A Review of Deep Learning based Image Captioning Models

Dev Kumar[1], Sneh Gehani[2], Pranali Oza[3]

[1]U.G. Student, Department of Information Technology, Thadomal Shahani Engineering College, Maharashtra, India
[2]U.G. Student, Department of Computer Engineering, Thadomal Shahani Engineering College, Maharashtra, India
[3]U.G. Student, Department of Information Technology, Thadomal Shahani Engineering College, Maharashtra, India

---***---

Abstract - *Image captioning, primarily means giving a suitable caption to an image. The task of Image captioning needs to evaluate an image, with respect to the subjects and objects in the image, the relationship between these semantic details needs to be determined accurately along with other attributes and features present in the image. Once this identification is done, a grammatically correct caption that best describes the image must be generated. Despite the advancement in technology, Image captioning remains a challenging task that employs both, Computer Vision for image identification and Natural Language Processing for generation of the image captions. However, with extensive research in this domain,several methods employing Deep Learning techniques have been adopted. In this paper, we present a survey on the several methods adopted for this task. We first briefly introduce methods that do not employ deep learning, primarily template and retrieval based. Then we move onto Deep Learning based methods that are further classified into categories based on the architecture they adopt. Each category is examined thoroughly and the most relevant models are compared on benchmark datasets. Finally, the future aspects of research in this domain are discussed.*

***Key Words*: Deep Learning, Computer Vision, Natural Language Processing, Image Caption Generation, Feature Extraction, Convolutional Neural Networks(CNN), Visual-Language, Long Short Term Memory(LSTM)**

1. INTRODUCTION

Images have become an unavoidable medium of communication in the modern era. Millions of images are generated everyday all around the world. In today's time, Artificial Intelligence is progressing at a really fast pace and is making its way into most of our daily activities. Researchers are leaving no stones unturned in order to examine the capabilities of AI in terms of solving most of the problems out there. Companies like Google are heavily invested in research pertaining to this domain. A great example of which can be seen in Google Lens. Google Lens is an image recognition technology that is capable of describing images and objects subjected to a smartphone's camera. It is used by travelers and people who are at new places and would want to know about the things they see but are somehow not able to ask around. In all, Google Lens is capable of generating information about a particular image and making sense out of it.

Images are capable of giving out a lot of information. It may have context to the events happening within the image but this not being detected by image captioning models is still a big research problem. For humans, it may still be an easy task to interpret the details of the image having had a context for the same; but enabling that for machines is a challenge. Social media platforms such as Facebook and Twitter are capable of describing the images and giving the captions. The captions may include minute details such as our location (e.g., beach, cafe), our appearance and importantly what activities are being captured in the image. Over the years, lots of work has been carried out to provide efficient solutions for this problem. A lot of models have been developed and a large number of articles have been published on image captioning with deep machine learning being popularly used. Deep learning algorithms can handle the complexities and challenges of image captioning quite well. Efforts are being made to improve the existing methods and take it on par with human intelligence when it comes to generating captions for images and using context in order to completely comprehend the image. As information systems generate more data that facilitate creation of more accurate models, the learning-based image captioning has become a sought-after research area. In this paper, we will be reviewing the recent work done with respect to the problem of image captioning. We will compare and contrast the recent findings in order to evaluate the efficiency of the models against each other. To provide access to an array of information on the central topic, we present a survey based on the deep learning-based papers on image captioning.

2. IMAGE CAPTIONING TECHNIQUES

There are a lot of ways by which image captioning models can be built and this section deals with the most common techniques known to us and a brief overview on the work that has been done in each technique. We will be discussing Retrieval-Based Image Captioning, Template-Based Image Captioning and Deep Learning-Based Image Caption.

2.1 Retrieval-Based Image Captioning

Retrieval-based Image Captioning has been a well-known approach for quite a long time. A lot of initial models for Image Captioning have been developed using this technique. As the name suggests, in this technique the caption is generated by selecting or retrieving the most probable caption from a predefined collection of captions. This technique involves finding visual similarities between the query image and the training dataset.

For a given image that is being queried and whose caption is to be generated, this technique involves plotting the image into the meaning space by solving a Markov Random Field, and the semantic distance between these images is deduced by Lin similarity measure [40] and each existing sentence is parsed with the help of Curran and Clark parser [41]. After which, the caption which is deduced to be closest to the given query image will be considered as a caption of the queried image. Ordonez et al. [42] firstly used global image computing to extract a group of images from a web-scale; basically, web-scale is a combination of captioned images.

In image captioning, according to Hodosh et al. [43] the problem of image captioning can be seen as a task of ranking. With respect to a particular image, the caption that correlates with the content of the image or the caption which is successful in accurately describing the content of the image will be given a higher rank. For this purpose, the authors propose the Analysis of Kernel Canonical Correlation method [44,45] that correlates maximum training images and their captions to align the images and the text as per their affinity. This method will be helpful in efficiently ranking the captions and therefore the caption with the highest relatability or correlation will be retrieved.

Notwithstanding the promising nature of the proposed model, there are several limitations with this method. First, the captions that are being assigned (due to the correlation or otherwise) are well-constructed sentences provided by humans. By default, this means that the assigned caption will be grammatically correct. Providing description of the images with sentences that have been predefined cannot help in generating captions for new object mixtures. The retrieved caption might not be relevant to a new change in the picture and the model may also be incapable of adapting to minute changes within the picture.

2.2 Template-Based Image Captioning

One other common method used in early image captioning is template based. The central idea here is to detect a set of visual attributes, objects and relationships with other objects first. Templates or specific grammar rules are then used for generating sentences that split sentences into its components like nouns, verbs, objects etc. These sentence fragments are then mapped with the target visual elements to predict the components of the sentences that can possibly be used to generate the final sentence and evaluate its correlation with the image components using various evaluation methods.

Yang et al. [10] first detects objects and scenes in the images using detection algorithms[11],[12]. Then a sentence template that uses a quadruplet consisting of Nouns-Verbs-Scenes-Prepositions is employed. A language model[14] is then utilized to predict the quadruplet that can be used to generate captions. Finally, the Hidden Markov Model (HMM) inference is used to obtain the quadruplet with the highest log-likelihood and the image caption is produced by filling the sentence structure by the chosen quadruplet. Similarly Kulkarni et al. [13] employ object detectors to determine the image contents and then send recognised image contents into attribute classifier and prepositional relation function to get some information on the attributes of the image components and information on prepositional relations between objects. Finally, a Conditional Random Field(CRF) is employed to determine the final description of the image contents that is then used as the image caption. The previous models both use words, however, phrases tend to give out more information as they are a combination of several words. Thus methods employing phrases are proposed as sentences produced using phrases rather than words tend to present more information. Ushiku et al. [15] propose a novel method called Common Subspace for Model and Similarity. The method withdraws phrases from training captions. The phrases extracted are then mapped with the image features into a single subspace where similarity based and model based classification are integrated to learn a classifier for each phrase. In the testing stage, phrases estimated from a query image are connected by using multi-stack beam search [16] to generate a description.

Despite template based captioning being capable of producing sentences that are syntactically correct and descriptions that are more relevant than retrieval-based methods, there are several disadvantages of using template-based methods. Firstly, due to the general lack of visual models and the captions generated using this method are dependent on the image content identified by visual models, the complexity, structure, novelty and creativity of the generated captions is severely limited. Secondly, following a strict template or structure for caption generation makes the generated captions seem less natural in comparison to the human-generated descriptions.

2.3 Deep Learning-Based Image Captioning

The crucial benefit of deep convolutional neural networks (CNN) is very useful. Image captioning has in recent years garnered more research focus in AI. It has many uses, since it mainly generates an automatic sentence description for an image. It allows computer systems to recognize images for mainly education purposes, sentiment analysis, an aid for the visibly impaired, etc. The model must be accurate enough to understand the various relations between various objects, and express that in a correct semantic manner in natural language. Image captioning methods primarily make use of the template-based methods, that requires describing the diverse elements (objects) in addition to their relationships and attributes.

These techniques are mainly based on the encoder-decoder methodology that includes two simple steps. Firstly, using CNN, Image features are deduced to encode the image into a hard and fast period embedding vector. Secondly, generating a language description usually a recurrent neural network is used as a decoder.

CNN-RNN framework based image captioning technique have two drawbacks in training phase:

1) Each caption gets equal importance without their individual importance
2) Objects may not be correctly recognized during caption generation

During the training phase, as per the relation between image and words different weights are assigned. In addition to maximizing the agreement-score among the captions produced through the captioning methods and the reference data from the adjoining images of the intentional images that can limit the issue of not recognizing correctly an image.

Despite the existence of several categories of deep learning methods including multimodal space, encoder-decoder architecture, attention based, novel object based, language models based on LSTM, we shall focus on three of the most relevant categories, Multimodal Space, Language Models and Encoder-Decoder Architecture.

3. IMAGE CAPTIONING METHODS BASED ON DEEP LEARNING

3.1 Multimodal Learning

Template based and retrieval based image captioning methods impose restrictions on generated sentences in generation phase. Methods using deep neural networks that do not depend on existing captions about structures of sentences can produce more communicative and adjustive sentences with more affluent structures. Using multimodal neural networks is one of the few methods that rely on pure learning to create image captions. Here, using deep convolutional neural networks, image features are first removed. Then, the extracted image feature is sent to a neural language model, which maps the image feature with the common word features and performs word predication trained on the image feature and previously generated context words. A general structure of image captioning methods employing multimodal learning is presented in Fig - 1.

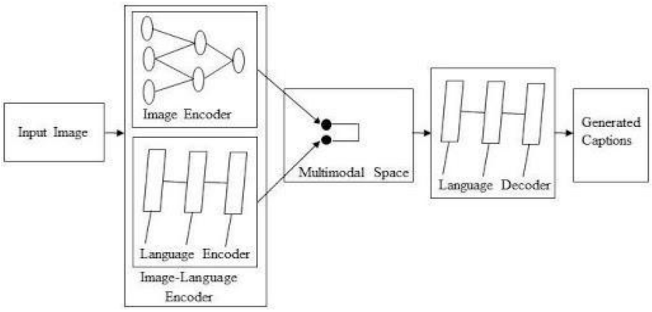


Fig -1: Multimodal space based image captioning, A Comprehensive Survey of Deep Learning for Image

A neural language model which is dependent on image inputs is suggested by Kiros et al. [1] for generating image captions. In their method, a log-bilinear language model[26] is adapted, where an image feature is added as an extra bias to help predict the probability of generating a word along with the support of previously generated words. Feature learning is employed by back propagating gradients from the loss function through the multimodal neural network model.

This model allows the generation of the captions word by word, with each individual word being generated by conditioning on both the previously generated words and the visual features.

To generate novel captions, Mao et al. [25] proposed a multimodal Recurrent Neural Network(m-RNN). This method extracts visual features by using a deep convolutional network(CNN) and sentences by using a deep recurrent neural network(RNN) with a multimodal part as the language model. The images and sentences are both used as input in this method where the CNN and RNN both interact with each other in the multimodal layer. For the generation of the next word the probability distribution is calculated where the new word is conditioned on the input image and the previously generated words. This RNN model consists of five layers in a single time frame consisting of two word embedding layers, a recurrent layer, a multimodal layer and a SoftMax layer[25]. Various other methods utilize predefined word embedding vectors for the initialization of their language model, however, this method randomly initializes the word embedding vectors which are later learnt from the training data.

In Schuster and Paliwal's method, image regions are aligned and represented by a CNN and sentence segments characterized by a Bidirectional Recurrent Neural Network[27] are used to train a multimodal Recurrent Neural Network model to generate descriptive captions for image regions[28]. After image region representation, visual and textual data are mapped into a mutual space and each region feature is linked to the textual feature that describes the region. The associated two modalities are then used to train a multimodal Recurrent Neural Network model, that can be used to find the probability of generating the next word given an image feature and context words. To assuage the weakness of learning long term dependencies[29,30] in image captioning in RNN, Chen and Zitnick suggest to dynamically shape a visual representation of the image while a caption is being generated for it, so that long term visual impressions can be evoked during this process[31]. This reverse projection is made possible due to the RNN having an additional recurrent visual hidden layer.

3.2 Encoder-Decoder Framework

Taking inspiration from the encoder-decoder framework in neural machine translation[100] which was originally used to translate sentences and phrases from one language to another, the encoder-decoder architecture has been adopted to perform the task of image captioning by giving an image as the input and receiving the output as a sentence. The general working of this architecture includes an encoder neural network that extracts global image features which are then fed to as input to a decoder that consists of a recurrent neural network to produce a caption word by word. The general structure of this framework is shown in Fig.-2.

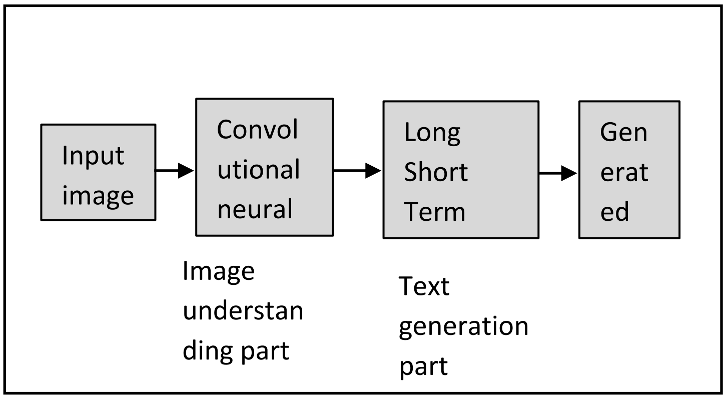


Fig -2: Encoder Decoder, A Comprehensive Survey of Deep Learning for Image Captioning

An encoder-Decoder framework that effectively unifies joint image-text embedding models with multimodal neural language models is introduced by Kiros et al.[1]. A deep Convolutional Neural Network (CNN) is used to encode the visual data whereas the textual data is encoded by employing a Long Short-Term (LSTM) Recurrent Neural Network. The image features from the deep CNN are projected into the embedding space of the hidden states of the LSTM. Then by minimizing a pairwise ranking loss, the ranking of the images and descriptions is learnt. This completes the encoder part of the framework. For the decoder, a novel structure-content neural language model is employed to decode image features conditioned on context word feature vectors, thus resulting in generation of novel captions word by word.

Vinyals et al. [2] also inspired by neural machine translation put forth a method called the Neural Image Caption Generator (NIC). The NIC uses a novel method for batch normalization of the encoder which is a Convolutional Neural Network(CNN). The image features extracted from the last hidden layer of this CNN are then fed as input into the decoder which is an LSTM capable of keeping track of objects that have previously been recognized or described. The model is trained by maximizing the likelihood of sentence image pairs in the training set.

Once the model is trained, either sampling or beam search can be used to make the predictions of possible word sequences that can be used as captions.

In the previous models, the image information was fed just once, in the initial state of the LSTM thus leading to the issue of vanishing gradient thus leading to difficulty in producing long length sentences[3, 4]. To solve this issue of vanishing grading Jia et al.[5] proposed a guided LSTM(gLSTM). Global textual information is added to every gate and cell state of the LSTM. The textual information is extracted using several different methods. A multimodal embedding space can be used to extract the semantic information. Or textual information can be extracted from image captions retrieved by a cross-modal retrieval task.

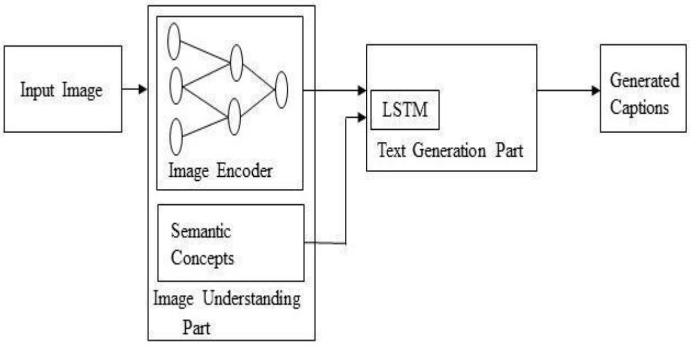


Fig -3: A block diagram of a semantic concept-based image captioning, A Comprehensive Survey of Deep Learning for Image Captioning

The issue with unidirectional sentence generation models is while they may include past textual context, they are still limited to retain future context in case of forward direction and vice versa in case of backward direction. Thus unidirectional models cannot produce contextually rich sentences. A bidirectional model tries to overcome this issue and utilise past and future dependence to give a prediction. Furthermore, certain object detection and classification methods [6, 7] have demonstrated that deep hierarchical models perform better learning in comparison than relatively shallower models. Thus Wang et al.[8] propose a deep bidirectional LSTM as the decoder in the encoder decoder framework. The bidirectional model is fed sentences from both forward and backward order to make use of past and future context information. The proposed model consists of three modules; to begin with, a CNN is used for encoding image inputs. The second module is a Text-LSTM(T-LSTM) for encoding the sentences provided as inputs. The third module consists of a Multimodal LSTM (M-LSTM) for fusing visual and textual feature vectors into one single semantic space and then decoding it to a sentence. Two separate LSTM layers are used to implement the bidirectional LSTM(bi-LSTM) for the computation of forward and backward hidden sequences.

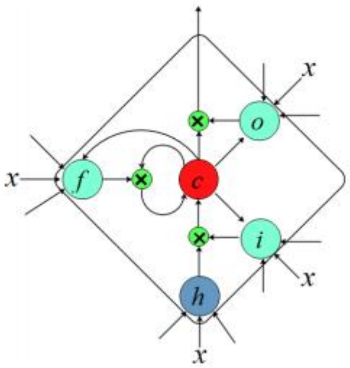
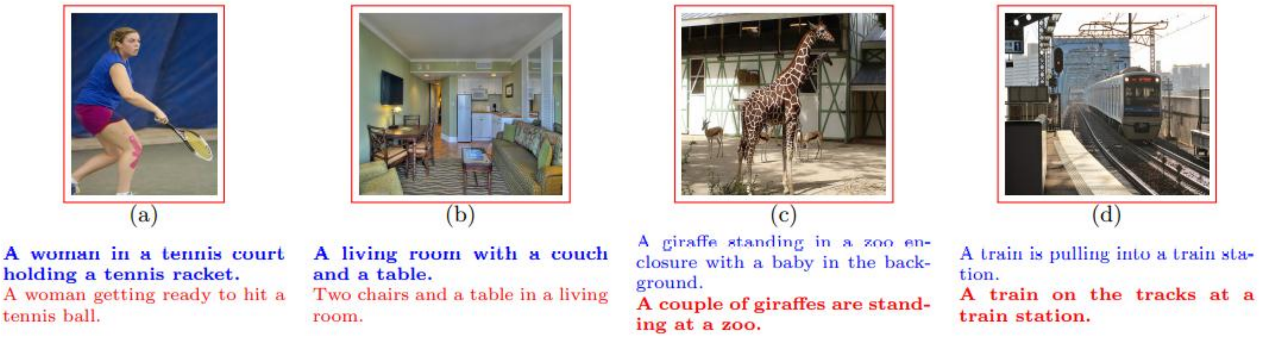


Fig -4: LSTM cell, Image Captioning with Deep Bidirectional LSTMs

Additionally, the paper proposes two variants of the bi-LSTM to make the model deeper. For the first variant, multiple LSTMs are stacked on top of each other and is called the Bi-S-LSTM. The second variant proposes using a fully connected multilayer perceptron (MLP) as an intermediate transition layer and the model is called the Bi-F-LSTM. This prevents the parameter size from growing dramatically while increasing the network depth of the LSTM. Thus long term visual interactions can be easily learnt.



(a)
A woman in a tennis court holding a tennis racket.
A woman getting ready to hit a tennis ball.

(b)
A living room with a couch and a table.
Two chairs and a table in a living room.

(c)
A giraffe standing in a zoo enclosure with a baby in the background.
A couple of giraffes are standing at a zoo.

(d)
A train is pulling into a train station.
A train on the tracks at a train station.

Fig -5: Examples of generated captions for Wang et al. [8] for given query image on MSCOCO validation set. Blue-colored captions are generated in forward direction and red-colored captions are generated in backward direction

Most encoder-decoder models used in image captioning are designed using a single LSTM (Long Short Term Memory) whose textual encoder and decoder are embedded in one layer limiting its capacity to perform a complex task such as image captioning. Moreover, increasing the 'vertical depth' of encoder decoder networks is an issue that remains to be unsolved. To solve this issue, Xiao et al.[9] propose a model that fuses the visual and textual semantics before decoding. The model separates the encoder and decoder to separate LSTMs to create a Deep Hierarchical Encoder-Decoder Network (DHEDN).

The base model consists of four modules, to begin with they have a deep CNN for encoding image features. The second module is a Sentence-LSTM (S-LSTM) encoder used for encoding sentence inputs. The third one is the most crucial one, the Vision-Sentence Embedding LSTM (VSE-LSTM) used for fusing the CNN visual features and the S-LSTM sentence feature into a single joint semantic space. Finally, the image features, sentence encoded features and vision sentence embedded vector are decoded using a Semantic Fusion LSTM (SF-LSTM) decoder into the target sentence.

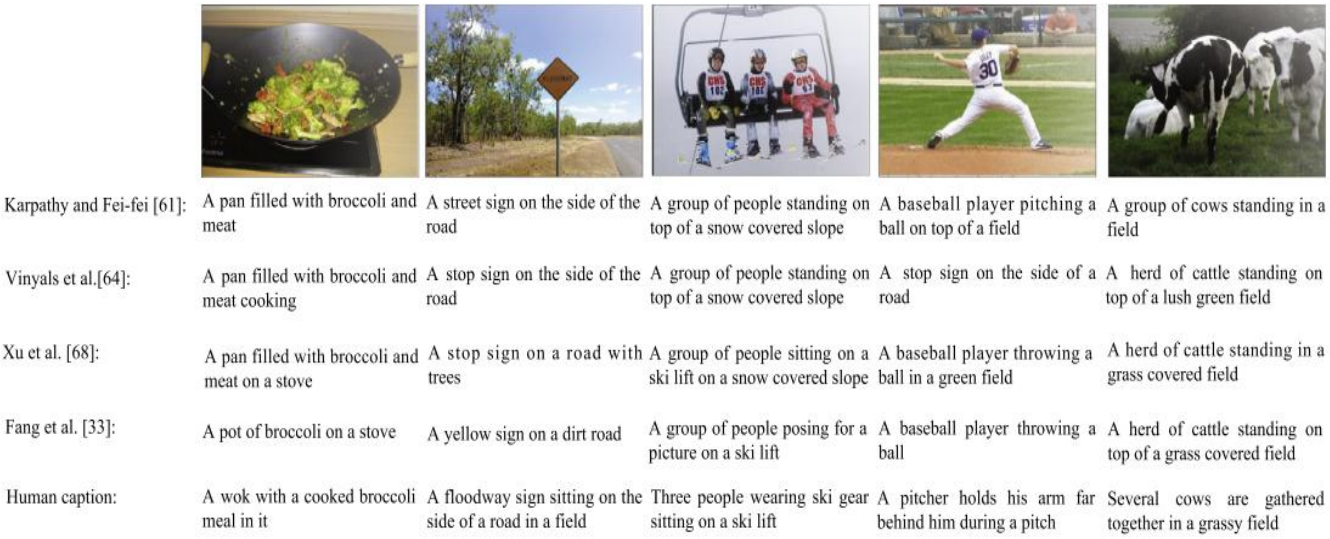
3.3 Language Models

As previously may have been discussed, Image Captioning is an interdisciplinary problem that requires solutions from computer vision as well as natural language processing(NLP). Using computer vision techniques, we are able to derive insights from a particular image but to use them in order to generate the caption, which is our final output, is possible using NLP. NLP tasks, in general, can be formulated as a sequence to sequence learning. In order to fulfill this task, various neural language models have been proposed. Few of which include neural probabilistic language model [20], log-bilinear models [21], skip-gram models [22], and recurrent neural networks (RNNs) [23]. Up until the recent past, RNNs have commonly been utilized in different Sequence Learning Tasks. Nonetheless, traditional RNNs experience the ill effects of vanishing and exploding gradient problems and are unable to sufficiently deal with long-term temporal dependencies.

In order to overcome this, LSTM[31] networks can be useful. LSTM networks are a type of RNN that consists of special units along with standard units. To store and preserve information in memory for longer periods of time, LSTM units use memory cells. This advanced version of RNNs has widely been used over the years for fulfilling tasks that involve sequence to sequence learning. Alternatively, we have a Gated Recurrent Unit(GRU) [32] which has a similar structure to LSTM but has a few minor differences. GRUs use fewer gates to control the flow of information. Additionally, GRUs do not use separate memory cells. However, LSTMs do not take into account the underlying hierarchical structure of a sentence. Due to long-term dependencies, they also require significant storage through a memory cell.

On the other hand, CNNs are able to learn the internal hierarchical structure of the sentences. They are also able to process faster than LSTMs. Gu et. al. [33] proposed an image captioning method which is based on the CNN language model. However, it needs to be combined with a recurrent neural network (RNN) to model the temporal dependencies properly as language-CNN alone cannot fulfill the dynamic temporal behavior of the language model.

CNN architectures are used in another sequence to sequence tasks. For example, conditional image generation [34] and machine translation [36, 37, 35] where the models have been able to overcome previously mentioned dependencies. On account of the immense success of CNNs in sequence learning tasks, Aneja et. al. [38] proposed a convolutional image captioning technique that uses a feed-forward network without any recurrent function unlike the above-mentioned technique[33]. The proposed convolutional architecture consists of four components viz. Input embedding layer, image embedding layer, convolutional module, and the output embedding layer (consisting of classification and training modules). This architecture is evaluated on the remarkable MSCOCO dataset. The results indicate that the convolutional approach performs on par with LSTM-RNN based architectures. It also suggests that adding an attention mechanism[37, 35] to the preexisting CNN architecture improves its performance and outperforms the LSTM-Attention-based line[39]. The results also support the fact that CNN models with additional 50% parameters can be trained in a comparable time since the sequential processing that takes place in RNN need not happen here.



Karpathy and Fei-fei [61]:	A pan filled with broccoli and meat	A street sign on the side of the road	A group of people standing on top of a snow covered slope	A baseball player pitching a ball on top of a field	A group of cows standing in a field
Vinyals et al.[64]:	A pan filled with broccoli and meat cooking	A stop sign on the side of the road	A group of people standing on top of a snow covered slope	A stop sign on the side of a road	A herd of cattle standing on top of a lush green field
Xu et al. [68]:	A pan filled with broccoli and meat on a stove	A stop sign on a road with trees	A group of people sitting on a ski lift on a snow covered slope	A baseball player throwing a ball in a green field	A herd of cattle standing in a grass covered field
Fang et al. [33]:	A pot of broccoli on a stove	A yellow sign on a dirt road	A group of people posing for a picture on a ski lift	A baseball player throwing a ball	A herd of cattle standing on top of a grass covered field
Human caption:	A wok with a cooked broccoli meal in it	A floodway sign sitting on the side of a road in a field	Three people wearing ski gear sitting on a ski lift	A pitcher holds his arm far behind him during a pitch	Several cows are gathered together in a grassy field

Fig -7: Examples of image captioning results obtained based on different methods.

4. EVALUATION METRICS

In order to determine the effectiveness of the models in terms of generating image captions, we will juxtapose the results of different models that have been a part of this study using the evaluation metrics available for the said purpose. It is necessary to examine the competence of the system-generated captions and those described by humans. For this purpose, the globally accepted and available evaluation metrics are described in detail in the subsequent paragraphs.

4.1 Bilingual Evaluation Understudy(BLEU)[16]

In this metric, a basic approach is followed where the generated caption is matched against a set of predefined texts interpreted by humans. The main purpose of this metric is to determine the affinity of the system-generated caption with the expected output, usually given by humans, and calculate a score based on the affinity. However, syntactical correctness is not a deal breaker in this metric in order to calculate the score. Lastly, the comprehensive quality of the system-generated text is determined by an average score. The BLEU metric heavily depends on the number of expected interpretations i.e. reference texts provided and the size of the system-generated text.

4.2 Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [17]

ROUGE can be termed as a collection of metrics which matches pairs and sequences of words (basically, n-gram) with human-generated summaries and reference texts in order to calculate a score. There are different ROUGE metrics based on the intended task. Some of which include ROUGE-N, ROUGE-W, ROUGE-S, ROUGE-L, and ROUGE-SU. Each of these previously mentioned metrics will be used for evaluating a different set of characteristics within a sentence. In one of the analyses, you may see ROUGE-L being used which is based on the Longest Common Subsequence and evaluates the score by identifying the longest co-occurring sequence of n-grams in the sentence.

4.3 Metric for Evaluation of Translation with Explicit ORdering (METEOR) [18]

Another distinct metric that helps in computing and scrutinizing system-generated language is METEOR. The system-generated captions and human interpretations are both matched under a generalized unigram. A score is then calculated based on the similarity between the two counterparts. In the case of multiple interpretations or possibilities, the best score will be chosen from the distinctly calculated ones.

4.4 Consensus-Based Image Description Evaluation (CIDEr) [19]

As the name suggests, this evaluation metric is based on a consensus, that is relevant to the most number of candidates. For a particular image, this metric will require a set of human interpretations intended to work as a caption for that image. With the availability of abundant human descriptions for a single image, this metric will measure the closeness or similarity of these references to the system-generated caption and will give a score based on the consensus achieved i.e. similarity with the majority of the references provided by humans.

5. RESULTS

Table-1: Performance of different image captioning methods on three datasets and commonly used evaluation metrics.

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Karpathy et al. 2015[28]	0.579	0.383	0.245	0.160	-
	Mao et al. 2015 [25]	0.565	0.386	0.256	0.170	-
	Jia et al. 2015 [5]	0.647	0.459	0.318	0.216	0.202
	Wang et al. 2016[8]	0.655	0.468	0.320	0.215	-
	Gu et al. 2017[33]	-	-	-	-	-
	Aneja et al. 2018 [38]	-	-	-	-	-
	Xiao et al. 2019[9]	0.651	0.470	0.326	0.220	0.201
Flickr30k	Karpathy et al. 2015[28]	0.573	0.369	0.240	0.157	-
	Mao et al. 2015 [25]	0.600	0.410	0.280	0.190	-
	Jia et al. 2015 [5]	0.646	0.466	0.305	0.206	0.179
	Wang et al. 2016 [8]	0.621	0.426	0.281	0.193	-
	Gu et al. 2017[33]	0.714	0.540	0.395	0.282	0.211
	Aneja et al. 2018 [38]	-	-	-	-	-
	Xiao et al. 2019[9]	0.653	0.468	0.329	0.229	0.190
MSCOCO	Karpathy et al. 2015[28]	0.625	0.450	0.321	0.230	0.195
	Mao et al. 2015 [25]	0.670	0.490	0.350	0.250	-
	Jia et al. 2015 [5]	0.670	0.491	0.358	0.264	0.227

	Wang et al. 2016[8]	0.672	0.492	0.352	0.244	-
	Gu et al. 2017 [33]	0.726	0.554	0.411	0.308	0.246
	Aneja et al. 2018 [38]	0.711	0.538	0.394	0.287	0.244
	Xiao et al.2019[9]	0.728	0.560	0.423	0.321	0.255

Table 1 shows the results of the most relevant models on Flickr8k, Flickr30k, MSCOCO datasets using evaluation metrics like BLEU-1,2,3,4 and METEOR. As the table suggests Mao et al.[25] performs significantly better on the MSCOCO dataset as compared to Flickr8k and Flick30k possibly due to the larger size of the dataset consisting of comprehensive representation of various scenes, more data, complexities and more. Jia et al.[5] also follows Mao et al.[25] and significantly does better on the MSCOCO dataset. Wang et al.[8] demonstrating the bidirectional LSTM approach however performs better on the Flickr8k dataset in comparison to the other models.Xiao et al.[9] consistently demonstrates the best or the second best performance on all three datasets on most of the evaluation metrics.

Table-2: Performance of encoder-decoder image captioning methods on MSCOCO dataset and commonly used evaluation metrics.(Bold indicates the best result; Underlined indicates the second best result).

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Jia et al. 2015[5]	0.670	0.491	0.358	0.264	0.2274	-	0.8125
Mao et al. 2015[25]	0.670	0.490	0.350	0.250	-	-	-
Wang et al. 2016 [8]	0.672	0.492	0.352	0.244	-	-	-
Xiao et al. 2019[9]	**0.731**	**0.563**	**0426**	**0.323**	**0.256**	**0.537**	**0.993**

Table 2 shows the results of the encoder-decoder models on MSCOCO datasets using evaluation metrics like BLEU-1,2,3,4,METEOR ,ROUGE-L and CIDEr. As the table suggests Xiao et al.[9] demonstrates the best performance on all the evaluation metrics followed by Wang et al.[8] performing well on BLEU-1 and BLEU-2 and Jia et al. [5]delivering the second best performance on BLEU-3 and BLEU-4.

Table-3: Performance of Language (LSTM) Based image captioning methods on MSCOCO dataset and commonly used evaluation metrics. (Bold indicates the best result; Underlined indicates the second best result).

METHOD	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Gu et al. 2017[33]	**0.726**	**0.554**	**0.411**	**0.303**	**0.246**	-	**0.961**
Aneja et al. 2018[38]	0.711	0.538	0.394	0.287	0.244	**0.522**	0.912

Table 3 shows the results of the Language Models(LSTM) models on MSCOCO datasets using evaluation metrics like BLEU-1,2,3,4,METEOR,ROUGE-L and CIDEr. As the table suggests Gu et al.[33] demonstrates the best performance on all the evaluation metrics.

Table-4: Performance of Multimodal Space image captioning methods on MSCOCO dataset and commonly used evaluation metrics.(Bold indicates the best result; Underlined indicates the second best result).

METHOD	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDEr**
Karpathy et al. 2015[28]	0.625	0.450	0.321	0.230	**0.195**	-	**0.660**
Mao et al. 2015[25]	**0.670**	**0.490**	**0.350**	**0.250**	-	-	-

Table 4 shows the results of the Multimodal Space models on MSCOCO datasets using evaluation metrics like BLEU-1,2,3,4,METEOR ,ROUGE-L and CIDEr. As the table suggests Mao et al.[25] demonstrates the best performance on all BLEU (1 through 4) evaluation metrics.

6. FUTURE RESEARCH DIRECTIONS

Despite remarkable progress shown in the automatic image captioning generation domain over the last few decades, there is still room for large scale improvement. Supervised Learning methods can generate novel captions however these generated captions are heavily dependent on the training sets which come from existing datasets. Thus, new open domain datasets can be an intriguing direction for future research in this area. Furthermore, these methods still often fail to correctly recognise several objects and interpret the relationships of these objects and attributes accurately. Additionally, the language models must be robust and sophisticated in order to generate captions that are not just novel but also syntactically correct all while accurately describing the relationships of the objects and attributes in the images. Existing methods are able to produce factual captions that may provide a brief description of the image, however, providing human-like novel captions is still a far fetched task as supervised learning methods depend on datasets that require copious amounts of labelled data. Thus, unsupervised learning and reinforcement learning techniques can be an interesting domain to further explore in the future.

7. CONCLUSION

In this paper, we have surveyed several deep learning-based image captioning methods. We have given a general overview of the various image captioning techniques developed over the years and then further elaborated on deep learning based techniques. General block diagrams of the framework of the major categories of deep learning methods are provided along with a few details of the most relevant models under each category. We have also elaborate on several of the most commonly used evaluation metrics and datasets. We have also provided a few of the results of the most relevant models on Flickr8k, Flickr30k and MSCOCO datasets.A brief discussion on the possible future research in this domain is also presented. Despite extensive research being carried out in this area, a model that is able to generate human-like captions for all images is yet to be developed. With continued improvements and development of several deep learning networks particularly unsupervised and reinforced learning models, image caption generation will be an area of ongoing research for quite some time.

REFERENCES

[1] Kiros R, Salakhutdinov R, Zemel R (2014) Multimodal neural language models. In: International conference on machine learning, pp 595–603

[2] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition. 3156–3164.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In the International Conference on Learning Representations (ICLR).

[4] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder

approaches. In Association for Computational Linguistics. 103–111.

[5] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE International Conference on Computer Vision. 2407–2415.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems. 1097–1105.

[6] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In the International Conference on Learning Representations (ICLR).

[7] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional LSTMs. In Proceedings of the 2016 ACM on Multimedia Conference. ACM, 988–997.

[8] X. Xiao, L. Wang, K. Ding, S. Xiang and C. Pan, "Deep Hierarchical Encoder–Decoder Network for Image Captioning," in IEEE Transactions on Multimedia, vol. 21, no. 11, pp. 2942-2956, Nov. 2019, doi: 10.1109/TMM.2019.2915033.

[9] Y. Yang, C. Teo, H. III, and Y. Aloimonos, “Corpus-guided sentence generation of natural images,” in Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 444–454, 2011.

[10] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.

[11] A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, Int. J. Comput. Vis. 42 (3) (2001) 145–175.

[12] G. Kulkarni et al., "BabyTalk: Understanding and Generating Simple Image Descriptions," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2891-2903, Dec. 2013, doi: 10.1109/TPAMI.2012.162.

[13] T. Dunning, Accurate methods for the statistics of surprise and coincidence, Comput. Linguist. 19 (1) (1993) 61–74.

[14] Y. Ushiku, M. Yamaguchi, Y. Mukuta, T. Harada, Common subspace for model and similarity: phrase learning for caption generation from images, in: IEEE International Conference on Computer Vision, 2015, pp. 2668–2676.

[15] Y. Ushiku, T. Harada, Y. Kuniyoshi, Efficient image annotation for automatic sentence generation, in: Proceedings of the Twentieth ACM International Conference on Multimedia, 2012.

[16] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Meeting on Association for Computational Linguistics, Vol. 4.

[17] C.-Y. Lin, F. J. Och, Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, in:Meeting on Association for Computational Linguistics, 2004.

[18] A. Lavie, A. Agarwal, Meteor: An automatic metric for mt. evaluation with improved correlation with human judgments, in: The Second Workshop on Statistical Machine Translation, 2007, pp. 228–231.

[19] R. Vedantam, C. L. Zitnick and D. Parikh, "CIDEr: Consensus-based image description evaluation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 4566-4575, doi: 10.1109/CVPR.2015.7299087.

[20] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. Journal of machine learning research 3, Feb, 1137–1155.

[21] Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In Proceedings of the 24th international conference on Machine learning. ACM, 641–648.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[23] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In Eleventh Annual Conference of the International Speech Communication Association.

[24] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In International Conference on Learning Representations (ICLR).

[25] Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In Proceedings of the 24th international conference on Machine learning (ICML '07). Association for Computing Machinery, New York, NY, USA, 641–648. DOI:https://doi.org/10.1145/1273496.1273577

[26] M. Schuster, K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (11) (1997) 2673–2681.

[27] A. Karpathy, F. Li, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3128–3137.

[28] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Netw. 5(5).

[29] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, S. Khudanpur, Recurrent neural network based language model, in: Proceedings of the Conference of the International Speech Communication Association, 2010, pp. 1045–1048.

[30] X. Chen, C. Zitnick, Mind's eye: a recurrent visual representation for image caption generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2422–2431.

[31] Sepp Hochreiter and JÃijrgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.

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

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

[34] Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. 2016. Conditional image generation with pixel CNN decoders. In Advances in Neural Information Processing Systems. 4790–4798.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems. 5998–6008.

[36] Jonas Gehring, Michael Auli, David Grangier, and Yann N Dauphin. 2016. A convolutional encoder model for neural machine translation. arXiv preprint arXiv:1611.02344.

[37] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. arXiv preprint arXiv:1705.03122.

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

[39] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3137–3146.

[40] D. Lin, An information-theoretic definition of similarity, in: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 296–304.

[41] J. Curran, S. Clark, J. Bos, Linguistically motivated large-scale nlp with cc and boxer, in: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 33–36.

[42] V. Ordonez, G. Kulkarni, T. L. Berg., Im2text: Describing images using 1 million captioned photographs, in: Advances in Neural Information Processing Systems, 2011, pp. 1143–1151.

[43] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: data, models and evaluation metrics, Journal of Artificial Intelligence Research 47 (2013) 853–899.

[44] F. R. Bach, M. I. Jordan, Kernel independent component analysis, Journal of Machine Learning Research 3 (2002) 1–48.

[45] D. R. Hardoon, S. R. Szedmak, J. R. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, Neural Computation 16 (2004) 2639–2664.