Academic Science Journal

Image Captioning Generator Using Deep Learning Models: An Abbreviated Survey

Yasir Hameed Zaidan and Jumana Waleed

Department of Computer Science, College of Science, University of Diyala

scicompms2132@uodiyala.edu.iq

Received: 5 January 2023 **Accepted: 4 April 2023**

DOI: https://doi.org/10.24237/ASJ.02.02.733B

Abstract

Captioning an image is the process of using a visual comprehension system with a model of language, by which we can construct sentences that are meaningful and syntactically accurate. These accurate phrases can explain the natural language (The seen content of the image). As a relatively young field of study, it is gaining growing attention. To accomplish image caption, semantic information about the images must be gathered and conveyed in natural language. Computer vision and natural language processing are both used in the difficult task of image captioning. That issue has received a lot of proposals for solutions. An abbreviated survey of image captioning studies is given in this paper. We concentrate our efforts on neural network-based approaches that deliver current outcomes. Neural network-based techniques are broken down into subcategories in accordance with the implementation architecture. The most recent methodologies are then compared to normative data sets. Methods based on neural networks are classified into subcategories according to the framework being used.

Keywords: Image Caption, Deep Learning Models, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM)

Academic Science Journal

منشئ تعليق الصور باستخدام نماذج التعلم العميق: مسح موجز

ياسر حميد زيدان و جمانة وليد

قسم علوم الحاسوب - كلية العلوم - جامعة ديالى

الخلاصة

التسمية التوضيحية للصورة هي عملية استخدام نظام الفهم البصري مع نموذج للغة ، يمكننا من خلاله بناء جمل ذات معنى ودقيقة نحويًا. يمكن لهذه العبارات الدقيقة أن تشرح اللغة الطبيعية (المحتوى المرئي للصورة). باعتباره مجالًا حديثًا للدراسة نسبيًا ، فإنه يحظى بالاهتمام. لإنجاز تعليق الصور ، يجب جمع المعلومات الدلالية حول الصور ونقلها باللغات الطبيعية. يعد تعليق الصورمهمة صعبة تجمع بين خبراء في رؤية الكمبيوتر ومعالجة اللغة الطبيعية. تم تقديم العديد من الحلول لهذه المشكلة. تم تقديم مسح مختصر لدراسات التعليق على الصور في هذه الورقة. نحن نركز جهودنا على النهج القائمة على الشبكة العصبية التي تقدم النتائج الحالية. يتم تصنيف الأساليب القائمة على الشبكة العصبية إلى فئات فرعية بناءً على إطار العمل الذي يتم تنفيذها فيه. ثم تتم مقارنة أحدث المنهجيات مع مجموعات البيانات المعيارية. يتم تصنيف الأساليب القائمة على الشبكات العصبية إلى فئات فرعية وفقًا للإطار المستخدم.

الكلمات المفتاحية : تعليق على الصورة, نماذج التعلم العميق, الشبكة العصبية التلافيفية, الشبكة العصبية المتكررة, الذاكرة طويلة المدى

Introduction

Automatically producing captions or descriptions for images is a challenging subject that involves a visual combination & language input. In other words, it requires both comprehensive visual understanding and advanced natural language creation. Because of this, the fields of computer vision and natural language processing now regard it as a fascinating topic [1]. The practice of using a computer to automatically produce a natural-language description of an image is known as image captioning. Image captioning is a discipline that integrates CV with NLP research. It requires the capacity to completely comprehend the semantic contents of an image as well as the ability to express that knowledge in human-like sentences. The difficulty of the endeavor is increased by the complexity of framing a phrase to convey such information [2].

The current focus on academics' attempts to summarize the contents of an image using a believable phrase in English or another language has piqued the interest of the computer vision sector. The objects, semantic features, object connections, and expected behaviors of an image are frequently transformed into a representation vector by a model for automatic picture caption synthesis in order to explain it. Furthermore, the caption can be generated word for word using a word generator. Recent developments in deep learning techniques have led to a great variety of approaches to that topic being offered to utilize that paradigm [3].

Model development has seen considerable advancements in the scientific community during the last few years, starting with the earliest ideas based on deep learning and employing advancements in Transformers, self-attention, as well as single-stream BERT-like algorithms. Methods have been improved by feeding recurrent neural networks (RNNs) with global visual descriptors and attentive procedures. In order to compare results to human-produced real-world scenarios. Researchers in both (NLP) and (CV) have developed standards and procedures for assessment. Image captioning is still a problem that needs to be resolved after years of research and advancements [4].

The fundamental contribution of this paper is to give an abbreviated survey of image captioning using deep learning models. The organization of this paper is a general structure of an image caption, image caption-based deep learning models, performance analysis, and main conclusions.

Image Caption

Captioning images has steadily grabbed the interest of many artificial intelligence experts, and it has evolved into an interesting and time-consuming procedure. The capabilities of NLP and CV are merged in photo captioning, which is crucial for understanding situations, depending on the information provided in an image. The implementation of human-computer interaction demonstrates the ubiquitous and crucial use of image captions. To accurately and completely describe the main semantic content of the picture. In order to properly caption a picture, background information about the target objects, the item, and the scene's surrounds must be used, along with the identification of objects and the justifications for their relationships. As a result, captioning images is a difficult task [5]. Because encoder-decoder models provide more accurate captions than template and retrieval-based models, they are increasingly commonly employed for picture captioning. Convolutional neural networks (CNN) are often utilized by the encoder to extract information, whereas (RNN) are frequently employed by the decoder to produce captions (CNN). Additionally, a strategy for focusing attention has been developed to reduce irrelevant data and boost the precision of visual descriptions [6]. Figure1 depicts the architecture of the encoder-decoder paradigm.

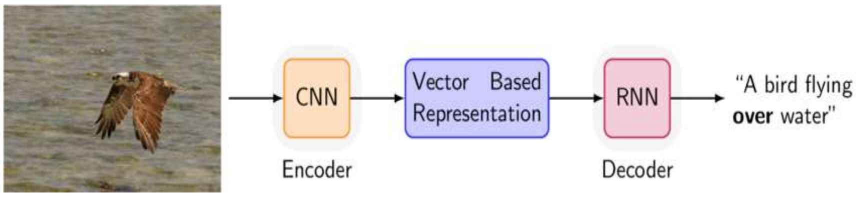

Figure 1: Example of the architectural design of the encoder-decoder for captioning images.

A. Convolution Neural Network

In several computer vision applications, the value of CNN models has been demonstrated. LeCun was the first to propose convolutional neural networks for image processing. CNN models have two fundamental characteristics: spatially shared weights and spatial aggregation. When the supplied data is generally in two dimensions. Processing of natural language and speech recognition are two examples of sequential data that CNN has been used to address.

By switching and stacking convolutional kernels and using pooling techniques, CNN uses abstract features to learn them. Contrary to pooling layers, which extract the most important features from sliding windows of the raw input data with a fixed length, convolutional layers (also known as convolutional kernels) in CNN integrate a number of local filters to produce invariant local features. Due to the fact that 2D-CNN has been widely demonstrated to be superior to 1D-CNN in past studies, just the mathematical aspects underlying it will be discussed [7]. 1D-CNN are provided:

Starting off, let's assume that an input sequence data is displayed as $x = [x1, ..., xT]$, with T being a series a length and xi denoting *Rd* value during each of time step.

Convolutional is defined as follows: Using the dot product of a filter vector Rmd and a concatenation vector representation, the convolution process is defined xi:i+m1:

$$ci = \phi(u^{T} xi{:}i+m-1 + b) \qquad (1)$$

Where ∗T is a divert of the matrix ∗, b and ϕ are a bias term and non-linear activation functions, respectively. xi:i+m−1 is the m-length window that begins with the i^{th} time step, which is defined as follows:

$$xi : i + m - 1 = xi \oplus xi+1 \oplus \cdots \oplus xi+m-1 \qquad (2)$$

The output scale *ci*, as defined in Eq. (1), can be thought of as the subsequence xi:i+m-1 is triggered through the filter u. Dragging the filtering window from the start of the time step to its end will result in a vector representation of the feature map.:

$$c_j = [c_1 , c_2 ,..., c_{l-m+1}] \qquad (3)$$

The j-th filter is represented through the index j. It's the same as multi-windows. { x1: m , x2:m+1 ,..., xl−m+1:l }.

The pooling layer may be used to further reduce the length of the feature map, hence reducing the number of model parameters. The pooling length is represented through the symbol s, it is a hyper-parameter of the pooling algorithm layer. In feature map cj, the MAX operation takes the maximum of s consecutive values. Hence, the following formula may be used to obtain the compressed feature vector:

$$h = [\, h1 , h2 ,..., h_{\frac{l-m}{s}+1} \,] \qquad (4)$$

Hj = max(c(j-1) s,c (j-1)s+1,...,cjs-1). To produce predictions, fully linked layers and the soft max layer are often placed as top layers, alternately with the convolution and max-pooling layers [8]. To illustrate, Fig. 2 depicts the structure of a one-layer CNN.

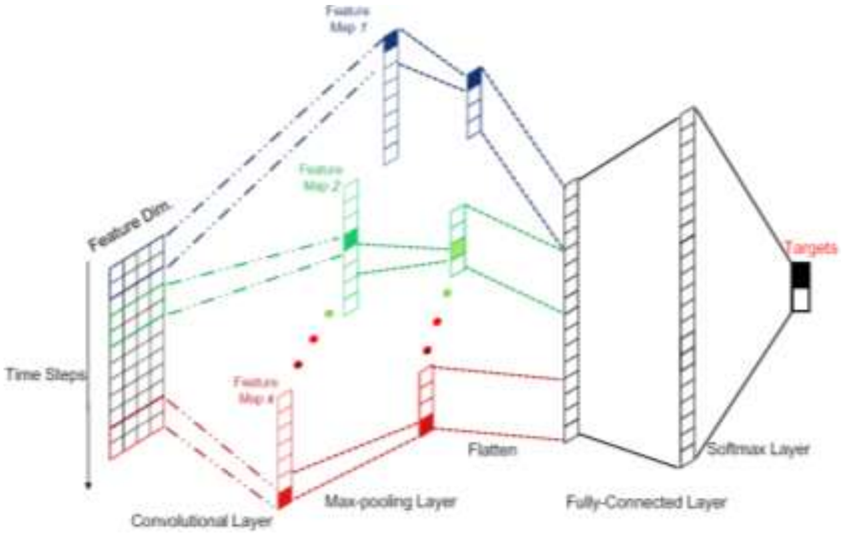

Figure 2: A one-layer CNN consists of a convolutional layer, a pooling layer, a fully connected layer, and a softmax layer. [8].

B. Recurrent Neural Network (RNN)

Write your material first and save it as a separate text file before starting to format your paper. Keep your visual and text files separate until the text has been styled and formatted. Avoid using hard tabs, and only use one hard return per paragraph to complete a sentence. There should be no pagination added to the paper. No need to number text heads; the template will take care of that.

RNNs are the most sophisticated neural networks, capable of generating and addressing arbitrary idea lengths of the input pattern. RNN may produce directed cycles by joining units together. All past inputs may be mapped to target vectors by an RNN, and the network's internal state has a memory of previous inputs. A perceptron can only map incoming data to target vectors; in contrast, the fundamental neural network is multi-layered. For operations that need supervision and include sequential input data and outputs, RNNs may be trained through back propagation across time.

As seen in Fig. 3, the RNN may use its internal memory to handle sequential data. In Fig. 3 (a), With the help of the current time data xt, the transition function modifies the current concealed output at step time t, and the prior output is hidden (ht)-1.

$$(ht) = H (xt , (ht)-1) \qquad (5)$$

The function of transformation In the last time step of processing the learned input sequence data of length T, (ht) is a hidden output and H is a nonlinear differentiable function. The resulting representation (ht) is layered with a standard multilayer perceptron (MLP) to map it to targets.

A number of transition functions may be used to build RNN models. Vanilla RNN, which is defined as follows, is the most fundamental.

$$(ht) = \phi (W_{xt} + H(ht)_{-1} + b) \qquad (6)$$

The transformation matrices are W and H, while the bias vector is b. The functions sigmoid and tan h are examples of nonlinear activation functions. Standard RNNs might not be able to recognize long-term relationships because of the vanishing gradient problem discovered during backpropagation during model training. To stop backpropagated errors from dissipating or inflating, (LSTM) and recurrent neural networks with gated inputs (GRU) were created. The core concept underlying these advanced RNN implementations is the use of gates to circumvent the problem of long-term dependence and let each recurrent unit to independently operate to adaptively store dependencies over time spans.

Transitional algorithms like LSTMs and GRUs, multilayer and bidirectional recurrent structures, and other techniques may be utilized in addition to these suggested improvements.

As seen in Fig. 3(b), the buried output of one recurrent layer can be used as input data for the subsequent recurrent layer by being transferred across time. As illustrated in Fig. 3(c), the structure of bidirectional recurrence has two unique hidden layers that may process sequence input in both forward and backward directions. The hidden layer function is described by the following equations, where → and← represent forward and backward processes, respectively.

$$\overrightarrow{\mathrm{h_t}} = \overrightarrow{\mathrm{H}}\left(\mathrm{x_t} \,.\, \overrightarrow{\mathrm{h}}_{\mathrm{t-1}}\right) \qquad (7)$$

$$\overleftarrow{\mathrm{h_t}} = \overleftarrow{\mathrm{H}}\left(\mathrm{x_t} \,.\, \overleftarrow{\mathrm{h}}_{\mathrm{t-1}}\right) \qquad (8)$$

The last vector (h_t) is then concatenated vector forwarding, backward operation outputs [8], as follows:

$$h_t = \overrightarrow{h_t} \oplus \overleftarrow{h_1} \qquad (9)$$

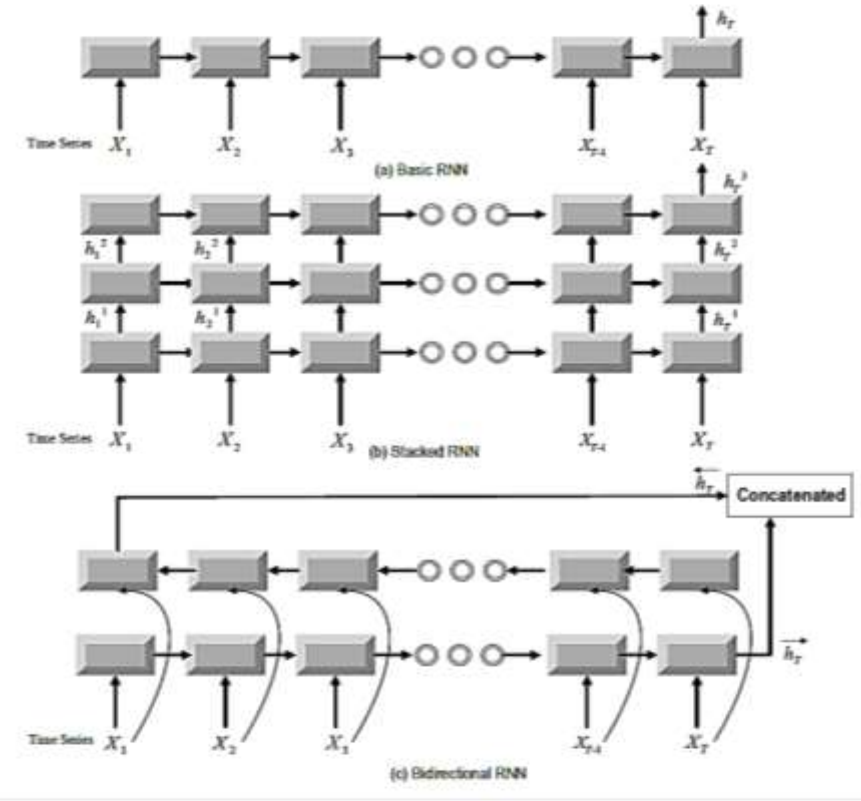

Figure 3: The illustration of normal, stacked, and bidirectional RNNs [8].

C. The Long Short-Term Memory

The RNN called an LSTM makes use of the results connection to function as a universal computer. It may be used to identify patterns and sequences and to analyze images. LSTMs are composed of input, output, and forgetting units. For example, the LSTM may remember what was computed at a previous time step and can control when the input is permitted into the neuron. If you want to understand more about LSTMs, here are a few things you should know. Figure 4 shows the LSTM approach's architectural layout [9]. Many different applications, especially picture caption generators, have demonstrated LSTM to be quite promising.

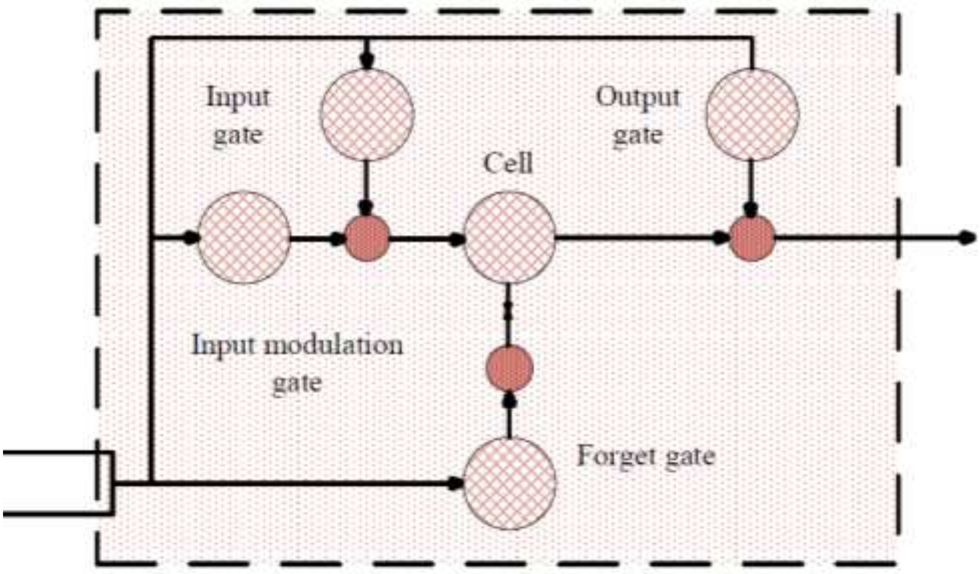

Figure 4: LSTM Architecture [9].

Image Caption Based on Deep Learning Models

Deep learning (DL) is a type of machine learning (ML) that uses several nonlinear processing layers and is based on neural networks. Finding hierarchical representations of data is the goal of deep learning [10]. Today, many deep-learning architectures are available, and that field of research is expanding at a breakneck pace with new models developed on a weekly basis. Additionally, the community is fairly receptive, as seen through the abundance of high-quality deep-learning lessons and books. As a result, just a summary of some significant deep-learning studies in image captioning is presented.

In 2014, Junhua Mao et al. [11] developed a model that uses a multimodal RNN to provide creative image descriptions. A deep RNN for texts and a deep CNN for pictures make up the model's two sub-networks. The whole multimodal RNN is made up of these two sub-networks interacting in a multimodal layer. Four benchmark datasets were used to assess the model's effectiveness (IAPR TC-12 & Flickr 8K & Flickr 30K & MS COCO). Oriol Vinyals et al. [12] introduced a deep recurrent architecture-based generative model that creates natural descriptions of an image using advances in computer vision and machine translation. The model is trained to maximize the likelihood that the goal description language will appear given the training image. In such an approach, a CNN compresses a picture into a little representation, and an RNN creates the sentence that goes with it. A variety of datasets were used to evaluate that model (Pascal, Flickr8k, Flickr30k, MSCOCO, and SBU).

In 2015, Kelvin Xu et al. [13] introduced an attention-based model that automatically learns to describe the content of images. In the encoder stage, In order to extract a collection of feature vectors known as annotation vectors, this model used a CNN. and in the decoder stage, it used an LSTM to generate a caption. Flickr9k, Flickr30k, and MS COCO datasets were used to evaluate that system.

In 2016, Philip Kinghorn et al. [14] developed a hierarchical deep networks with the goal of producing highly detailed image descriptions. For image description generation, it combines multiple CNNs with specific types of RNNs such as LSTM and GRU. That proposal included several stages: Firstly, feature extraction from the generated region of interest, Then combining the features and word vectors to generate region attributes, And finally, generating object labels and scene labels using the re-added VGG layers. That model has been hierarchically trained on a variety of datasets from various domains. It was evaluated using the IAPR TC-12 dataset. The accuracy obtained was around 80%. To generate image descriptions, Ying Hua Tan & Chee Seng Chan [15] proposed an LSTM model with a phrase-based hierarchy. The proposed model encodes sentences as a sequence of phrases and words rather than a sequence of words alone as in previous approaches. That model was tested using the Flickr8k and Flickr30k datasets. Through a semantic attention model, combining the best of both worlds, Quanzeng You et al. [16] developed an innovative new technique to image captioning. One of the top-down visual features was extracted using a CNN in that model. Additionally, to combine the visual feature with visual ideas in an RNN and create the image caption, a semantic attention model is utilized. Visual ideas are detected by looking for areas, objects, characteristics, and other features. That method was tested using the Microsoft COCO dataset and the Flickr30K dataset.

In 2017, Liang Yang and Haifeng Hu [17] proposed a novel parallel RNN with time-varying parameters. This model combined two conventional CNNs (VggNet and Inception v3) to extract global image characteristics and an RNN to produce a time-varying feature at each time step, which was used to represent the current word. The strategy was evaluated using the Flickr8k, Flickr30k, and MSCOCO datasets. Xinwei He et al. [18] suggested a method for constructing photo captions that used POS tags to drive training and testing processes. The encoder was a VGG16 CNN, while the decoder was an LSTM network in that architecture. The model's performance was evaluated using two well-known and demanding datasets, MS COCO and Flickr30K.

In 2018, Aihong Yuan et al. [19] produced a model for the problem of image caption creation using a three-gated model that combined global and local image characteristics. The model had three gated buildings. First Gate, which can decide when and how much of the global picture feature to adaptively send into the phrase generator. The phrase generator, on the other hand, is a gated RNN. To enhance the capacity of RNN stacking for nonlinearity fitting, the gated feedback approach is adopted. The system was evaluated on the Flickr8K, Flickr30K, and MS COCO datasets. Philip Kinghorn et al. [20] with the aim of describing and annotating several objects and people in a single image, created a local-based deep learning architecture for image description generation. This model combines encoder-decoder RNNs for phrase production with object recognition and identification, scene categorization, attribute prediction using RNNs, and attribute-based attribute prediction. IAPR TC-12 dataset was used to assess it.

In 2019, Ying Hua Tan & Chee Seng Chan [21] created an LSTM model based on phrases for providing picture descriptions in a hierarchical manner, where NPs that characterize the image's most important items are created first, followed by a comprehensive caption constructed from the NPs. The CNN-encoded picture was used as context in this model's LSTM decoding of the image caption. We tested that method using the Flickr8k, Flickr30k, and MS-COCO datasets. Rehab Alahmadi et al. [22] introduced an encoder-decoder machine translation paradigm-based sequence-to-sequence model that uses RNN as an image encoder. The model receives a sequence of pictures as input that represent the objects in the image. This model uses RNN with CNN as an encoder instead of only using CNN. To model the encoder and decoder, a form of RNN was utilized, called LSTM which solves the gradient exploding and the vanishing problem in the regular RNN. There have been used three LSTMs, two as the encoder and one as the decoder. The Flickr30K dataset was utilized to evaluate this model. Priyanka Kalena et al. [23] developed a deep learning model that creates image captions automatically with the aim of not just explaining the scene but also helping those with visual impairments to better comprehend their surroundings. Within this model, CNN was utilized as a feature extraction model that converts an image into a vector representation, then an RNN decoder model that produces related phrases based on the image characteristics learnt. Flickr30K and MSCOCO datasets were utilized to evaluate the model. Xiangrong Zhang et al. [24] A model with an attribute attention mechanism was provided for the description construction of remote-sensing pictures. The remote sensing picture's description was generated using the LSTM, while its image characteristics were generated using the CNN. The datasets from Sydney, UCM, and RSICD were used to evaluate that model. Fen Xiao et al. [25] proposed a captioning algorithm that used an adaptive semantic attention model to integrate two separate LSTM networks. The encoder module in that model is CNN, LSTM A and LSTM L are the language generation and visual attention modules, respectively, that make up the decoder module. The effectiveness of the system was evaluated using two image-sentence datasets, Flickr30k and MSCOCO.

In 2020, Dicong Qiu et al. [26] suggested combining the responsibilities of onboard picture priority for data downlink and local image search for planetary image data servers to address the issue of image captioning and text similarity evaluation. This proposed approach included several stages, Firstly, feature extraction, then visual attention, and finally, text decoder. In order to evaluate that approach, the Martian Image Caption Dataset (MICD) was utilized. From experiments, the obtained accuracy was about 80%. Junbo Wang et al. [27] introduced an approach for image captioning with two novel elements: a context-aware attention mechanism and graph-based visual connection modeling. A GNN learns the implicit visual relationships among the visual objects or areas in an image, a deep CNN extracts image features, a GNN picks relevant relationship representations, an LSTM-based language model creates sentences, and a GNN learns implicit visual relationships. This method was tested using MS COCO and Flickr30K. Shiyang Yan et al. [28] proposed using a hierarchical attention strategy to improve performance by taking into consideration both detected object features and global picture features. The CNN encoder and object detector, respectively, extract global and local information in such a case. Global and local attention techniques are used to send both of these feature classes to the LSTM models. Words are decoded from the concatenated outputs of the two LSTM models. The MS COCO dataset was used to test that model. Shan Cao et al. [29] proposed that the IGGAN is a powerful cycle-consistent technique that employs object-object interactions and multi-scale feature representation to adversarially train the model for unsupervised picture captioning without the need for labeled image caption pairs. Feature extraction from object-object interactions, picture encoding, and cycle-consistent adversarial creation were the three main building blocks of this technique. This system was evaluated using the MS COCO dataset. Yan Chu et al. [30] combined ResNet50 and LSTM with software attention to create a single joint model for automatic image captioning. This design consists of a single encoder-decoder pair. ResNet50, a CNN-based encoder, embeds an image into a fixed-length vector in order to construct a complete representation of it. The decoder uses LSTMs and a soft attention mechanism to focus attention on specific areas of a picture in order to predict the following text. This system was validated using MS COCO. to improve the attention mechanism and the extraction of picture characteristics, Zhenrong Deng et al. [31] developed a model for picture captioning based on DenseNet and an attentional adaptive mechanism. DenseNet was used to extract the image's overall features throughout the model's stage of encoding. The LSTM network was employed as a language generation model for picture captioning tasks during the decoding phase to increase the quality of the generated captions while the adaptive attention mechanism simultaneously decides whether to use image feature information for word production. Using the COCO and Flickr30k datasets, the model was verified.

In 2021, Zongjian Zhang et al. [32] introduced a visual connection attention model based on parallel attention and learning spatial restrictions. A faster R-CNN serves as the model's image encoder, while a two-layer LSTM serves as the model's language decoder. Both attention models are present between the LSTM-2 (fine decoder) and the LSTM-1 (coarse decoder) (fine decoder). We tested this model using the MSCOCO dataset. Peng Tian et al. [33] For utilizing mutual connections and extracting context information between the three different semantic layers in order to jointly solve the three vision tasks of accurately and completely describing the scene image, a multi-level semantic context information network with a symmetrical overall structure has been proposed. By grouping and tying together the object regions, the relationship regions are produced. The features for the object, relationship, and caption region suggestions were extracted using the ROI-pooling approach. Following message transmission, the object, relationship, and caption area features are fed into the relationship context network and scene context network, respectively. Feature messages are then sent to the GGNN for passing and updating. After extracting relationship context data and updating the object's attributes, object and relationship detection was carried out using the modified object's attributes and relationship context data. Last but not least, the model feeds the obtained relationship and scene context data to an attention mechanism, which creates information and caption features that are then integrated with the decoder to create caption sentences. Jiesi Li et al. [34] introduced a multi-level similarity-guided semantic matching technique for image captioning that can learn the latent semantic connection between images and create captions through merging local and global semantic similarities. This technique comprised of visual semantic unit extraction and encoding using Faster R-CNN, as well as language decoding using LSTM and an attention mechanism. MS COCO dataset was utilized in the evaluation process. Yeong-Hwa Chang et al. [35] developed a more effective image captioning model that produces textual descriptions of pictures automatically and incorporates object identification, color analysis, and image captioning. The encoder in that model was VGG16, and the decoder was an LSTM network with attention. MS COCO was utilized to evaluate This model. Xian Zhong et al. [36] proposed an adaptive spatial information image captioning technique based on attention for retrieving a sequence of spatial data points about notable objects in a tiny region of a photograph or a whole photograph. In that technique, the encoding stage records both the object-level visual properties of the highlighted items and their spatial bounding box. The global feature maps of an entire image are acquired and combined with the local features prior to being transmitted to the LSTM-based language decoder. That model was evaluated using MS-COCO and Flickr30k. Priyanka Meel and Dinesh Kumar Vishwakarma [37] developed a method of detecting false news across several media by fusing hidden pattern extraction abilities from text with visual picture attributes via image captioning and forensic analysis. This system focused on four techniques for multimodal data analysis in specifically, the HAN deep model for text, producing picture captions and headlines that correlate to news text, noise variance inconsistency, and error level analysis. The Fake News Samples dataset was used to test that model and the obtained accuracy was 95.90%. Emre Boran et al. [38] a dense video captioning approach that incorporates image captions as an additional input in addition to video data. In that way, it is helpful to take advantage of the image captions' additional diversity and richness in order to generate more coherent descriptions for the videos. The ActivityNet Captions video captioning dataset was used to evaluate that model. The ActivityNet Captions dataset contains 10 000 videos for training and 4.9000 videos for validation. Imad Afyouni et al.[39] presented a hybrid object-based, attention-enriched image captioning architecture with an emphasis on the Arabic language. CNN was used to extract image features, and LSTM was used to generate captions in that model. The MSCOCO and Flickr30k datasets were used to train and then tested that method by creating an Arabic version of a subset of the COCO dataset and the obtained accuracy was greater than 60%.

Performance Analysis

A comparison was made among different image caption approaches based on deep learning models by using BLEU, METEOR, and accuracy measurements, as shown in Table I.

Most of these approaches utilize similarity-based measures between ground truth and machine-generated sentences. The score of BLEU can be utilized as an evaluation measure, in spite of possessing some shortcomings, this standard measure is commonly utilized in the tasks of machine translation. BLEU provides *m*-gram precision between reference and candidate sentences, and the greater m indicates a more suitable understanding at the sentence level instead of a greater similarity between words within the sentence. BLEU measure is given as follows [40]:

$$BLEU = B_p \times \exp(\textstyle\sum_{m=1}^{M} w_m log\, p_m) \qquad (10)$$

$$B_p = \begin{cases} e^{(-\frac{r}{c}+1)} \text{ ، } if\ c \le r \\ 1 \text{ ، } if\ c > r \end{cases}$$

Where p_m indicates the adjusted *m*-gram precision, w_m indicates the uniform weights which is equal to $1/M$ (M equal to 4), B_p indicates the brevity penalty, and c and r indicate the length of the candidate and efficient reference corpus, respectively.

The METEOR measure specifies the whole matches between sentences using specific criteria of matching, like paraphrase, synonym, and exact word matching. The score of METEOR is given as follows [41]:

$$METEOR = \frac{10U_P U_R}{U_R + 9U_P} \times \left(1 - 0.5\left(\frac{K}{Um}\right)\right) \qquad (11)$$

Where U_P indicates the precision of the unigram which is calculated as the proportion of the number of unigrams in the mapped candidate translation to the entire number of unigrams in the candidate translation, U_R indicates the recall of the unigram which is calculated as the proportion of the number of unigrams in the mapped candidate translation to the entire number of unigrams in the reference translation, K indicates the chunks obtained from grouping the whole unigrams in the mapped candidate translation, and Um indicates the unigram matches.

Table 1: A comparison between various image caption approaches based on deep learning models

			BLEU				
Author/(s); Ref; Year	**Deep Learning Models**	**Datasets**	***B-1***	***B-2***	***B-3***	***B-4***	**METEOR**
Junhua Mao et al.[10]; 2014	CNN; RNN	IAPR TC12	0.482	0.357	0.269	0.208	-
		Flickr 8K	0.565	0.386	0.256	0.170	-
		Flickr 30K	0.54	0.36	0.23	0.15	-
		MS COCO	0.668	0.488	0.342	0.239	0.221
Oriol Vinyals et al. [11]; 2014	CNN; RNN	Pascal	0.59	-	-	-	-
		Flickr8k	0. 63	-	-	-	-
		Flickr30k	0.66	-	-	-	-
		MSCOCO	-	-	-	0.277	0.237
		SBU	0.28	-	-	-	-
Junhua Mao et al.[12]; 2014	CNN; RNN	IAPR TC12	0.3951	0.1828	0.1311	-	-
		Flickr 8K;	0.5778	0.2751	0.2307	-	-
		Flickr 30K	0.5479	0.2392	0.1952	-	-
Kelvin Xu et al. [13]; 2015	CNN; LSTM	Flickr8k	0. 67	0. 457	0. 314	0.213	0. 2030
		Flickr30k	0. 669	0.439	0. 296	0. 199	0. 1846
		MS COCO	0.718	0.504	0.357	0.250	0.2304
Philip Kinghorn et al.[14]; 2016	CNN; RNN	IAPR TC-12	0.201	0.105	0.053	0.024	0.073
Ying Hua Tan and Chee Seng Chan [15]; 2016	CNN; LSTM	Flickr8k	0.636	0.436	0.276	0.166	-
		Flickr30k	0.666	0.458	0.282	0.170	-
Quanzeng You et al. [16]; 2016	CNN; RNN	Microsoft COCO	0.910	0.786	0.654	0.534	0.341
		Flickr30K	0.824	0.679	0.534	0.412	0.269
Liang Yang and Haifeng Hu [17]; 2017	CNN ; RNN	Flickr8k	0.591	0.399	0.267	0.171	0.178
		Flickr30k	0.607	0.417	0.298	0.188	0.179
		MSCOCO	-	-	-	-	0.234
Xinwei He et al.[18]; 2017	CNN ; RNN	Flickr30K	0.638	0.446	0.307	0.211	-
		MS COCO	0.711	0.535	0.388	0.279	0.239
Aihong Yuan et al.[19]; 2018	CNN; RNN	Flickr8K	0.699	0.485	0.344	0.235	0.223
		Flickr30K MS COCO	0.694	0.457	0.332	0.226	0.230
			0.719	0.529	0.387	0.284	0.243
Philip Kinghorn et al. [20]; 2018	CNN; RNN	IAPR TC-12	0.231	0.099	0.046	0.024	0.067
Ying Hua Tan and Chee Seng Chan [21]; 2019	CNN; LSTM	Flickr8k	0.670	0.457	0.314	0.213	0.203
		Flickr30k	0.669	0.439	0.296	0.199	0.185
		MS-COCO	0.718	0.504	0.357	0.250	0.230
Rehab Alahmadi et al. [22]; 2019	CNN; LSTM	Flickr 30K	0.675	0.467	0.313	0.207	0.190
Priyanka Kalena et al. [23]; 2019	CNN; RNN	Flickr30k	-	-	-	-	-
		MSCOCO					
	CNN; LSTM	Sydney	0.8143	0.7351	0.6586	0.5806	0.4111

Xiangrong Zhang et al. [24]; 2019		UCM	0.8154	0.7575	0.6936	0.6458	0.4240
		RSICD	0.7571	0.6336	0.5385	0.4612	0.3513
Fen Xiao et al. [25]; 2019	CNN; LSTM	Flickr30k	0.686	0.507	0.368	0.266	0.215
		MSCOCO	0.758	0.594	0.455	0.346	0.271
Dicong Qiu et al. [26]; 2020	CNN; RNN	MICD	0.547	0.482	0.431	0.388	0.309
Junbo Wang et al. [27]; 2020	CNN; GNN; LSTM	MS COCO	0.759	0.603	0.465	0.358	0.278
		Flickr30K	0.698	0.517	0.378	0.277	0.215
Shiyang Yan et al. [28]; 2020	CNN ; LSTM	MSCOCO	0.73036	0.53688	0.39069	0.28551	0.25324
Shan Cao et al.[29]; 2020	RMCNet; LSTM	MS COCO	-	-	-	21.9	21.1
Yan Chu et al.[30]; 2020	ResNet50; LSTM	MS COCO 2014	-	-	-	0.326	0.261
Zhenrong Deng et al. [31]; 2020	DenseNet; LSTM	Flickr30k	0.667	0.486	0.321	0.224	0.214
		COCO	0.739	0.570	0.422	0.326	0.270
Zongjian Zhang et al. [32]; 2021	Faster R-CNN; LSTM	MSCOCO	0.788	0.614	0.472	0.363	0.279
peng Tian et al.[33]; 2021	Faster R-CNN; GGNN ; LSTM	MSCOCO	0.705	0.503	0.374	0.294	0.238
Jiesi Li et al. [34]; 2021	Faster R-CNN; LSTM	MSCOCO	0.812	-	-	0.390	0.285
Yeong-Hwa Chang et al.[35]; 2021	VGG16; LSTM	MS COCO	-	-	-	-	-
Xian Zhong et al. [36]; 2021	R-CNN; ResNet; LSTM	MS-COCO	0.785	0.622	0.485	0.378	0.277
		Flickr30k	0.732	0.532	0.375	0.277	0.221
Priyanka Meel and Dinesh Kumar Vishwakarma. [37]; 2021	(HAN)	Fake News Samples	-	-	-	-	-
Emre Boran et al.[38]; 2021	ResNet; LSTM	ActivityNet	-	-	-	0.888	0.1446
Imad Afyouni et al.[39]; 2021	CNN; LSTM	MS COCO Flickr30k	-	-	-	-	-

In Table 2, Nine datasets are demonstrated in detail, the bulk of which contains over 10,000 images. These datasets may be viewed as a continual endeavor on the part of researchers to offer the massive quantities of diverse data required for the most advanced deep-learning neural networks.

Table 2: A comparison between various benchmark datasets of image captioning

Dataset	Ref.	No. Image	Description

IAPR TC-12	Guillaumin et al. [42]	20,000 Images	Features photographs of various sports and activities, people, animals, cities, landscapes, and so on.
Flickr 8K	Hodosh et al. [43]	8,000 images	Humans and animals are the most common subjects in that dataset.
Flickr 30K	Peter Young et al.[44]	31,783 Images	Mostly about people going about their daily lives and events.
MSCOCO	Xinlei Chen et al.[45]	123,287 Images	To produce that project, images of complicated everyday scenarios with ordinary things in their natural contexts were gathered.
Sydney	Bo Qu et al.[46] , Fan Zhang et al. [47]	613 Images	The Sydney Dataset has seven classes: residential, airport, grassland, rivers, ocean, industrial, and runway.
UCM	Bo Qu et al.[46] , Yi Yang et al. [48]	2,100 Images	Agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, highway, golf course, harbour, junction, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, and tennis court are among the 21 classifications offered.
RSICD	Xiaoqiang Lu et al.[49]	10,921 Remote Sensing Images	The photographs in the collection were collected from Google Earth and scaled to 224*224 pixels at various resolutions.
MICD	Dicong Qiu et al. [26]	12.500 Images	Capture Martian geologic characteristics, particularly landscape features, in images that contain several item or feature categories.
VRD	Cewu Lu et al.[50]	5000 Images	Images can be used to capture a wide range of interactions between pairs of things (e.g. "man riding bicycle" and "man pushing bicycle").

Conclusions

A crucial and essential area of artificial intelligence that combines computer vision and natural language processing is picture captioning, or the process of automatically creating descriptions for images. Image captioning not only provides descriptive information for multimedia content, but it also aids in the finding of patterns, trends, and significant events. This brief assessment looked at the most prominent papers on image caption creation challenges using deep learning models published in the last decade, as well as the most extensively used related datasets, all of which are based on effective deep learning models, mainly the CNN model and LSTMs.

Extraction of knowledge and semantic information from multimodal data allows for the development of a wide range of applications, including security, law enforcement, and social media.

Acknowledgment

We want to thank the Department of Computer Science, College of Science, University of Diyala for giving sufficient funding and support for our research.

References

1. Y. H. Tan, C. S. Chan, Phrase-based image caption generator with hierarchical LSTM network, Neurocomputing, 333, 86-100(2019)
2. S. Bai, S. An, A survey on automatic image caption generation, Neurocomputing, 311, 291-304(2018)
3. X. He, B. Shi, X. Bai, G. S. Xia, Z. Zhang, W. Dong, Image Caption Generation with Part of Speech Guidance, Pattern Recognition Letters, 119, 229–237(2019)
4. M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, R. Cucchiara, From Show to Tell: A Survey on Deep Learning-based Image Captioning, 2021. arXiv:2107.06912 [cs.CV].
5. P. Tian, H. Mo, L. Jiang, Image caption generation using multi-level semantic context information, Symmetry, 13(7), (2021)
6. Y. H. Chang, Y. J. Chen, R. H. Huang, Y. T. Yu, Enhanced Image Captioning with Color Recognition Using Deep Learning Methods, Applied Sciences, 12(1), 209(2021)
7. J. Waleed, S. Albawi, H. Q. Flayyih, A. Alkhayyat, An Effective and Accurate CNN Model for Detecting Tomato Leaves Diseases, In: 2021 4th International Iraqi Conference on Engineering Technology and Their Applications (IICETA), 33-37(2021)
8. R. Zhao, R. Yan, Z. Chen, K. Mao, P. Wang, Robert X. G., Deep learning and its applications to machine health monitoring, Mechanical Systems and Signal Processing, 115, 213-237(2019)

9. A. Mosavi, S. Faizollahzadeh ardabili, A. R. Várkonyi-Kóczy, List of Deep Learning Models, Preprints, (2019)

10. H. Q. Flayyih, J. Waleed and S. Albawi, ASystematic Mapping Study on Brain Tumors Recognition Based on Machine Learning Algorithms, In: 2020 3rd International Conference on Engineering Technology and its Applications (IICETA), 2020, 191-197.

11. J. Mao, W. Xu, Y. Yang, J, Wang, Z. Huang, A. Yuille, Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN), (2014)

12. O. Vinyals Google, A. Toshev Google, S. Bengio Google, D. Erhan Google, Show and Tell: A Neural Image Caption Generator, Computer Vision and Pattern Recognition, (2015)

13. K. Xu, J. Lei Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, Y. Bengio, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, (2015)

14. P. Kinghorn, L. Zhang, L. Shao, A hierarchical and regional deep learning architecture for image description generation, Pattern Recognition Letters, 119, 77-85(2019)

15. Y. H. Tan, C. S. Chan, phi-LSTM: A Phrase-based Hierarchical LSTM Model for Image Captioning, (2016)

16. Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image Captioning with Semantic Attention, In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 4651-4659(2016)

17. L. Yang, H. Hu, TVPRNN for image caption generation, Electronics Letters, 53(22), 1471-1473(2017)

18. X. He, B. Shi, X. Bai, G. S. Xia, Z. Zhang, W. Dong, Image Caption Generation with Part of Speech Guidance, Pattern Recognition Letters, 119, 229-237(2019)

19. A. Yuan, X. Li, X. Lu, 3G structure for image caption generation, Neurocomputing, 330, 17-28(2019)

20. P. Kinghorn, L. Zhang, L. Shao, A region-based image caption generator with refined descriptions, Neurocomputing, 272, 416-424(2018)

21. Y. H. Tan, C. S. Chan, Phrase-based image caption generator with hierarchical LSTM network, Neurocomputing, 333, 86-100(2019)

22. R. Alahmadi, C. H. Park, J. Hahn, Sequence-to-sequence image caption generator, Proc. SPIE 11041, In: Eleventh International Conference on Machine Vision (ICMV 2018), 110410C, March 2019.

23. G. Sharma, P. Kalena, N. Malde, A. Nair, S. Parkar, Visual Image Caption Generator Using Deep Learning, In: 2nd International Conference on Advances in Science & Technology (ICAST-2019), (2019)

24. X. Zhang, X. Wang, X. Tang, H. Zhou, C. Li, Description generation for remote sensing images using attribute attention mechanism, Remote Sensing, 11(6), (2019)

25. F. Xiao, X. Gong, Y. Zhang, Y. Shen, J. Li, X. Gao, DAA: Dual LSTMs with adaptive attention for image captioning, Neurocomputing, 364, 322-329(2019)

26. D. Qiu, B. Rothrock, T. Islam, A. K. Didier, V. Z. Sun, C. A. Mattmann, M. Ono, SCOTI: Science Captioning of Terrain Images for data prioritization and local image search, Planetary and Space Science, 188, (2020)

27. J. Wang, W. Wang, L. Wang, Z. Wang, D. D. Feng, T. Tan, Learning visual relationships and context-aware attention for image captioning, Pattern Recognition, 98, (2020)

28. S. Yan, Y. Xie, F. Wu, J. S. Smith, W. Lu, B. Zhang, Image captioning via hierarchical attention mechanism and policy gradient optimization, Signal Processing, 167, (2020)

29. S. Cao, G. An, Z. Zheng, Q. Ruan, Interactions Guided Generative Adversarial Network for unsupervised image captioning, Neurocomputing, 417, 419-431, (2020)

30. Y. Chu, X. Yue, L. Yu, M. Sergei, Z. Wang, Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention, Wireless Communications and Mobile Computing, (2020)

31. Z. Deng, Z. Jiang, R. Lan, W. Huang, X. Luo, Image captioning using DenseNet network and adaptive attention, Signal Processing: Image Communication, 85, (2020)

32. Z. Zhang, Q. Wu, Y. Wang, F. Chen, Exploring region relationships implicitly: Image captioning with visual relationship attention, Image and Vision Computing, 109, (2021)

33. P. Tian, H. Mo, L. Jiang, Image caption generation using multi-level semantic context information, Symmetry, 13(7), (2021)

34. J. Li, N. Xu, W. Nie, S. Zhang, Image Captioning with multi-level similarity-guided semantic matching, Visual Informatics, 5(4), 41-48(2021)

35. Y. H. Chang, Y. J. Chen, R. H. Huang, Y. T. Yu, Enhanced Image Captioning with Color Recognition Using Deep Learning Methods, Applied Sciences, 12(1), 209(2021)

36. X. Zhong, G. Nie, W. Huang, W. Liu, B. Ma, C. W. Lin, Attention-Guided Image Captioning With Adaptive Global and Local Feature Fusion, Journal of Visual Communication and Image Representation, 78, (2021)

37. P. Meel, D. K. Vishwakarma, HAN, Image Captioning, and Forensics Ensemble Multimodal Fake News Detection, Information Sciences, 567, 23-41(2021)

38. E. Boran, A. Erdem, N. Ikizler-Cinbis, E. Erdem, P. Madhyastha, L. Specia, Leveraging auxiliary image descriptions for dense video captioning, Pattern Recognition Letters, 146, 70-76(2021)

39. I. Afyouni, I. Azhar, A. Elnagar, AraCap: A hybrid deep learning architecture for Arabic Image Captioning, Procedia CIRP, 189, 382-389(2021)

40. K. Papineni, S. Roukos, T. Ward, W.J. Zhu, BLEU: a method for automatic evaluation of machine translation, 40th Annual Meeting on Association for Computational Linguistics, ACL, 311-318(2002)

41. S. Banerjee, A. Lavie, METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 65-72(2005)

42. M. Guillaumin, J. Verbeek, C. Schmid, Multiple Instance Metric Learning From Automatically Labeled Bags of Faces, In ECCV, 634-647(2010)

43. M. Hodosh , P. Young , J. Hockenmaier, Framing Image Description as A Ranking Task: Data, Models and Evaluation Metrics, Journal Artificial Intelligent Research, 47, 853-899(2013)

44. P. Young , A. Lai , M. Hodosh , J. Hockenmaier, From Image Descriptions To Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions, In: Proceedings of the Meeting on Association for Computational Linguistics, 67-78(2014)

45. X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollar, C. Zitnick, Microsoft COCO Captions: Data Collection and Evaluation Server, 2015

46. B. Qu, X. Li, D. Tao, X. Lu, Deep Semantic Understanding of High-Resolution Remote Sensing Image, In: Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (CITS), Kunming, China, 1-5(2016)

47. F. Zhang, B. Du, L. Zhang, Saliency-Guided Unsupervised Feature Learning for Scene Classification, In: IEEE Transactions on Geoscience and Remote Sensing, 53(4), 2175-2184(2015)

48. Y. Yang, S. Newsam, Bag-of-Visual-Words And Spatial Extensions For Land-Use Classification, In: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 270-279(2010)

49. X. Lu, B. Wang, X. Zheng, X. Li, Exploring Models and Data for Remote Sensing Image Caption Generation, IEEE Trans. Geosci. Remote Sens. 56, 2183-2195(2018)

50. C. Lu, R. Krishna, M. Bernstein, Visual Relationship Detection with Language Priors, In: Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016, 852–869