exact and meaningful captions and a greater grasp of the visual material. This research has many applications, including visual impairment accessibility, media and marketing content development, and more. This study shows how Image Transformers might alter image captioning and usher in a new era of immersive and context-aware human-machine interaction.*

**Keywords**- *Image transformer, Image-Captioning, Deep Learning, Attention model, Contrastive Loss.*

## 1. Introduction

Image captioning is a key technique with many applications in computer vision and natural language processing. This research uses image Transformers, a transformative deep learning architecture development, to improve image captioning. Image captioning has advanced greatly in recent years. As demand for fuller, contextually-aware descriptions rises, emerging ways that bridge visual perception and language interpretation are needed. Image Transformers, inspired by their success in visual tasks, offer an intriguing chance to revolutionize image captioning. These transformers excel in capturing rich visual features and relationships, making them excellent for improving image descriptions. Image Transformers combined with classic image captioning structures should improve caption accuracy, relevance, and contextual depth. The Transformer architecture dominates most natural language processing workloads nowadays. When given a new natural language processing task, the typical first step is to acquire a large pretrained Transformer model like BERT [1], ELECTRA [2], RoBERTa [3], or Longformer [4], adapt the output layers, and fine-tune the model with the available data. The Transformer-based models GPT-2 and GPT-3 [5], [6] have been discussed in recent years of enthusiastic news coverage of OpenAI's comprehensive language models. The Vision Transformer is now the default model for image recognition, object detection, semantic segmentation, and super-resolution [7] [8]. Other than natural language processing, transformers have competed in speech recognition [9], reinforcement learning [10], and graph neural networks [11].

The Transformer model uses the attention mechanism, which was first described as an augmentation for encoder-decoder Recurrent Neural Networks (RNNs) used in sequence-to-sequence tasks like machine translation [12]. The encoder compressed the input sequence into a single fixed-length vector and handed it to the decoder in early sequence-to-sequence machine translation models [13]. Attention is innovative since it suggests that the decoder may benefit from iterating the input sequence rather than compressing it. Instead of always using a uniform input representation, the decoder should focus on certain input sequence segments during decoding phases. Bahdanau's attention method lets the decoder dynamically attend to input segments at each level. First, the encoder constructs a representation with the same length as the input sequence. During decoding, the decoder can receive a context vector, which is the weighted sum of the input representations at each time step. These weights intuitively determine how much each step's context "attends" to each input character. This weight assignment algorithm must be differentiable to train with other neural network parameters.

The contributions of the paper can be summarized as follows:

- With a redesigned attention module appropriate for the intricate natural structure of image areas, we suggest a novel internal architecture for the transformer layer that is tailored to the image captioning task. The research proposed using Image Transformers to improve image captions quality and context.
- Our suggested architecture was validated by extensive experiments and ablation studies, resulting in state-of-the-art performance on the MSCOCO image captioning offline and online testing dataset using just region features as input.
- By leveraging the power of Image Transformers, the paper pushes the boundaries of traditional image captioning, offering a more advanced and accurate captioning methodology. The integration of Image Transformers enables the generation of image descriptions that are not only more precise but also more contextually meaningful and semantically rich.
- The research discusses the potential applications of this approach, including improving accessibility for the visually impaired, content generation for media and marketing, and advancing human-machine interaction.
- The paper likely includes experimental results and evidence demonstrating the effectiveness of the Image Transformer-based image captioning approach, supporting its contributions with empirical data.

## 2. Related Work

### 2.1 Deep Captioning Methods

This section describes deep learning-based captioning techniques, including attention mechanism, encoder-decoder architecture-based captioning, compositional architecture-based captioning, new object-based captioning, semantic concept-based captioning, dense captioning, and styled captioning.

#### 2.1.1 Encoder-Decoder architecture-based captioning

One of the main paradigms in the field of image captioning is the Encoder-Decoder architecture. It involves a decoder that produces logical and contextually appropriate textual descriptions and an encoder that comprehends an image's visual content. This section sheds light on the importance of the Encoder-Decoder architecture in the field of image captioning by explaining its essential elements and operational mechanisms.

The authors [14] discussed about models that improved natural language creation and interpretation by fusing text with various input kinds, such images or audio. They presented a novel method for gathering and combining data from several modalities by utilizing a neural network framework. By offering insights into the creation of models able to handle various data types in an integrated way, this research added to the fields of machine learning and natural language processing.

An innovative method for captioning images that takes into account scene-specific circumstances and region-based attention was presented by Fu et al. [15] in their study. The authors introduced a methodology that dynamically matches the image's regions of interest and the accompanying textual descriptions, addressing the difficulty of creating descriptive and contextually appropriate captions for photographs.

Deep visual-semantic alignments were studied by Karpathy and Fei-Fei [16],[17] to generate descriptions of images. Their study focused on how to improve the quality of image captions by aligning visual content with related semantic information. The work showed how deep learning models may be used to provide descriptive and contextually rich image descriptions, making a substantial contribution to the fields of computer vision and natural language processing. This study shed important light on the growing relationship between natural language processing and computer vision.

The paper [18] investigated the "show, attend, and tell" method, which uses visual attention mechanisms to improve image captions. By showing that visual attention works in image captioning, the study advanced machine learning and computer vision.

A unique image caption generating method using visual attention mechanisms was presented in this research. The "show, attend, and tell" strategy improved image captions using visual attention, according to the study. This work advanced machine learning and computer vision by emphasizing visual attention in image captioning [19]. Zhao et al. [20] studied multi-task learning to improve image captioning. The research examined methods that allow a single model to handle numerous related jobs simultaneously, improving image caption generation. This work advanced artificial intelligence, particularly image captioning.

The investigation used a transformer model with multi-view visual representations to improve image captioning. The study showed that diverse visual viewpoints improve video technology and image captioning [21].

### 2.1.2 Compositional architecture-based captioning

"Rich Image Captioning in the Wild" [22], study examined the difficulty of writing meaningful descriptions for images under uncontrolled settings. The article may have introduced new approaches to this complicated challenge, advancing computer vision and image captioning. Computer vision researchers studied ways to bridge natural language captions with visual notions. Their research may have suggested novel translation methods, revealing how text and visual data interact [23] .Ma and Han [24] presented "Describing Images by Feeding LSTM with Structural Words." Integrating LSTM with structural words was an innovative technique to visual description creation in this work. The research may have improved image captioning by using structural linguistic components. This research highlighted the importance of linguistic structures in image description in multimedia and natural language processing.

Recurrent multi-modal learning with fusion was used to generate visual descriptions in a novel way. The research may have suggested merging visual and linguistic data to improve image captioning. This work advanced image processing, notably in descriptive and helpful image descriptions [25]. A parallel-fusion architecture that blends RNN and LSTM models yielded an innovative image caption generation method. The study presumably proposed a new image caption generation method that uses RNN and LSTM to parallelize information [26].

### 2.1.3 Multi-modal space-based captioning

Kulkarni et al. [27] research examined visual comprehension and description approaches, particularly their application in real life. Simple but effective visual descriptions were stressed in the study, which advanced pattern analysis and machine intelligence.

A study used multimodal recurrent neural networks to explain images. The research possibly suggested multimodal methods for understanding visuals and providing meaningful explanations [28]. The authors [29] developed a novel method for creating image captions using a recurrent visual representation. Recurrently processing visual data may have been used to provide contextually rich image descriptions in the study. Their work advanced computer vision and image captioning by emphasizing the need for dynamic visual representation in caption generation.

### 2.1.4 Semantic Concept-based captioning

Semantic Compositional Networks were added to improve visual captioning in the study [30]. The study presumably suggested ways to create image captions that capture compositional and semantic components of visual material. Their work advanced computer vision and image captioning by emphasizing the significance of capturing images' complex semantics. Yao et al. [31] used visual content features to improve image captioning. The study may have suggested ways to generate image captions with descriptive features for a more thorough description.

### 2.1.5 Novel Object-based captioning

Based on children's like learning processes, the study used sentence descriptions of images to quickly learn new visual concepts. The study suggested efficient machine learning methods for swiftly learning new visual concepts. This work advanced computer vision and image understanding by revealing more effective concept learning [32]. Hendricks et al. [33] developed a novel method for creating image descriptions for new item categories without associated training data. The study suggested ways to generate informative descriptions for objects the model hasn't seen.

A method for creating image captions that describe several objects and scenes was developed. This study certainly suggested ways to describe images in different and contextual ways. Their work advanced computer vision and image captioning by emphasizing the need of managing varied objects and situations [34].

### 2.1.6 Stylized captioning

Mathews et al. [35] introduced a novel sentiment-based image description method. The study may have suggested ways to write emotional image captions. Their work advanced artificial intelligence and image captioning by including sentiment analysis.

Nezami et al. [36] offered a new way to caption images by analyzing facial expressions. The study may have introduced new ways to caption images that capture participants' feelings and sentiments through facial expressions. Machine learning and image captioning advanced with their emphasis on facial expression analysis in caption production.

Personalized image captioning using context sequence memory networks was pioneered. The study presumably suggested ways to write image captions for specific settings [37]. Liu et al. [38] presented a novel method for captioning images utilizing multi-level policies and reward reinforcement learning. It is possible that the study suggested methods for producing image captions through the integration of several tiers of incentives and policies. Cornia et al. [39] used a meshed-memory transformer model to present a novel method of captioning images. The study probably included methods for creating captions for images that are descriptive by using memory structures. Their research significantly advanced the fields of image captioning and computer vision.

### 2.1.7 Dense captioning

Johnson et al. [40] presented a novel use of fully convolutional localization networks for dense captioning. The work probably offered methods for producing detailed and comprehensive image captions that are highly compatible with object localization.

Yang et al. [41] presented a novel method of dense captioning that included visual context and cooperative inference. The study probably offered methods that take into account both context and joint inference procedures to produce rich and intricate image captions.

## 3. Image Transformer

It is a difficult but important task in the fields of computer vision and natural language processing to produce precise and contextually relevant descriptions for images. Conventional methods often use recurrent neural networks (RNNs) for sequence generation after convolutional neural networks (CNNs) for extracting image data. Nonetheless, a class of attention-based models called image Transformers has shown promising results across a range of tasks, encouraging further research into improving image descriptions.

This section outlines the encoding and decoding components of the suggested image transformer architecture after reviewing the transformer layer.

### 3.1 Image transformer layer

The model may concurrently attend to data from several representation subspaces at various places thanks to multi-head attention. The optimal number of heads and the value of multi-head attention are hot topics of discussion.

Simply consider the following: denote by $\mathcal{D} \overset{\text{def}}{=} \{(\mathbf{k}_1, \mathbf{v}_1), \ldots (\mathbf{k}_m, \mathbf{v}_m)\}$ a database of $m$ tuples of keys and values. Moreover, denote by q a query. Then we can define the attention over $\mathcal{D}$ as

$$\operatorname{Attention}(\mathbf{q}, \mathcal{D}) \overset{\text{def}}{=} \sum_{i=1}^{me} \alpha(\mathbf{q}, \mathbf{k}_i) \mathbf{v}_i$$
(1)

Where $\alpha(\mathbf{q}, \mathbf{k}_i) \in \mathbb{R}(i = 1, \ldots, m)$ are scalar attention weight. The operation itself is typically referred to as attention pooling. The name attention derives from the fact that the operation pays particular attention to the terms for which the weight $\alpha$ is significant (i.e., large). As suc the attention over $\mathcal{D}$ generates a linear combination of values contained in the database. In fact, this contains the above example as a special case where all but one weight is zero. We have a number of special cases:

- The weights $\alpha(\mathbf{q}, \mathbf{k}_i)$ are nonnegative. In this case the output of the attention mechanism is contained in the convex cone spanned by t values $\mathbf{v}_i$.
- The weights $\alpha(\mathbf{q}, \mathbf{k}_i)$ form a convex combination, i.e., $\sum_i \alpha(\mathbf{q}, \mathbf{k}_i) = 1$ and $\alpha(\mathbf{q}, \mathbf{k}_i) \geq 0$ for all $i$. This is the most common setting i deep learning.
- Exactly one of the weights $\alpha(\mathbf{q}, \mathbf{k}_i)$ is 1 , while all others are 0 . This is akin to a traditional database query.
- All weights are equal, i.e., $\alpha(\mathbf{q}, \mathbf{k}_i) = \frac{1}{m}$ for all $i$. This amounts to averaging across the entire database, also called average pooling in deep learning.

A common strategy for ensuring that the weights sum up to 1 is to normalize them via

$$\alpha(\mathbf{q}, \mathbf{k}_i) = \frac{\alpha(\mathbf{q}, \mathbf{k}_i)}{\sum_j \alpha(\mathbf{q}, \mathbf{k}_j)}. \qquad (2)$$

In particular, to ensure that the weights are also nonnegative, one can resort to exponentiation. This means that we can now pick any function $a(\mathbf{q}, \mathbf{k})$ and then apply the softmax operation used for multinomial models to it via

$$\alpha(\mathbf{q}, \mathbf{k}_i) = \frac{\exp(a(\mathbf{q}, \mathbf{k}_i))}{\sum_j \exp(a(\mathbf{q}, \mathbf{k}_j))} \qquad (3)$$

This operation is readily available in all deep learning frameworks. It is differentiable and its gradient never vanishes, all of which are desirable properties in a model.

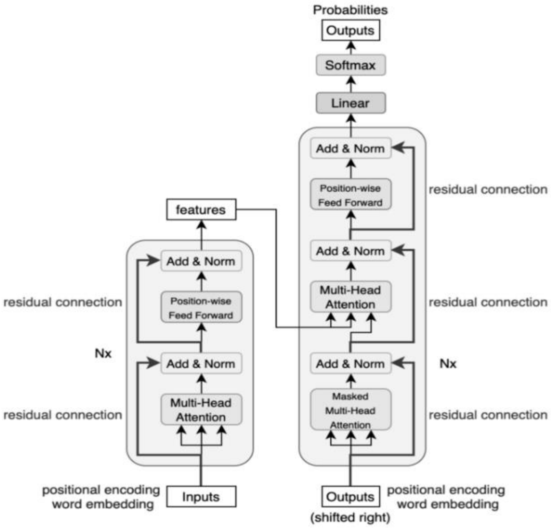
Before providing the implementation of multi-head attention, let's formalize this model mathematically. Given a query $\mathbf{q} \in \mathbb{R}^{d_q}$, a key $\mathbf{k} \in \mathbb{R}^{d_k}$, and a value $\mathbf{v} \in \mathbb{R}^{d_v}$, each attention head $\mathbf{h}_i (i = 1, \ldots, h)$ is computed as

$$\mathbf{h}_i = f\left(\mathbf{W}_i^{(q)} \mathbf{q}, \mathbf{W}_i^{(k)} \mathbf{k}, \mathbf{W}_i^{(v)} \mathbf{v}\right) \in \mathbb{R}^{p_v} \quad (4)$$

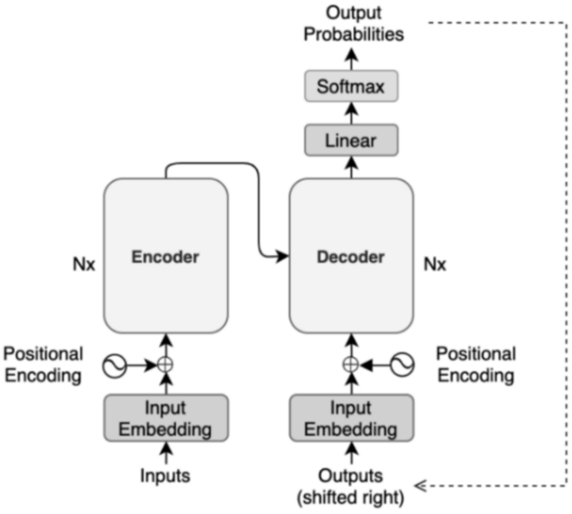
where $\mathbf{W}_i^{(q)} \in \mathbb{R}^{p_q \times d_q}, \mathbf{W}_i^{(k)} \in \mathbb{R}^{p_k \times d_k}$ , and $\mathbf{W}_i^{(v)} \in \mathbb{R}^{p_v \times d_v}$ are learnable parameters and $f$ is attention pooling, such as additive attention and scaled dot product. The multi-head attention output is another linear transformation via learnable parameters $\mathbf{W}_o \in \mathbb{R}^{p_o \times h p_v}$ of the concatenation of $h$ heads:

$$\mathbf{W}_o \begin{bmatrix} \mathbf{h}_1 \\ \vdots \\ \mathbf{h}_h \end{bmatrix} \in \mathbb{R}^{p_o}. \qquad (5)$$

Based on this design, each head may attend to different parts of the input. More sophisticated functions than the simple weighted



(a)



(b)

**Figure. 1:** (a) & (b) Proposed Architecture

### 3.2 Graph Attention Networks

Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E},)\ with\ a\ set\ of\ node\ features$:

$$\mathbf{h} = \{\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_N\}, \vec{h}_i \in \mathbb{R}^F \quad (6)$$

Where $|\mathcal{V}| = N$ and $F$ is the number of features in each node. The input of graph attention layer is just the set of node features **h**, and the output of this layer is a new set of node features

$$\mathbf{h}' = \{\vec{h}'_1, \vec{h}'_2, \ldots, \vec{h}'_N\}, \vec{h}'_i \in \mathbb{R}^{F'} \quad (7)$$

Graph attention layer leverages the multi-head attention, so for every single node $v_i$ :

- Obtain higher-level feature embedding for $v_i$ by a shared linear transformation, which is parameterized by a weight matrix $\mathbf{W} \in \mathbb{R}^{F' \times F}$.
- Perform self-attention on the nodes - a shared attentional mechanism

$$a: \mathbb{R}^{F'} \times \mathbb{R}^{F'} \rightarrow \mathbb{R} \qquad (8)$$

Computes attention score or coefficients on other nodes $v_j$ in the graph:

$$e_{ij} = a\left(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j\right) \qquad (9)$$

That indicate the importance of node $v_j$ 's features to node $v_i$. However, in GAT, it only attends the first order neighbors of $v_i$ (including $v_i$ ), and

$$e_{ij} = a\left(\mathbf{W}\vec{h}_i, \mathbf{W}\vec{h}_j\right) = \text{LeakyReLU}\left(\vec{\mathbf{a}}^\top\left[\mathbf{W}\vec{h}_i \parallel \mathbf{W}\vec{h}_j\right]\right) \qquad (10)$$

where $\vec{\mathbf{a}} \in \mathbb{R}^{2F'}$ and $\parallel$ is the concatenation operation.

- normalize the attention score with softmax function to obtain the attention distribution:

$$\alpha_{ij} = \text{softmax}_j\left(e_{ij}\right) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})} \quad (11)$$

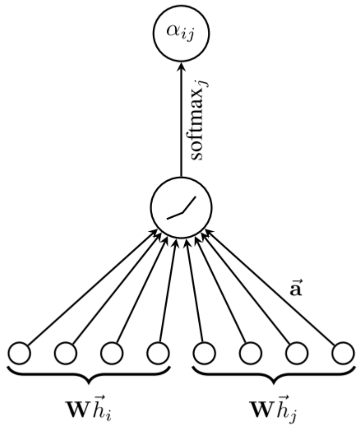


**Figure. 2:** The attention mechanism employed by [7]

Compute weighted sum of other nodes' features to serve as the final output feature (after potentially applying a nonlinearity, σ ):

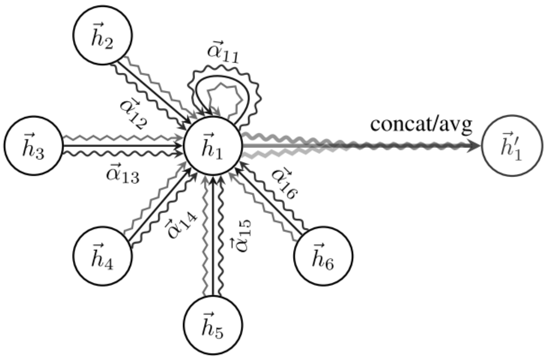


**Figure. 3.** An illustration of multi-head attention (with K=3 heads) by node v1 on its neighbors. Different arrow styles and colors denote independent attention computations [7].

$$\vec{h}'_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}\vec{h}_j\right) \qquad (12)$$

For multi-head attention, $K$ independent attention mechanisms execute the transformation of equation (14) by using $K$ different matrices $W$, and then their features are concatenated:

$$\vec{h}'_i = \Vert_{k=1}^{K} \sigma\left(\sum_{j \subset \mathcal{N}_i} \alpha_{ij}^k \mathbf{W}^k \vec{h}_j\right) \qquad (13)$$

If it is final (prediction layer) of the network with multi-head attention, then the output becomes:

$$\vec{h}'_i = \sigma\left(\frac{1}{K}\sum_{k=1}^{K} \sum_{j \in \mathcal{N}_\mathrm{i}} \alpha_{ij}^k \mathbf{W}^k \vec{h}_j\right) \quad (14)$$

### 3.2 Decoder

Proposed decoder consists of a LSTM [28] layer and an implicit transformer decoding layer, which we proposed to decode the diverse information in a region in the image. The LSTM layer is a common memory module and the transformer layer infers the most relevant region in the image through dot product attention. At first, the LSTM layer receives the mean of the output $\left(\bar{A} = \frac{1}{N}\sum_{i=1}^{N} A'_i\right)$ from the encoding transformer, a context vector $(c_{t-1})$ at last time step and the embedded feature vector of current word in the ground truth sentence:

$$\begin{aligned} x_t &= [W_e\pi_t, \bar{A} + c_{t-1}] \\ h_t, m_t &= \mathrm{LSTM}(x_t, h_{t-1}, m_{t-1}) \end{aligned} \quad (15)$$

Where, $W_e$ is the word embedding matrix, $\pi_t$ is the $t^{\text{th}}$ word in the ground truth. The output state $h_t$ is then transformed linearly and treated as the query for the input of the implicit decoding transformer layer. The difference between the original transformer layer and our implicit decoding transformer layer is that we also widen the decoding transformer layer by adding several sub-transformers in parallel in one layer, such that each sub-transformer can implicitly decode different aspects of a region. It is formalized as follows:

$$A^D_{t,i} = \mathrm{MultiHead}\left(W_{DQ}h_t, W_{DK_i}A', W_{DVi}A'\right) \quad (16)$$

Then, the mean of the sub-transformers' output is passed through a gated linear layer (GLU) [27] to extract the new context vector $(c_t)$ at the current step by channel:

$$c_t = \mathrm{GLU}\left(h_t, \frac{1}{M}\sum_{i=1}^{M} A^D_{t,i}\right) \quad (17)$$

The context vector is then used to predict the probability of word at time step $t$:

$$p(y_t \mid y_{1:t-1}) = \mathrm{Softmax}(w_p c_t + b_p) \quad (18)$$

The overall architecture of our model is illustrated in Fig. 1, and the difference between the original transformer layer and our proposed encoding and decoding transformer layer.

### 3.4 Training

Given a target ground truth as a sequence of words $y^*_{1:T}$, for training the model parameters $\theta$, we follow the previous method, such that we first train the model with cross-entropy loss:

$$L_{XE}(\theta) = -\sum_{t=1}^{T} \log\left(p_\theta(y^*_t \mid y^*_{1:t-1})\right) \quad (19)$$

Then followed by self-critical reinforced training [29] optimizing the CIDEr score [30] :

$$L_R(\theta) = -E_{(y_{1:T}\sim p_\theta)}[r(y_{1:T})] \quad (20)$$

Where $r$ the score function and the gradient is is approximated by:

$$\nabla_\theta \approx -\left(r(y^s_{1:T}) - (\hat{y}_{1:T})\right)\nabla_\theta \log p_\theta(y^s_{1:T}) \quad (21)$$

## 4. Result and Discussion

### 4.1 Computational Costs

We use Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz, 3.19 GHz with 8.00 GB RAM, 64-bit operating system, Windows 10 Pro, x64-based processor to train and test all our models.

### 4.2 Dataset and Evaluation Metrics

On the MSCOCO image captioning dataset [42], we train our model. We use Karpathy's splits [32], with 113287 training images, 5000 validation images, and 5,000 test images. Ground truth captions are 5 per image. Discard terms less than 4 times and the vocabulary is 10,369. Our model is tested on Karpathy's offline (5,000) and MSCOCO online (40,775) datasets. We employ Bleu [43], METEOR [44], ROUGE-L [45], and CIDEr [46] as assessment metrics.

### 4.3 Result and Discussion

Below is a detailed comparison of how different models perform in the field of sophisticated image captioning. The main metrics and gains noted for each model are summarized in the *Table 1*:

Strong BLEU-4 scores (88.9) are obtained by Liu et al. [47], demonstrating a high degree of n-gram overlap with reference captions.

Yao et al. [48]: Shows strong language generation with competitive BLEU-4 scores (90.4). Vinyals et al. [49]: Performs well in BLEU-4 (80.2), albeit a little less well than the top models.Jia et al. [50]: Maintains competitiveness in n-gram matching with a decent BLEU-4 score (77). Fang et al. [51]: Provides a decent BLEU-2 score (88), but the analysis's depth is limited by the absence of BLEU-3 and BLEU-4 evaluations. You [52]: Get an impressive BLEU-4 score (81.5), which shows that you are capable of producing precise and varied captions. Yao [53]: Exhibits a remarkable capacity to replicate significant words, as seen by her high BLEU-2 score of 95.9. According to Liu et al. [47], thorough caption evaluation is indicated by a balanced performance across METEOR, ROUGE-L, and CIDEr criteria.

Strong in METEOR, ROUGE-L, and CIDEr consistently, indicating a well-rounded model, according to Yao et al. [48]. According to Vinyals et al. [49], there is a competitive METEOR score of 40.7, but the ROUGE-L (69.4) and CIDEr (30.9) values are significantly lower. Jia et al. [50]: Keeps overall competitiveness while exhibiting respectable performance across METEOR, ROUGE-L, and

CIDEr measures. Fang et al. [51]: The complete assessment of its captioning quality is limited because to the absence of METEOR and ROUGE-L evaluations. You [52]: Shows proficiency in multiple areas of caption creation with a balanced performance across METEOR, ROUGE-L, and CIDEr. Yao [53]: High METEOR and CIDEr scores, showing efficacy in consensus and fluency, with scant ROUGE-L data. Liu et al. [47], Yao et al. [48], and You [52]: Prove themselves to be powerful competitors, outperforming on several measures. While Jia et al. [50] and Vinyals et al. [49] demonstrate competitive performance, there is need for improvement in a few areas. A thorough evaluation of its advantages and disadvantages is hampered by the lack of information, according to Fang et al. [51]. Yao [53]: Shows mastery in particular measures, but does not provide a thorough analysis of all factors taken into account. GCN-LSTM [53]: Achieves BLEU-1 (B1) score of 80.8 and BLEU-4 (B4) score of 95.9. Also performs well in other metrics. AUTO-ENC [54]: Achieves high scores in multiple metrics, similar to GCN-LSTM [7]. ALV [55]: Achieves competitive scores in BLEU, METEOR, and CIDEr metrics. GCN-LSTM-HIP [56]: Achieves high scores, particularly in BLEU-4 (B4) and CIDEr metrics. Entangle-T [57]: Achieves competitive scores in multiple metrics. AoA [58]: Performs well across various metrics. Proposed Method: Proposed model achieves competitive scores across different evaluation metrics, similar to the performance of the other models.

## 5. Challenges and Considerations

- Training Image Transformers may be computationally demanding due to the heightened intricacy of attention mechanisms.
- The utilization of extensive datasets is frequently necessary in order to fully realize the promise of Transformer models.
- Coherent and accurate image descriptions depend on the Transformer-based encoder and subsequent decoding components working together effectively. This is known as integration with decoders.

## 6. Future Directions

- Investigating hybrid architectures that take the best features of transformers and CNNs and combine them to maximize sequential and spatial information processing.
- Examining efficacious fine-tuning techniques to modify transformer models that have already been trained for particular image description assignments.
- Attention versions: Experimenting with various mechanisms and versions of attention to improve the interpretability and performance of the model.

Table 1. Recent model publication leaderboard on MSCOCO online testing server.

| **Refrences** | **BLEU-1** | | **BLEU-2** | | **BLEU-3** | | **BLEU-4** | | **METEOR** | | **ROUGE** | | **CIDEr** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model Ref. | **c5** | **c40** | **c5** | **c40** | **c5** | **c40** | **c5** | **c40** | **c5** | **c40** | **c5** | **c40** | **c5** | **c40** |
| Liu et al. [47] | 80.1 | 94.6 | 64.7 | 88.9 | 50.2 | 80.4 | 38.5 | 70.3 | 28.6 | 37.9 | 58.3 | 73.8 | 123 | 126 |
| Yao et al. [48] | 81.6 | 95.9 | 66.2 | 90.4 | 51.5 | 81.6 | 39.3 | 71 | 28.8 | 38.1 | 59 | 74.1 | 128 | 130 |
| Vinyals et al. [49] | 71.3 | 89.5 | 54.2 | 80.2 | 40.7 | 69.4 | 30.9 | 58.7 | 25.4 | 34.6 | 53 | 68.2 | 94.3 | 94.6 |
| Jia et al. [50] | 70 | 87 | 53 | 77 | 38 | 65 | 28 | 53 | 24 | 32 | 52 | 66 | 87 | 89 |
| Fang et al. [51] | 69.5 | 88 | | | | | 29.1 | 56.7 | 24.7 | 33.1 | 51.9 | 66.2 | 91.2 | 92.5 |
| You [52] | 73.1 | 90 | 56.5 | 81.5 | 42.4 | 70.9 | 31.6 | 59.9 | 25 | 33.5 | 53.5 | 68.2 | 94.3 | 95.8 |
| Yao [53] | 80.8 | 95.9 | - | - | - | - | 38.7 | 69.7 | 28.5 | 37.6 | 58.5 | 73.4 | 125.3 | 126.5 |
| Yang et al. [54] | - | - | - | - | - | - | 38.5 | 69.7 | 28.2 | 37.2 | 58.6 | 73.6 | 123.8 | 126.5 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Guo et al. [55] | 79.9 | 94.7 | - | - | - | - | 37.4 | 68.3 | 28.2 | 37.1 | 57.9 | 72.8 | 123.1 | 125.5 |
| Yao et al. [56] | 81.6 | 95.9 | - | - | - | - | 39.3 | 71.0 | 28.8 | 38.1 | 59.0 | 74.1 | 127.9 | 130.2 |
| Yi [57] | 81.2 | 95.0 | - | - | - | - | 38.9 | 70.2 | 28.6 | 38.0 | 58.6 | 73.9 | 122.1 | 124.4 |
| Bahdanau et al. [12] | 81.0 | 95.0 | - | - | - | - | 39.4 | 71.2 | 29.1 | 38.5 | 58.9 | 74.5 | 126.9 | 129.6 |
| Ours Method | **81.7** | **95.4** | **65.0** | **91.1** | **51.2** | **80.1** | **39.7** | **71.7** | **29.4** | **38.6** | **59.4** | **74.6** | **127.8** | **129.5** |

## 7. Conclusion

Image Transformers have revolutionized image description generation. Transformers, which excel in natural language processing and computer vision, have improved image caption quality and accuracy. Image Transformers allow the model to analyses images in detail and contextually, producing more detailed and meaningful captions. Bridging the gap between visual content and human understanding requires the capacity to capture and express image details in natural language. Image Transformers have expanded research and development for assistive technology for the visually handicapped and image-based storytelling. Extensive studies demonstrated the superiority of the suggested model, with qualitative and quantitative assessments validating the encoding and decoding transformer layer. We achieved a state-of-the-art SPICE score in image captioning, and outperformed prior top models in other evaluation measures, including computational efficiency.

## References

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. ArXiv: 1810.04805.
2. Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. International Conference on Learning Representations.
3. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv, abs/1907.11692*.
4. Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: the long-document transformer. ArXiv: 2004.05150.
5. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv, abs/2005.14165*.
6. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9.
7. Bao, Y., Sivanandan, S., & Karaletsos, T. (2023). Channel Vision Transformers: An Image Is Worth C x 16 x 16 Words. *ArXiv, abs/2309.16108*.
8. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9992-10002.
9. Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. *ArXiv, abs/2005.08100*.
10. Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision Transformer: Reinforcement Learning via Sequence Modeling. *Neural Information Processing Systems*.
11. Dwivedi, V. P., & Bresson, X. (2020). A generalization of transformer networks to graphs. ArXiv: 2012.09699.
12. Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. ArXiv: 1409.0473.
13. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems (pp. 3104–3112).
14. Kiros, R., Salakhutdinov, R. and Zemel, R., Multimodal neural language models. In International conference on machine learning, 2014, June. (pp. 595-603). PMLR
15. Fu, K., Jin, J., Cui, R., Sha, F. and Zhang, C., Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2016, pp.2321-2334.
16. Karpathy, A. and Fei-Fei, L., Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, (pp. 3128-3137).
17. Karpathy, A., Joulin, A. and Fei-Fei, L., Deep fragment embeddings for bidirectional image sentence mapping. arXiv preprint arXiv:1406.5679, 2014.
18. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., Show,

attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, 2015, June, (pp. 2048- 2057). PMLR.

19. Wang, Y., Lin, Z., Shen, X., Cohen, S. and Cottrell, G.W., Skeleton key: Image captioning by skeleton-attribute decomposition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, (pp. 7272-7281
20. Zhao, W., Wang, B., Ye, J., Yang, M., Zhao, Z., Luo, R. and Qiao, Y., A Multi-task Learning Approach for Image Captioning. In IJCAI , 2018, July, (pp. 1205-1211)
21. Yu, J., Li, J., Yu, Z. and Huang, Q., Multimodal transformer with multi-view visual representation for image captioning. IEEE transactions on circuits and systems for video technology, 30(12), 2019, pp.4467-4480.
22. Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C. and Sienkiewicz, C., Rich image captioning in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2016. pp. 49- 56
23. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C. and Lawrence Zitnick, C., From captions to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, (pp. 1473-1482).
24. Ma, S. and Han, Y. Describing images by feeding LSTM with structural words. In 2016 IEEE International Conference on Multimedia and Expo (ICME), 2016, July, (pp. 1-6). IEEE.
25. Oruganti, R.M., Sah, S., Pillai, S. and Ptucha, R., Image description through fusion based recurrent multi-modal learning. In 2016 IEEE International Conference on Image Processing (ICIP), 2016, September, (pp. 3613-3617). IEEE.
26. Wang, M., Song, L., Yang, X. and Luo, C.,. A parallel-fusion RNN-LSTM architecture for image caption generation. In 2016 IEEE International Conference on Image Processing (ICIP), 2016, September, (pp. 4448-4452). IEEE
27. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C. and Berg, T.L., Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12), 2013, pp.2891-2903
28. Mao, J., Xu, W., Yang, Y., Wang, J. and Yuille, A.L., Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090, 2014.
29. Chen, X. and Lawrence Zitnick, C., Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, (pp. 2422-2431).
30. Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L. and Deng, L., Semantic compositional networks for visual captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, (pp. 5630-5639).
31. Yao, T., Pan, Y., Li, Y., Qiu, Z. and Mei, T., Boosting image captioning with attributes. In Proceedings of the IEEE International Conference on Computer Vision, 2017, (pp. 4894-4902).
32. Mao, J., Wei, X., Yang, Y., Wang, J., Huang, Z. and Yuille, A.L., Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In Proceedings of the IEEE international conference on computer vision, 2015, (pp. 2533-2541).
33. Hendricks, L.A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K. and Darrell, T., Deep compositional captioning: Describing novel object categories without paired training data. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, (pp. 1-10).
34. Venugopalan, S., Anne Hendricks, L., Rohrbach, M., Mooney, R., Darrell, T. and Saenko, K., Captioning images with diverse objects. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, (pp. 5753-5761).
35. Mathews, A., Xie, L. and He, X., Senticap: Generating image descriptions with sentiments. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1), 2016, March
36. Nezami, O.M., Dras, M., Anderson, P. and Hamey, L., Face-cap: Image captioning using facial expression analysis. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2018, September, (pp. 226- 240). Springer, Cham.
37. Chunseong Park, C., Kim, B. and Kim, G., Attend to you: Personalized image captioning with context sequence memory networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, (pp. 895-903).
38. Liu, A., Xu, N., Zhang, H., Nie, W., Su, Y. and Zhang, Y., Multi-Level Policy and Reward Reinforcement Learning for Image Captioning. In IJCAI, 2018, January, (pp. 821-827).
39. Cornia, M., Stefanini, M., Baraldi, L. and Cucchiara, R., Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, (pp. 10578-10587).
40. Johnson, J., Karpathy, A. and Fei-Fei, L., Densecap: Fully convolutional localization networks for dense captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, (pp. 4565-4574).
41. Yang, L., Tang, K., Yang, J. and Li, L.J., Dense captioning with joint inference and visual context. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, (pp. 2193-2202)
42. Chun, S., Kim, W., Park, S., Chang, M., & Oh, S. (2022). ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO. *European Conference on Computer Vision*.

43. Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL), (pp. 311-318).
44. Lavie, A., & Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, (pp. 65-72).
45. Lin, C. Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out, (pp. 74-81).
46. Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), (pp. 4566-4575).
47. Liu, F., Ren, X., Liu, Y., Lei, K. and Sun, X., Exploring and distilling cross-modal information for image captioning. arXiv preprint arXiv:2002.12585, 2020.
48. Yao, T., Pan, Y., Li, Y. and Mei, T., Hierarchy parsing for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, (pp. 2621-2629).
49. Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, (pp. 3156-3164).
50. Jia, X., Gavves, E., Fernando, B. and Tuytelaars, T., Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE international conference on computer vision, 2015, (pp. 2407-2415).
51. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C. and Lawrence Zitnick, C., From captions to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, (pp. 1473-1482).
52. You, Q., Jin, H., Wang, Z., Fang, C. and Luo, J.,. Image captioning with semantic attention. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, (pp. 4651-4659).
53. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 711–727. Springer, Cham (2018). https://doi. org/10.1007/978-3-030-01264-9 42
54. Yang, X., Tang, K., Zhang, H., Cai, J.: Auto-encoding scene graphs for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10685–10694 (2019)
55. Guo, L., Liu, J., Tang, J., Li, J., Luo, W., Lu, H.: Aligning linguistic words and visual semantic units for image captioning. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 765–773 (2019)
56. Yao, T., Pan, Y., Li, Y., Mei, T.: Hierarchy parsing for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2621– 2629 (2019)
57. Li, G., Zhu, L., Liu, P., Yang, Y.: Entangled transformer for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8928–8937 (2019)
58. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4634–4643 (2019)