A Survey on Attention-Based Models for Image Captioning

Asmaa A. E. Osman[1], Mohamed A. Wahby Shalaby[2], Mona M. Soliman[3], Khaled M. Elsayed[4]
Information Technology Department-Faculty of Computers and Artificial Intelligence,
Cairo University, Giza, Egypt[1,2,3,4]
Smart Engineering Systems Research Center (SESC), Nile University, Giza, Egypt[2]

Abstract—**Image captioning task is highly used in many real-world applications. The captioning task is concerned with understanding the image using computer vision methods. Then, natural language processing methods are used to produce a description for the image. Different approaches were proposed to solve this task, and deep learning attention-based models have been proven to be the state-of-the-art. A survey on attention-based models for image captioning is presented in this paper including new categories that were not included in other survey papers. The attention-based approaches are classified into four main categories, further classified into subcategories. All categories and subcategories of the attention-based approaches are discussed in detail. Furthermore, the state-of-the-art approaches are compared and the accuracy improvements are stated especially in the transformer-based models, and a summary of the benchmark datasets and the main performance metrics is presented.**

Keywords—Image captioning; attention model; deep learning; computer vision; natural language processing

I. Introduction

Image captioning is targeted to represent an image with a sentence that should be accurate and summarized. The problem of image captioning is similar to using a machine to translate a sentence, but in image captioning, the machine task will be translating an image into a sentence. So, it is necessary to visually understand the image before producing the caption. The caption of the image should be expressive through detecting the objects of the image and their attributes, finding the relationship between the detected objects and the place/activity where the objects are included.

The task of image captioning is very necessary for that it can be as an assistant to the impaired people by providing a brief description for the image while exploring the internet. Image captioning can be used in implementing self-driving cars by providing the agent with the ability to drive in a safe, fast and accurate way. Also, generating a caption for medical images automated the process of diseases diagnosis and treatment. In addition, it can be used to generate captions for the images included in the news articles. There are many other applications for image captioning, like in service robotics, military, education and image indexing.

In order to generate a sentence with reasonable linguistics and true semantics, Computer Vision (CV) methods are used to visually understand the image. In addition, Natural Language Processing (NLP) models are employed to generate a correct sentence. The power of Deep Learning (DL) approaches in CV [1-7] and NLP [8-12] makes it the first choice for many approaches in image captioning. Convolutional Neural Network (CNN) was most commonly used in the vision part to get the image features. Then, Recurrent Neural Network (RNN) was used as a language model [13-16].

According to [17], deep neural network approaches in image captioning task can be categorized based on:

- Type of learning: (Supervised [18,19], Unsupervised [20,21] and Reinforcement Learning [22,23])
- Architecture: (Encoder-Decoder [24,25] and Compositional [26,27])
- Feature Mapping: (Visual Space [28,29] and Multimodal Space [30])
- Number of Captions: (Dense Captioning [31], Whole Scene Captioning [32])
- Language Model: (LSTM and others)

For the purpose of generating high quality captions, it was helpful to use advanced visual processing by considering the most salient features in the images while generating the caption words which is called attention model. The attention mechanism takes inspiration from the human visual system, which does not focus on all the scene parts but only on small parts of the scene. The salient features included in the image take precedence in encoding the image instead of the whole image. Attention has been used in different tasks, like machine translation and object identification. Moreover, many image captioning approaches employed the attention model and achieved a very good enhancement [33-37].

In this paper, a detailed survey for the attention-based approaches employed in image captioning is presented. In addition, a taxonomy of these attention-based models is provided including two new categories for categorizing the attention-based approaches. Most of the state-of-the-art articles for image captioning using attention-based models are included and compared with respect to the benchmark datasets and metrics.

The organization of this survey paper is as follow: In Section II, Literature review is presented. The attention mechanism and its taxonomy is presented in Section III, including four main categories of the attention models and their subcategories. The benchmark datasets in addition to the popular performance metrics are introduced in Section IV. State-of-the-art models are compared in Section V. Finally, this survey paper is concluded in Section VI.

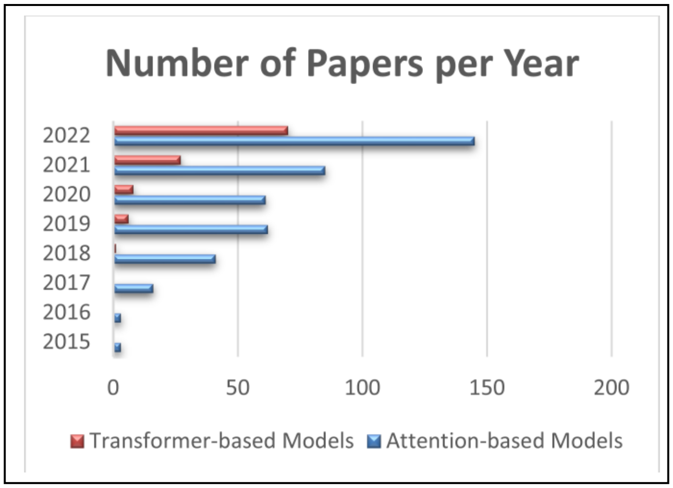

Fig. 1. Number of attention-based papers in image captioning per year

II. Literature Review

Many surveys have been published for the deep learning techniques in image captioning [17, 38-42]. Some of these surveys [17, 38, 40-42] considered few attention-based approaches in image captioning because most of the attention approaches were issued after publishing these deep learning surveys. A comparative study for attention-based techniques was published by Khaing and Phyu [43]. Their survey presented a good comparative study of the attention-based models but without any categorization and moreover, the most recent reference in their survey was in the year 2018, and there is big progress in the attention-based methods starting from the year 2019 as shown in Fig. 1. The newest survey for attention-based models was presented by Zohourianshahzadi and Kalita [44]. In [44], they presented an evolution path of the attention models including hard and soft attention, semantic attention, spatial attention, adaptive attention, and bottom-up and top-down attention.

As per our knowledge, there is no detailed survey with a good taxonomy for the attention-based approaches employed in image captioning. Motivated by this gap in the existing image captioning survey papers, especially for the attention-based approaches, a detailed survey for the attention-based approaches employed in image captioning is presented in this paper by introducing new categories.

III. Taxonomy of Attention-Based Models

Employing the attention mechanism in image captioning was motivated by the successful work achieved in neural machine translation [45] and object recognition [46, 47]. The attention was employed in the decoder part of the translation task to mitigate the encoder from the need to model all input sentence information [45]. Xu et al. [48] proposed captioning approach by exploring the attention technique to consider the significant regions in the process of caption generation.

According to [48], the attention was applied at the decoder so that at every time step (t), LSTM produced a new word depending on the hidden state (h_{t-1}), the words produced at the previous steps and a vector called context vector $(\hat{z}_t)$. The context vector $(\hat{z}_t)$ represents the information of an appropriate location of the image at specific time step t. The context vector $\hat{z}_t$ can be calculated using the annotation vectors, which are the features related to the image regions, and their assigned weights $\propto$. The weights $\propto$ are assigned to every annotation vector a_i, $(i = 1,..L)$ using Multilayer Perceptron depending on the previous step hidden state h_{t-1}. The attention model f_{att} used for calculating the weights had two variants either soft or hard attention depending on how the weights will be interpreted.

Variants of the attention model were proposed in image captioning research area, some researchers enhanced the model by employing the attention as multi-stages or by inserting information to guide the attention. However, the most notable variant of the attention is the transformer-based models as can be seen from Fig. 1, there is a big interest in applying the transformer-based models in comparison with the other categories.

In image captioning, the attention mechanisms can be categorized into four categories, as demonstrated in Fig. 2. According to Chen et al. [49], the visual attention-based approaches may concentrate on the spatial features or the semantic features, so visual attention is added as a category for characterizing the attention-based models. In addition, according to He et al. [50], the attention-based methods can be categorized based on applying the attention as single-stage in the decoder, two-stages, two-stages with scene graph or based on the transformer. This classification is added as subcategories into the category named Attention Blocks. In addition to these main two categories, in this survey paper, two new categories that were not included in other survey papers for characterizing the attention-based models are added, which are Number of Attention Layers and Guided-Attention.

A. Visual Attention

Visual Attention [51, 52] is a significant technique in the human visual system. The brain targets a region or an object using computational capabilities with the guidance of low-level image features in a time step. The visual attention models can be divided into spatial and semantic attention.

1) Spatial attention: For spatial attention, the attention is demonstrated spatially at a specific region [48, 49, 53-56]. For each fixed location, attention weights are calculated related to this location at each iteration. Several approaches apply soft attention, which models the feature maps with the computed weights.

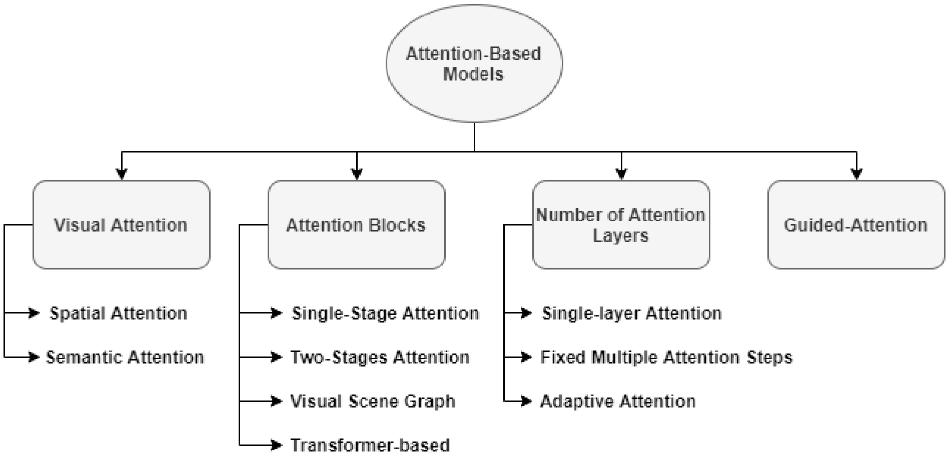

Fig. 2. Taxonomy of attention-based image captioning models

While other approaches use hard attention by selecting a set of regions which are salient from the feature map and concealing the other regions. Through applying the weighted pooling, some of the important spatial data may be lost. In addition, regularly the spatial attention is computed in the last convolutional layer, which leads to some analogous feature results for distinct regions because of the big size of the filter, resulting in ineffective spatial attention.

2) Semantic attention: Instead of attending to the fixed resolutions in spatial attention, other approaches proposed attending to the image's semantic concepts [57, 58]. Semantic attention is more like the human description of the image because people describe the most important objects and do not talk about all regions in the image. Attributes can be utilized from any image location even if there is no actual existence of these attributes within the image. For the purpose of attending to semantically necessary attributes, You et al. [57] employed a semantic attention framework that used top-down and bottom-up models. Bottom-up was used to select the semantic attributes, and top-down was used to decide when and where to apply the attention. Another approach was proposed by Gan et al. [58] in which they recognized the semantic tags and computed the probability of the tags to be utilized in forming the LSTM parameters. LSTM weight matrices were expanded to a group of weight matrices that are tag-dependent.

Using semantic attention requires extra resources that are important for detecting the relationship among the semantic concepts and the image.

B. Attention Blocks

The attention-based models can also be categorized according to the block where the attention is applied. The attention can be applied as a single-stage in the decoder block, two-stages by obtaining bottom-up and top-down attentions, two-stages with injecting a graph network, or Transformer-based models.

1) Decoder-based attention (single-stage attention): In decoder-based attention models, the attention is employed at the decoder. In the process of producing the caption words, the informative regions [59] are targeted in the attention by the decoder. Depending on the LSTM hidden states and the previously predicted caption words, Xu et al. [48] proposed to use the attention module in the decoder of the captioning approach while generating the sentence words. A weighting matrix is introduced for each feature map receptive field then this weighted map and the last predicted word were forwarded to the language model for the purpose of predicting the next word.

2) Two-stage attention: Rather than attending to the salient regions like Decoder-based attention, Anderson et al. [53] presented a model that contains two-stage attention. Faster R-CNN [60] was employed in the bottom-up attention module. Then, the attention was distributed among the image regions using a top-down attention mechanism. They used two LSTM layers for the purpose of applying attention to the selected spatial features, the first layer was for the top-down attention, and the other was for the language layer. The drawback of their model is that it cannot handle object-object relationships.

An approach that is similar to [53] was introduced by Lu et al. [61]. Their proposed decoder determines whether the word will be visual and predicted according to a certain image region or the word will be predicted from the textual vocabulary. The essential advantage of their approach is in its availability to have additional object detectors, which can lead to producing different image captions. The main gap in two-stage attention models is that the models are lacking for getting the relationship between the image regions.

3) Two-stage attention with graph: To enhance the two-stage attention models, graph networks can be employed to discover the relationship among the detected regions which can result in enhanced features and accordingly improve the caption generation. Similar to [53, 62], Yao et al. [55] employed the attention mechanism for attending to the informative image regions. The key novelty in (GCN-LSTM) [55] is that they used two graphs for detecting the relationship between the image regions. A semantic graph was employed with the nodes representing the image regions and the edges representing the relationship between these detected image regions. While the geometrical relations between the regions 'vertices' were demonstrated by the spatial graph. Then, Graph Convolutional Network (GCN) [63] was utilized to output relation-aware region representations.

The approaches presented in [55, 64] employed Faster R-CNN to identify the image objects and thus explore the relationships between regions of interest. Faster R-CNN was trained on the Visual Genome dataset [65]. While in [66], the visual relationships were modelled on Flickr30K [67] and MS COCO [68] and so the pre-established classes of the relations are not required.

The authors in [55] extended the approach to (GCN-LSTM-HIP) [69] to include a hierarchical tree of three levels which have the image as the root, the detected regions as the first layer and the instances/foreground of the regions at the leaf layer. Then, a Tree-LSTM [70] was employed for modelling the dependency structure and improving the features.

Another model presented by Guo et al. [71] which detected a set of visual semantic units 'VSUs' where the units represent the objects, attributes and the object's relationships. Semantic and geometry graphs were employed while the vertices representing the semantic units and the edges representing the connections between them differed from [55] that presented the relationships as edges. GCN was then introduced in [71] to output context-aware embeddings for the visual semantic units. Attention for the different kinds of units was applied via context gated attention (CGA). Another scene graph approach was presented by Yang et al. [72] that used the edges to represent the relationships in the graph. Language inductive bias was integrated into the captioning framework, and its features are represented via a scene graph auto-encoder (SGAE).

The main drawback in the graph scene-based models is that, however, the models made an enhancement to the performance compared to the two-stage models, but the need for additional models for scene graph construction is still a problem. Also, with respect to the computational cost, having two graphs is ineffective.

4) Transformer-based: Unlike the graph-based models, the transformer models don't include any graphs and thus don't need additional models for the graph construction. The transformer was originally designed for text translation [73]. The transformer is able to avoid any duplication by employing the attention in a comprehensive way between the input and the output. Extensive approaches were proposed to employ the transformer models in image captioning [74-85].

Huang et al. [86] proposed Attention on Attention (AoA) approach, which adds attention over the traditional attention. "Information vector" and "attention gate" were produced by the query and the attended results, then second attention was produced by element-wise multiplication between them. AoA was applied in the encoder to detect the relations between the objects. While in the decoder, AoA was employed for holding the relevant attention output and ignoring the deceptive results.

Captioning transformer with stacked attention module was proposed by Zhu et al. [76]. A multi-level observation was proposed in such a way that all transformer layers had the opportunity for generating the sentence word. Average pooling was then employed to find the probability of the word by merging all the contributions.

Cornia et al. [74] proposed a transformer approach to consider low and high-level relationships by modelling them as multi-level. They utilized persistent memory vectors while encoding the relationships with prior information. In addition, rather than applying the attention only to the last encoding layer, all the encoder layers contributed to the sentence generation process and connected to the decoder layers in mesh-like connectivity.

A Multimodal transformer was proposed by Yu et al. [75], which is able to model three different relations, which are: word-to-word, word-to-object and object-to-object. Self-attention in the same modality and co-attention in distinct modalities were acquired. In addition, multiple views were employed in two designs: aligned and unaligned multiple views.

The conventional transformer was expanded with the addition of EnTangled Attention (ETA) and Gated Bilateral Controller (GBC) [77]. ETA gave the transformer the ability to use semantic concepts and visual information. The interconnection between the multimodal information was controlled by the GBC. Object relation transformer [78] was proposed in which geometrical information for the relationship between each pair of objects was included within the transformer through spatial attention.

He et al. [50] proposed a model with the idea of changing the internal structure of the transformer that was originally proposed to handle text. They introduced an expanded transformer that includes three parallel sub-transformer layers to handle three different relationships: parent, child, and neighbor.

C. Number of Attention Layers

The attention models can be characterized according to the number of required attention steps either to attend once per word, attend with fixed steps or adaptively determine the number of required attention steps.

1) Single-layer attention: The attention operation is connected to the word generation procedure in the traditional attention-based framework [48]. The framework attended once to the image prior to generating the following word. The model attended to selected image regions in each iteration, and the computed attention features were sent to the RNN as input. The problem of attending once per word is that some important information may be lost, especially if the model attended to an incorrect region.

2) Fixed multiple attention steps: In order to enhance the single attention process by avoiding predicting incorrect words, several approaches attended multiple times to enhance the attended region and get the lost data [87, 88]. Du et al. [87] proposed a model that attended more times to the image per word and showed that it could improve image captioning without adding extra parameters. Two LSTMs model were utilized, which have the ability to attend for arbitrary times and enable the flexibility of the attention operation.

Triple attention approach was proposed by Zhu et al. [88]. The attention is utilized to the input phase of the previous step LSTM hidden states. In addition, attention was also utilized in the output phase of present hidden states. Conditional embedding was used in addition to the word/image embedding at every input stage of LSTM. This way, the prior text information was coupled with image information, and accordingly, text and image information appeared in the input of the word generation procedure.

For the purpose of getting attention to different semantic abstractions, Chen et al. [49] applied the attention in a multi-layer since the lower layers are the dependent layers for the feature maps. The attention in their approach was approached to each entry of the feature maps, which are multi-layer. They also proposed channel-wise attention for applying the re-weighting process in every channel through the word generation process. The channel-wise attention could be viewed as the procedure of choosing the semantic concepts by paying more attention to the channels produced by filters indicated by the semantics.

A hierarchical approach (CNN+CNN) [89] was proposed such that they employed the CNN as a decoder besides being the encoder. Their hierarchical attention model learns the relationship of the attributes for all image regions and all levels. The dot-product operation used in their framework results in reducing the parameters and can be faster than Multi-layer Perceptron attention used in [48, 54]. The idea of hierarchical attention was also employed in [90-92]. Yan et al. [90] proposed a mechanism made up of global and local attention modules which related to the global CNN features, extracted by CNN encoder, and local object features, extracted by object detector, respectively.

Sequential attention was presented by Fang et al. [93] to take into consideration the sequential attention relationships in several time steps at word generation and correspondingly improve the visual data in caption generation. Another sequential attention was proposed by Liu et al. [94], in which the image was represented as a sequence of objects, and the attention was employed to consider all objects information during sentence word generation.

3) Adaptive attention: According to the previously presented approaches, sometimes there are no image regions corresponding to each sentence word. So, an adaptive attention approach [54] was proposed that includes a sentinel gate and spatial attention to determine where and when to attend in the caption generation. They presented an extension to LSTM that, rather than having one hidden state, they added a visual sentinel vector. In addition, a sentinel gate was proposed to determine whether the attention will be targeted to the visible sentinel or the image. Another adaptive attention approach was proposed by Deng et al. [95] that adaptively determine whether it is needed to depend on the language model or the visual signals. Their proposed approach can make the image captioning task more flexible by enhancing the obligatory correlation between image regions and sentence words.

Adaptive semantic attention framework [96] was proposed to incorporate dual-LSTMs; the first LSTM works as a visual sentinel to acquire fine-grained representations. The second LSTM serves as a language model that produces the sentence words depending on the updated attended vector and first LSTM output.

Huang et al. [97] presented an adaptive attention time model (AAT). The model was learned to determine the number of required attention steps in each step of the decoder in order to produce the next word. Using AAT, the mapping between the image regions and caption words can be applied arbitrarily such that a caption word may attend to multiple regions and vice versa. Their approach doesn't add parameters gradient noise.

D. Guided Attention

For the purpose of enhancing the performance of image captioning approaches and generating accurate captions, some approaches inserted additional information guidance [98, 99] like the concept features that make a connection between the input image and the caption. In [100], the model was guided through semantic information acquired from the images and sent as extra input for the LSTM units. While in [101], the approach could be guided through concept features which are obtained from predicting the recurrent word existence in the captions. Another way for learning the features is by adding a network for guidance [102]. More similar to [102], Sow et al. [103] inserted a network for guidance, but rather than obtaining one vector for guidance, [103] obtained a sequential network for guidance which was able to adjust the guided vectors in the sentence generation process. They also utilized the Luong attention mechanism [104] that is an enhanced style of the attention technique.

Text-guided attention approach was presented by Mun et al. [105]. Related sample captions, namely guidance captions, were employed to get visual attention and produce appropriate captions. The related sample captions were obtained through the similar training images that participate in equivalent related regions with the input image. Topic-guided attention was proposed by Zhu et al. [106], which picked up the significant features by the information guidance through incorporating the topics within the image with the attention mechanism.

IV. Datasets and Performance Metrics

A. Datasets

Different datasets have been presented in the research area of image captioning. The popular datasets, which are Flickr8K [107], Flickr30k [67], Microsoft COCO [68] and Visual Genome [65] are presented.

1) Flickr8K [107]: Dataset consists of about eight thousand images selected from six groups on Flickr.com and does not have a tendency to famous locations or people; instead, various situations and locations are represented. The dataset includes five captions for each image through human annotations.

2) Flickr30K [67]; Extension to Flickr8K, consists of 31,783 images. Flickr30k contains 8.7 objects per image, 44,518 object categories, 6.2 objects per category, 5 sentences per image and 16.6 expressions per image.

3) Microsoft COCO dataset [68]: A large-scale dataset that broadly used in image captioning task. MS COCO includes 328,000 images, 7.7 objects per image, 91 object categories, 2.5 million labelled instances, 27,473 objects per category and five sentences per image.

4) Visual genome dataset [65]; It is an image captioning dataset that considers the relationship modelling between objects. It generates captions for different image regions, unlike the other datasets, which generate the caption to the entire scene. The dataset includes more than 100 thousand images, 18 attributes, 21 objects per image, and 18 objects relationships.

B. Performance Metrics

For the purpose of evaluating the image captioning techniques, different metrics were proposed to compare the output generated caption with the original caption. In this section, the main used performance metrics which are BLEU [108], ROUGE [109], METEOR [110], CIDEr [111] and SPICE [112] are presented.

1) BLEU "Bilingual Evaluation Understudy" [108]: It is originally introduced by IBM for the evaluation of machine translation. This metric measures the quality of the generated sentence by calculating its similarity with the original reference translations. N-grams of the machine-generated sentence are compared to those of the reference sentences and get the matching counter. The output score is higher, and the quality of the generated sentence is better when there are more reference sentences and there is a higher number of matches. The range of BLEU values is from zero to one, and a small number of generated captions can get one only if it is identical to the ground truth caption.

2) ROUGE [109]: It is originally introduced for the evaluation of text summarization. ROUGE metric calculates the quality of the text generated summary by counting the number of its n-gram, sequences of words, and pairs of words that overlapped with the reference summaries created by experts. ROUGE-N (N-grams), ROUGE-L, ROUGE-W, and ROUGE-S are the types of the ROUGE metric.

3) METEOR [110]: A metric utilized for evaluating the machine-generated texts by matching the unigrams of the machine-generated sentence and the reference sentences. Once this matching is computed, recall and precision of unigram and a measure of fragmentation were utilized for computing a METEOR score.

4) CIDEr [111]: A metric utilized for evaluating the image descriptions. The five available captions of the dataset used in the other metrics are not enough for finding the consensus among the judgment of the human and the output captions. A consensus is a measurement for counting the mutual n-grams between the ground truth and predicted captions and assigning low weights for the common n-grams.

5) SPICE [112]: The previously explained metrics depend on the n-grams and SPICE metric overcomes this restriction by employing a scene graph in which the reference and generated captions are converted to a graph-based semantic representation. SPICE is measuring if the objects and attributes are represented in the generated caption in an effective way in addition to their relationships.

V. Comparison and Discussion

In this section, the performance of different state-of-the-art approaches is presented and discussed. In Table I, different approaches are compared with respect to their experimental results on the benchmark MS COCO dataset and the commonly used performance metrics BLEU-4 (B@4), METEOR (MT), ROUGE-L (R), and CIDEr (C).

From the beginning of using the attention mechanism in image captioning by Xu et al. [48], it has been shown that their approach obtained better performance on Flick8k, Flickr30k and MS-COCO. The reason behind the better performance is that their approach considered the most relevant objects when generating the image caption. Moreover, they showed that the hard attention variant of their mechanism outperforms the soft attention on these benchmark datasets. After that, You et al. [57] showed that attending to the semantic attributes instead of attending to the spatial attention [48] can improve the results by generating semantically rich captions.

Further improvement in the results was obtained by introducing multiple attention layers, which can be used in a hierarchical structure or by using either a fixed or adaptive number of attention layers. Du et al. [87] achieved 38.1, 28.3, 58.0, 126.1 and 22.0 on BLEU-4, METEOR, ROUGE, CIDEr and SPICE, respectively. These results are higher than the results of hierarchical attention [89]. The hierarchical structure in [89] used the CNN as decoder; however, Du et al. [87] used two LSTMs model to enable attention at arbitrary times and make the attention operation more flexible.

The adaptive attention approach of Huang et al. [97] achieved 38.7, 28.6, 58.5, 128.6 and 22.2 on BLEU-4, METEOR, ROUGE, CIDEr and SPICE, respectively, which are higher than that of both Wang and Chan [89] and Du et al. [87]. The reason for their better performance is that their model was learned to determine the number of required attention steps in each decoder step, and the mapping between the image regions and caption words can be applied arbitrarily.

Anderson et al. [53] achieved a good performance by employing a two-stage decoder containing bottom-up attention and top-down attention. Yao et al. [55], Guo et al. [71] and Yang et al. [72] further enhanced the results of the two-stage decoder by introducing scene graphs for detecting the relationship between image regions. Yao et al. [69] achieved better results than [55, 71, 72] by introducing a hierarchical tree and using a tree-LSTM to model the dependency structure.

The best performance in Table I was achieved by Yu et al. [75] and Pan et al. [114]. In [75], Yu et al. used a multimodal transformer that can model three different relations, and the model was designed in two views aligned and unaligned multi-view visual representation. However, Pan et al. [114] modelled second order interactions through proposing X-linear attention module plugged into transformer. Both of [75] and [114] are Transformer-based attention models which proves that Transformer-based models can achieve better results in comparison with other attention-based mechanisms. The big interest in applying the transformer, as can be seen from Fig. 1, comes from its ability to weight the importance of every input region and its ability to avoid any duplication by employing the attention in a comprehensive way between the input and the output. In addition, it can be parallelized in an effective way.

Employing the attention mechanism in image captioning started from the year 2015 [48] and it is getting more attention from that time since the number of research papers employed the attention is increasing every year as explained in Fig. 1. In addition, the authors have a tendency for using the scene graph with attention models and also great attention is going towards applying the transformer in the image captioning task due to its parallelization nature and better performance. In addition, part of the research in image captioning task recently is going towards applying the attention as multi-layer in order to enhance the predicted words or adaptively determine the number of required attention steps.

TABLE I. COMPARISON BETWEEN THE STATE-OF-THE-ART ATTENTION-BASED CAPTIONING APPROACHES

Ref.	Year	Category of the Attention	Results (C5)			
			B@4	MT	R	C
[48]	2015	Spatial Single Stage	25.0	23.04	-	-
[57]	2016	Semantic Single Stage	31.6	25.0	53.5	94.3
[49]	2017	Spatial Multi-Layer	30.2	24.4	52.4	91.2
[94]	2017	Multi-Layer	32.0	25.8	54.0	102.9
[54]	2017	Adaptive	33.6	26.4	55.0	104.2
[89]	2018	Multi-Layer	26.7	23.4	51.0	84.4
[76]	2018	Transformer	33.3	-	54.8	108.1
[88]	2018	Multi-Layer	33.8	27.0	55.4	106.4
[61]	2018	Two-Stages	34.7	27.1	-	107.2
[93]	2018	Multi-Layer	34.9	26.7	-	108.1
[102]	2018	Guided	35.3	26.7	55.5	107.8
[53]	2018	Two-Stages	36.9	27.6	57.1	117.9
[87]	2018	Multi-Layer	38.1	28.3	58.0	126.1
[55]	2018	Two-Stages with Graph	38.7	28.5	58.5	125.3
[103]	2019	Guided	34.0	26.3	55.2	103.6
[71]	2019	Two-Stages with Graph	37.4	28.2	57.9	123.1
[72]	2019	Two-Stages with Graph	38.5	28.2	58.6	123.8
[97]	2019	Adaptive	38.7	28.6	58.5	128.6
[77]	2019	Transformer	38.9	28.6	58.6	122.1
[69]	2019	Two-Stages with Graph	39.3	28.8	59.0	127.9
[86]	2019	Transformer	39.4	29.1	58.9	126.9
[75]	**2019**	**Transformer**	**40.4**	**29.4**	**59.6**	**130**
[90]	2020	Multi-Layer	28.5	25.3	56.5	92.4
[95]	2020	Adaptive	32.6	27.0	-	-
[66]	2020	Two-Stages with Graph	34.3	27.0	55.5	106.1
[113]	2020	Transformer	38.8	29.0	58.7	126.3
[74]	2020	Transformer	39.7	29.4	59.2	129.3
[50]	2020	Transformer	39.6	29.1	59.2	127.4
[114]	**2020**	**Transformer**	**40.3**	**29.6**	**59.5**	**131.1**
[81]	2021	Transformer	40.0	29.1	59.4	129.4
[82]	2021	Transformer	38.5	28.9	58.6	129.6
[80]	2021	Adaptive	36.3	27.2	56.8	113.3
[98]	2021	Guided	39.8	28.8	59.4	128.3
[99]	2022	Guided	35.9	28.4	57.3	115.9
[85]	2022	Transformer	37.9	28.8	58.1	126.7
[115]	2022	Transformer	39.2	28.8	58.7	125.6
[116]	2022	Transformer	39.9	29.1	59.1	127.8
[117]	2022	Adaptive	38.5	28.3	57.5	120.7

VI. CONCLUSION AND FUTURE WORK

In this paper, a survey was presented for the attention-based image captioning approaches. Four main categories of the attention-based approaches and their subcategories are summarized. Furthermore, the attention-based approaches were compared on benchmark datasets and popular performance metrics. As discussed in the paper, there is a great improvement in the image captioning task due to using the attention-based models especially using Transformer-based approaches. Although there is an impressive effect of using the attention-based models in image captioning, there is still room for improvement. Faster R-CNN is extensively employed as an encoder because of its ability to get effective detection results. However, training of Faster R-CNN is not a simple task, and it gives unsatisfied results in some cases, like when having images of low resolutions or when the objects are deformed or of small size. So, it may be better if other image encoders are used or when an enhanced version of Faster-RCNN is employed. In addition, another room for improvement can be found in the transformer-based models with introducing new transformer architectures, which may help in improving the quality of the result description.

REFERENCES

[1] Xinlei Chen, and C. Lawrence Zitnick. “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.

[2] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. “On the properties of neural machine translation:

Encoder-decoder approaches," In Association for Computational Linguistics, 2014, pp. 103–111.

[3] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.

[4] Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler. "Towards diverse and natural image descriptions via a conditional GAN," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), 2017, pp. 2989–2998.

[5] Soad Samir, Eid Emary, Khaled El-Sayed, and Hoda Onsi. "Optimization of a pre-trained AlexNet model for detecting and localizing image forgeries," Information, 2020, 11(5): 275.

[6] Ahmed Ali Hammam, Mona M. Soliman, Aboul Ella Hassanein. "Real-time multiple spatiotemporal action localization and prediction approach using deep learning," Neural Networks, 2020, 128: 331–344.

[7] Elham S. Salama, Reda A.El-Khoribi, Mahmoud E. Shoman, Mohamed A. Wahby Shalaby. "A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition," Egyptian Informatics Journal, 2021, 22(2): 167–176.

[8] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection," In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'05, 2005, 1: 886–893.

[9] Marie-Catherine De Marneffe, Bill MacCartney, and Christopher D. Manning. "Generating typed dependency parses from phrase structure parses," In Proceedings of LREC, 2006, 6: 449–454.

[10] Etienne Denoual and Yves Lepage. "BLEU in characters: Towards automatic MT evaluation in languages without word delimiters," In Companion Volume to the Proceedings of the 2nd International Joint Conference on Natural Language Processing, 2005, pp. 81–86.

[11] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. "Language models for image captioning: The quirks and what works," arXiv preprint arXiv:1505.01809, 2015.

[12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. "Long-term recurrent convolutional networks for visual recognition and description," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.

[13] Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru 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.

[14] A. Karpathy and L. Fei-Fei. "Deep visual-semantic alignments for generating image descriptions," CVPR, 2015.

[15] Junhua Mao,Wei Xu, Yi Yang, JiangWang, Zhiheng Huang, and Alan Yuille. "Deep captioning with multimodal recurrent neural networks (m-RNN)," In International Conference on Learning Representations (ICLR'15), 2015.

[16] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. "Explain images with multimodal recurrent neural networks," arXiv preprint arXiv:1410.1090, 2014.

[17] Hossain, MD Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. "A comprehensive survey of deep learning for image captioning," ACM Computing Surveys (CsUR), 2019, 51(6): 1-36.

[18] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. "Unifying visual-semantic embeddings with multimodal neural language models," In Workshop on Neural Information Processing Systems (NIPS'14), 2014.

[19] Andrej Karpathy, Armand Joulin, and Fei Fei F. Li. "Deep fragment embeddings for bidirectional image sentence mapping," In Advances in Neural Information Processing Systems, 2014, pp. 1889–1897.

[20] Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, and Qi Ju. "Improving image captioning with conditional generative adversarial nets," In Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 8142-8150.

[21] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. "Speaking the same language: Matching machine to human captions by adversarial training," In IEEE International Conference on Computer Vision (ICCV'17), 2017, pp. 4155–4164.

[22] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. "Deep reinforcement learning-based image captioning with embedding reward," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), 2017, pp. 1151–1159.

[23] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. "Self-critical sequence training for image captioning," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17), 2017, pp. 1179–1195.

[24] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. "An empirical study of language CNN for image captioning," In Proceedings of the International Conference on Computer Vision (ICCV'17), 2017, pp. 1231–1240.

[25] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. "Improved image captioning via policy gradient optimization of spider," In Proceedings of the IEEE International Conference on Computer Vision (ICCV'17), 2017, 3: 873–881.

[26] Shubo Ma and Yahong Han. "Describing images by feeding LSTM with structural words," In 2016 IEEE International Conference on Multimedia and Expo (ICME'16), IEEE, 2016, pp. 1–6.

[27] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. "Captioning images with diverse objects," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1170–1178.

[28] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. "Image captioning and visual question answering based on attributes and external knowledge," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6), pp. 1367–1381.

[29] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M. Hospedales. "Actor-critic sequence training for image captioning," arXiv preprint arXiv:1706.09601, 2017.

[30] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. "Multimodal neural language models," In Proceedings of the 31st International Conference on Machine Learning (ICML'14), 2014, pp. 595–603.

[31] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4565–4574.

[32] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. "Convolutional image captioning," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5561–5570.

[33] Y. Bin, Y. Yang, J. Zhou, Z. Huang, and H.T. Shen. "Adaptively Attending to Visual Attributes and Linguistic Knowledge for Captioning," In Proceedings of the 2017 ACM on Multimedia Conference, 2017, pp. 1345-1353.

[34] S. Qu, Y. Xi, and S. Ding. "Visual Attention Based on Long-Short Term Memory Model for Image Caption Generation," Control and Decision Conference (CCDC), 2017 29th Chinese, 2017, pp. 4789-4794.

[35] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian. "GLA: Global-Local Attention for Image Description," IEEE Trans. on Multimedia, 2018, 20(3): 726-737.

[36] S. Ye, J. Han, and N. Liu. "Attentive Linear Transformation for Image Captioning," IEEE Trans. on Image Processing, 2018, 27(11): 5514-5524.

[37] Cornia, Marcella, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. "Paying more attention to saliency: Image captioning with saliency and context attention," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2018, 14(2): 1-21.

[38] Bai, Shuang, and Shan An. "A survey on automatic image caption generation," Neurocomputing 311, 2018, pp. 291-304.

[39] Bernardi, Raffaella, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. "Automatic description generation from images: A survey of models, datasets, and evaluation measures," Journal of Artificial Intelligence Research 55, 2016, pp. 409-442.

[40] Liu, Xiaoxiao, Qingyang Xu, and Ning Wang. "A survey on deep neural network-based image captioning," The Visual Computer, 2019, 35(3): 445-470.

[41] Staniūtė, Raimonda, and Dmitrij Šešok. "A systematic literature review on image captioning," Applied Sciences, 2019, 9(10): 2024.

[42] Wang, Yiyu, Jungang Xu, Yingfei Sun, and Ben He. "Image Captioning based on Deep Learning Methods: A Survey," arXiv preprint arXiv:1905.08110, 2019.

[43] Khaing, Phyu Phyu. "Attention-Based Deep Learning Model for Image Captioning: A Comparative Study," International Journal of Image, Graphics and Signal Processing, 2019, 10(6): 1.

[44] Zohourianshahzadi, Zanyar, and Jugal K. Kalita. "Neural attention for image captioning: review of outstanding methods," Artificial Intelligence Review, 2022, pp. 1-30.

[45] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. "Neural machine translation by jointly learning to align and translate," arXiv:1409.0473, September 2014.

[46] Ba, Jimmy Lei, Mnih, Volodymyr, and Kavukcuoglu, Koray. "Multiple object recognition with visual attention," arXiv:1412.7755, December 2014.

[47] Mnih, Volodymyr, Hees, Nicolas, Graves, Alex, and Kavukcuoglu, Koray. "Recurrent models of visual attention," In NIPS, 2014.

[48] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y. "Show, attend and tell: Neural image caption generation with visual attention," In International conference on machine learning, 2015, pp. 2048-2057.

[49] Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S. "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5659-5667.

[50] He, Sen, Wentong Liao, Hamed R. Tavakoli, Michael Yang, Bodo Rosenhahn, and Nicolas Pugeault. "Image captioning through image transformer," In Proceedings of the Asian Conference on Computer Vision, 2020.

[51] C. Koch and S. Ullman. "Shifts in selective visual attention: towards the underlying neural circuitry," In Matters of intelligence, Springer, 1987, pp. 115–141.

[52] M.W. Spratling and M. H. Johnson. "A feedback model of visual attention," In Journal of cognitive neuroscience, 2004, 16(2): 219–237.

[53] Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei 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.

[54] Lu, Jiasen, Caiming Xiong, Devi Parikh, and Richard Socher. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 375-383.

[55] Yao, Ting, Yingwei Pan, Yehao Li, and Tao Mei. "Exploring visual relationship for image captioning," In Proceedings of the European conference on computer vision (ECCV), 2018, pp. 684-699.

[56] Zhou, Dongming, Jing Yang, and Riqiang Bao. "Collaborative strategy network for spatial attention image captioning," Applied Intelligence 52, no. 8 (2022): 9017-9032.

[57] You, Quanzeng, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. "Image captioning with semantic attention," In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4651-4659.

[58] Gan, Zhe, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. "Semantic compositional networks for visual captioning," In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5630-5639.

[59] Luo, W., Li, Y., Urtasun, R., Zemel, R. "Understanding the effective receptive field in deep convolutional neural networks," In Advances in neural information processing systems, 2016, pp. 4898-4906.

[60] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural information processing systems, 2015, 28: 91-99.

[61] Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh. "Neural baby talk," In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7219-7228.

[62] Jin, Junqi, Kun Fu, Runpeng Cui, Fei Sha, and Changshui Zhang. "Aligning where to see and what to tell: image caption with region-based attention and scene factorization," arXiv preprint arXiv:1506.06272, 2015.

[63] Kipf, T.N., Welling, M. "Semi-supervised classification with graph convolutional networks," In: ICLR, 2017.

[64] Y. Li, W. Ouyang, B. Zhou, K. Wang, X. Wang. "Scene graph generation from objects, phrases and region captions," In Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1261–1270.

[65] Krishna, Ranjay, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International journal of computer vision, 2017, 123(1): 32-73.

[66] Junbo Wang, Wei Wang, Liang Wang, Zhiyong Wang, David Dagan Feng, Tieniu Tan. "Learning Visual Relationship and Context-Aware Attention for Image Captioning," Pattern Recognition, 2020.

[67] Plummer, Bryan A., Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," In Proceedings of the IEEE international conference on computer vision, 2015, pp. 2641-2649.

[68] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft Coco: Common objects in context," In European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[69] Yao, Ting, Yingwei Pan, Yehao Li, and Tao Mei. "Hierarchy parsing for image captioning," In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2621-2629.

[70] Kai Sheng Tai, Richard Socher, and Christopher D Manning. "Improved semantic representations from tree-structured long short-term memory networks," In ACL, 2015.

[71] Guo, Longteng, Jing Liu, Jinhui Tang, Jiangwei Li, Wei Luo, and Hanqing 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.

[72] Yang, Xu, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. "Auto-encoding scene graphs for image captioning," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10685-10694.

[73] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need," In Advances in neural information processing systems, 2017, pp. 5998-6008.

[74] Cornia, Marcella, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. "Meshed-memory transformer for image captioning," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10578-10587.

[75] Yu, Jun, Jing Li, Zhou Yu, and Qingming Huang. "Multimodal transformer with multi-view visual representation for image captioning," IEEE transactions on circuits and systems for video technology, 2019, 30(12):4467-4480.

[76] Zhu, Xinxin, Lixiang Li, Jing Liu, Haipeng Peng, and Xinxin Niu. "Captioning transformer with stacked attention modules," Applied Sciences, 2018, 8(5): 739.

[77] Li, Guang, Linchao Zhu, Ping Liu, and Yi Yang. "Entangled transformer for image captioning," In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 8928-8937.

[78] Herdade, Simao, Armin Kappeler, Kofi Boakye, and Joao Soares. "Image captioning: Transforming objects into words," arXiv preprint arXiv:1906.05963, 2019.

[79] Jiangyun Li, Peng Yao, Longteng Guo, and Weicun Zhang. "Boosted transformer for image captioning," Applied Sciences, 2019, 9(16): 3260.

[80] Chenggang Yan, Yiming Hao, Liang Li, Jian Yin, Anan Liu, Zhendong Mao, Zhenyu Chen, and Xingyu Gao. "Task-adaptive attention for

image captioning," IEEE Transactions on Circuits and Systems for Video technology, 2021, 32(1): 43-51.

[81] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. "CPTR: Full transformer network for image captioning," arXiv preprint arXiv:2101.10804, 2021.

[82] Weitao Jiang, Xiying Li, Haifeng Hu, Qiang Lu, and Bohong Liu. "Multi-gate attention network for image captioning," IEEE Access, 2021, 9: 69700-69709.

[83] Kumar, Deepika, Varun Srivastava, Daniela Elena Popescu, and Jude D. Hemanth. "Dual-Modal Transformer with Enhanced Inter-and Intra-Modality Interactions for Image Captioning," Applied Sciences 12, no. 13 (2022): 6733.

[84] Sarto, Sara, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Retrieval-augmented transformer for image captioning," In Proceedings of the 19th International Conference on Content-based Multimedia Indexing, pp. 1-7. 2022.

[85] Dubey, Shikha, Farrukh Olimov, Muhammad Aasim Rafique, Joonmo Kim, and Moongu Jeon. "Label-attention transformer with geometrically coherent objects for image captioning," Information Sciences, 2022.

[86] Huang, Lun, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. "Attention on attention for image captioning," In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4634-4643.

[87] Du, Jiajun, Yu Qin, Hongtao Lu, and Yonghua Zhang. "Attend more times for image captioning," arXiv preprint arXiv:1812.03283, 2018.

[88] Zhu, Xinxin, Lixiang Li, Jing Liu, Ziyi Li, Haipeng Peng, and Xinxin Niu. "Image captioning with triple-attention and stack parallel LSTM," Neurocomputing, 2018, 319: 55-65.

[89] Wang, Qingzhong, and Antoni B. Chan. "Cnn+ cnn: Convolutional decoders for image captioning," arXiv preprint arXiv:1805.09019, 2018.

[90] Yan, Shiyang, Yuan Xie, Fangyu Wu, Jeremy S. Smith, Wenjin Lu, and Bailing Zhang. "Image captioning via hierarchical attention mechanism and policy gradient optimization," Signal Processing, 2020, 167: 107329.

[91] Wang, Qingzhong, and Antoni B. Chan. "Gated hierarchical attention for image captioning" In Asian Conference on Computer Vision, Springer, Cham, 2018, pp. 21-37.

[92] Wang, Weixuan, Zhihong Chen, and Haifeng Hu. "Hierarchical attention network for image captioning," In Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 8957-8964.

[93] Fang, Fang, Qinyu Li, Hanli Wang, and Pengjie Tang. "Refining attention: a sequential attention model for image captioning," In 2018 IEEE international conference on multimedia and expo (ICME), 2018, pp. 1-6.

[94] Liu, Chang, Fuchun Sun, Changhu Wang, Feng Wang, and Alan Yuille. "MAT: A multimodal attentive translator for image captioning," arXiv preprint arXiv:1702.05658, 2017.

[95] Deng, Zhenrong, Zhouqin Jiang, Rushi Lan, Wenming Huang, and Xiaonan Luo. "Image captioning using DenseNet network and adaptive attention," Signal Processing: Image Communication, 2020, 85: 115836.

[96] Xiao, Fen, Xue Gong, Yiming Zhang, Yanqing Shen, Jun Li, and Xieping Gao. "DAA: Dual LSTMs with adaptive attention for image captioning," Neurocomputing, 2019, 364: 322-329.

[97] Huang, Lun, Wenmin Wang, Yaxian Xia, and Jie Chen. "Adaptively aligned image captioning via adaptive attention time," Advances in Neural Information Processing Systems, 2019, 32: 8942-8 951.

[98] Ziwei Tang, Yaohua Yi, and Hao Sheng. "Attention-Guided Image Captioning through Word Information," Sensors, 2021, 21(23): 7982.

[99] Murad Popattia, Muhammad Rafi, Rizwan Qureshi, and Shah Nawaz. "Guiding Attention using Partial-Order Relationships for Image Captioning," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4671-4680.

[100] Jia, Xu, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. "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.

[101] Yao, Ting, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. "Boosting image captioning with attributes," In Proceedings of the IEEE international conference on computer vision, 2017, pp. 4894-4902.

[102] Jiang, Wenhao, Lin Ma, Xinpeng Chen, Hanwang Zhang, and Wei Liu. "Learning to guide decoding for image captioning," In Thirty-second AAAI conference on artificial intelligence, 2018.

[103] Sow, Daouda, Zengchang Qin, Mouhamed Niasse, and Tao Wan. "A sequential guiding network with attention for image captioning," In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 3802-3806.

[104] Thang Luong, Hieu Pham, and Christopher D. Manning. "Effective approaches to attention-based neural machine translation," in EMNLP, 2015.

[105] Mun, Jonghwan, Minsu Cho, and Bohyung Han. "Text-guided attention model for image captioning," In Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1).

[106] Zhu, Zhihao, Zhan Xue, and Zejian Yuan. "Topic-guided attention for image captioning," In 2018 25th IEEE International Conference on Image Processing (ICIP), 2018, pp. 2615-2619.

[107] Hodosh, Micah, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics," Journal of Artificial Intelligence Research, 2013, 47: 853-899.

[108] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zh. BLEU: "A method for automatic evaluation of machine translation," In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2002, pp. 311–318.

[109] Chin-Yew Lin. "Rouge: A package for automatic evaluation of summaries," In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 2004, vol. 8.

[110] Satanjeev Banerjee and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," In Proceedings of the ACLWorkshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, 29: 65–72.

[111] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4566–4575.

[112] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. "Spice: Semantic propositional image caption evaluation," In European Conference on Computer Vision. Springer, 2016, pp. 382–398.

[113] Guo, Longteng, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. "Normalized and geometry-aware self-attention network for image captioning," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10327-10336.

[114] Pan, Yingwei, Ting Yao, Yehao Li, and Tao Mei. "X-linear attention networks for image captioning," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10971-10980.

[115] Wei, Yiwei, Chunlei Wu, Guohe Li, and Haitao Shi. "Sequential Transformer via an Outside-In Attention for image captioning," Engineering Applications of Artificial Intelligence 108 (2022): 104574.

[116] Wang, Chi, Yulin Shen, and Luping Ji. "Geometry Attention Transformer with position-aware LSTMs for image captioning," Expert Systems with Applications 201 (2022): 117174.

[117] Wang, Changzhi, and Xiaodong Gu. "Image captioning with adaptive incremental global context attention," Applied Intelligence 52, no. 6 (2022): 6575-6597.