IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.0092316

Semantic Representations with Attention Networks for Boosting Image Captioning

DEEMA ABDAL HAFETH[1], STEFANOS KOLLIAS[2] (Fellow, IEEE), MUBEEN GHAFOOR[1]
[1]School of Computer Science, University of Lincoln, Lincoln LN6 7TS, U.K.
[2]School of Electrical Computer Engineering, National Technical University of Athens, 15780, Greece.

Corresponding author: Deema Abdal Hafeth (e-mail: dabdalhafeth@lincoln.ac.uk).

ABSTRACT Image captioning has shown encouraging outcomes with Transformer-based architectures that typically use attention-based methods to establish semantic associations between objects in an image for caption prediction. Nevertheless, when appearance features of objects in an image display low interdependence, attention-based methods have difficulty in capturing the semantic association between them. To tackle this problem, additional knowledge beyond the task-specific dataset is often required to create captions that are more precise and meaningful. In this article, a semantic attention network is proposed to incorporate general-purpose knowledge into a transformer attention block model. This design combines visual and semantic properties of internal image knowledge in one place for fusion, serving as a reference point to aid in the learning of alignments between vision and language and to improve visual attention and semantic association. The proposed framework is validated on the Microsoft COCO dataset, and experimental results demonstrate competitive performance against the current state of the art.

INDEX TERMS Attention, Image Captioning, Knowledge Base, Semantic Feature, Transformer.

I. INTRODUCTION

IMAGE captioning techniques, which automatically create a natural language description from an image, are an important aspect of multimedia content analysis. They have attracted much attention since they offer insight into the relationships between the multi-modal mapping of vision and natural language tasks. It is, however, a challenging research subject and entails a great deal of deep knowledge regarding how to process reasoning over vision and language understanding [1]–[3], in particular the act of recognising an image's objects, understanding their interactions, and expressing them in human language. Image captioning, which aims to describe the image in continuous natural language, also has a variety of practical applications. Some examples of image captioning's various use cases involve helping blind people by explaining images and their surroundings [4] , improving service robotics in visual assistant applications [5], and providing reports for medical images to doctors to support diagnoses and boost productivity [6].

Image captioning methods have been intensively investigated and improved since the advent of deep learning [7]; they go beyond prediction tasks and are increasingly capable of reporting in natural languages. Allowing for these capabilities, however, requires a solid linguistic knowledge and coherent understanding of the given image. In this paper, the semantic representations of given images within our attention model is explained, demonstrating the value of this architecture on the image captioning problem.

Existing approaches to image captioning have evolved an encoder-decoder structure [8]–[10], based on the sequence-to-sequence paradigm for machine translation [11]. Typically, Convolutional Neural Networks (CNNs) are utilised as encoders, converting image data into usable visual features. Alongside this, Recurrent Neural Networks (RNNs) are typically utilised as decoders for generating a language description. The standard encoding and decoding structure works on the input image to provide a related description of the scene, objects, and their relationships. The majority of current methods investigate mapping relationships between words in a sentence and specific regions of an image [9], [12]–[14]. The main challenge faced by researchers in the application of vision-and-language models is data-related; the majority of image captioning models are trained on a large amount of paired image and caption data, but these datasets typically have only a few ground truth captions per image, which are insufficient to provide a clear description of the contents of each image. It's common knowledge in this area of research that not all captions hold equivalent impor-

tance in describing the contents of a given image. Another limitation is that many image captioning models use just the visual characteristics of an image to direct the encoder, while the decoder typically relies on the textual information from the training set – this can result in difficulties accurately identifying objects in a given image. When several objects are present in an image, the described structure may not be able to identify all objects present and may especially struggle to identify any relationships between objects. These weaknesses can result in the model missing tiny objects, providing incorrect object relationships, or producing incorrect text representations, which go on to affect the quality of the resulting caption. This shows that it may be beneficial to introduce more knowledge sources to the network during training in an aims to increase contextual understanding and further caption generation accuracy.

To overcome the aforementioned restrictions, a semantic-guided attention model is suggested as an image captioning technique to enhance visual understanding. In order to improve visual understanding, information from external knowledge bases is used to detect semantic features and embed them into a continuous vector space. These are then used to link and control the image visual representation, also taking into account the relationship between key objects in the image. In terms of improving visual comprehension, an external knowledge network is introduced, which aids in the generation of more flexible description sentences by utilising information other than the basic content of the image provided. In this manner, the proposed model can gather data from external sources other than the image's basic content to use in creating more flexible description sentences. We believe that embedding this type of knowledge in a model is a necessary step to enable progress on complex multi-modal image captioning problems.

Our contributions are summarised as follows:

• We propose a semantic-guided attention network based on the Transformer model. The solution employs the semantic representation of input images to develop a more comprehensive image understanding and generate high-quality natural language captions.

• We use semantic features of the image's main elements to link and guide visual features, such as spatial relationships between objects, so that the information in the image is more highly integrated.

• We insert information from an auxiliary knowledge base source to increase our model's reasoning capabilities. This involves gathering information from outside of an image's basic content and allows for the creation of more appropriate image descriptions.

• We do an in-depth investigation that includes ablations and analytical tests on the Microsoft COCO dataset. The results demonstrate how we have enhanced the capability of deep semantic understanding, which is absent in common vision and language models.

II. RELATED WORK

Image captioning is a research area that has received a lot of attention in recent years, and there are many works that have proposed different approaches to solve this problem. In this section, we will discuss some of the important works in this area.

A. IMAGE CAPTIONING

According to the history of the field, automatic image captioning development approaches can be classified into three groups: template-based methods, retrieval-based methods, and neural network-based methods. The template-based methods [15], [16] define all the properties of the images using object and image classification techniques. These approaches construct captions by filling in pre-defined templates with data obtained from the detected images. The use of template has the advantage of producing captions that are more likely to be clear and precise. They are, however, inflexible and suffer from output variety limitations [13]. Retrieval-based methods have been widely applied to image captioning [17]–[19]. A collection of query related images is built from an image database, rank them based on how similar they are, and then modify the descriptions of the identified images to construct a new description for the requested image. However, the utility of this strategy is severely limited, particularly when dealing with pictures that are unseen, not in the dataset, or not classified. The neural network-based methods are influenced by deep neural networks' success in machine learning and are used in an encoder-decoder architecture. A CNN encoder retrieves image features, which an RNN decoder is used for language modelling and constructing image captions. These methods are more flexible and produce higher-quality image captions than the aforementioned two methods. Vinyals *et al.* [3] presented a neural image caption (NIC) model. It is the first effort to incorporate the encoder-decoder paradigm into image captioning. It acts as a foundation for later upgrades and a benchmark model for performance comparisons between models. The CNN is commonly used as the encoder to extract image information expressed as fixed-length vectors via matrix transformation, decoding visual information using Long Short-Term Memory (LSTM). This method increases the probability of the correct description given the image. However, visual input is only available during the first decoder update and employs a more complex CNN.

According to the history of the field, automatic image captioning development approaches can be classified into three groups: template-based methods, retrieval-based methods, and neural network-based methods. The template-based methods [15], [16] define all the properties of the images using object and image classification techniques. These approaches construct captions by filling in pre-defined templates with data obtained from the detected images. The use of template has the advantage of producing captions that are more likely to be clear and precise. They are, however, inflexible and suffer from output variety limitations [13].

Retrieval-based methods have been widely applied to image captioning [17]–[19]. A collection of query related images is built from an image database, rank them based on how similar they are, and then modify the descriptions of the identified images to construct a new description for the requested image. However, the utility of this strategy is severely limited, particularly when dealing with pictures that are unseen, not in the dataset, or not classified. The neural network-based methods are influenced by deep neural networks' success in machine learning and are used in an encoder-decoder architecture. A CNN encoder retrieves image features, which an RNN decoder is used for language modelling and constructing image captions. These methods are more flexible and produce higher-quality image captions than the aforementioned two methods. Vinyals *et al.* [3] presented a neural image caption (NIC) model. It is the first effort to incorporate the encoder-decoder paradigm into image captioning. It acts as a foundation for later upgrades and a benchmark model for performance comparisons between models. The CNN is commonly used as the encoder to extract image information expressed as fixed-length vectors via matrix transformation, decoding visual information using long short-term memory (LSTM). This method increases the probability of the correct description given the image. However, visual input is only available during the first decoder update and employs a more complex CNN.

The design of the visual description framework has research a new stage, particularly after applying the attention mechanism [20]. Employing an attention mechanism for image's regions can reduce problems with missing image objects or spatial information about objects. This is because it learns which regions to focus on by swiftly scanning them, then devoting more attention resources to those areas to gain a more specific understanding of the target of attention. In 2015, Xu *et al.* [20] proposed the first attention mechanism approach applied to the encoder-decoder framework for image captioning. They used visual attention to automatically focus on different areas and learn the alignments between words in the provided sentences and image regions, making their framework new and state-of-the-art for attention image captioning. However, encoding image systems that rely only on visual features from the whole image do not always succeed in extracting all the information that requires attention. Consequently, that may generate inaccurate and limited descriptive statements, since incorrect visual components may be retrieved. Huang *et al.* [21] proposed the Attention-on-Attention network (AOANet) for the encoder and decoder architecture. The network determined the relevance between attention results and queries in the transformer model [11]. The transformer model has an attentive mechanism in which each element of a set is connected to every other element in the set. The AOANet outperformed other methods in terms of output quality and training time for image captioning [9], [22]–[24].

However, the encoder-decoder attention transformer modelling technique cannot effectively capture the semantic relationship between image items, especially if their appearance attributes have a weak dependency. In this paper, semantic links among image items is defined by modelling a visual relationship. This is done by using a self-attention technique to weigh the importance of the appearance features of query objects in relation to candidate visual relationships.

B. OBJECT DETECTION

It is one of the foundational processes for visual comprehension and reasoning that enhance captioning performance by capturing semantic relationships between image and language. Several image captioning methods [8], [25], [26] extracted visual information from multiple areas in an image using the Faster R-CNN object detector approach [27]. Anderson *et al.* [8] proposed a combined bottom-up and top-down attention mechanism using a Faster R-CNN object detector to attend more naturally at the level of objects and other visual regions. Their approach enhanced visual attention but did not consider semantic connections between the identified regions in the image. To come up with further information that is not included in the provided captions, other researchers have proposed advanced models that incorporate visual features with high-level semantic information for image object instance-level concepts as another form of image representation. Most recent works [8], [28] unified object instance-level concepts probability with region features as the visual input for image captioning and visual question answering models, while other methods [13], [29]–[31] used level concepts as supplementary information to enhance image-text semantic alignments. Unfortunately, in some works, the object concepts are not related to both the object areas and the caption supplied, resulting in a lack of grounding. These motivate us to spread semantics in visual attention across the total proposed regions. A semantic guided attention model is proposed, that uses object instance-level concepts [8]. To align the object-region features in the pre-trained linguistic semantic space. That exploits not only an overview understanding of the input image but also visual semantic attributes for attention calculation.

C. IMAGE CAPTIONING WITH COMMONSENSE KNOWLEDGE BASE

Knowledge Bases (KBs) can be selected as an external source of information to enhance many deep neural network processes. Several approaches have been developed in the natural language processing and computer vision fields to benefit image captioning via the utilisation of KBs [32]–[34]. Motivated by the importance of object instance-level concepts' semantic relations, in this paper we focus on integrating common sense KBs from external resources and integrating them with other attributes to achieve better model performance. ConceptNet [35] is used as an external KB rather than several other alternatives due to its broad coverage of concepts and accompanying semantic embedding of useful features. A type of new attention method is developed that embeds common sense KBs from ConceptNet and share the

common space embedding feature with other input image information in the encoder transformer. This provides an additional advantage and information for an image's object representations.

III. BACKGROUND

Attention mechanisms have proven to be a powerful tool for modelling complex dependencies in both natural language processing and image captioning. By allowing models to focus on specific parts of the input when making predictions, attention-based models can produce more accurate and detailed results. In this section, we describe relevant background on attention on transformers models and image caption generation with attention mechanism.

A. ATTENTION ON TRANSFORMERS

1) Scaled Dot-Product Attention

Scaled dot-product attention is the transformer's key component. The input consists of a set of three inputs, i.e., keys K, values V, and queries Q. This attention is shown as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V. \quad (1)$$

where $Q \in R^{L\times d_K}$ is the query, $K \in R^{L\times d_K}$ is the key, $V \in R^{L\times d_v}$ is the value, and the L is the length of the sequence. The $QK^T \in R^{L\times L}$ is a product operation. This could make the result become too large or small, and influence the precision of the variable. So $\sqrt{d_k}$ is used to scale QK^T. At last, we get $Attention(Q, K, V) \in R^{L\times d_v}$. In practice, dot-product attention is faster and more space-efficient. When the keys, values, and queries are the same matrices, this mechanism is called self-Attention.

2) Multi-Head Attention

The image captioning model uses an encoder-decoder structure with stacking layers of attention blocks. Multi-Head Attention (MHA) and Feed-Forward Networks (FFN) are present in each attention block, Figure 1. Generally, The CNN is used to encode the given image I to the image region feature vector as input, $V = \{V_1, V_2, ..., V_{K\times k}\}$, $V_i \in R^{d_{model}}$, where $K \times K$ is the number of regions, and V_i represents a region of the image. The decoder generates the target caption $y = \{y_1, y_2, ..., y_m\}$, where m is the maximum length of the generated sentence.

The MHA applies scaled dot-product attention multiple times n in parallel to manage the mixing of input across parts of an input vector. This is resulting in richer representations and higher performance. Then the outputs of the separate attention are concatenated. The following equations illustrate MHA:

$$head_i = Attention(QW_i{}^Q, KW_i{}^K, VW_i{}^V). \quad (2)$$

$$H = Concat(head_1, ..., head_n). \quad (3)$$

$$X = HW_h. \quad (4)$$

where the projections are parameter matrices $W_i{}^Q \in R^{d_{model}\times d_k}$, $W_i{}^k \in R^{d_{model}\times d_k}$, $W_i{}^V \in R^{d_{model}\times d_v}$, $Q \in R^{L\times d_{model}}$, $K \in R^{L\times d_{model}}$, $V \in R^{L\times d_{model}}$ are the inputs of the multi-head attention and d_{model} is the dimensional image feature. $head_i \in R^{L\times d_v}$ is the output of the scaled dot-product attention. n scaled dot-product attentions are concatenated $Concat$ to generate $H \in R^{L\times(n\times d_v)}$. A $W_h \in R^{(n\times d_v)\times d_{model}}$ is used to project H into the output $X \in R^{L\times d_{model}}$.

Another main component in the encoder transformer network is a FFN. It takes input from MHA output and consists of two fully connected layers with a ReLU activation function and dropout function.

$$FFN(x) = FC(Dropout(ReLU(FC(x)))). \quad (5)$$

where x is the previous MHA module's output. The FFN modules learn further non-linearly for the attended feature vector. To reduce information loss, FFN is encapsulated within a residual connection and layer normalisation. It is applied to the outputs of the multi-head attention and the FFN to produce an encoded feature A'. where A' denotes the output of the transformer encoder module, which describes the visual information of the input image.

$$A' = Layernorm(x + FFN(x)). \quad (6)$$

B. IMAGE CAPTIONING AND ATTENTION MECHANISM

Traditional CNN-RNN image captioning framework-based approaches usually concatenate image area and language information as input. Despite the state-of-the-art performance associated with the use of object detectors, the drawbacks of these models are that they construct unneeded areas and extract visual features from excessively overlapped, noisy, and ambiguous regions, which makes the task difficult. If feature vectors do not include meaningful information, then the attention model generates feature vectors unrelated to the proper caption. A lack of grounding learning is also caused by the absence of unambiguous semantic alignments between areas or objects in an image and words or phrases in the appropriate caption. To overcome the aforementioned difficulties, an aligning vision and language features in a shared semantic space are suggested by detecting input object semantic level concepts as a point of reference. The training examples are divided into three triplets: each image contains a word sequence, a collection of object classes, and a set of image region features used for image captioning.

IV. DESIGN OF THE PROPOSED FRAMEWORK

Semantic guided-attention networks can adaptively perform the image encoder procedure to describe a given image. Figure 1 depicts the proposed framework for image captioning. First, an object detection model is used, i.e., Faster R-CNN, to extract the feature of the original image. Then, an attention module is needed to encode the visual features and output an attentive feature. Following this, common sense embedding features are extracted from external KBs

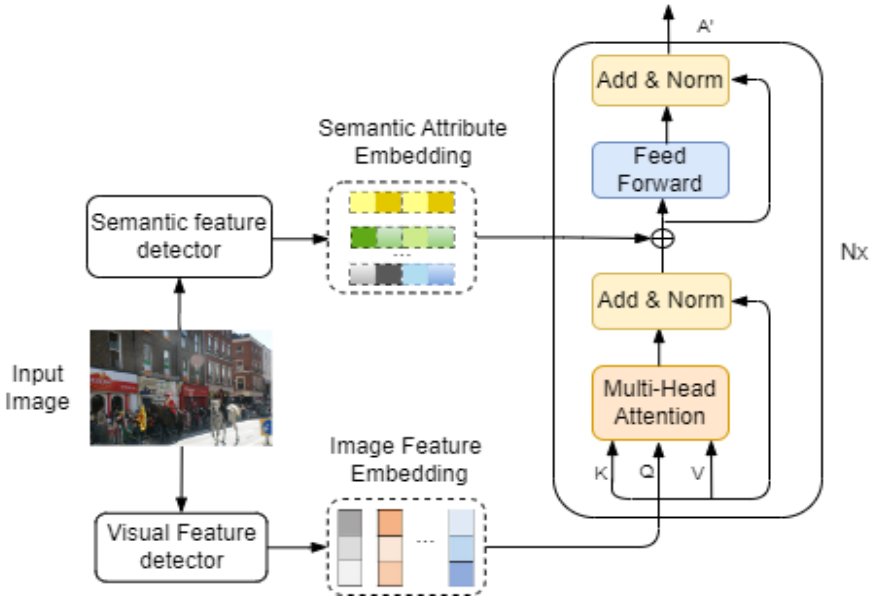


FIGURE 1. **The architecture of the Semantic Representations with Attention Networks for Boosting Image Captioning Image Encoder.**

ConceptNet and share the common feature space with other input image information in the encoder transformer to depict relationships between different objects and scenes in image. Finally, a language decoder is applied to generate language descriptions.

A. OBJECT DETECTION

Following [8], a Faster R-CNN [27] in conjunction with ResNet-101 [36] is adopted, which has a CNN base, for object detection and feature extraction. The Faster R-CNN model is pre-trained on the Visual Genome dataset [37] and outputs object classes. Its first stage is a Region Proposal Network (RPN) that uses intermediate feature maps from ResNet-101 as inputs and generates bounding boxes for proposed objects. Intersection-Over-Union (IOU) is metric that measures the overlap between two bounding boxes. The reference boxes n that have an IoU more than 0.7 are selected. In the second stage, Region-Of-Interest (ROI) pooling layer is used to convert all proposal bounding boxes to the same spatial size feature map (e.g., 14 x 14 x 2048). For simplicity, the top 36 ROIs are only used. These are followed by a softmax distribution to predict the bounding box object classes and refinements for each box proposal.

After using Faster R-CNN, each image can be represented as a set of object semantic concepts $S = \{s_1, s_2, ..., s_n\}$ and a set of object visual features $V = \{v_1, v_2, ..., v_n\}$, in which $v_i \in R^{d_1}$, $N = 36$, and $d_1 = 2048$. For each selected region i, the feature after the average pooling layer is extracted to serve as object visual feature v_i. So far, we have extended the training example to include the number of objects n, their corresponding semantic classes S, and visual features V. All will be used later by the encoder transformer.

B. IMAGE ENCODER

To improve the image understanding capability of the image encoder, we construct the merged box as a semantic relationship guide to direct attention and enhance visual feature representation. Based on the detection results of Faster R-CNN, self-attention layers contain two kinds of inputs: the object's visual features V_i of previously detected objects and their semantic classes S_i. First, the visual object's attention is founded. This is done by employing layers of MHA. Secondly, the ConceptNet knowledge base [35] is used to generate objects with semantic concepts embedded $E^{se}(S)$. Thus, semantic concepts $S = \{s_1, s_2, ..., s_n\}$ are embedded to semantic concepts features $O = \{o_1, o_2, ..., o_n\}$, where $o_i \in R^{d_2}$, and $d_2 = 300$. Notably, towards a specific object (e.g., car, door, lion), visual features may vary from object to object while semantic features always remain unchanged. The merged box combines the visual attention features $Attention(Q, K, V)$ and semantic vectors to get visual semantic representation for input image. The semantic guided attention representations A^{GA} for Self-attention layers are calculated as follows:

$$A^{GA} = Attention(Q, K, V) + E^{se}(S). \quad (7)$$

where $E^{se}(S)$ object semantic vectors. The attention block is followed by a FFN that takes the input from the output of MHA and transforms each feature vector using two linear layers with ReLU activation and dropout in between as follows:

$$FFN(A^{GA}) = FC(Dropout(ReLU(FC(A^{GA})))). \quad (8)$$

The output for the final transformer layer A^{GA} will feed to the LSTM decoder and give an overall understanding of the image. The encoder output is identical in size to the decoder

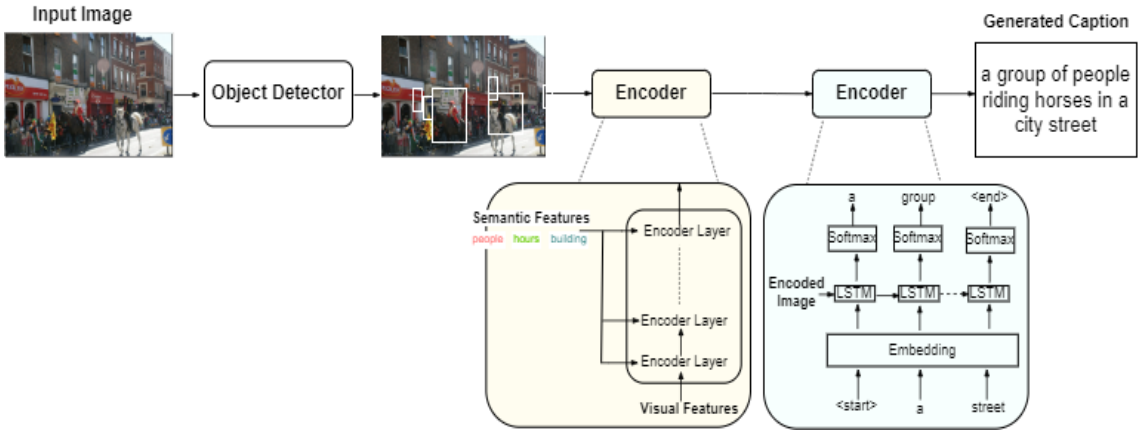


FIGURE 2. The Architecture of the Semantic Representations with Attention Networks for Boosting Image Captioning System.

input. In the end, the proposed encoder model learns visual relations attentively at each time step, guided by semantic features at the object level.

C. LANGUAGE DECODER

In this paper, the LSTM units are selected as the decoder. The decoding component is used to decode visual features to iteratively generate descriptive text sequences. The improvement of the decoder mainly focuses on enriching the information in both the visual and text [3]. As shown in Figure 2, we jointly integrate both the attention features from encoder output A' and the word embedding vectors W_t in one fusion representation into the LSTM unit. The LSTM network has a state cell and several gates, such as a forget gate and an input gate. These gates ensure the effective memory and updating of information at each time step $t \in [1, T]$, and generate output word y_t. The LSTM outputs the hidden state h, that saves the decoding state, and then uses it to compute the conditional probabilities on the vocabulary:

$$p(y_t|y_{1:t-1}, A) = softmax(W_p h_t). \tag{9}$$

where $W_p \in R^{D\times|\sum|}$ is the weight parameters to be learnt and $|\sum|$ the size of the vocabulary.

In this way, the LSTM performs step-by-step decoding to generate the final word sequence $Y = (y_1, y_2, ..., y_m)$ where m is the maximum length of the generated sentence.

V. MATERIALS AND EVALUATION METRICS

In this section, we first provide a detailed description for the dataset used to evaluate the performance of the proposed model. Then we discuss the implementation details followed by description of evaluation measures used in our study.

A. DATASET

We trained and evaluated our algorithm on the Microsoft COCO (MS-COCO) 2014 dataset [38], where images are labelled with five human annotated captions manually. The offline "Karpathy" data split [39] is used for offline performance comparisons, using 82,783 training images and 5,000 images are used for testing and validation, respectively. The descriptions included in the dataset were pre-processed to be lowercase, tokenize sentences, remove punctuation, and drop words occurring less than 5 times in total. After pre-processing, there were 10,369 unique words present in the set.

B. IMPLEMENTATION DETAILS

Our algorithm was developed in PyTorch, and we employ a pre-trained Faster R-CNN [27] model (subsection IV-A) on ImageNet [40] and Visual Genome [37] to extract bottom-up feature vectors of images [8] and ConceptNet 5 [35]. We follow the practice in [21] and set the dimension of the original vectors to 2048, and project them to a new space with the dimension D = 1024, which is also the hidden size of the LSTM in the decoder. Our best performing model was pre-trained for 30 epochs with a softmax cross-entropy loss using the ADAM optimiser [41] and batch size of 10 with a 2e-4 learning rate. The number of attention blocks was set to six, and the number of parallel attention heads was set to eight. To minimise over-fitting, drop-out was applied at the inputs and outputs of all layers. The beam search approach [42] is used in the testing stage to locate the sentence with the best probability.

C. EVALUATION MEASURES

In this paper, we adopt the BLEU@N(B@N (BiLingual Evaluation Understudy) [43], METEOR (Metric for Evaluation of Translation with Explicit ORdering) [44], ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation-Longest) [45], and CIDEr-D (Consensus-based Image Description Evaluation) [46] metrics against all ground truth to evaluate the generated sentences. The common practice in the literature is to report all the aforementioned metrics for image caption generation.

Each metric has its own calculating technique and special advantages. BLEU is a precision-based metric. The main component of BLEU is the n-gram precision of the generated

caption with respect to the ground-truth captions. METEOR is an automatic machine translation evaluation metric. It looks at the precision, recall, and alignments between the related tokens. This is accomplished by performing generalised uni-gram matches between a candidate sentence and references and then computing a score depending on the findings of the matches. As a result, the captions generated by these approaches exhibit high precision and recall accuracy, as well as good word-level similarity. While ROUGE-L can assess the sufficiency and fluency of machine translation, for the purpose of evaluating image captioning, precision and recall are determined using the longest subsequence of tokens that exist in both the candidate and reference captions in the same relative order, maybe with additional tokens in between. CIDEr is a method that measures the effectiveness of image captioning using human consensus. This measure takes into account information content and grammatical accuracy. It assesses how closely an image captioning sentence resembles the vast majority of ground truth sentences written by humans.

VI. EVALUATION RESULTS AND DISCUSSION

Extensive experiments were conducted to assess the proposed models for image captioning. In this section, we present all the results and discussions using the Microsoft COCO caption evaluation tool. A comparison and analysis of the proposed model's performance is made with other state-of-the-art models. Next, we evaluate the model architecture, followed by a qualitative analysis. Finally, we assess the proposed model's effectiveness.

A. COMPARISON OF THE EFFECTIVENESS WITH RELATED STUDIES

In this section, we evaluate the effectiveness of semantic guided attention on the MS-COCO dataset. Three state-of-the-art methods are compared as follows: (*i*) Show and Tell [3] uses CNN layers; the convolution layer, the pooling layer, and full connection layers as the encoder to obtain fixed-length vectors for the image features represented. To successively construct descriptive text sequences, visual characteristics are decoded using an RNN/LSTM decoder.
(*ii*) Up-Down [8] integrates Faster R-CNN to obtain object regions and other salient image regions and enables attention to be calculated at a high level of semantic information.
(*iii*) AoA [21] introduces a method of using an extra attention layer with a Transformer to encode regions into hidden states, and connects with an extra attention layer with LSTM based decoders for image captioning.
The results in Table 1 show that our proposed method achieves higher scores in terms of BLEU@1, ROUGE-l, and CIDEr-D at 78.6%, 57.7% ,and 120.98% , respectively. The results indicate the advantage of semantic feature selection and KBs ConceptNet embedding in the proposed framework, which generates better image captions suitable for different contexts. Moreover, ROUGE evaluates the appropriateness and fluency of the generated captions, whereas CIDEr focuses on grammar and relevance. These results demonstrate that our method outperforms the other approaches in terms of appropriateness, fluency, and relevance.

TABLE 1. The performance of our model and other the state-of-the-art methods on MS-COCO. All values are reported as a percentage.

Method	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Show and tell [3]	-	29.6	25.2	52.6	94.0
Up-Down [8]	77.2	36.2	27.0	56.4	113.5
AoANet [21]	77.4	**37.0**	**28.4**	57.5	119.8
Semantic Guided Attention	**78.6**	36.0	27.6	**57.7**	**120.9**

However, the performance of the proposed method is slitly worse than Up-Down and AoANet under the BLEU@4 and METEOR metrics, because Up-Down simultaneously comprises two LSTM to train the captioners. Additionally, AoANet outperforms the others in terms of BLEU@4 and METEOR, it proposed a better language model that employs an attention strategy in the encoder transformer and LSTM decoder. Comparatively, this study focuses on improving the encoder block but not the language decoder block, which can somewhat degrade the outcomes while also reducing the model complexity and producing a lightweight model.
Furthermore, it can be observed that the proposed semantic attention method performs differently across BLEU@n metrics. It achieves good performance on B@1 by 78.6% compared to B@4 by 36.0%. These results further support the idea that in some circumstances, an increase in the BLEU@n score does not imply that the generated text is good, like when the text is short [47]. This is because the BLEU@n metric is used to assess the quality of machine generated text. It uses n-grams, where n denotes the number of overlapping words. The final scores are determined by comparing a set of reference texts to the generated text and taking the average. The BLEU@n scores, on the other hand, do not account for syntactical accuracy.
The results demonstrate that the proposed method improves image captioning performance. The key advantage of our model is that it can understand the relationships among detected objects, and uses the external ConceptNet KBs module to enhance the semantic association between objects with a weak dependence on appearance features. It is reasonable to expect that the performance of the proposed method can be further improved by utilising more information, such as appearance, motion, and attribute features.

B. EVALUATION OF MODEL ARCHITECTURE

The encoder Transformer is a semantic guided attention model that incorporates multi modal information at the same time to produce a visual representation. The proposed model is evaluated by varying the number of multi head attentions. We set variants with different numbers of heads, $H \in \{4, 8, 16, 32\}$. They are trained using cross-entropy loss. The results of the quantitative comparisons of BLEU@1, METEOR, ROUGE-L, and CIDEr-D, are shown in Figure 3. It

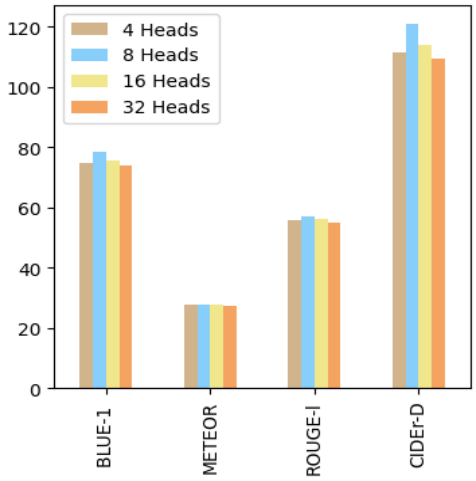


FIGURE 3. Experimental results for the different number of parallel attention heads.

is clearly seen that the models with more attention heads do not necessarily exhibit better performance. Meanwhile, when H is set to 8, the model exhibits the best performance, providing a more accurate understanding of the image content. However, increasing the number of heads would result in over fitting, which would make the incorporation of multi modal information more difficult. Furthermore, when a critical point has been reached, the difficulty of training increases as the number of parameters increases. This results in a performance decline, rather than an increase in performance.

C. QUALITATIVE ANALYSIS

Table 4 shows a few images from the MS-COCO dataset with their generated descriptions, the state-of-the-art model descriptions [21], and ground-truth descriptions. It is clearly seen that the generated sentences well describe the contents of the image. Specifically, our model has advantages in that it can better understand the visual relationship that characterises semantic association between objects by simultaneously using the semantic features of the objects. This can improve the visual relationship between objects with a weak dependence on appearance features and enhance the semantic association between them.

For example, in the first image, our method recognises the main objects in the image, even when the visual objects are ambiguous. It differentiates the main objects in the image better than the baseline model, which benefits from using external knowledge to introduce semantic relations between image objects. Our model can further describe the visual relationship between a *man* and *table* in the generated text 'a man sitting at the table with wine glasses and a bottle of wine'. Similarly, the second example illustrates how our model can more accurately infer the concept of marine from the words *ocean* and *kite*, plus create more extensive and descriptive image content than the baseline model. The caption generated by the baseline model is logically correct, but it might not accurately describe the image content. As illustrated in the third example, although the baseline method can capture the relative position between objects *dog* and *room*, it fails to further model the semantic association between other objects with a weak dependence on the appearance features. Our method first models the visual relationship that characterises the semantic association between objects, and then it is assigned to measure the importance of appearance features to the modelled visual relationship between *top* and *sofa*, especially for objects with a weak dependence on appearance features. In summary, our approach has proven its ability to understand image content by taking advantage of enhanced image understanding abilities.

TABLE 2. The performance comparison between transformers attention model and transformers semantic guided attention model. All values are reported as a percentage.

Method	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr-D
Transformers attention model	75.8	35.4	**27.9**	56.5	114.4
Transformers semantic guided attention model	**78.6**	**36.0**	27.6	**57.7**	**120.98**

D. EVALUATION OF THE PROPOSED MODEL

In this section, the validation and robustness of our proposed method are reported and the results are compared to a basic attention structure that employs only the attention method and does not employ the suggested semantic guided attention network. All the experiments are trained on the MS-COCO dataset and use semantic guided attention network hyperparameters and structure to ensure fairness. The results are shown in Table 2.

The proposed method can significantly improve the majority of the evaluation measures. The experimental results based on the basic attention model is not as good as those of the proposed method. The larger performance improvement of the semantic guided attention method supports our belief that incorporating external knowledge offers additional benefits and serves as supplemental knowledge for predicting the final image caption.

VII. CONCLUSION

In this paper, an attention network that leverages semantic representation data network based on the transformer's self-attention mechanism is introduced. It boosts a comprehensive understanding and guidance of visual features. It maps data from vision and language and applies it to tasks such as image caption generation. Our experiments demonstrate that some types of commonsense and KBs, like ConceptNet, can capture factual knowledge from input data. Furthermore, extensive experiments on the MS COCO dataset are performed to investigate the mechanics behind how and why our semantic direct attention model with external knowledge functions works. We conclude that combining information

Semantic Guided Attention model:
a man is flying a kite in the ocean.

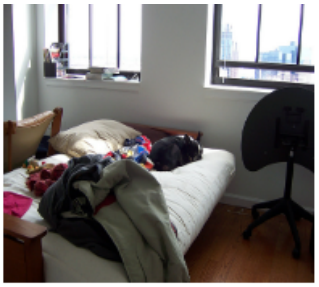
Baseline model:
a group of people standing around a table.

Ground-truth annotations:
. a man in glasses is holding a wine glass.
. a man holds a glass in a room with many other people.
. a man holds a glass as others mill around behind him.
. a group of people standing around in a room.
. a man holding a glass speaking to someone.

Semantic Guided Attention model:
a man is flying a kite in the ocean.

Baseline model:
a man riding a kiteboard on top of a body of water.

Ground-truth annotations:
. a man kite boarding over a large body of water.
. a man in the water kite surfing on a board.
. a surfer wrangles a parachute with a scenic mountain background.
. a paraglider who has just landed in the ocean.
. a person riding a surf board with a parachute in a body of water.



Semantic Guided Attention model:
a dog laying on top of a sofa in a room and chair.

Baseline model:
a dog laying in a room.

Ground-truth annotations:
. there is a dog sleeping on a futon.
. messy bed and only two chairs in a small room.
. a very cute dog laying on a big bed.
. a very cute dog laying on a big bed.
. a small black and white dog sleeping on a small cot.

FIGURE 4. Examples images and captions results by semantic guided attention model and baseline model, coupled with the corresponding ground truth sentences.

from many sources is a helpful strategy for improving the applicability of existing machine learning models. Furthermore, by increasing the availability of supporting knowledge, this strategy provides the foundation for future developments in reasoning methods to analyse this information. For example, it may lead to improved skills in addressing problems that require a high degree of comprehension. In future research, we will continue to enhance the accuracy of image captioning network generation. This may be accomplished by increasing reasoning and text recognition, as well as by improving the semantic comprehension of image and text information.

REFERENCES

[1] X. Jia, Y. Wang, Y. Peng, and S. Chen, "Semantic association enhancement transformer with relative position for image captioning," *Multimedia Tools and Applications*, vol. 81, no. 15, pp. 21 349–21 367, 2022.

[2] D. Wang, Z. Hu, Y. Zhou, R. Hong, and M. Wang, "A text-guided generation and refinement model for image captioning," *IEEE Transactions on Multimedia*, 2022.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[4] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning images taken by people who are blind," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 417–434.

[5] R. C. Luo, Y.-T. Hsu, Y.-C. Wen, and H.-J. Ye, "Visual image caption generation for service robotics and industrial applications," in *2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS)*. IEEE, 2019, pp. 827–832.

[6] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos, "A survey on biomedical image captioning," in *Proceedings of the second workshop on shortcomings in vision and language*, 2019, pp. 26–36.

[7] R. Sasibhooshan, S. Kumaraswamy, and S. Sasidharan, "Image caption generation using visual attention prediction and contextual spatial relation extraction," *Journal of Big Data*, vol. 10, no. 1, pp. 1–18, 2023.

[8] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "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, pp. 6077–6086.

[9] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 578–10 587.

[10] J. Zhang, Y. Xie, W. Ding, and Z. Wang, "Cross on cross attention: Deep fusion transformer for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[12] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Image-text embedding learning via visual and textual semantic reasoning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 641–656, 2022.

[13] F. Liu, Y. Liu, X. Ren, X. He, and X. Sun, "Aligning visual regions and textual concepts for semantic-grounded image representations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[14] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," *Advances in neural information processing systems*, vol. 32, 2019.

[15] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*. Springer, 2010, pp. 15–29.

[16] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple

image descriptions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.
[17] P. Kuznetsova, V. Ordonez, T. L. Berg, and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 351–362, 2014.
[18] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Generalizing image captions for image-text parallel corpus." in *ACL (2)*. Citeseer, 2013, pp. 790–796.
[19] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *Advances in neural information processing systems*, vol. 24, 2011.
[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.
[21] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4634–4643.
[22] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8928–8937.
[23] X. Zhu, L. Li, J. Liu, H. Peng, and X. Niu, "Captioning transformer with stacked attention modules," *Applied Sciences*, vol. 8, no. 5, p. 739, 2018.
[24] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE transactions on circuits and systems for video technology*, vol. 30, no. 12, pp. 4467–4480, 2019.
[25] L. Huang, W. Wang, Y. Xia, and J. Chen, "Adaptively aligned image captioning via adaptive attention time," *Advances in neural information processing systems*, vol. 32, 2019.
[26] Y. Qin, J. Du, Y. Zhang, and H. Lu, "Look back and predict forward in image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8367–8375.
[27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.
[28] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 041–13 049.
[29] J. Li, P. Yao, L. Guo, and W. Zhang, "Boosted transformer for image captioning," *Applied Sciences*, vol. 9, no. 16, p. 3260, 2019.
[30] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 685–10 694.
[31] L. Guo, J. Liu, J. Tang, J. Li, W. Luo, and H. Lu, "Aligning linguistic words and visual semantic units for image captioning," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 765–773.
[32] Y. Zhou, Y. Sun, and V. Honavar, "Improving image captioning by leveraging knowledge graphs," in *2019 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2019, pp. 283–293.
[33] Y. Zhang, M. Jiang, and Q. Zhao, "Query and attention augmentation for knowledge-based explainable reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 576–15 585.
[34] Y. Li, Y. Pan, T. Yao, and T. Mei, "Comprehending and ordering semantics for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 990–17 999.
[35] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, 2017.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[37] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.
[38] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.
[39] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.
[40] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[42] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.
[43] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
[44] S. Banerjee and A. Lavie, "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*, 2005, pp. 65–72.
[45] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
[46] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
[47] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of bleu in machine translation research," in *11th conference of the european chapter of the association for computational linguistics*, 2006, pp. 249–256.

DEEMA ABDAL HAFETH is currently pursuing her Ph.D. in Computer Science within the School of Computer Science at the University of Lincoln in the UK. Her research interests involve Deep Learning, Image Captioning, and Data Mining.

STEFANOS KOLLIAS is an IEEE Fellow, Professor within the ECE School at the National Technical University of Athens (NTUA) in Greece. He has published over 100 journal articles and over 300 conference papers alongside supervising 43 Ph.D. students. His research interests include Machine/Deep Learning, Multimedia Analysis, Search, Retrieval and Recognition, Vision, Medical Informatics, Cultural Heritage, HCI, and Affective Computing.

MUBEEN GHAFOOR is a Senior Lecturer in Computer Science within the School of Computer Science at the University of Lincoln in the UK. He holds a Ph.D. in Image Processing from Mohammad Ali Jinnah University in Pakistan. He has vast research and industrial experience in the fields of Data Science, Image Processing, Machine Vision Systems, Signal Analysis, GPU-Based Hardware Design, and Software System Designing.