DOI: 10.24996/ijs.2024.65.7.38

ISSN: 0067-2904

A Survey on Image Caption Generation in Various Languages

Haneen Siraj Ibrahim*, Narjis Mezaal Shati

Department of Computer Science, College of Sciences, Mustansiriyah University, Baghdad, Iraq

Received: 19/2/2023 Accepted: 30/6/2023 Published: 30/7/2024

Abstract

The image caption is the process of adding an explicit, coherent description to the contents of the image. This is done by using the latest deep learning techniques, which include computer vision and natural language processing, to understand the contents of the image and give it an appropriate caption. Multiple datasets suitable for many applications have been proposed. The biggest challenge for researchers with natural language processing is that the datasets are incompatible with all languages. The researchers worked on translating the most famous English data sets with Google Translate to understand the content of the images in their mother tongue. In this paper, the proposed review aims to enhance the understanding of image captioning strategies and to survey previous research related to image captioning while examining the most popular databases in different languages, mostly English, translating into other languages using the latest models for describing images, summarizing evaluation measures, and comparing them.

Keywords: CNN, Computer Vision, Image Captioning, LSTM, Natural Language Processing.

استطلاع لتوليد التسميات التوضيحية للصور بلغات مختلفة

حنين سراج ابراهيم *, نرجس مزعل شاتي

قسم علوم الحاسوب، كلية العلوم، جامعة المستنصرية، بغداد، العراق

الخلاصة

التسمية التوضيحية للصورة هي عملية إضافة وصف واضح ومتماسك لمحتويات الصورة. يتم ذلك باستعمال أحدث تقنيات التعلم العميق ويتضمن رؤية الحاسوب مع معالجة اللغة الطبيعية. لفهم محتويات الصورة وإعطاء التسمية التوضيحية المناسبة لها. تم اقتراح مجموعات بيانات متعددة مناسبة للعديد من التطبيقات. يتمثل التحدي الأكبر للباحثين في معالجة اللغة الطبيعية في أن مجموعات البيانات غير متوافقة مع جميع اللغات. عمل الباحثون على ترجمة مجموعات البيانات الإنجليزية الأكثر شهرة باستخدام Google Translate لفهم محتوى الصورة بلغتهم الأم. في هذه الورقة ، تهدف المراجعة المقترحة إلى تعزيز فهم استراتيجيات التسميات التوضيحية للصور ومسح الأبحاث السابقة المتعلقة بتسميات الصور أثناء فحص قواعد

*Email: haneenserag9@uomustansiriyah.edu.iq

البيانات الأكثر شيوعًا بلغات مختلفة ، ومعظمها باللغة الإنجليزية مترجمة إلى لغات أخرى باستعمال أحدث
النماذج لوصف الصور ، تلخيص مقاييس التقييم ومقارنتها.

1. Introduction

In our electronic age, we deal with visual images or videos largely every day. It facilitates inter-person communication, news search, and information sharing. The percentage of social media usage in life is**:** Facebook (68% monthly and 45% daily), YouTube (63% monthly and 21% daily), Twitter (12% monthly and 3% daily), Instagram (16% monthly and 6% daily), and Snapchat (6% monthly and 2% daily) [1]. We need to create a caption for visuals in general. The importance of the topic appears in many areas, the most important of which are people who suffer from visual impairment, especially color blindness. My point of view is that image caption systems give information about colors, for example; a man wears an orange hat and glasses from the Flickr8k database. It is also important for indexing and classifying images by title [2]. Also, in answering visual questions [3].

An image caption generator is a description of the content of the image, and it needs one of the methods of deep learning to do this difficult work that brings the computer closer to the human view of things and their interpretation. The human brain can receive the image and understand it easily, and the computer can do this by using convolutional neural networks (CNN) and generating natural language. Text pre-processing is the most important stage in building the image captioning model, where the text data spoken by the human is easily converted to a format in some form so that the machine can understand it, and the language varies from one country to another; for example, the English language needs to convert uppercase letters to lowercase letters, while the Arabic language does not contain uppercase or lowercase letters, and the form of the word is connected letters and also contains diacritics. The Arabic language needs to remove diacritics, prefixes, and suffixes. Researchers have been interested in the topic of image caption generators and needed an adequate dataset that simulates various natural languages. The English language was achieving success and high accuracy compared to the rest of the languages [4].

CNN's strengths are in the areas of image processing and pattern recognition. Deep learning methods, particularly CNN technology, have produced outstanding accuracy rates in the field of face recognition in recent years [5]. and the power of neural convolutional networks in image retrieval [6] The image captioning generator is built by building an intricate neural network model using convolutional neural networks (CNN) and is tested on and trained on a different dataset. The datasets differ in terms of language differences, number of images, and number of captions for each image.

This paper shows a survey of previous research related to image captions, an explanation of the techniques used in various languages, and a dataset used to address the most prominent problems of image caption searches, which is a database that is not compatible with all languages and Summarizing the most popular evaluation metrics in the field of image captioning and comparing them in terms of the most used The well-known image captioning datasets used in this study include: Arabic Flickr8k [7], English Flickr8k [8], Flickr30k [9], Flickrstyle10k [10], MS Coco [11], India Visual Genome [12], Visual Genome [13], Image Paragraph Captioning [14], #Pracegover [15], and VizWiz-Captions [16] are discussed.

2. Related Work

Image caption systems primarily aim to describe and understand visual objects and convert them into sentences and words in multiple languages that suit human understanding. Earlier studies have suggested a variety of concepts and strategies. No matter the techniques, several previous similar works are selected.

Al-Muzaini 2018 [22] presented two new datasets in Arabic depending on Flickr8K and MS COCO: the first containing 5358 captions and 1176 photos by means of a crowd-flower, and the second containing 150 images and 750 captions from human-translated descriptions using RNN. A larger dataset would yield positive outcomes, according to the review model's score of BLEU-1 = 46.2. Jindal [21] used other models (not RNN) for creating Arabic image captions. With a larger dataset, the suggested model performs better. The image dataset with Arabic captions will therefore be enlarged and made available to the public in order to support future research.

Rahman et al. 2019 [19] introduced a captioning system for the images named "Chittron" in the Bangla language by creating a dataset containing 16k images with manual captioning in the Bangla language. They used the deep learning network VGG16 to obtain the features of the images and the LSTM network, and they showed weaknesses in the BLEU scores because the data set contains one caption for each image. To improve this in the future, working on a large and diverse data set with more than one caption for each image is required.

ElJundi et al. 2020 [7] developed a complete model for Arabic Image Captioning (AIC) using VGG16 to extract features from the image, and the LSTM natural language model added a fresh open dataset for AIC as well. They discovered that translating captions to Arabic from datasets in English based on models created from those datasets was less effective than creating captions directly from an Arabic Flicker8k dataset, which improved the BLEU-1 = 33. The results in Arabic image captions are lower compared to English because of the complexity of the Arabic language and the small database. They suggested improvement mechanisms in the future to increase the size of the database, and vocalization has been added to the Arabic training set.

Wang, J. 2020 [23], introduced a new technique for captioning images that makes use of links between graph neural networks, visual regions, and a context-sensitive attention mechanism. The model's competitive advantage was its ability to remember prior visual content. According to the authors, the model can outperform cutting-edge attention-based techniques after being tested and trained on the Flickr30K and MS-COCO datasets. The model scored BLEU-1 = 74.0. They introduced the processing of both the specific visible things in the image as well as the implicit visual relationship between the visible elements in the image. The mechanism for future optimization is to incorporate explicit visual relationships into our methodology. Our suggested approach can also be used for additional vision-to-language tasks, like visual dialogue and visual question-answering.

Gawde et al. 2020 [24] used CNN and LSTM models with a variety of hyperparameters that identify the image's features, then map them to the appropriate parameters to make dynamic and relevant captions and hashtags for social media for the classified image. Descriptive keywords are employed. They applied the findings to a sizable photo dataset, and after that, they used the Flickr8K and Flickr30K datasets as well as their own collection of several image categories to generate captions with greater result accuracy (BLEU-1 = 75.39). The model provides accurate results for external images that represent nature, such as sunsets and the sea, without humans. In the future, additional comparisons to strategies that take into account visual and temporal attention are possible. For consumers, we also plan to develop an Android app.

Sabri 2021 [4] introduced an entirely new transformer-based architecture model for processing using an improved word processing pipeline segmented to mitigate a portion of the complexity of Arabic morphology using newer image models such as EfficientNet and MobileNetV2, including attention mechanisms. Two datasets are commonly used by researchers to train and measure image suspension models. The first data set is Objects Shared in Context (COCO), and the second data set is Arabic Flicker8k, which improved the BLEU-1 = 44.3 and the BLEU-4 = 15.6.

The performance of LSTM/GRU-based models was faster compared to the performance of models based on transformers and more accurate. It clarified the mechanism for improving the transformer-based model in the future by inventing more complex architectures, showed the need for Arabic image caption searches in a strong database such as the COCO dataset, and decreased some of the morphological complexity by using a new pre-processing method and including methods that weren't previously used for picture captioning models.

Mishra et al. 2021 [26] manually translated the well-known MSCOCO dataset from English to Hindi and used the encoder RESNet-101 and the decoder GRU. Their model scored BLEU-1 = 67.0. The model's weakness is tweeting on the image comprehensively for the exact details of the situation or any disaster. In the future, expand the work to create more than one caption for each dense image.

Hejazi 2022 [17] and [18] created an Arabic image caption model using the Arabic Flickr8K dataset and tested 32 combinations of factors affecting the creation of captions using deep learning techniques for pre-processing (GRU and LSTM) and feature extraction (VGG16, Inception V3). The results showed the values of BLEUs 1 through 4 are 36.5, 21.4, 12, and 6, respectively. It showed the best results. The researcher faced limitations due to the small size of the database available in Arabic. Developing a solution mechanism in the future will require increasing the size of the database or including other methods of pre-processing and the latest methods of deep learning.

Lasheen and Barakat 2022 [20] proposed Arabic image captioning using an effective deep learning model that relies on the architecture of the encoder-decoder, RESNet-101 in an encoder and LSTM in a decoder, using the Arabic Flickr8k dataset for training. Back-propagation has been used to develop soft attention in a comprehensive approach, and they got results of BLEU-1 = 58.708, BLEU-2 = 46.523, BLEU-3 = 35.712, and BLEU-4 = 27.12. They did not improve the utilization of transformers for larger training data sets from the BLEU-N results. They explained the mechanism for improvement in the future. More effective models may be looked at to enhance outcomes in this field using generative adversarial networks. To deal with morphologically complex languages like Arabic, new text pre-processing techniques and extra evaluation techniques are being suggested.

Emami et al. 2022 [21] developed an encoder-decoder architecture (CNN and RNN) model and evaluated several Arabic image captions using GigaBERT and AraBERT as pre-trained models. They conducted training with two public datasets in Arabic (COCO and Flickr8k). Regarding the image caption standard, the model-scored results for BLEUs 1 through 4 are 0.39, 0.25, 0.15, and 0.092, respectively. They explained some ways in the future to create a strong and rich Arabic database through translation and verification, similar to the COCO database, and make it available to the public.

Tiwari (2022) [25] built an image caption model using CNN and RNN and trained it on two datasets, MSCOCO and Stanford-Paragraph. He achieved promising results: BLEU-1 = 37.74. Weakness in the results for flat caption models compared to hierarchical caption models. In the future, the attention learning approach could be combined with flat and hierarchical captioning models to improve results.
Mishra et al. 2021 [26] manually translated the well-known MSCOCO dataset from English to Hindi. They used the RESNet-101 encoder and the GRU decoder. The model scored a result of BLEU-1 = 67.0. Weaknesses are tweeting on the image comprehensively for the exact details of the situation or any disaster. In the future, it will expand the work to create more than one caption for each dense image.

Table 1 shows a summary of the literature. Survey containing Five columns are used: the researcher's first column lists the approach or algorithm used, publication year, and citations; the second, third, and fourth columns represent the language and techniques used and list the datasets used for testing and training the results according to the criteria used in each research; and the last column lists some drawbacks.

Table 1: Review of the literature.

Author(s), Year, reference,	**Language**	**Main Techniques**	**Dataset**	**results**	**Drawbacks**
HEJAZI, H.D. 2022.[17]	• AR	• (LSTM, GRU, dropout, (Inception V3, VGG16)for features extraction.	• Arabic Fliker8K	• BLEU 1 = 36.5, BLEU 2 = 21.4, BLEU 3 = 12, BLEU 4 = 6.6	• Since AIC only has one publicly accessible Dataset, the drawback was the short Dataset size.
• Sabri, S.M. (2021).[4]	• AR	• (LSTM-based model, GRU-based model, Transformer-based model, Use newer pre-trained models (Efficient Net And MobileNetV2) instead of VGG19	• Arabic Fliker8K	• BLEU 1= 44.3, BLEU 4 = 15.6	• The performance of LSTM/GRU- based models was faster compared to the performance of models based on Transformer more accurate
• ElJundi et al. 2020[7]	• `AR	• CNN, LSTM, Neural Machine Translation (NMT) For dataset Translation,	• Arabic Fliker8K	• BLEU 1= 33 , BLEU 2=6	• The results in Arabic image captions are less compared to English because of the complexity of the Arabic language and the small database.
• Lasheen, M. T., & Barakat, N. H. 2022[20]	• AR	• Encoder-Decoder architecture Encoder: The RESNet-101, The decoder: (LSTM)	• Arabic Flickr8k	• BLEU 1= 58.708 BLEU 2= 46.523 BLEU 3= 35.712 and BLEU 4=	• did not improve the utilization of transformers or larger training data sets from the BLEU-N results

• Emami 2022[21]	• AR	• Encoder-decoder architectures (CNN and RNN), OSCAR	• Arabic-COCO and Flickr8k.	27.12 • BLEU 1= 0.39 BLEU 2= 0.25 BLEU 3= 0.15 BLEU 4= 0.092	• found some problems with AIC , including the lack of a large, balanced, and well-explained database
• Al-Muzaini 2018[22]	• AR	(CNN, RNN LSTM)	• Arabic (Flickr8K, MS COCO)	• BLEU 1=46.2	• Except Jindal's paper [21], is not used (RNN) other models for creating Arabic image captions.
• Matiur Rahmana 2019[19]	• Bangla language	• (VGG16), (LSTM),	• BanglaLekha-ImageCaptions data set	• -	• weaknesses in the BLEU scores because the data set contains one caption for each image
• Mishra, S. K 2022[12]	• Hindi	• Faster (R-CNN, LSTM, GRU)	• Hindi Genome dataset	• BLEU 1= 35.77 BLEU 2=17.96 BLEU 3=9.81 BLEU 4=5.83	• A blockage happens when things in photos are not clearly visible or when several objects in an image overlap, and the suggested model is unable to address this issue. The model combines non-dominant characteristics with dominant features when some features—in this case, pixel values—make up the majority.
• Mishra, S. K 2021[26]	• Hindi	• Encoder: The RESNet-101, The decoder: GRU	• Hindi MS COCO	• BLEU 1= 67.0	• Weaknesses are tweeting on the image comprehensively for the exact details of the situation or any disaster
• Wang, J. (2020)[23]	• EN	• GNN, LSTM	• MS COCO , Flickr30K	• BLEU 1= 74,0	• The suggested solutions take into account what has already been addressed and address both the specific visible things in the image as well as the implicit visual relationship between the visible elements in the image.
• RishikeshGawde 2020[24]	• EN	• GNN, LSTM	• Flickr30K Flickr8K	• BLEU 1= 75.39	• The model provides accurate results for external images that represent nature, such as sunsets and the sea, without humans
• Tiwari, A. (2022)[25]	• EN	• (CNN, RNN)	• MSCOCO, Stanford-Paragraph dataset,	• BLEU 1=37.74	• The model was evaluated using CIDEr -D, Meteor, and BLEU -1,2,3,4 and revealed

		Genome	minute differences between (LSTM) and methods GRU and did not find any difference between the two models. This work did not achieve good results compared with the hierarchical models it is based upon

3. Image Captioning Generation

The objective of image captioning is to produce a natural statement that explains the content of the image in as many languages as possible. The researchers used LSTM and CNN models to develop a model while focusing on natural language processing and computer vision. This area is important in text detection in natural images too [27]. CNN serves as an attribute extractor for an encoder from images, and LSTM serves as a decoder to produce words that characterize the image in order using a pre-trained model on the image net dataset ([28] and [29]).

See an example in Figure 1. One of the deep learning models uses the image as an input. In this model, convolutional, pooling, and fully connected layers make up a typical CNN, with each layer's output being a function output from the layers before it. The output needs to be resistant to semantically irrelevant changes in order to be employed in picture captioning. The power of CNN in feature extraction has been the subject of several recent works. After removing the fully connected layer, the features are taken from the final layer and supplied into the captioning model, as in the works in [18] and [4].

The best pre-trained (CNN) models used in [18] and [4] are:

- VGG16
- VGG19
- Inception v3
- ResNet50
- Efficient Net, etc.

Utilize recurrent neural networks (RNN) and deep learning networks to perform tasks like text synthesis, voice recognition, and handwriting-to-text conversion (translation, image captioning, summary, etc.) that require the input of arbitrary-length sequences and the output of a set length. RNN has the ability to take input sequences of arbitrary length without causing a model's size to increase. Prediction takes into account the input history sequence (memory), and the same weights are applied to the input sequence throughout time. However, it still retains a short-term memory, which causes previous memories to fade away over time [18].

Long-Short-Term Memory (LSTM) introduction by S. Hochreiter and J. Schmidhuber to solve a problem in RNN. The primary distinction between LSTM cell architecture and traditional RNN is the presence of gates to control the propagation of earlier states (memory). The gate of forget is used to choose what information to preserve or forget after receiving the input and earlier hidden state (by setting the values close to 1 to keep or zero to forget). An earlier concealed state is transmitted to the input gate, and the values are controlled by a tan function and sigmoid. The values are then delivered to the next gate as input (cell state) [30].

The Gated Recurrent Unit (GRU) was introduced by J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. It is similar to the LSTM network as an updated-generation recurrent neural network but has a more straightforward design with only two gates: a reset gate and an update gate. Cell state was also discarded, and information is now sent to subsequent cells using the hidden state instead [31].

Finally, produce words that describe an image in a language we speak, according to specific databases dedicated to captioning the image. The sizes and languages of the databases iffer, and each of them is compatible with many applications. We will explain the databases in detail below.

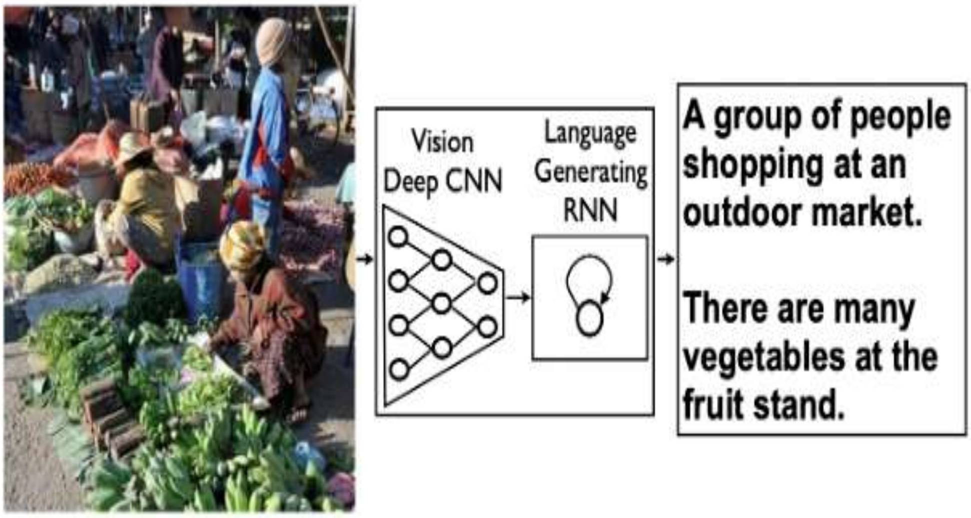
Vision
Deep CNN
Language
Generating
RNN
A group of people shopping at an outdoor market.
There are many vegetables at the fruit stand.

Figure 1: An overview of Vinyal et al.'s end-to-end Model for Image Captioning [34].

4. Image Captioning Datasets

A dataset is formatted in a way to benefit from object recognition and consists of digital images and feedback used for performance testing, training, and assessment of artificial intelligence and machine learning. The data set helps in learning to identify the information in the image. Artificial intelligence algorithms can be trained for image captioning or image-based VQA [32]. Such models help the visually impaired identify things. The data set size plays a large role in the accuracy of the results when using deep learning compared to machine learning [33].

There are many data sets in the area of processing natural language with computer vision in various languages. The English language was the first to compare the precision of the rest of the languages. The researchers worked on translating the databases into the language they speak. We will present some datasets and classify them according to the language used. Table 2 compares these datasets, and Table 3 shows the image-caption model used in the dataset.

Table 2: public datasets for captioning images.

Data sets	No. of image	Train	Test	Validate	No. Of Captions
Arabic Flickr8K[7]	• 8.000	• 6.000	• 1.000	• 1.000	• 3
• English Flickr8K[8]	• 8.000	• 6.000	• 1.000	• 1.000	• 5
• Flickr30K[9]	• 31.783	• 29.783	• 1.000	• 1.000	• 5
• FlickrStyle10K[10]	• 10.000	• 7.000	• 1.000	• 2.000	• 3
• MSCOCO[11]	• 328.000	• 113.287	• 5.000	• 5.000	• 5
• Visual Genome[13]	• 108.249	• -	• -	• -	• Additionally, each image has 35 objects, 26 properties,

					21 relationships, 50 region, and 17 question-answer pairs.
• IndiaVisual Genome[12]	• 87.398	• 77.398	• 5.000	• 5.000	• 4.100.000
• ImageParagraph Captioning[14]	• 19.561	• 14.575	• 2.489	• 2.487	• -
• # PraCegoVer[15]	• 500.000	• 60%	• 20%	• 20%	• 1
• VizWiz-Captions[16]	• 39.181	• 23.431	• 8.000	• 7.750	• 5

4.1 Arabic Flickr8K:

It is the first publicly available Arabic dataset developed by [7] regarding captions for images in Arabic and is the basis for English Fliker8K, containing 8,000 images each paired with three different captions. It contains 6,000 training images, 1,000 test images, and 1,000 verification images. This data set suffers from its small size because deep learning leads to greater accuracy in the results as the size of the data increases.

4.2 Flickr8K:

A new standard collection in the English language of sentence-based image captions and searches consists of 8000 images, each associated with 5 distinct captions that clearly describe significant people and places. It includes 6,000 training images, 1,000 test images, and 1,000 verification images. The photographs chosen from six distinct groups do not feature any known subjects and feature a variety of scenes [8].

4.3 Flickr30K:

It is one of the most popular datasets in the English language and is used to characterize an image's subject matter and produce a suitable caption. It consists of 31,783 images taken from Flickr that are all from daily life and contains 158,915 captions, each paired with five different captions. The Flickr30K dataset also includes object detectors and classifiers for identifying large objects and colors with bias. There is no fixed number for training and testing [9].

4.4 FlickrStyle10K:

It is a monolingual explanatory text dataset that provides a romantic and humorous caption on the image, built on the Flickr30K dataset. It contains 10,000 images based on Flickr, divided as follows: 7,000 training, 2,000 verification, and 1,000 test [10].

4.5 MSCOCO:

It is a large multilingual dataset, containing 328,000 images with five captions each. The dataset has a segmentation feature that helps to easily identify objects. It contains images of 91 different object categories, with 2.5 million instances of objects. The dataset is used to train object detection, segmentation, and captioning algorithms [11].

The MSCOCO Arabic dataset version was created by [21], [22]. It was completely translated using Google's advanced cloud translation API but is not publicly available.

4.6 *India Visual Genome Dataset*

It is the first dataset used for Hindi translation of image captions by translating Visual Genome in English using Google Translator to convert to Hindi, and it contains 87,398 images (77398 training images, 5000 verification images, 5000 testing images, and 4,100,000 captions) [12].

4.7 Visual Genome Dataset:

It is a data set that is used to describe the picture's content and link the objects of the image itself to provide a comprehensive caption for the image. It contains 108,000 images, and each image contains 21 features, 35 elements, and 21 pair-wise associations between the various objects in the dataset [13].

4.8 Image Paragraph Captioning Dataset:

It is a subset of the Visual Genome Regions (VG Regions), containing 20K Each image contains one paragraph, split into training = 14575, validation = 2487, and test = 2489. It is important to describe the image paragraph, and the total sentence average length is 11.91 words [14].

4.9 #PraCegoVer:

It is the first sizable dataset with Portuguese image captions. based on Instagram posts, has been compared to the MSCOCO dataset, and is divided into two parts:

First: #PraCegoVer-63K includes (dataset size = 62.935, train size = 37.881, validation size = 12.442, and test size = 12.612).

Second: #PraCegoVer-173K includes a dataset size of 173.337, a train size of 104.004, a validation size of 34.452, and a test size of 34.882 [15].

4.10 VizWiz-Captions dataset:

This is a new data set containing approximately 40k images, each paired with 5 captions, to assist visually impaired people in navigating and carrying out daily duties who rely on image-captioning services.

The VizWiz-Captions dataset contains 23k training images, 17k training captions, 7k validation images, 38k validation captions, 8k test images, and 40k test captions [16].

Table 3: Image-caption model used dataset (model img-cap represents image-caption model)

model img-cap	AR Flickr8K	AR MSCOCO	Flickr8K	Flickr30K	FlickrStyle10K	MSCOCO	India Visual Genome	Visual Genome	Image Paragraph Captioning	#PraCegoVer	VizWiz
model img-cap[25]						• √		• √	• √		
• model img-cap[4]	• √										
• model img-cap[13]								• √			
• model img-cap[17]	• √										
• model img-cap[10]					• √						
• model img-cap [7]	• √										
• model img-cap[15]										• √	
• model img-cap[12]							• √				
• model img-cap[22]	• √	• √									
• model img-cap[21]	• √	• √									
• model img-cap[14]									• √		

• model img-cap[20]	• √					
• model img-cap[34]				• √		
• model img-cap[35]				• √		
• model img-cap[36]				• √		
• model img-cap[37]				• √		
• model img-cap[38]				• √		
• model img-cap[39]		• √	• √			
• model img-cap[40]			• √	• √		
• model img-cap[41]		• √	• √	• √		
• model img-cap[42]			• √	• √		
• model img-cap[43]		• √	• √			
• model img-cap[23]		• √	• √	• √		
• model img-cap[44]			• √			
• model img-cap[45]						• √
• model img-cap[46]					• √	
• model img-cap[47]					• √	
• model img-cap[48]				• √		
• model img-cap[49]				• √		
• model img-cap[18]	• √					
• model img-cap[50]		• √	• √	• √		
• model img-cap[51]				• √		
• model img-cap[52]			• √	• √		
• model img-cap[53]		• √	• √	• √		
• model img-cap[54]				• √		
• model img-cap[55]				• √		
• model img-cap[56]				• √		
• model img-cap[57]				• √		
• model img-cap[58]				• √		
• model img-cap[59]			• √	• √		

• model img-cap[60]			• √	• √	
• model img-cap[61]	• √				
• model img-cap[8]	• √				
• model img-cap[9]		• √			
• model img-cap[16]					• √
• model img-cap[62]					• √
• model img-cap[63]			• √		• √
• model img-cap[64]	• √				
• model img-cap[65]	• √	• √	• √		
• model img-cap[66]			• √		
• model img-cap[67]				• √	

AR: Arabic, img-cap: image-caption

5. Evaluation Metrics

Evaluation metrics are important in examining the quality of a statistical or machine-learning model. Higher scores indicate better sentences. There are many different types of evaluation scales that are widely applied. The models are BLEU [68], ROUGE [69 ,[METEOR [70], CIDER [71], and SPICE [72]. Based on the researcher's statistics about the evaluation metrics used in the field of image captions, it was found that in Figure 3, the BLEU and METEOR are the most used due to their efficiency in short captions.[73]

6.1 Bilingual evaluation understudy (BLEU):

The most popular algorithm is inexpensive and widely applied in the field of NLP for determining the level of a sentence that has been translated into another language by Google Translate. It works by comparing the text to a set of captions and calculating scores that are averaged to determine the best evaluation for short captions [73]. This benchmark was introduced in 2002 by [68].

6.2 Recall-Oriented Understudy for Gisting Evaluation (ROUGE):

One of the measures used to evaluate machine translation in natural language processing is calculating the number of word sequences or other overlapping units as a result of machine translation. The sentence contains word pairs, n-grams, and the real sentence [69]. This measure has several different types, such as ROUGE-N (ROUGE-1, ROUGE-2), ROUGE-L, ROUGE-S, and ROUGE-SU.

6.3 Metric for Evaluation of Translation with Explicit ORdering (METEOR):

It is a measure for evaluating the language resulting from a translation, such as Google Translate, by calculating the match result of the word between machine translation and human translation. This measure is designed to solve problems or weaknesses in bilingual evaluation understudy (BLEU) and works to compare reference sentences with parts of the standard word and also uses synonyms of words or parts of the text for comparison [70].

6.4 Consensus-based Image Description Evaluation (CIDEr):

It is a novel automated metric for assessing image captions that determines the degree of similarity (cosine distance) by using the word frequency-inverse document frequencies (TF-IDF) of a genuine sentence [71].

6.5 Semantic Propositional Image Caption Evaluation (SPICE):

The researchers were interested in creating captions for the image using a computer. To assess the model's effectiveness, the evaluation scale represents a challenge. This scale depends on the semantic concept and works on converting the result into a middle representation that combines the original sentence and a sentence to measure the degree of accuracy of the model in generating captions for objects [72].

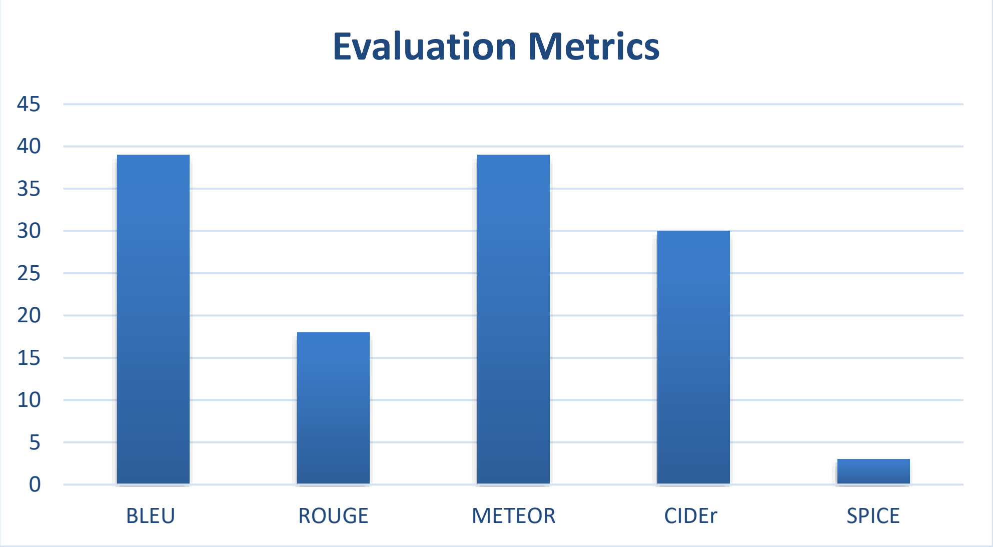

Figure 3: Comparison between Evaluation Metrics

6. Conclusions

This paper discusses the techniques and databases used, the drawbacks included, the different results reached by the researchers in the area of deep learning, and the different ways of describing the image. We provided descriptions of the datasets that used image captioning in different languages. We conclude that the biggest hurdle was providing a dataset in different languages. Works in English were the most accurate compared to other languages. They translate the dataset through Google Translator into their own language. It was concluded that the MSCOCO database is the most used database because of its large size, as it contains approximately 328k images, which helps increase the accuracy of deep learning results. The larger the data set, the more accurate the results. Evaluation metrics used for image captions were summarized, and it was found that BLEU and METEOR are the most commonly used in this field for being good at short captions.

Acknowledgments

For their assistance with this study, the authors would like to thank Mustansiriyah University in Bagdad, Iraq (HTTP://UOMUSTANSIRIYAH.EDU.IQ.(

7. Reference

[1] G. L. Krishna, “Social Media Application for Recruitment Using Pythagorean Fuzzy,” *Iraqi Journal of Science*, vol. 63, no. 4, pp. 1786–1801, 2022.

[2] A. Attai and A. Elnagar, “A survey on Arabic Image Captioning Systems Using Deep Learning Models,” in *Proceedings of the 14th International Conference on Innovations in Information*

Technology (IIT), pp. 114-119, 2020.

[3] A. Salaberria, G. Azkune, O. Lopez de Lacalle, A. Soroa, and E. Agirre, "Image captioning for effective use of language models in knowledge-based visual question answering," *Expert Systems with Applications*, vol. 212, no. 2023, p. 118669, 2023.

[4] S. M. Sabri, "Arabic Image Captioning using Deep Learning with Attention," *M.Sc. thesis, Institute for Artificial Intelligence, University of Georgia*, 2021.

[5] H. K. Dhahir and N. H. Salman, "A Review on Face Detection Based on Convolution Neural Network Techniques," *Iraqi Journal of Science*, vol. 63, no. 4, pp. 1823–1835, 2022.

[6] N. M. Khassaf and S. H. Shaker, "Image Retrieval based Convolutional Neural Network," *Al-Mustansiriyah Journal of Science*, vol. 31, no. 4, pp. 43–54, 2020.

[7] O. ElJundi, M. Dhaybi, K. Mokadam, H. Hajj and D. Asmar, "Resources and End-to-End Neural Network Models for Arabic Image Captioning," in *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, Vol. 5, pp. 233-241, no. 2184-4321 ,2020.

[8] M. Hodosh, P. Young and J. Hockenmaier, "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[9] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 74–93, 2016.

[10] C. Gan, Z. Gan, X. He, J. Gao and L. Deng, "StyleNet: Generating Attractive Visual Captions with Styles," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , pp. 955-964, 2017.

[11] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *Proceedings of the Computer Vision–ECCV the 13th European Conference*, vol 8693 , pp. 740–755, 2014.

[12] S. K. Mishra, Harshit, S. Saha, and P. Bhattacharyya, "An Object Localization-based Dense Image Captioning Framework in Hindi," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 2, pp. 1–15, 2022.

[13] R. Krishna, Y. Zhu, O.Groth, J. Johnson, K.Hata, J.Kravitz, S.Chen, Y.Kalantidis, L. Li, David, A. Shamma, M.S. Bernstein and F.Li, "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[14] X. Yang, C. Gao, H. Zhang and J. Cai, "Hierarchical Scene Graph Encoder-Decoder for Image Paragraph Captioning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, doi: https://doi.org/10.1145/3394171.3413859.

[15] G. O. Dossantos, E. L. Colombini and S. Avila, "#PraCegoVer: A Large Dataset for Image Captioning in Portuguese," *Data*, vol. 7, no. 2, p. 13, Jan. 2022.

[16] D. Gurari, Y. Zhao, M. Zhang and N. Bhattacharya, "Captioning Images Taken by People Who Are Blind," in *Proceedings of Computer Vision – ECCV 2020*, pp. 417–434, 2020.

[17]H. D. Hejazi, "Arabic Image Captioning (AIC): Utilizing Deep Learning and Main Factors Comparison and Prioritization.," *M.Sc. thesis, The British University in Dubai (BUiD)*, 2022.

[18] H. Hejazi and K. Shaalan, "Deep Learning for Arabic Image Captioning: A Comparative Study of Main Factors and Preprocessing Recommendations," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 11,pp. 37-44, 2021.

[19] M. Rahman, N. Mohammed, N. Mansoor, and S. Momen, "Chittron: An Automatic Bangla Image Captioning System," *Procedia Computer Science*, vol. 154, pp. 636–642, 2019.

[20] M. T. Lasheen and N. H. Barakat, "Arabic Image Captioning: The Effect of Text Pre-processing on the Attention Weights and the BLEU-N Scores," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7,pp. 413- 423, 2022.

[21] J. Emami, P. Nugues, A. Elnagar and I. Afyouni, "Arabic Image Captioning using Pre-training of Deep Bidirectional Transformers," in *Proceedings of the 15th International Conference on Natural Language Generation*, pp. 40–51, 2022.

[22] H. A. Al-muzaini, T. N. Al-yahya, H. Benhidour, "Automatic Arabic Image Captioning using RNN-LSTM-Based Language Model and CNN," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 6,pp. 67-73, 2018.

[23] C. Wang and X. Gu, "Learning joint relationship attention network for image captioning" *Expert Systems with Applications*, vol. 211,no. C, p. 118474, 2023.

[24] O. N. Shinde, R. Gawde and A. Paradkar, "Social Media Image Caption Generation Using

Deep Learning," *International Journal of Engineering Development and Research (IJEDR)*,Vol.8, no 4, pp.222-228, 2020.

[25] A.Tiwari, "IMAGE PARAGRAPH GENERATION USING DEEP LEARNING," *M.Sc. Thesis, Department Of Electronics And Communication Engineering, Delhi Technological University*, 2022.

[26]S. K. Mishra, R. Dhir, S. Saha, and P. Bhattacharyya, "A Hindi Image Caption Generation Framework Using Deep Learning," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 2, pp. 1–19, 2021.

[27] Z. A. Ramadhan and D. Alzubaydi, "Text Detection in Natural Image By Connected Component Labeling," *Al-Mustansiriyah Journal of Science*, vol. 30, no. 1, p. 111, 2019.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2009.

[29] V. D. Shinde, M. P. Dave, A. M. Singh, and A. C. Dubey, "Image Caption Generator using Big Data and Machine Learning," *International Research Journal of Engineering and Technology (IRJET)*, vol. 07, no. 04,pp. 6197-6201, 2020.

[30] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[31] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[32]S. Zhang, Q. Wei, Y. Li, Y. Chen and L. Jiao, "Visual Question Answering of Remote Sensing Image Based on Attention Mechanism," in *Proceedings of IFIP Advances in Information and Communication Technology*, vol 659, pp. 228–238, 2022.

[33] A. Panwar, G. Semwal, S. Goel and S. Gupta, "Stratification of the Lesions in Color Fundus Images of Diabetic Retinopathy Patients Using Deep Learning Models and Machine Learning Classifiers," in *Proceedings of Lecture Notes in Electrical Engineering*, pp. 653–666, 2022.

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

[35] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of International conference on machine learning*, pp. 2048–2057, 2015.

[36] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

[37] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10578–10587, 2020.

[38] X. Hu et al., "Vivo: Visual vocabulary pre-training for novel object captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, pp. 1575–1583, 2021.

[39] H. Sharma and A. S. Jalal, "Incorporating external knowledge for image captioning using CNN and LSTM," *Modern Physics Letters B*, vol. 34, no. 28, p. 2050315, 2020.

[40] N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, "Topic-oriented image captioning based on order-embedding," *IEEE Trans. Image Process*., vol. 28, no. 6, pp. 2743–2754, 2018.

[41] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van den Hengel, "Image Captioning and Visual Question Answering Based on Attributes and External Knowledge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1367–1381, Jun. 2018.

[42] X. Yang and C. Xu, "Image captioning by asking questions," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 15, no. 2s, pp. 1–19, 2019.

[43] S. Srivastava, H. Sharma and P. Dixit, "Image Captioning based on Deep Convolutional Neural Networks and LSTM," in *Proceedings of 2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, Mathura, India, pp. 1-4, 2022.

[44] R. Das and T. D. Singh, "Assamese news image caption generation using attention mechanism," *Multimedia Tools and Applications*, vol. 81, no. 7, pp. 10051-10069, 2022.

[45] L.-C. Yang, C.-Y. Yang and J. Y.- jen Hsu, "Object Relation Attention for Image Paragraph Captioning," *AAAI*, vol. 35, no. 4, pp. 3136-3144, 2021.

[46] Q. Jiao and W. Yu, "Generating Interested Captions in Image with Visual Semantic Concepts," in *Proceedings of 2022 IEEE 10th Joint International Information Technology and Artificial*

Intelligence Conference (ITAIC), Chongqing, China, pp. 1758-1763, 2022.

[47] A. M. Bhagawan, Pava, A. M. Gudihal, D. G. Mangalagatti and Savitha , “IMAGE CAPTION GENERATOR USING DEEP NEURAL NETWORKS,” *International Research Journal of Modernization in Engineering Technology and Science*, vol. 04, no. 07,pp.2881-2889 ,2022.

[48] J. Ji, Y. Ma, X. Sun, Y. Zhou, Y. Wu, and R. Ji, “Knowing What to Learn: A Metric-Oriented Focal Mechanism for Image Captioning,” in *Proceedings of the IEEE Transactions on Image Processing*, vol. 31, pp. 4321–4335, 2022.

[49] Z. Fei, “Attention-Aligned Transformer for Image Captioning,” *AAAI*, vol. 36, no. 1, pp. 607-615, 2022.

[50] L. Yang and H. Hu, “TVPRNN for image caption generation,” *Electronics Letters*, vol. 53, no. 22, pp. 1471–1473, 2017.

[51] Y. Ming, N. N. Hu, C. X. Fan, F. Feng, J. W. Zhou, and H. Yu, “Visuals to text : A comprehensive review on automatic image captioning,” *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 8, pp. 1339–1365, 2022.

[52] G. Ding, M. Chen, S. Zhao, H. Chen, J. Han, and Q. Liu, “Neural image caption generation with weighted training and reference,” *Cognit. Comput.*, vol. 11, no. 6, pp. 763–777, 2019.

[53] Y. Peng, X. Liu, W. Wang, X. Zhao, and M. Wei, “Image caption model of double LSTM with scene factors,” *Image and Vision Computing*, vol. 86, pp. 38–44, 2019.

[54] Y. Chen and M.-C. Chang, “On multimodal semantic consistency detection of news articles with image caption pairs,” in *Proceedings of the IEEE International Conference on Consumer Electronics-Taiwan*, pp. 355–356, 2022.

[55] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, “Improved image captioning via policy gradient optimization of spider,” in *Proceedings of the IEEE international conference on computer vision*, pp. 873–881, 2017.

[56] H. R. Tavakoli, R. Shetty, A. Borji, and J. Laaksonen, “Paying attention to descriptions generated by image captioning models,” in *Proceedings of the IEEE International Conference on Computer Vision* , pp. 2487–2496, 2017.

[57] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, “Areas of attention for image captioning,” in *Proceedings of the IEEE international conference on computer vision*, pp. 1242–1250, 2017.

[58] Y. Wang, Z. Lin, X. Shen, S. Cohen, and G. W. Cottrell, “Skeleton key: Image captioning by skeleton-attribute decomposition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7272–7281, 2017.

[59] B. Dai, S. Fidler, R. Urtasun, and D. Lin, “Towards diverse and natural image descriptions via a conditional GAN,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2970–2979, 2017.

[60] S. Chen, Q. Jin, P. Wang, and Q. Wu, “Say as you wish: Fine-grained control of image caption generation with abstract scene graphs,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9962–9971,2020.

[61] A. D. Martin, E. Ahmadzadeh and I. Moon, “Privacy-Preserving Image Captioning with Deep Learning and Double Random Phase Encoding,” *Mathematics*, vol. 10, no. 16, p. 2859, 2022.

[62] P. Dognin et al., “Image Captioning as an Assistive Technology: Lessons Learned from VizWiz 2020 Challenge,” *Journal of Artificial Intelligence Research*, vol. 73, pp. 437–459, 2022.

[63] C. Chen, “Image captioning and visual question answering with external knowledge,” M.Sc. thesis, science in information studies, University of Texas at Austin, 2020.

[64] M. Bhalekar and M. Bedekar, “D-CNN: A New model for Generating Image Captions with Text Extraction Using Deep Learning for Visually Challenged Individuals,” *Eng. Technol. Appl. Sci. Res.*, vol. 12, no. 2, pp. 8366–8373, 2022.

[65] D. Kumar, V. Srivastava, D. E. Popescu, and J. D. Hemanth, “Dual-Modal Transformer with Enhanced Inter- and Intra-Modality Interactions for Image Captioning,” *Applied Sciences*, vol. 12, no. 13, p. 6733, 2022.

[66] T. Xian, Z. Li, Z. Tang and H. Ma, "Adaptive Path Selection for Dynamic Image Captioning," *in IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5762-5775, 2022.

[67] Z. Wang, Y. Luo, Y. Li, Z. Huang, and H. Yin, “Look deeper see richer: Depth-aware image paragraph captioning,” in *Proceedings of the 26th ACM international conference on Multimedia*, pp. 672–680, 2018.

[68] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a Method for Automatic Evaluation of

Machine Translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL),* pp. 311–318, 2002.

[69] C.Y. Lin, "Rouge: A package for automatic evaluation of summaries," In *Proceedings of the Text summarization branches out*, pp. 74-81, 2004.

[70] S. Banerjee and A. Lavie, "METEOR: An automatic metric for mt evaluation with improved correlation with human judgments," In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization,* pp. 65–72, 2005.

[71] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566-4575, 2015.

[72]P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," In *Proceedings of Computer Vision – ECCV 2016*, pp. 382–398, 2016.

[73] H. Sharma, "A Survey on Image Captioning datasets and Evaluation Metrics," *IOP Conference Series: Materials Science and Engineering*, vol. 1116, no. 1, p. 012184, 2021.