Image Captioning through Cognitive IOT and Machine-Learning Approaches

Tarun Jaiswal[a], Manju Pandey[b], and Priyanka Tripathi[c]
a,b,c
Department of Computer Applications, National Institute of Technology (NIT),
Raipur, India

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: Text generation is the technique of producing a textual explanation of an image. Both natural-language processing (NLP) and cognitive science (CS) techniques are used to create the captions of images. Subtitles and descriptions are both texts shown on a video that delivers extra or interpretive details for observers who are deaf and hard of hearing or require extra clues than just the sound. Many times, the shown text comprises an interpretation or conversion of the pronounced language in the video. Other usages of text-generation have been originated based on the requirements of various spectators. For instance, the caption for the deaf and people who have severe hearing impairments contain an explanation of other acoustic details that audiences with hearing problems might miss, such as the description of the music, proof that the narrator is now offscreen, etc. Caption-generation is a very inspiring Cognitive Internet of things (CIoT) and artificial intelligence (AI) task where the description of text must be originated from a given image. It requires the approach of computer vision to recognize image details or content and a language model from the NLP region to translate the image interpretation into the words in the right order. In this survey paper, we describe the comprehensive overview of prevalent deep-learning based text generation methods. Besides this, we describe the various datasets and the famous evaluation metrics used in the deep-learning based automatic text generation.

Keywords: Image captioning, deep learning, NLP, Cognitive IoT, Machine Learning, Attention Mechanism, Metrics.

1. Introduction

The word "Caption" is primarily used in the North-America. Caption generally refers to text descriptions that are in the analogous language as spoken in the video. And when we talk about translated videos, they are called subtitles, and they are used worldwide. Captions appear on the computer screen, TV, smart devices, and movie screen, and they describe the text version of the audio or video. This captioning technique is very beneficial for people who cannot hear or who have difficulty in hearing so that they can also enjoy watching the movie. For people experiencing severe deafness who are also not deaf, subtitles can even make the spoken sentences relatively easy to understand since the hearing, analogous to vision, is affected by desires, that is once we have thought what others speak, dialogues are perfectly obvious. Captions give valuable details such as who is talking or what is the background in the scene (see Fig.1), which is very crucial for understanding a new event/scenario. Captions are generated through the transcript of the program. The captioner divides the conversation into captions and ensures that the words come along in harmony with the audio. Computer programs convert the captioning-detail and integrate it with the sound and video to generate novel electronic or digital documents. The captions must, preferably, display at the lower part of the screen. Image captioning is a complex problem in which artificial intelligence plays an important role. Through artificial intelligence, we can get the computer to do all the functions that we think. A costly and time-consuming method for machine learning (ML) and natural language (NL) scientists is annotating and marking machine learning data sets. Although, a modern deep learning (DL) technique is often used to translate, find, and restore picture & video captions, making the system created captions more accurate and reliable.

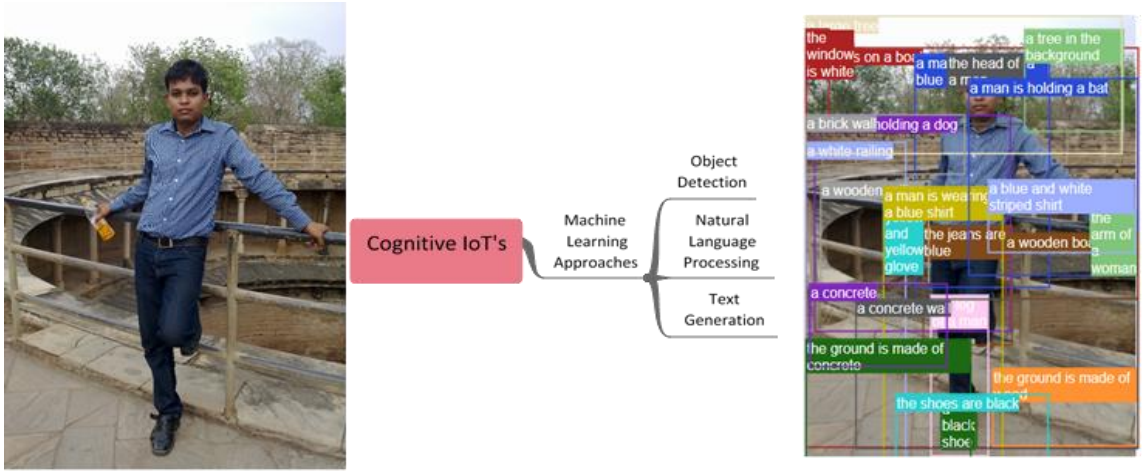

Fig. 1. Text Generation through Cognitive IoT's and Machine-Learning Approaches.

Currently, image captioning is divided into 03 categories-

Open captioning is usually shown explicitly in the video frame, and this can not be disabled. At first, open captioning was the only option available for television viewers; and only for a small number of programs.

Closed captioning allows audiences to switch captions open or closed. The audiences of television had to compete requirement:-Some required the subtitles to access the information, and others needed them to understand the scene.

Real-time captions are generated as activity occurs. A Captioner, i.e., frequently trained as a court-reporter or stenographer, employs a stenotype m/c with a phonetic keyboard and unique S/W. A machine immediately converts the phonetic symbols into english-subtitles. The small interruption is based on the captioner's requirement to listen and code the word and computer processing time. The immediate captioning approach is used for the applications which do not have a script; real-time activities, together with congressional proceedings; live webinars; news-services; and nonbroadcast seminars, for example, global gatherings of technical societies. Though several real-time text generation is more than 98% correct, the viewer can see only rare mistakes. The captioner can misunderstand a phrase, listen to an unknown word, or have an S/W dictionary problem.

Despite the recent promising development in neural-image-captioning, automated image captioning seems complicated.

In image captioning, an image is provided to an algorithm and tasked with creating a sensible caption. It is a challenging task for several reasons, not the least because it involves a notion of saliency or relevance. Recent deep learning approaches mostly include some "attention" mechanism (sometimes even more than one) to focus on relevant image features.

Visual attention mechanisms are an essential part of the advanced computer vision system. In nearly all areas, it is also an integral component of most advanced achievements in almost all areas, such as detecting objects, image captioning, and many more. Many traditional visual attention systems employ image captioning & VQA from the top-down perspective, a task-oriented technique that assigns caption based on the selectively determined weight of image characteristics. While the bottom-up technique is constructed on the visual feedforward attention system that first defines the targeted image area and allots the features vector to those areas.

Here we describe the basis upon which existing image captions are constructed. We have obtained remarkable progress in various critical problematic areas such as object/image classification, speech analysis, etc. However, image captioning has not yet achieved the milestone that other genres have achieved. But with the advent of DL in object recognition/detection, the identification of fine-grained features has led to an incredible improvement in image captioning precision. Previous approaches include shifting of most crucial explanation from the indexed database by identifying the semantically related images via image retrieval.

Existing image captioning techniques may be classified into 03 groups-

The 1st set of techniques:- Recognize objects & attributes and after that compiles the image definition from the phrases that include those objects.

The 2nd set of techniques:- Embed the images and compatible captions in the equivalent vector space. For a particular inquiry, captions closest to the image in the embedding-space are obtained, and those captions are updated to produce the latest caption for the provided image. However, these techniques do not produce a new description of a provided test image as the descriptions from the almost identical images are used for the caption.

The 3rd set of techniques:- Explicitly create the order of the phrases related to the image conditioned on the image and earlier produced words. So, they are able to generate new arrangements of words that might never have occurred in the training data. It becomes the standard procedure in most advanced algorithms for image captioning and is defined in a very efficient manner in section 2. Let's deep dive:

There are 03 critical issues while filling the semantic difference among the visual scenes & language in order to generate different, innovative, and human-like captions.

The main obstacle derives from the accumulated nature of the NL & visual-scenes. While the training dataset comprises co-incidents of particular objects in their perspective, a captioning system must be generalized by composing-objects in other circumstances.

The conventional captioning approach has some drawbacks, such as it suffers from understandability & genuineness since it produces the caption in serial order, i.e. Subsequently created word relies on the former word and the image characteristic. This can often contribute to syntactically consistent, but semantically inappropriate language arrangements, also an absence of variety in the created captions. The compositionality problem is addressed with a context-aware captioning model that allows the captioner to define phrases that are dependent on the part of the visual scenes.

The 2nd problem is the discrimination in the dataset that impacts modern captioning schemes. The trained models overfit the typical objects that co-occur in a common environment (e.g. curtain and windows), contributing to a dilemma in which certain structures fail to generalize to scenarios in unknown environments where the same objects exist (e.g., curtain and playground). Although lowering the bias of the dataset is a complex and complicated work.

The 3rd problem is in the assessment of the quality of produced captions. Automated metrics, however somewhat useful, is still inadequate because they do not consider the image into consideration. In several instances, their grading stays insufficient and confusing on certain occasions — particularly when scoring dissimilar and vivid captions.

Development in automated captioning of pictures and interpretation of scenes allows more accurate machine vision systems for personal trainers for individuals with visual impairments and better their everyday lives. The semantic gap in language bridging & vision points needs to incorporate the cognitive approach for better scene understanding.

2. Cognitive Internet-Of-Thing (CIOT)

CIoT is a new extended functionality of an IoT architecture that works via practical cognitive AI with computer models. CIoT learns, can give the sense, make decisions, and deliver to humans very smartly. The ability of CIoT is very similar to the cognitive ability of a human being. To take full advantage of the IoT, we must add the CIoT as an essential ingredient. The objective of the CIoT is to remove the boundary between the human being and the physical worlds that are sharing the object and information with us. The objective of the CIoT is to achieve performance enhancement and more intelligent or smart IoT with the help of supportive instruments and cognitive computing [1]. Some of the features of the CIoT are as - learning ability can be increased by giving CIoT a short time, or in other words, it is capable of self-learning via minimum training. CIoT is adaptive in nature. CIoT dynamically tackles the real data as a process and analysis. CIoT is constructed to be collaborative with peoples, m/c, and other mechanical devices. CIoT is particularly significant for its capability to cooperate with persons in a fully human way [1].

3. Image Based Visual Captioning Approaches

The complete classification is described in Fig.2 for DL based image captioning approaches.

3.1 Standard Image Caption Methods

The authors described the novel methods to pre-learn great sensitive filters via convolutional layers known as FRAME and abbreviated as Filters, Random field, & Maximum Entropy [2]. The learning algorithm produces intense images, and via the CNN component, describes the learned model. The proposed CNN-FRAME generative model activates CNN nodes by learning from small No.'s of training samples in a generative custom.

3.2 Deep Learning

The authors surveyed an image captioning technique via deep learning [3]. In this comprehensive survey, the authors discussed image captioning classification, framework, and its advantage & disadvantages. This reviewed paper also concentrated on numerous datasets, evaluation metrics with their demonstrated outcomes.

In this paper, the authors introduced an innovative automates algorithm for image captioning to produce the textual performance of lifelogging camera collection-based capture images [4]. The authors developed and discovered new methods based on DL to produce image streams via temporal consistency-constraints to build summaries. The experiment outcomes exhibit that the suggested method outperformed state-of-the art captioning methods on several measurable metrics.

Fig. 2. Image Captioning Approaches

The 02-challenging task -hashtag prediction and post generation (Instagram dataset) handle via a new image captioning method known as Context Sequence Memory Network (CSMN) [5]. CSMN substantial benefits in modelling the information from both the older and future. CSMN can produce multiple types of contextual information, capture long-term information without failure, and better context understanding via CNN memory structure. The CSMN uses Amazon Mechanical Turk to illustrate the efficiency of the method, consistently outperforming all baseline methods.

In an image captioning model, predicting the next-word based on a prior word and concealed status is difficult. This problem can be overcome by providing a novel CNN language model. CNN language model fit for statistical language modelling scheme[6]. Considerable studies are carried out on 02 datasets: Flickr30K and MS COCO, and the CNN language model demonstrate clear enhancements compared to other advanced models. CNN language model achieves superior performance than the vanilla recurrent neural network-based language models.

Semantic Deep Learning

The suggested technique is able to generate the image and video-based text-descriptions; hence the method is named image parsing to text description (I2T) [7]. The I2T follows 03 major phases- first, via image parsing engine, images and videos are divided into visual forms with natural language construction. Second, the image-parsing outcomes are converted into the semantic description in the form of Web-ontology-language(OWL). Third, the text-generation engines provide facilities to convert encoded transmitted data into a displayable form for query-able text study. The I2T method's objective is to construct and-or-graph (AoG) for visual-knowledge demonstration via the semiautomatic method. The maritime and urban-scene video monitoring system is the best example of an automatic I2T system.

The authors described the novel captioning model based on images & sentences. In this method, generated words follow the previous one and are aligned via experience [8]. The authors also proposed scene-specific context modelling that is able to retrieve higher-order semantic information. The authors designed a benchmark with available outcomes on several popular datasets. The authors concentrate on region-based attention or scene-appropriate contexts improvement systems. The experimental result shows that via mixing 02 modelling techniques, significant performance enhancement can be achieved.

Convolutional-neural-networks (CNNs) & Recurrent-neural networks (RNNs) are combined and utilized for the Vision-to-Language (V2L) problems. The authors proposed a new CNN-RNN method that combines high-rank semantic ideas and also improves the approaches performance through the visual-question-answering (VQA)[9]. The authors observed both speed and accuracy enhancements over the baseline method.

The novel method, SPICE: Semantic-propositional-image- caption-evaluation (SPICE) is based on the hypothesis that semantic-propositional subject plays a significant role in human caption assessment [10]. The authors demonstrated that the SPICE model has the ability to define new ideas by empirically estimating its performance on BLEU, METEOR, ROUGE-L, and CIDEr and shows qualitative results. SPICE achieves 0.88 system-level correlation on MSCOCO, 0.43 for CIDEr and 0.53 for METEOR. SPICE is able to answer text-generation queries.

For describing huge no. of object categories, the novel object Captioner (NOC) was proposed by the authors [11]. NOC object Captioner is based on the deep visual semantic captioning scheme. There are many benefits associated with the NOC, such as semantic embedding and generalization. The NOC can extract and generate captions. Experimental validation shows that the NOC outperformed the other advanced methods.

The integration of SPICE & CIDEr is known as SPIDEr [12]. SPIDEr is optimization via a policy gradient (PG). In the SPIDEr framework, the SPICE score confirms that the captions are semantical, whereas the CIDEr score ensures that the captions are syntactically fluent. The policy gradient (PG) approach is applied for enhancements via Monte Carlo rollouts. The authors tested the proposed method on COCO metrics and achieved better performance than the other method.

The authors suggested a novel approach for detailed wording and matching in image captioning, known as Conditional Generative Adversarial Networks (CGAN) [13]. CGAN can learn to produce conditioned descriptors and an evaluator to assess these conditions. CGAN generates an image descriptor that is more semantic, natural, and diverse. Policy gradient is used for solving the nontrivial problems from reinforcement learning. Policy gradient can handle early feedback. The CGAN model outperforms the most advanced method on the benchmark dataset MS COCO & Flickr30k.

Supervised/Unsupervised/Reinforcement Deep Learning

Most advanced techniques of image captioning need supervised training data that contains caption with paired image data. Usually, these approaches are unable to use unsupervised information such as textual data without accompanying photos, which is a much more abundant commodity. The proposed novel way uses textual data by artificially filling the lost values. The evaluation of this learning method on a newly designed model detects the visual concept present in the image and feeds them to a reviewer-decoder framework with an attention-mechanism [14]. Different from the earlier techniques that encode the visual concept by utilizing the word embeddings, the proposed techniques use the regional image structures that acquire more inherent details. The key advantage of this architecture is that it incorporates important thought vectors that collect prominent image features and then employ a soft, attentive decoder to decode the thought vectors and produce image captions. The evaluation of the suggested model on both the dataset (MS COCO & Flickr30K) shows that when this model is integrated with

semi-supervised-learning approaches, it greatly increases the performance and helps the model produce a perfect caption.

An unsupervised approach is used for image captioning training. And for this purpose, it needed a sentence dataset and visual concept detector [15]. The sentence dataset demonstrates the image captioning framework and produces the possible sentences for a further process, whereas the visual concept detector monitors the framework to identify the visual ideas in an image. Generated captions are semantical with the image and projected via common latent space. Detailed research and analytical studies on a large-scale image explanation dataset of 02-million natural sentences demonstrate the suggested model's superiority in contrast to the most advanced techniques.

Call self-critical-sequence-training (SCST) [16] approach worked on image optimization-based captioning systems using reinforcement learning. The proposed method is a form of a famous REINFORCE algorithm. By using SCST, the reward signal is assessed and avoiding normalization assessment. SCST optimized the CIDEr metric and enhanced the result from 104.9 to 114.7.

LSTM Deep Learning

The authors described the modification of the long-short-term-memory (LSTM) framework known as guiding Long Short-Term Memory (gLSTM) [17]. sIn gLSTM semantic detail is obtained from the image, which is more strongly coupled to the image content. The extensive study demonstrates that the proposed approach achieves a better outcome than the previous methods.

Deep convolutional-neural-network (CNN) and other 02 distinct LSTM net are combined via a novel model known as end-to-end trainable deep bi-directional LSTM for image-captioning [18]. This arrangement can learn via past and upcoming context data at high-level semantic space. The authors introduced 02 novel deep bi-directional variant models for learning hierarchical visual-language embeddings and multicrop, multiscale & vertical mirrors to avoid overfitting in training deep-models. The proposed method is applied on standard benchmarks: Flickr8K, Flickr30K, and MSCOCO. Detailed assessments across several models and databases show that the bidirectional LSTM models on caption generation and other numerous tasks such as detection, attention, etc., perform much better than the other prevalent most advanced models.

There is 02 generalization standard for short structured representations- first, standard generalization to novel image with the same scenes and second, generalization to novel groupings of known objects [19]. The 02 generalization standard is applied to the MSCOCO dataset. The extensive experiments show that the generalization standard better than the LSTM.

Visual encoder & language decoder comprehensibly collaborate in a recurrent custom via a new image-based captioning framework called Recurrent Image Captioner (RIC) [20]. Through CNN-based visual encoder to allow the spatially invariant transformation of visual-signals.In the attention filter phase situated among the encoder and decoder. To preprocess for producing great textual representations used bidirectional LSTM. Comprehensive studies on the most prevalent dataset like Flickr8k, Flickr30k & MS COCO exhibit the suggested method's greatness compared to the other most advanced techniques.

Together, CNN and RNN make a new network for image-based captioning, known as Long Short-Term Memory with Copying Mechanism (LSTM-C) [21]. The LSTM-C smartly joints the word by word sentences. LSTM-C is also capable of producing via RNN decoder with a copying tool used to put proper places of output sentence. Extensive experiments and analyses on the MSCOCO and ImageNet datasets show that LSTM architecture works superior to the other methods across various assessment metrics.

CNNs and RNNs are combined thru Long short-term memory with attributes (LSTM-A). To incorporat inter-attribute correlations into Multiple-instance-learning (MIL) [22]. LSTM-A builds a framework via a fuzzy relationship between the image representation and attributes. The effectiveness of the LSTM-A is validated on MS COCO.

Attention-based Neural Encoder-Decoder Frameworks

Given an image I with its matching description y which is characterized as a sequence of words {y_1,y_2,….,y_(t-1) }, using an RNN as the decoder, the neural-based encoder-decoder framework enhances the log-likelihood of the RNN joint probability of each time-step t which is achieved by utilizing the chain-rule:

$$p(y|I) = \prod_{l=1}^{T} p(y_t|y_{1:t-1},\ I, \theta) \qquad (1)$$

Together with the outstanding efficiency of LSTMs and their good ability to catch dependencies in the long term, the RNN decoder may be presented as an LSTM, where the concealed state at each time-step t is described as:

$$h_t = LSTM(x_t, h_{t-1}, m_{t-1}) \qquad (2)$$

Where x_t represent the input to the LSTM at timestep t,h_(t-1) is the earlier unseen state, and m_(t-1) is the prior memory state at timestep t-1.

Along with the recommendation of attention structure, in sequence modelling systems, the context vector plays a key role [23] and considerably enhances efficiency. The context-vector is sometimes considered as a focus

component, which governs the network when the forecast is produced. In image captioning systems, the context vector offers visual information like where to search in the picture at each stage to forecast a word. Instead of depending strictly on a single secret state. The decoder would join to specific areas in the image throughout the caption-generation practice through the context vector, which is calculated as a weighted-sum of each pixel in the spatial image acquired by the encoder CNN.

Dense Captioning Deep Learning

The authors suggested a dense caption scheme to localize and designate salient regions in NL [24]. The authors also suggested a fully convolutional localization network (FCLN) framework that combines both localization and description task. FCLN works based on the concept of end-to-end optimization, well-organized forward pass, processing of an image with one round, and it does not require extra proposal regions. FCLN is combined with convolution network, dense localization layer, and RNN model that can produce the label-sequences. The suggested scheme is applied to a very famous dataset, i.e. visual genome that contains 94000 images and 4100000 regions grounded captions. The experiment's outcomes validate that the suggested scheme significantly outperformed the current most advanced methods in both generation and retrieval-settings.

The dense visual annotations are connected via huge overlapping target regions for finding the exact localization. The proposed method works on 02 important problems of dense captioning [25]. By the usage of joint inference and context fusion, authors avoid the problem of dense captioning. The usefulness of the suggested method is verified on Visual Genome. The detailed result shows that the suggested technique is competitive with advanced methods with a relative gain of 73%.

Dense captioning has some limitations, i.e. not able to generate a coherent story for an image. Authors overcome these limitations by decomposing both images and sections into their basic portions, detecting the semantic region, and presenting a hierarchical method that controls the compositional structure of images and language [26]. Extensive experiment outcomes show that the suggested technique is competitive with the most advanced approaches and presented how region-level information can be efficiently transferred to paragraph-captioning.

Stylized Deep Learning

MemCap is able to generate an image caption in a linguistic manner and explicitly encodes the information [27]. MemCap learns to memorize elements during training. MemCap is memory segments that contain a group of embedding vectors for encoding style-related phrases in the training corpus. For obtaining the style-related phrases, the authors described the sentence decomposing algorithm. In the sentence, the decomposing algorithm contains 02 parts: style-related part and content-related part. Extensive experiments on SentiCap and FlickrStyle10K dataset demonstrate that the suggested MemCap model significantly outperforms the compared methods for image captioning.

Compositional Architecture Deep Learning

For large-scale visual learning, the authors proposed a new convolution-based framework. The proposed framework end-to-end trainable and check the outcome rate on standard recognition tasks, description and retrieval glitches, and video description challenges. In this proposed method, recurrent convolution frameworks are "doubly deep" accumulated in spatial&temporal "layers". The proposed framework has beneficial when target notions are difficult and/or training data are inadequate. The experimental outcome shows that the proposed model has numerous advantages over the prevalent method as defined separately or optimized [28].

The authors solved the generative descriptive problem based on a no-paired image sentence benchmark via Deep-compositional-captioner (DCC) [29]. DCC framework is capable of combining sentences that define novel objects and their contacts with other objects. The authors demonstrated that the DCC framework is assessed on the MSCOCO dataset to obtain greater competitive success as compared to the state-of-the art outcomes and qualitative outcomes.

MultiModal Space Based Deep Learning

The authors proposed a framework that produces NL descriptions. The proposed method is beneficial for inter-modal correspondences among language and visualdata. The proposed framework represents a CNN group over image-region, bidirectional RNN over the sentence, and multimodal embedding[30]. The authors discussed the multimodal recurrent neural network scheme for producing image regions via learning. The proposed method applied on standard datasets Flickr8K, Flickr30K, and MSCOCO and produces descriptions that expressively outperform retrieval-baselines on together full images and a novel dataset region-level annotations.

The authors proposed an innovative image captioning architecture based on a multimodal recurrent neural network (m-RNN) [31]. m-RNN contains 02 sub-net: First, a deep-RNN for sentences and a deep convolutional network for image generation. Image captions produced via sampling and word produces via directly models probability distribution. The m-RNN is applied on 04 standard datasets, namely IAPR TC-12, Flickr-8K, Flickr-30K, and MSCOCO. The experimental outcome validates the effectiveness of the suggested method.

The authors described the method for producing the natural sentences in sequences for a series of images. The authors construct a multimodal scheme, i.e. coherence recurrent convolutional network (CRCN) [32]. CRCN comprehends convolutional neural networks, bi-directional recurrent neural networks, and an entity-based local coherence model. The proposed method is able to learn the blog in the form of text image. The suggested approach

is experimentally described in numerous datasets and shows that the suggested technique outperformed the other prevalent methods.

The authors suggested a recurrent highway network with language CNN for producing image caption [33]. The suggested scheme contains 03 sub-net. First, for image representation, deep convolutional neural network is used. Second, the convolutional neural network is used in language modelling. Third, for sequence prediction Multimodal Recurrent Highway Network is used. The proposed scheme is able to use the hierarchical & temporal construction of past words. The 02 dataset MSCOCO and Flickr30K are used for validation. Extensive experiment outcomes on 02 datasets MSCOCO and Flickr30K demonstrate the suggested scheme's advantage compared to the most advanced methods.

Attention Mechanism

The authors proposed the methods that identify an image's info via automatic learning, which is known as the attention-based model [34]. The proposed model is trained via deterministic mode. The proposed method experimentally authenticates the practice of attention thru state-of-the-art performance on 03 standard datasets, namely Flickr8k, Flickr30k & MS COCO.

The attention framework has been designed to reproduce simple human behaviour. Prior to image summarization, people prefer to pay attention to the particular areas of the picture and then describe the interaction between objects in those areas. A similar method is adopted in the attention framework. There are many methods that the researcher has used to repeat it that are generally recognized as hard or soft attention system [35]. Other researchers have explored top-down and bottom-up attention framework. He S et al. [36] at the latest announced that the improved solution is still a top-down attention scheme as the outcomes from the studies with humans and m/c exhibited identical outcomes. In the top system, the method begins as input from a given image and then transforms it into words. In addition, the latest multimodel dataset is generated from individual fixations and scene details for the largest number of new cases.

X-linear Attention Networks (X-LAN)

A unique unified design for the attention module, known as X—linear-attention block, is completely capitalized on the bi-linear pooling to capture the 2nd order characteristic interface along with spatial and streamwise bilinear attention. In addition, particular incorporation of the X-Linear attention block into the image encoder and sentence decoder to obtain increased intra & inter model correlation boosts visual knowledge and carries out intricate multi-modal analysis for captioning images. Fig.3 shows how to combine such blocks with the encoder & decoder framework through (X-LAN) X-Linear-attention networks.

X- Linear-Attention-Block

Although the traditional focus module activates the relationships among various modalities perfectly, only the 1st order interaction function is used, reflecting the restricted potential of dynamic multi-modal logic in image captioning. Influenced by bilinear pooling's new achievements in fine-grained visual identification [37, 38] or visual query response [39, 40], The entire capitalization on bi-linear pooling methods for creating a unified-attention module (X-linear-attention block) for image captioning, as described in Fig. 4. The X-linear-attention block framework enhanced the intake capability of the output attended characteristic by using higher-order relation among the input single-modal or multi-modal.

In general, assume the query Q∈R^(D_q), a set of keys K={k_i }_(i=1)^N, and a set of values V={v_i }_(i=1)^N, where k_i∈R^(D_k) and V_i∈R^(D_v) notify the i-th value/ key group. First, the X-linear-attention block conducts low-rank bilinear pooling to obtain a bilinear combined query-key representation B_i^k∈R^(D_B) among the query Q and each key k_i：

$$B_i^k = \sigma\left(W_k k_i\right) \odot \sigma\left(W_q^k Q\right), \quad (2)$$

Where W_k∈R^(D_B×D_k),and W_q^k∈R^(D_B×D_q) are embedding matrices, σ represents ReLu unit, and ⊙ denotes element-wise multiplication. As such, the learned bi-linear query-key representation B_i^k coveys the second-order characterization connections among query and key.

Next, Centered on all bi-linear query-key representations {B_i^k }_(i=1)^N, 02 types of bilinear attention deliveries are achieved to combined both spatial & channel-wise information inside all attributes. Particularly, the spatial bilinear attention distribution is presented by displaying each bilinear query-key interpretation using 02 embedding layers into the associated weights of attention, accompanied by a softmax layer, for normalization.

$$B'^k_i = \sigma\left(W_B^k B_i^k\right), b_i^s = W_b B'^k_i, \beta^s = softmax(b^s), \quad (3)$$

Where W_B^k∈R^(D_c×D_B) and W_b∈R^(1×D_c) describe the embedding matrices, ⟦B^'⟧_i^k represent the transformed bi-linear query-key representation, while b_i^s is the i-th element in b^s. Here every element β_i^s in β^sindicates the normalized spatial attention weight for each key/value pair. In the meantime, the executed squeeze-excitation operation [41] overall altered bi-linear query-key representations {⟦B^'⟧_i^k }_(i=1)^N or channelwise attention measurement.

Specifically, all modified bilinear query-key interpretations through average pooling are combined by the squeeze activity, heading to a global channel descriptor$\bar{B}$:

$$\bar{B} = \frac{1}{N}\sum_{i=1}^{N} B'^k_i \quad (4)$$

Thereafter the excitation action generates channel-wise interest distribution β^c through exploiting the self-gating process over the global channel descriptor ($\bar{B}$) with a sigmoid.

$$b^c = W_e\bar{B}, \beta^c = sigmoid\ (b^c), \quad (5)$$

Where W_e∈R^(D_B×D_c), represents the embedding matrix.

Lastly, the X-linear-attention block produces the attended value characteristic V^ by collecting the improved bilinear values with spatial & channel-wise bilinear attention:

$$\hat{V} = F_{X-Linear}(K, V, Q) = \beta^c \odot \sum_{i=1}^{N} \beta_i^s\, B_i^v, \qquad (6)$$

$$B_i^v = \sigma(W_v v_i) \odot \left(W_q^v Q\right),$$

Where B_i^v represents the improved value of bi-linear pooling on query Q, and each value V_i, W_v∈R^(D_B×D_q) denotes the embedding matrices. Therefore, relative to traditional attention frameworks that merely investigate the relationship between question and key in the 1st order, the X-linear-attention block generates the more representative attended characteristic. However, relationships with higher-order features are utilized through the bilinear-pooling.

Extension with higher order interactions.

For utilization of the higher-order characteristic interactions, the further iteration of the above procedure of b-linear attention measurement and characteristic combination using a stack of our X-linear-attention blocks. Formally, for the m-th X-linear-attention block, first of all, the consideration is given to the prior output attended characteristic V^((m-1)) as input query, combined with recent input Keys

$K^{(m-1)} = \left\{k_i^{(m-1)}\right\}_{i=1}^{N}, and\ values\ V^{(m-1)} = \left\{v_i^{(m-1)}\right\}_{i=1}^{N}$:

$$\hat{V}^{(m)} = F_{X-Linear}\left(K^{(m-1)}, V^{(m-1)}, \hat{V}^{(m-1)}\right), \qquad (7)$$

Where $\hat{V}^{(m)}$ represents the output of new attended characteristics. $\hat{V}^{(0)}, K^{(0)}, and\ V^{(0)}$ represents $Q, K\ and\ V, respectively$. Afterwards, all keys/values are subsequently changed based on the performance of the new attended characteristic $\hat{V}^{(m)}$:

$$k_i^{(m)} = LayerNorm\left(\sigma\left(W_m^k\left[\hat{V}^{(m)}, k_i^{(m-1)}\right]\right) + k_i^{(m-1)}\right),$$

$$v_i^{(m)} = LayerNorm\left(\sigma\left(W_m^v\left[\hat{V}^{(m)}, V_i^{(m-1)}\right]\right) + V_i^{(m-1)}\right), \qquad (8)$$

Where W_m^k and W_m^v represents compositional matrices. Observe that each key/value is combined with the new attended characteristic, followed by a remaining association & layer normalization as in the [42]. The repetition of the procedure (Eq. (7) and Eq.(8)) M times through stacking M in X-linear-attention blocks, that reflect higher (2Mth) order characteristic interactions.

First of all, Faster R-CNN is utilized to recognize the range of image sections. After, a stack of X-linear-attention blocks is supplied in an image encoder to encode the region-level characteristics with the higherorder intra-modal interaction in among, leading to a set of improved region-level and image-level features. Relies on the improved visual characteristics, the X-linear-attention block is even more adopted in sentence decoder to execute multi-modal reasoning. This promotes the investigations of highorder inter-modal interactions among visual content & natural-sentence to increase sentence generation.

3.2.9.2 Local and Global Attention Model

Policy net and value net both are jointly produce decision-making architecture for image-based captioning. For predicting the next word (local guidance), policy net is used, whereas evaluating all possible additions of the recent state (global-guidance), value net is used. Both the net are trained via the actor-critic reinforcement-learning framework [43]. Experimentally, the authors demonstrated the advantages of the proposed approach over conventional image captioning methods and shown that the suggested framework outperforms most advanced methods across diverse evaluation metrics by using the Microsoft COCO dataset.

3.2.9.3 Adaptive Attention Model

The authors extract significant sequential word creation details via a novel adaptive-attention framework with visual-sentinel for image-based captioning [44]. The proposed frameworks decided whether to attend the image and where automatically. Broad studies are carried out on both the COCO image captioning 2015 challenge dataset and Flickr30K. The proposed frameworks and their efficiency enhancement for customized captioning of images over advanced captioning models.

3.2.9.4 Semantic-Attention Model

The top-up and bottom-up techniques joined via semantic-attention algorithm for image captioning with tight

coupling of RNN [45]. The suggested approach for image captioning obtains advanced performance through prevalent standard benchmarks (Microsoft COCO and Flickr30K) across numerous evaluation metrics.

Caption generators emphasize the feature of an image that is generated previously via text-conditional attention model [46]. To acquire text-based images for the text-conditional attention model, the authors hire a gLSTM with CNN fine-tuning framework. The proposed models can learn jointly, i.e. embedding text/image, text-conditional attention language framework under the one umbrella. For this purpose, the authors used MSCOCO dataset. The extensive experiment outcome shows that the proposed techniques outperform than the prevalent advanced captioning techniques.

3.2.9.5 Spatial and Channel wise Attention Model

Gan et al. [47] suggested a new image captioning system known as SCA-CNN effectively employed on structural prediction tasks (Image-based VQA). This proposed system can generate a visual attention model spatially. SCA-CNN integrates spatial and channel-wise attention, and it is used for sentence generation in a multilayer features map dynamically. The SCA-CNN attention mechanism and spatial transformer attention regions produce advanced outcomes on the Flickr8K, Flickr30K, and MSCOCO datasets. Extensive experiments and analysis demonstrate that the suggested technique works better than the other advanced methods across different evaluation metrics.

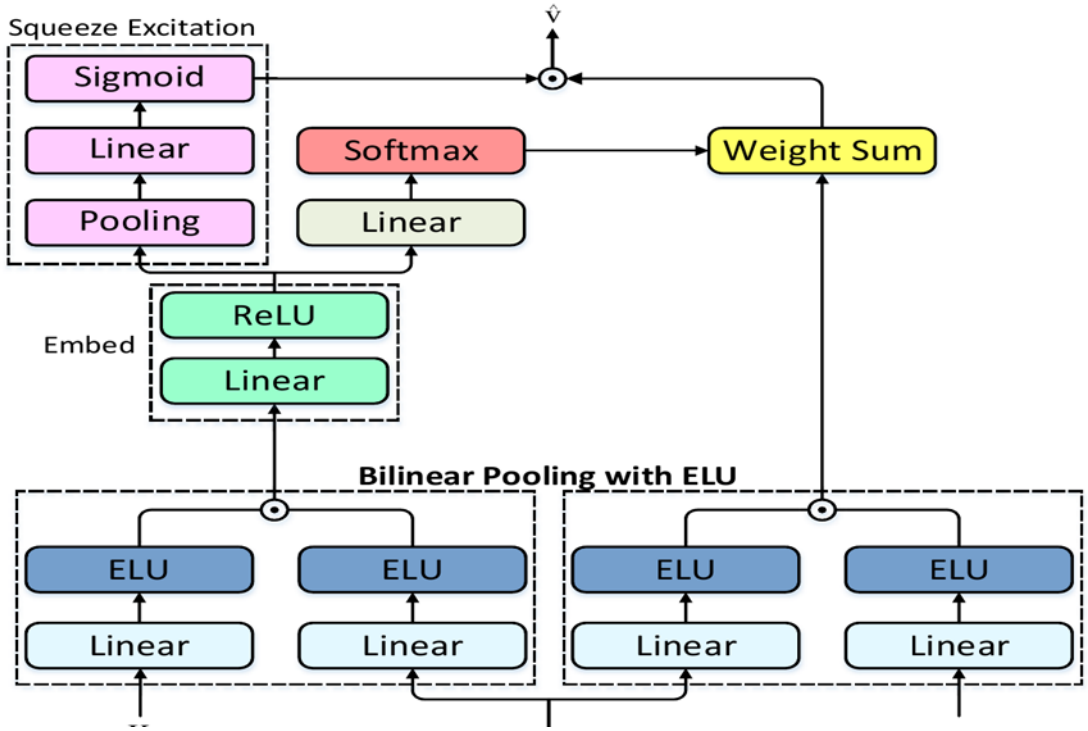

Fig. 3. A simplified illustration of X-linear-attention block & ELU to catch infinity-order characteristic connections.

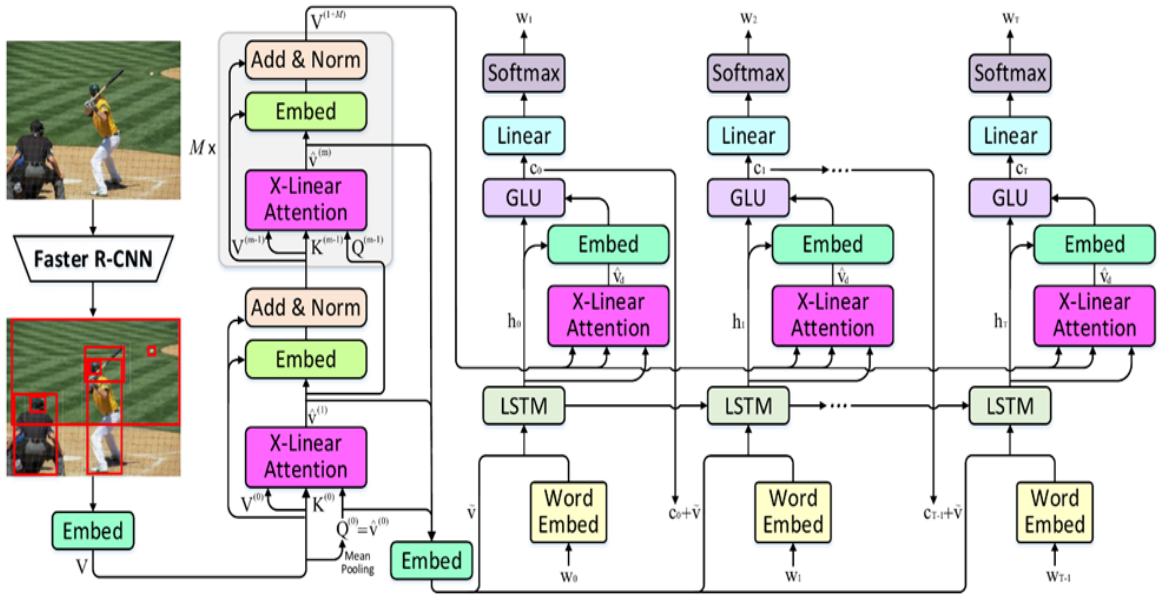

Fig. 4. Overview of X-linear-attention networks (X-LAN) for image-captioning.

3.2.9.6 Areas of Attention Model

Integrate the region of image, words caption, and showing the RNN language model via Areas of Attention framework [48]. Areas of attention enhanced image captioning thru confining the related areas during the experiment. Areas of Attention works on CNN activation grids, object-proposals, and spatial transformers nets implemented in a convolutional technique. Spatial transformers networks are tight together with Areas of Attention for great outcomes. Areas of Attention and spatial transformer attention combinable validate on MSCOCO Captioning benchmark and obtain the advanced performance in the standard metrics.

Top-Down attention Model

Caption guided visual-saliency [49] presents region to word mapping for the encoder-decoder net. During caption training, the presented model can learn implicitly. For predicated captioning and arbitrary query sentences, the spatiotemporal heatmap is used. The presented model improves saliency and its design. Evaluation & person decisions show that the presented model considerably outperformed previous work and is able to designate many more classifications of objects.

Bottom-Up attention Model

The Faster R-CNN is used with the ResNet-101 [50] CNN. To produce the required set of image characteristic, the VQA is used.The final result of the model is taken, and then it executes the non-maximum suppression. Each object class employs an IoU threshold. After this, entire regions are selected, and class discovery likelihood increased the confidence threshold. For each chosen area i, v_i is known as the mean-pooled convolutionary characteristic in this area, and the size D of the image characteristic is 2048. In this way, the faster R-CNN work efficiently as the 'Hard' attention mechanism, since only a small amount of image bounding box characteristic is chosen from the wide variety of potential combination.

Visual Space Based Deep Learning

For images and sentence-based descriptions, the authors discovered bidirectional mapping. The proposed method follows an RNN that can construct visual representation. The suggested technique is demonstrated on a dissimilar task: sentence generation and image/sentence retrieval. The outcomes of the suggested method are better than the other prevalent methods [51].

The forward and backward time-bounded scenario uses the 1st estimated inference process for 1-Best (and M-Best) decoding, and the bi-directional neural sequence is represented by increased beam search (BS) [52]. BS is famous for the approximate inference process for decoding sequences from unidirectional neural sequence frameworks. Bidirectional beam search (BiBS) is suggested for allowing the use of bi-directional models. The authors used a Fill-in-the-Blank Captioning job which needs reasoning about both previous and upcoming sentence scheme to rebuild sensible-image based descriptions. The proposed BiBS technique can produce more appropriate sentences and obtain excellent success on the famous Visual Madlibs dataset with several standard metrics.

The specific algorithm for the Bidirectional beam-search is presented below:

$Data$: $Given\ initial\ set\ of\ sequences\ Y_{[1:B],[1,T]}$

$\vec{\theta}_{[1:B],[1;T|} = \overleftarrow{\theta}_{[1:B],[1:T|} = 0$

While $not\ converged\ do$

// Update beams left-to-right

$for\ t = 1, \dots, T\ do$

$\vec{\theta}_{[1:B],t}, \overleftarrow{h}_{[1:B],t} = \overrightarrow{URNN}\left(\vec{h}_{[1:B],t-1}, Y_{[1:B],t-1}\right)$

$$Y_{[1:B],t} = top - B \sum_{i=1}^{t} \vec{\theta}_{b,i}\left(y_{b,i}\right) + \sum_{j=t}^{T} \overleftarrow{\theta}_{b'j}\left(y_{b',j}\right)$$

end

//Update beams right –to –left

$for\ t = 1, \dots, T\ do$

$\overleftarrow{\theta}_{[1:B],t}, \overleftarrow{h}_{[1:B],t} = \overleftarrow{URNN}\left(\overleftarrow{h}_{[1:B],t-1}, Y_{[1:B],t-1}\right)$

$$Y_{[1:B],t} = top - B \sum_{i=1}^{t} \overleftarrow{\theta}_{b,i}\left(y_{b,i}\right) + \sum_{j=t}^{T} \vec{\theta}_{b'j}\left(y_{b',j}\right)$$

end

Algorithm 1: Bidirectional Beam Search (BiBS).

3.2.11 Other Deep Learning Methods

The earliest approach using the coarse-to-fine method for image captioning is suggested by Wang, Y. et al. [53], which produces the skeleton-based sentence and phrases independently. Detailed study and research on the famous Microsoft COCO dataset demonstrate that the suggested technique performs better than the other advanced methods across distinct evaluation metrics, particularly on SPICE. SPICE has a higher-correlation than the traditional method.

Beam Search

The authors described the generative model with the deep recurrent framework that is able to process natural sentences for images [54]. The proposed model described Beam Search as the last phase to produce a sentence and enhance the chance of a specified interpretation sentence given in the training image. The chosen algorithm is based on the best first searching, which is worked iteratively. The algorithm selects k best sentences w.r.t. time

t as the candidate to produce sentences. Beam search avoids selecting the maximum probability word at each step, i.e. local maximum and alternatively chose the sequence of words with great overall probability score, i.e. global maximum. The specific algorithm for the beam search is shown below. The proposed method is applied to numerous datasets that prove the proposed method's accuracy and efficiency. The prior advanced BLUE-01 score on the pascal dataset is 25, whereas the proposed method achieves 59, based on human performance around 69. The proposed method shows the BLEU-1 score enhancements on Flickr30k, from 56 to 66, and on SBU, from 19 to 28. Finally, on the recently updated COCO dataset, the proposed method yields a BLEU-4 of 27.7, which is the latest most advanced technique.

Algorithm 2: K-Beam Search(K)

```
k ← K;
KBeamScores ← list(1);
bestKCaption list(" < begin > ");
bestKCaptionSequence ← [ ];
bestKScores ← [ ];
while True do
predicted possibility ← Pred(image, prevWords);
KBeamScores ← TopK(KBeamScores + predicted possibility);
nextWords ← TopK(reverse dic(KBeamScores));
bestKCaption. append(nextWords);
endIndex list(index(nextWord = " < end > "));
if len(endIndex) > 0 then
bestKCaptionSequence.
append(bestKCaption[endIndex]);
bestKScores.
append(KBeamScores[endIndex]);
k ← k − 1;
end
if k = 0 then
break;
end
end
bestIndex ← max(bestKScores);
return bestKCaptionSequence[bestIndex]
```

3.2.12 Attribute-based Representation

In several computer vision work, the attribute-based framework as a high-level demonstration has described the interest in several computer vision work like image recovery &image annotation and object identification. Farhadi et al. suggested a set of visual semantic for nonfamiliar objects recognition [55]. In [56], Vogel et al.used the visual attributes describing the scene for image area characterization and joined these native semantic with global image depiction. The 06 groups [57] of attributes are used to construct intermediate level characterization that is used for the classification of the image. An objectbank that allows the objects to be used as scene descriptions attribute [58, 59].

3.3 Novel Image Caption Generation Methods

Im2text method generate automatically image description via a great captioned photo collection. The Im2text executes on a great collection of Flickr queries and applies a filter over one-million images for removing noise. The authors compare the suggested technique with the other prevalent technique and find that the proposed method is a very basic non-parametric method and generate operative/pleasant outcomes [60]. The authors suggested an approach capable of enhancing the performance of captioning on an image and suitable for the task produces new concept learning [61]. The proposed method is based on the transposed weight sharing task to prevent overfitting the novel concepts. The 03 datasets are developed via novel concepts. The authors achieved results that are on par with or improved than the current state-of-the-art.

For creating images from NL descriptions, the authors suggested a generative framework known as alignDRAW [62]. The alignDRAW model, a grouping of recurrent variational-autoencoder and alignment-model over words. The proposed model trained on Microsoft COCO. The authors validate it on numerous baseline generative methods. The experimental outcome shows that the alignDRAW generates great quality samples than other prevalent methods and produces the picture with new unseen captioning.

The author proposed the novel method for image-based captioning and described a noun translation scheme [63]. The proposed method acquire great outcome by translating from a dataset of nouns captions. The proposed method also represents the lower and upper bounds of word categories. The extensive experiment shows that a blind-noun translation model can produce better captions compared to the advanced captioning techniques.

3.4 Retrieval Based Image Captioning

The author suggested hypernymy prediction to test the capability of the scheme to learn partial-order from incomplete data [64]. So the authors used hypernym pairs that is a pair of an idea is a speciality or an instance of the second. The authors worked on two main task-hypernym predictions and caption-image retrieval. The proposed methods outperform than all previous work.

3.5 Template Based Image Captioning

Neural Baby Talk can generate natural language explicitly grounded in entities object detectors discover in the image [65]. Neural Baby Talk is a two-stage approach. A two-stage approach comprises a mixture of words from a vocabulary of text and its corresponding slots to image regions and fills those slots based on recognized categories. The Neural Baby Talk framework's whole arrangement is end to end and tied together via sentence template creation and slot filling with object detectors. Neural Baby Talk evaluates on 02 standardized datasets, i.e., MSCOCO & Flickr30K. Valued comparisons to multiple baseline methods illustrate the feasibility of the Neural Baby Talk system.

3.6 Cross Lingual and Multilingual Image Captioning

Multimodal pivots are used for the enhancement of statistical machine translation of image descriptions [66]. The main objective of multimodal pivots is to retrieve the image over a huge database image captioning in the objective language. The employment of image captioning in the comparable images for cross-lingual re-ranking of translation outcomes. The authors demonstrated that the multimodal pivots into a target-side retrieval model enhanced the statistical machine translation (SMT) performance compared to baseline METEOR, BLEU, and TER on the suggested proportionate dataset obtained from the MSCOCO.

4. Dataset Used in Image Captioning Methods

4.1 nocaps (novel object captioning at scale)

nocaps dataset contains 166,100 men created captions that describe 15,100 images from the open image validation and test sets. The corresponding training data has an open-image image-level label, a boundingbox for the object, and contains sets of image caption from COCO. However, there are several other classes apart from COCO in an open image, approximately four hundred classes of an object in the test image have indeed no or little significant training caption(hence,nocaps)[67].

4.2 MS COCO

Currently, the MS COCO database [68] includes the 123,287 images with 05 distinct explanations. For 80 object classes, images in this dataset are annotated, meaning that bounding boxes are available for all cases of each of these types for all images. The MS COCO dataset is utilized extensively to describe images, which is made possible by the standard estimation server currently accessible. MS COCO extensions and the inclusion of questions and answers are currently in the development process [69].

4.3 Flickr 8k

The Flickr8K dataset [70] and it's enhanced version Flickr30K dataset [71] comprises images from Flickr, near about 8000 and 30000 images, correspondingly. The images in these 02 databases are chosen for unique objects and activities by user inquiries. These databases have 05 explanations for each image obtained from the AMT employs a technique equivalent to that of the Pascal1K dataset.

4.4 Flickr 30k

Flickr30K [72] is an automated image interpretation and grounded language comprehension dataset. It comprises 30K Flickr images and has 158K-captions generated by person annotators. It doesn't have a specified division of images for analysis and justification of instruction. For preparation, measurement, and evaluation, investigators may select their own choice of numbers. The dataset also includes typical object detectors, a colour classifier, and a tendency against more significant object collection.

4.5 Visual Genome

Visual genome dataset [73] can model the relationship and gathers object annotation, relation, and attributes inside each image to study these models. This dataset comprises around 108K pictures, where each picture has an average of 26-attributes, 35-objects, and twenty-one relationships among the objects. This dataset has the standardized attributes, objects; noun phrases, and relationship in area depiction and question-answer pairs to wordnet synsets. These annotations together constitute the densest and greatest dataset of pairs of image specifications, artifacts, properties, relationships, and query response pairs.

4.6 SBU Captioned Photo Dataset

Ordonez et al. [74] implemented the SBU1M Captions dataset that varies from the earlier datasets in that it is a web-scale dataset comprising about 01-million captioned photos. It is constructed from data accessible on Flickr with image details supplied by the user. The images are collected and processed from Flickr with the limitation that at least one verb & one noun is included in the prespecified control lists. The resultant dataset is supplied in the CSV format of URL.

4.7 IAPR TC12

Grubinger et al. [75] presented this dataset, and it contains about 20000 thousands of images with their descriptions. The images are obtained from various search engines like bing, google, and yahoo, and their description is also generated in multiple languages. Each and every image is related to the 01 to 05 description, while each description of the image denotes a diverse view of the image. The IAPR TC-12 dataset also contains the segmentation for the objects.

4.8 Instagram Dataset

This dataset [76, 77] contains images received from Instagram, a social media platform for photo sharing. This dataset contains approximately 10K photos, many of which come from famous people. However, this dataset

is used for hashtag prediction and post generation tasks. This dataset includes 1.1M posts from 6.3K members on a wide variety of subjects and a lengthy hashtag-list.

4.9 Stock3M Dataset

This dataset [53] is twenty-six times bigger than the famous MS COCO dataset and contains 3217654 uploaded images. There is a broad diversity of content in the images of this dataset.

4.10MIT-Adobe FiveK Dataset

This dataset [78] contains approximately five thousand images, and these images are about the person, environment, and human-made things.

4.11FlickrStyle10K Dataset

This dataset [79] contains ten thousand flicker images along with decorated captions. The training data comprises seven thousand images. The authentication and test data contain approximately one thousand and two thousand images.

4.12PASCAL 1K

A widely utilised dataset as a standard for assessing the description production mechanism's consistency is the Pascal1K sentence dataset[79]. This standard size dataset includes thousands of images. These images are obtained from the pascal 2008 object recognition dataset [80]. It also contains objects from different visual classes that include people, animals, and automobiles.

4.13STAIR

The STAIR is a Japanese dataset [81] for image descriptions, and it is grounded on the MSCOCO dataset. It contains 164062 images and 820310 total Japanese descriptions. This dataset is the most significant available Japanese image description dataset.

4.14UIUC PASCAL

UIUC PASCAL Sentences dataset [82] was one of the first datasets of image captions, comprising 1,000 photographs related to five distinct explanations gathered by crowdsourcing. It has been used for initial image captioning approaches, but it is seldom used because of its restricted domain, restricted magnitude, and relatively basic captions.

4.15AIC

The first largest dataset of Chinese description in the area of image caption development is the Chinese image description dataset, extracted from the AI Challenger [83]. This dataset contains 210000 pictures for preparation purposes and 30000 pictures for validation sets. Every single picture is followed by 05 Chinese explanations, identical to MSCOCO, highlighting essential details in the pictures, covers all the important types, views, acts, and supplementary material.

4.16VizWiz

This dataset [84] contains 117115 training captions, 23431 training images, 7750 validation images, 38750 validation caption,40000 test captions, and 8000 test images.

4.17Visual and Linguistic Treebank (VLT2K)

The Visual and Linguistic Treebank (VLT2K; [85]) utilizes the images that are obtained from the pascal2010 action-recognition dataset. It extends these images with a 02 to 03 sentence depiction per image. On the AMT, these explanations are obtained with detailed instructions for verbalizing the key activity seen in the image and the actor directly involved. At the same time, the most significant background items are also listed. For a subcategory of 341 images of the visual and verbal treebank, object annotation is obtainable (in the manner of polygons nearby all objects declared in the representations.) For this subcategory, individually generated visual dependency representations are also incorporated (03 VDRs for each image).

4.18Abstract Scenes dataset

Abstract Scenes dataset [86] includes the ten thousand clipart images and their description. This description is classified into 02 different groups. The number one group includes a description of the single sentence, while the number two group contains an alternative description per image. These 02 explanations consist of 03 simple sentences with a different element of the scene mentioned in each sentence. This dataset's key benefit is the freedom to discover the generation of image descriptions without the need for automated object detection while eliminating the related noise. The latest edition of this dataset has been generated as a portion of the visual question-answering (VQA) dataset [87]. It includes 50,000 separate pictures of the scene with more detailed human figures and 05 explanations of single sentences description.

4.19NYU dataset

One article [88] uses the NYU [89] dataset comprising 1,449 interior scenes along with 3-D object segmentation scenes. With 05 explanations per image, this dataset has been extended by Lin et al. [88].

4.20BBC News dataset

The BBC News dataset [90] was one of the initial set of images and co-occurring texts. Feng and Lapata [90] obtained 3,361 news stories from the british broadcasting-corporation news website, with the limitation that a picture and a description be included in the post.

4.21Deja-Image Captions dataset

The Deja-Image Captions dataset [91] includes four million images and 180000 captions which is obtained from the Flicker. The image captions are normalised by lemmatization and stopping word elimination from

constructing a dataset of almost equivalent texts. For example, the phrase plane flies in the blue-sky and a plane flying into the blue-sky is normalized as the plane fly in the blue-sky. Image caption pairs are maintained if more than one individual repeats the captions in a normalized form.

4.22MSVD

MSVD [92] gathers 1,970 video clips from youtube, with approximately forty available English explanations for each video. MSVD-dataset is split into training, validation, and testing set along with setting 1,200/100/670 for videos.

4.23M-VAD

This huge movie dataset M-VAD [93] includes forty-nine thousand movie extracts that are obtained basically from the dvd movies. Each of which can be followed by a single sentence from concise video service narrations (DSN) that are semi-automatically translated. This dataset contains 39000 videoclips for training purposes,5000 video clips for validation and testing purposes.

4.24MPII-MD

MPII-MD [94] is yet another series of large-scale movie snippets featuring 68,000 film clips from Hollywood films with accompanying sentences. MPII-MD is constructed in the same way as M-VAD [93], while its alignment among video clips and explanations is manually proofed. This dataset is generally categorized into training, validation, and test samples category.

4.25Visual Madlibs Dataset (VML)

Visual Madlibs Dataset (VML) [95] is a subset of the MS COCO dataset, and it contains 10,783 images that seek to go beyond defining which objects are in the image.

4.26Toronto COCO-QA

Thes dataset [96] is visual queries and response dataset, and in it, queries are created automatically from the image caption of the MS COCO dataset. This dataset contains 123,287 photos with 117,684 queries regarding objects, numbers, colours, or positions with a single word response.

4.27Microsoft Research Video Description Corpus (MS VDC)

MS VDC dataset [97] includes a parallel description of 2089 small video snippets. The explanations are single-sentence summaries of the video's actions or events. Paraphrase & multi-lingual substitutes are collected in this dataset so that the dataset can be useful for interpretation, paraphrasing, and video explanation.

4.28Short Videos Described with Sentences

Short videos [98] described with sentences include a sixty-one video clip, and each video clip is 35 seconds long and has 03 diverse outdoor environments. This dataset has several concurrent events among a subset of 04 items in this dataset: a human, a backpack, a chair, and a trash-can. In this dataset, multiple sentences explain what is happening in the video, and every video is manually annotated.

5. Evaluation Metrics

5.1 BLEU (Bilingual-evaluation-understudy)

BLEU [99] is a standard utilized to calculate the quality of text produced by a computer. A collection of reference texts is compared to individual text fragments, and scores are calculated for each of them.BLEU has 04 variants, i.e., BLUE1, BLUE2, BLUE3, and BLUE4. The calculated scores are averaged when calculating the overall quality of the output text. Syntactic exactness, however, is not measured here. Based on the count of reference translations and the extent of the produced text, the BLEU metric's output can vary. An updated precision metric was later presented by Papineni et al.[99]. This measure employs n-grams. BLEU is famous since it is a leader in automated machine-translated text assessment and has a fair connection with human quality evaluations [100,101]. Some boundaries, such as BLEU scores, are excellent only if the text produced is small [100]. In certain cases, an improvement in the BLEU score does not demonstrate that the text's content is perfect.

5.2 ROUGE

ROUGE [102] is a collection of metrics used for determining the value of text summaries. It performs the comparison of word sequences, word pairs, and n-grams with a collection of human-generated comparison summaries. There are 04 numerous versions of ROUGE such as ROUGE-1, 2, ROUGE-W, and ROUGE-SU4. ROUGE-1 & ROUGE-W are used for sole document estimation, while RPUGE-2 & ROUGE-SU4 are best for small sum-ups. However, ROUGE has some issues while assessing multi-document-text summaries.

5.3 METEOR

Another metric used to measure the machine's interpreted language is METEOR (Metric-for-evaluation-of-translation-with-explicit ORdering) [103]. The reference texts are compared to Standard word segments. Besides this, stems of an expression and alternative word of words are often conceived for matching. METEOR can create an improved association at the sentence or the segment-level.

5.4 CIDEr

CIDEr(Consensus-based-image-description-evaluation) [104] is another metric for estimating the image explanation. There are only 05 captions per image in most of the datasets. With this limited number of sentences, previous measurement metrics function were not adequate to quantify the consensus among captions produced and human decision. However, using the word frequency-inverse text frequency (TF-IDF) [105], CIDEr meets human consensus.

5.5 SPICE

SPICE(Semantic-propositional-image-caption-evaluation) [106] is a modern metric for caption assessment based on semantic principles. It is constructed on a graph-based semantic depiction entitled a scene graph. This graph can obtain information from diverse attributes, objects, and their relationship from the image description.

6. Discussion and Conclusion

In this research paper, we have described DL-based image-captioning approaches. We also provided the classification of image captioning methods, presented standard block diagrams of the major groups. Thus, we have analyzed and compared the different image captioning methods and provide a comprehensive review of these methods. We addressed several datasets and evaluation metrics. In this field, we described briefly possible opportunities for further research. Even though image captioning techniques focused on deep learning have made considerable growth in current years, a comprehensive image captioning approach that can create significantly better quality captions for almost all images is yet to be obtained. With the introduction of novel DL-network architectures, automated image-captioning will continue to be a popular field of research. In relation to human analysis, this research paper identified the new and most innovative image caption method and encouraged reproducibility and further studies.

Consent For Publication

Not Applicable.

Funding

None.

Conflict of Interest

The authors declare no conflict of interest, financial or otherwise.

Acknowledgements

I would like to thank the reviewers, whose comments and feedback have significantly improved the quality of this article.

References

1. Sathi, Cognitive (Internet of) Things: Collaboration to Optimize Action. Palgrave Macmillan US, 2016.
2. J. Kim, J. Jun, and B. Zhang, "Bilinear Attention Networks," NeurIPS, arXiv e-prints, arXiv:1805.07932, May 2018.
3. M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A Comprehensive Survey of Deep Learning for Image Captioning," ACM Comput. Surv., vol. 51, pp. 1–36, 2018
4. Fan and D. J. Crandall, "DeepDiary: Automatic Caption Generation for Lifelogging Image Streams," ArXiv, vol. abs/1608.03819, 2016.
5. C. Park, B. Kim, and G. Kim, "Attend to You: Personalized Image Captioning with Context Sequence Memory Networks," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 6432–6440, 2017.
6. J. Gu, G. Wang, J. Cai, and T. Chen, "An Empirical Study of Language CNN for Image Captioning," 2017 IEEE Int. Conf. Comput. Vis., pp. 1231–1240, 2017.
7. B. Z. Yao, X. Yang, L. Lin, M. Lee, and S. Zhu, "I2T: Image Parsing to Text Description," Proc. IEEE, vol. 98, pp. 1485–1508, 2010.
8. J. Jin, K. Fu, R. Cui, F. Sha, and C. Zhang, "Aligning where to see and what to tell: image caption with region-based attention and scene factorization," ArXiv, vol. abs/1506.06272, 2015.
9. Q. Wu, C. Shen, L. Liu, A. Dick, and A. V. D. Hengel, "What Value Do Explicit High Level Concepts Have in Vision to Language Problems?," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 203–212, 2016.
10. Q. Feng, Y. Wu, H. Fan, C. Yan, and Y. Yang, "Cascaded Revision Network for Novel Object Captioning," IEEE Trans. Circuits Syst. Video Technol., vol. 30, pp. 3413–3421, 2020.
11. S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning Images with Diverse Objects," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1170–1178, 2017.
12. S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, "Improved Image Captioning via Policy Gradient optimization of SPIDEr," IEEE Int. Conf. Comput. Vis., pp. 873–881, 2017.
13. B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards Diverse and Natural Image Descriptions via a Conditional GAN," IEEE Int. Conf. Comput. Vis., pp. 2989–2998, 2017.
14. W. Chen, A. Lucchi, and T. Hofmann, "Bootstrap, Review, Decode: Using Out-of-Domain Textual Data to Improve Image Captioning," ArXiv, vol. abs/1611.05321, 2016.
15. Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised Image Captioning," IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 4120–4129, 2019.
16. S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-Critical Sequence Training for Image Captioning," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1179–1195, 2017.

17. 17. X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding the Long-Short Term Memory Model for Image Caption Generation," IEEE Int. Conf. Comput. Vis., pp. 2407–2415, 2015.
18. Wang, H. Yang, C. Bartz, and C. Meinel, "Image Captioning with Deep Bidirectional LSTMs," Proc. 24th ACM Int. Conf. Multimed., 2016.
19. Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik, "Learning to generalize to new compositions in image understanding," ArXiv, vol. abs/1608.07639, 2016.
20. H. Liu, Y. Yang, F. Shen, L. Duan, and H. Shen, "Recurrent Image Captioner: Describing Images with Spatial-Invariant Transformation and Attention Filtering," ArXiv, vol. abs/1612.04949, 2016.
21. T. Yao, Y. Pan, Y. Li, and T. Mei, "Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 5263–5271, 2017.
22. T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting Image Captioning with Attributes," IEEE Int. Conf. Comput. Vis., pp. 4904–4912, 2017.
23. Sutskever, O. Vinyals, and Q. V Le, "Sequence to Sequence Learning with Neural Networks," NIPS, 2014, Online Available: https://arxiv.org/abs/1409.3215.
24. J. Johnson, A. Karpathy, and L. Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4565–4574, 2016.
25. L. Yang, K. D. Tang, J. Yang, and L. Li, "Dense Captioning with Joint Inference and Visual Context," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1978–1987, 2017.
26. J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, "A Hierarchical Approach for Generating Descriptive Image Paragraphs," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3337–3345, 2017.
27. W. Zhao, X. Wu, and X. Zhang, "MemCap: Memorizing Style Knowledge for Image Captioning," AAAI, vol. 34, no. 07, pp. 12984–12992, 2020.
28. J. Donahue, L.A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2625–2634, 2015.
29. L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1–10, 2016.
30. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, pp. 664–676, 2017.
31. J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille, "Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)," arXiv Comput. Vis. Pattern Recognit., 2015.
32. C. C. Park and G. Kim, "Expressing an Image Stream with a Sequence of Natural Sentences," NIPS, 2015.
33. J. Gu, G. Wang, and T. Chen, "Recurrent Highway Networks with Language CNN for Image Captioning," ArXiv, vol. abs/1612.07086, 2016.
34. H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han, "Show, Observe and Tell: Attribute-driven Attention Model for Image Captioning," IJCAI, pp. 606–612, 2018.
35. K. Xu, J. Ba, R. Kiros, K. Cho, A.C. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio,"Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," ICML, 2015.
36. S. He, H. Tavakoli, A. Borji, and N. Pugeault, "A Synchronized Multi-Modal Attention-Caption Dataset and Analysis," ArXiv, vol. abs/1903.02499, 2019.
37. Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact Bilinear Pooling," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 317–326, 2016.
38. C. Yu, X. Zhao, Q. Zheng, P. Zhang, and X. You, "Hierarchical Bilinear Pooling for Fine-Grained Visual Recognition," ECCV, pp. 595–610, 2018.
A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding," ArXiv, vol. abs/1606.01847, 2016.
39. P. Fang, J. Zhou, S. Roy, L. Petersson, and M. Harandi, "Bilinear Attention Networks for Person Retrieval," IEEE/CVF Int. Conf. Comput. Vis., pp. 8029–8038, 2019.
40. J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 7132–7141, 2018.
41. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," NIPS, pp. 7132–7141, 2017.

42. Z. Ren, X. Wang, N. Zhang, X. Lv, and L. Li, "Deep Reinforcement Learning-Based Image Captioning with Embedding Reward," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1151–1159, 2017.
43. J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3242–3250, 2017.
44. Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image Captioning with Semantic Attention," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4651–4659, 2016.
45. L. Zhou, C. Xu, P. A. Koch, and J. J. Corso, "Image Caption Generation with Text-Conditional Semantic Attention," ArXiv, vol. abs/1606.04621, 2016.
46. L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua, "SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 6298–6306, 2017.
47. M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of Attention for Image Captioning," IEEE Int. Conf. Comput. Vis., pp. 1251–1259, 2017.
48. V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-Down Visual Saliency Guided by Captions," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3135–3144, 2017.
49. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 770–778, 2016.
50. X. Chen and C. L. Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 2422–2431, 2015.
51. Q. Sun, S. Lee, and D. Batra, "Bidirectional Beam Search: Forward-Backward Inference in Neural Sequence Models for Fill-in-the-Blank Image Captioning," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 7215–7223, 2017.
52. Y. Wang, Z. L. Lin, X. Shen, S. Cohen, and G. Cottrell, "Skeleton Key: Image Captioning by Skeleton-Attribute Decomposition," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 7378–7387, 2017.
53. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3156–3164, 2015.
A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1778–1785, 2009.
54. J. Vogel and B. Schiele, "Semantic Modeling of Natural Scenes for Content-Based Image Retrieval," Int. J. Comput. Vis., vol. 72, pp. 133–157, 2006.
55. Y. Su and F. Jurie, "Improving Image Classification Using Semantic Attributes," Int. J. Comput. Vis., vol. 100, pp. 59–77, 2012.
56. L. Li, H. Su, E. Xing, and L. Fei-Fei, "Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification," NIPS, pp. 1378–1386, 2010.
57. L. Li, H. Su, Y. Lim, and L. Fei-Fei, "Objects as Attributes for Scene Classification," ECCV Workshops, pp. 57–69, 2010.
58. Elliott and M. Kleppe, "1 Million Captioned Dutch Newspaper Images," LREC, pp. 3054–3058, 2016.
59. J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Learning Like a Child: Fast Novel Visual Concept Learning from Sentence Descriptions of Images," IEEE Int. Conf. Comput. Vis., pp. 2533–2541, 2015.
60. Mansimov, E. Parisotto, J. Ba, and R. Salakhutdinov, "Generating Images from Captions with Attention," CoRR, vol. abs/1511.02793, 2016.
61. Heuer, C. Monz, and A. Smeulders, "Generating captions without looking beyond objects," ArXiv, vol. abs/1610.03708, 2016.
62. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-Embeddings of Images and Language," CoRR, vol. abs/1511.06361, 2016.
63. Lu, J. Yang, D. Batra, and D. Parikh, "Neural Baby Talk," IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 7219–7228, 2018.
64. Hitschler, S. Schamoni, and S. Riezler, "Multimodal Pivots for Image Caption Translation," ArXiv, vol. abs/1601.03916, 2016.
65. H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee and P. Anderson, "nocaps: novel object captioning at scale," Int. Conf. Comput. Vis., pp. 8947–8956, 2019.

66. T. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, "Microsoft COCO: Common Objects in Context," ECCV, pp. 740–755, 2014.
67. Agrawal, J. Lu, S. Antol, M. Mitchell, C.L. Zitnick, D. Parikh and D. Batra., "VQA: Visual Question Answering," Int. J. Comput. Vis., vol. 123, pp. 4–31, 2015.
68. M. Hodosh, P. Young, and J. Hockenmaier, "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics (Extended Abstract)," J. Artif. Intell. Res., vol. 47, pp. 853–899, 2013.
69. P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," Trans. Assoc. Comput. Linguist., vol. 2, pp. 67–78, 2014.
70. B. A. Plummer, L. Wang, C. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," Int. J. Comput. Vis., vol. 123, pp. 74–93, 2015.
71. R. Krishna, Y. Zhu, Z. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D.A. Shamma, M.S. Bernstein, and L. Fei-Fei, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," Int. J. Comput. Vis., vol. 123, pp. 32–73, 2016.
72. V. Ordonez, G. Kulkarni, and T. Berg, "Im2Text: Describing Images Using 1 Million Captioned Photographs," NIPS, pp. 1143–1151, 2011.
73. M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems," 2006.
74. Tran, X. He, L. Zhang, and J. Sun, "Rich Image Captioning in the Wild," IEEE Conf. Comput. Vis. Pattern Recognit. Work., pp. 434–441, 2016.
75. E. Mulyanto, E. I. Setiawan, E. M. Yuniarno, and M. H. Purnomo, "Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset," IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl., pp. 1–5, 2019.
76. V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," CVPR 2011, pp. 97–104, 2011.
77. T. Chen, Z. Zhang, Q. You, C. Fang, Z. Wang, H. Jin, and J. Luo,"Factual' or 'Emotional': Stylized Image Captioning with Adaptive Learning and Attention," in ECCV, pp. 527–543, 2018.
78. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," Int. J. Comput. Vis., vol. 88, pp. 303–338, 2009.
79. Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset," ArXiv, vol. abs/1705.00823, 2017.
80. D. Elliott and F. Keller, "Image Description using Visual Dependency Representations," EMNLP, pp. 1292–1302, 2013.
81. J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, and Y. Wang, "AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding," ArXiv, vol. abs/1711.06475, 2017.
82. D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning images taken by people who are blind," European Conference on Computer Vision, pp. 417-434, 2020.
83. C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting Image Annotations Using Amazon's Mechanical Turk," Mturk@HLT-NAACL, pp. 139—147, 2010.
84. C. L. Zitnick, D. Parikh, and L. Vanderwende, "Learning the Visual Interpretation of Sentences," IEEE Int. Conf. Comput. Vis., pp. 1681–1688, 2013.
85. Agrawal, A. Kembhavi, D. Batra, and D. Parikh, "C-VQA: A Compositional Split of the Visual Question Answering (VQA) v1.0 Dataset," ArXiv, vol. abs/1704.08243, 2017.
86. D. Lin, S. Fidler, C. Kong, and R. Urtasun, "Generating Multi-sentence Natural Language Descriptions of Indoor Scenes," BMVC, pp. 1-13,2015.
87. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," ECCV, pp. 746–760, 2012.
88. Y. Feng and M. Lapata, "Automatic Image Annotation Using Auxiliary Text Information," ACL, pp. 272–280, 2008.
89. J. Chen, P. Kuznetsova, D. Warren, and Y. Choi, "Déjà Image-Captions: A Corpus of Expressive Descriptions in Repetition," HLT-NAACL, pp. 504–514, 2015.

90. D. L. Chen and W. Dolan, "Collecting Highly Parallel Data for Paraphrase Evaluation," ACL, pp. 190–200, 2011.
91. Torabi, C. Pal, H. Larochelle, and A. C. Courville, "Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research," ArXiv, vol. abs/1503.01070, 2015.
92. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for Movie Description," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3202–3212, 2015.
93. L. Yu, E. Park, A. Berg, and T. Berg, "Visual Madlibs: Fill in the blank Image Generation and Question Answering," ArXiv, vol. abs/1506.00278, 2015.
94. Ren, R. Kiros, and R. Zemel, "Exploring Models and Data for Image Question Answering," NIPS, pp. 2953–2961, 2015.
95. F. Rahutomo and A. H. Ayatullah, "Indonesian Dataset Expansion of Microsoft Research Video Description Corpus and Its Similarity Analysis," Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control, pp. 319–326, 2018.
96. H. Yu and J. Siskind, "Grounded Language Learning from Video Described with Sentences," ACL, pp. 53–63, 2013.
97. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," ACL, pp. 311, 2002.
98. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluation the Role of Bleu in Machine Translation Research," EACL, pp. 249–256, 2006.
99. E. Denoual and Y. Lepage, "BLEU in Characters: Towards Automatic MT Evaluation in Languages without Word Delimiters," IJCNLP, 2005.
100. C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," ACL, pp. 74–81, 2004.
101. S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," IEEvaluation@ACL, pp. 65–72, 2005.
102. R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," IEEE Conf. Comput. Vis. Pattern Recognit., pp. 4566–4575, 2015.
103. S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," J. Doc., vol. 60, pp. 503–520, 2004.
104. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," ECCV, pp. 382–398, 2016.