Image Captioning using Deep Learning: A Systematic Literature Review

Murk Chohan[1], Adil Khan[2], Muhammad Saleem Mahar[3]
Saif Hassan[4], Abdul Ghafoor[5], Mehmood Khan[6]
Department of Computer Science
Sukkur IBA University
Pakistan

***Abstract*—Auto Image captioning is defined as the process of generating captions or textual descriptions for images based on the contents of the image. It is a machine learning task that involves both natural language processing (for text generation) and computer vision (for understanding image contents). Auto image captioning is a very recent and growing research problem nowadays. Day by day various new methods are being introduced to achieve satisfactory results in this field. However, there are still lots of attention required to achieve results as good as a human. This study aims to find out in a systematic way that what different and recent methods and models are used for image captioning using deep learning? What methods are implemented to use those models? And what methods are more likely to give good results. For doing so we have performed a systematic literature review on recent studies from 2017 to 2019 from well-known databases (Scopus, Web of Sciences, IEEEXplore). We found a total of 61 prime studies relevant to the objective of this research. We found that CNN is used to understand image contents and find out objects in an image while RNN or LSTM is used for language generation. The most commonly used datasets are MS COCO used in all studies and flicker 8k and flicker 30k. The most commonly used evaluation matrix is BLEU (1 to 4) used in all studies. It is also found that LSTM with CNN has outperformed RNN with CNN. We found that the two most promising methods for implementing this model are Encoder Decoder, and attention mechanism and a combination of them can help in improving results to a good scale. This research provides a guideline and recommendation to researchers who want to contribute to auto image captioning.**

Keywords*—Image Captioning; Deep Learning; Neural Network; Recurrent Neural Network (RNN); Convolution Neural Network (CNN); Long Short Term Memory (LSTM)

I. INTRODUCTION

Auto image captioning is the process to automatically generate human like descriptions of the images. It is very dominant task with good practical and industrial significance [62]. Auto Image captioning has a good practical use in industry, security, surveillance, medical, agriculture and many more prime domains. It is not just very crucial but also very challenging task in computer vision [1]. Traditional object detection and image classification task just needed to identify objects within the image where the task of Auto image captioning is not just identifying the objects but also identifying the relationship between them and total scene understanding of the image. After understanding the scene it is also required to generate a human like description of that image. Since the boost of automation and Artificial Intelligence lots of research is going on to give machine human like capabilities and reduce manual work. For machines acquiring results and accuracy as good as human in image captioning problem has always been a very challenging task.

Auto image captioning is performed by following key tasks in order. At first features are extracted after proper extraction of features different objects from an image are detected, after that the relationship between objects are to be identified (i.e. if objects are cat and grass it is to be identified that if cat in on grass). Once objects are detected and relationships are identified now it is required to generate the text description, i.e. Sequence of words in orderly form that they make a good sentence according to the relationship between the image objects.

To perform above key tasks using deep learning different deep learning networks are used. For Example to get visual features and objects CNN with different region proposing models like RCNN, Faster RCNN can be used and to generate text description in sequence RNN or LSTM can be used. Using these networks various different methods are developed to perform auto image captioning in various different domains. However, still, there is room for the machine to make capable enough to generate descriptions like a human [61]. . After training the Deep Learning network for image captioning to evaluate its performance various evaluation matrices like BLEU, CIDEr, and ROUGE-L exists.

The purpose of this Systematic Literature Review is to study all newest Articles from 2017 to 2019 to find different methods to achieve auto image captioning in different domains, what different datasets are used to achieve the task, In which different practical domains this task is used, which technique Outperforms others and finally attains to describe the technicalities behind different networks, methods and evaluation matrices. Our study will help new researchers who want to work in this domain to attain better accuracy. We specially focused and the collection of quality articles which have been published till now. We attempt to find our different techniques presented in [1- 60] articles, find their methods strengths and weakness. Finally we attempt to summarize them to explain which technique has better performance in its particular domain. Our work mostly focuses on identifying the most popular techniques. The areas in which yet there is

attention require and in result section we also attempt to explain the technical concepts behind the used approaches.

II. METHODOLOGY

The planning conducting and reporting of this Systematic literature review is done step by step. First in planning section we identified the need of conducting this research its importance. Identifying the research questions and design search strategy, designing quality assessment criteria and finally designing data extraction strategy is also planned during this stage. After proper planning we have conducted the research. In alignment with our research problem we have come up with research questions for which we try to find answers during this research.

A. Research Questions

Before conducting this study we kept the following research questions to measure the quality of our work. This study basically provide a detailed knowledge related to these research questions. Table I provides the list of research questions.

TABLE I. LIST OF RESEARCH QUESTIONS

RQ#	Research Question	Motivation
RQ 1	How image captioning recognizes the important objects, attributes of objects and their relationships in an image?	Identifying DL techniques for object detection and relation finding mechanism
RQ 2	How Deep learning-based techniques are capable of handling the complexities and challenges of image captioning?	Identifying DL methods to handle challenges of image captioning
RQ 3	What deep learning techniques are used for image captioning?	Identifying DL techniques for Language generation as well as object detection
RQ 4	Which techniques outperform other techniques?	Comparison between several techniques
RQ 5	What datasets are used for Image Captioning?	Identifying different datasets used for image captioning
RQ 6	What evaluation mechanisms are used in literature for image captioning?	Identifying different methods to evaluate image captioning models

B. Search Results

According to our research questions we came up with our search keywords and we categorized them in two different groups, shown in the Table II.

Using scientific approach for searching the results from different academic databases. We composed the query string from the keywords cited in Table II.

Query String: ("Image Captioning") AND ("Deep Learning" OR "Neural Network" OR "RNN" OR "LSTM" OR "CNN")

We applied the cited search query string on three well known academic databases namely IEEE Xplore, Web of Sciences and Scopus to search the articles. We adopted the most recent articles published during 2017-2019 from the journals, and our initial search results are illustrated in the Table III.

TABLE II. KEYWORDS IN TWO DIFFERENT GROUPS (GROUP 1 AND GROUP 2)

Keywords: Group 1	Keywords: Group 2
Image Captioning	Deep Learning, Neural Network, RNN, CNN, LSTM

TABLE III. INITIAL STAGE RESULTS FROM IEEE XPLORE, WEB OF SCIENCES AND SCOPUS

Database	*Original search results*
IEEE Xplore	247
Web of Science	167
Scopus	313
Total	**727**

Since an article can be indexed in many databases we removed the duplicate articles from either one of the database. After duplicate removal total number of studies from all three databases are shown in Table IV.

Abstract screening is also important to filter the searched studies to keep valuable studies that are more related to someone's work. We performed abstract screening on the 577 articles which were remained after duplicate removal to check out the relevance of studies with our work. We found many studies not relevant to our topic like some were about audio captioning or video captioning. After the abstract screening, we had a total of 308 studies out of 577 studies, Table V illustrates the total number of studies from each database after the abstract screening.

TABLE IV. NUMBER OF STUDIES OF DUPLICATE REMOVAL

Database	*Duplicate removal results*
IEEE Xplore	162
Web of Science	167
Scopus	248
Total	**577**

TABLE V. NUMBER OF STUDIES AFTER ABSTRACT SCREENING RESULTS

Database	*Abstract screening results*
IEEE Xplore	63
Web of Science	92
Scopus	143
Total	**308**

C. Quality Assessment Criteria

The quality of 308 articles was assessed for quality assessment criteria. We assessed the quality of selected 308 studies to ensure the quality assessment of our study. We went through the full text screening of those studies which were ambiguous and was not clear from abstract screening. The process of quality assessment criteria (QAC) was done with full text screening. All four authors agreed to make some quality assessment questions (QAQ) to ensure the quality of our work.

QA Q1 The article must be published in journal
QA Q2 Article has proposed a proper method to implement image captioning using deep learning.
QA Q3 The article must have clear and unambiguous results.
QA Q4 Article must discuss the applications and challenges of image captioning.
QA Q5 Article must discuss the evaluation strategy of the built model.

We assessed the quality of 308 studies on the basis of quality assessment criteria (QAC) questions and through full text screening, we found total 61 studies from all three databases. Number of each studies from all three databases shown in Table VI.

The result which we found above illustrated in PRISMA diagram (see Fig. 1). All this process we this dissipated in following diagram.

D. Data Extraction and Synthesis

After selection of final 61 primary studies we extracted data from those studies for performing final synthesis. We defined our data extraction strategy based on our research questions. We have extracted following parameters from our primary studies for further synthesis, year or article published, title, models use for language generation and object detection, methods use to implement models, datasets used, evaluation matrices used for evaluation purpose and finally accuracy of proposed model.

The purpose of synthesis is to summarize the facts extracted in data extraction and give a clear picture of work done in past and directions to new researchers.

TABLE VI. NUMBER OF STUDIES AFTER QUALITY ASSESSMENT CRITERIA

Database	*Quality assessment results*
IEEE Xplore	12
Web of Science	16
Scopus	33
Total	**61**

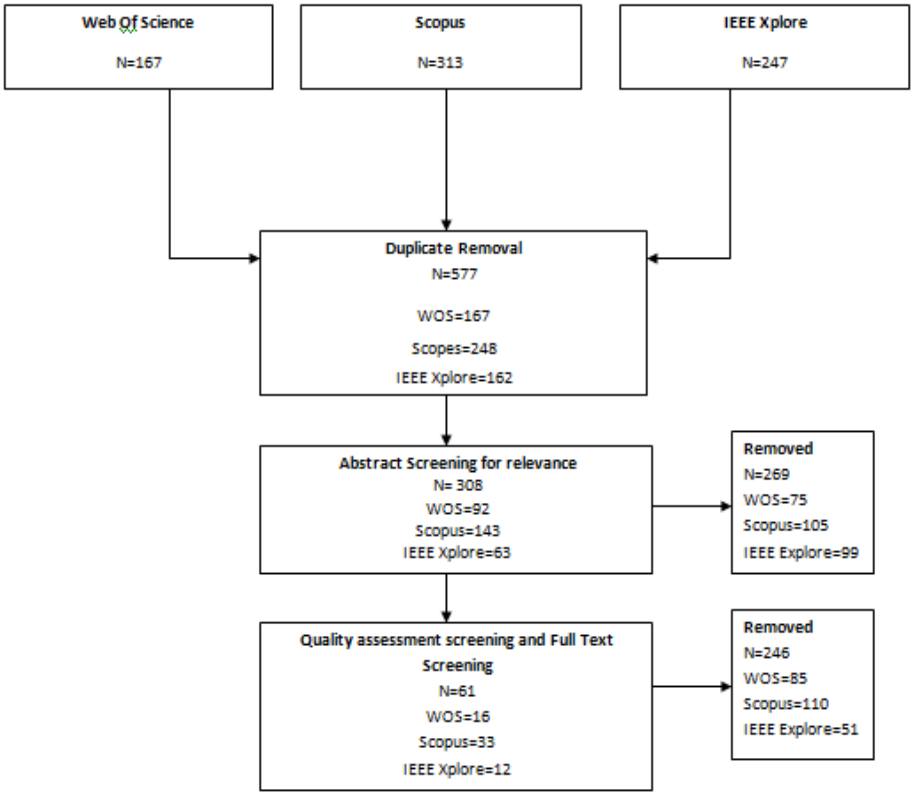

Fig. 1. PRISMA Diagram.

III. Results and Discussion

A. Datasets

There are many datasets available for performing image captioning. In literature most common used data sets are MS COCO and flicker 8k and 30k. Moreover for a text description of specific task like in medical or traffic movement description their own dedicated datasets are created. Fig. 2 below show the datasets along with their frequency in our selected studies.

1) MSCOCO: MS COCO stands for common object in context. It is very large dataset which contains 330k images, 1.5 object instances and 5 captions per image. MS COCO is found to be very widely used dataset in literature. It is very best suited for image captioning because unlike other datasets it contains non iconic images. Iconic images are those images which contains only one object with a background where as non-iconic images contains various objects overlapping. Object layout plays an important role in understanding context of scene and that is very carefully taken care of while labeling images. Fig. 3 shows some images taken from MS COCO dataset.

2) Deep learning networks: Deep learning network used for images is Convolution neural network. CNN has been proved best to map image data into output variable. There are various prebuilt model that take advantage from this feature of CNN i.e. RCNN faster RCNN etc. these models are used for object detection and localization in images which is very necessary task in image captioning since it's not just classification task and understanding image contents is necessary. Once image data is understood there is need of predicting the sequence of words to generate the text for that particular image. For sequence prediction two most famous networks are Recurrent Neural Network (RNN) and long short term memory (LSTM). For image captioning generation task CNN is either used with RNN or LSTM where CNN is used for understanding image contents and RNN or LSTM for text description generation. Fig. 4 and Table VII represents the number of studies that have used RNN or LSTM with CNN. Fig. 5 shows use of CNN and RNN networks for image captioning in year 2017 to 2019.

In terms of performance we Compared BLEU-1 performance of both text prediction networks and found out that LSTM outperforms RNN in terms of accuracy. Fig. 6 shows the result of top 5 highest accuracy achieving papers for both networks.

3) Convolution Neural Network (CNN): Convolution Neural Network is an algorithm of Deep Learning which is normally used to process images. CNN is an evolution of simple ANN that gives better result on images. Simple dense network is best for classification tasks where some features are used to classify the image. CNN performs best with more features in an image. It is used to process the local features as well. Because images contain repeating patterns of particular thing (any image). It takes images as an input and understands it to perform assigned tasks. Two main functions of CNN are convolution and pooling. Convolution is used in CNN to detect the edges of an image and pooling is used to reduce the size of an image. It is a method in which we take a small number matrix called kernel or filter then move it over our picture and convert it depending on the filter values. Following formula is used to calculate the feature map, where f is used to denote input image and h is used to denote filter. The outcome matrix rows and column indexes are labeled with m and n, respectively.

$$G[M,N] = (F * H)[M,N] = \sum J \sum K \ (H[J,K]F[M-J,N-K])$$

Calculation of convolution layer done in two steps. First step is used to calculate the intermediate value $\boldsymbol{Z}$, and its addition with bias. Second step is to apply non-linear activation $\boldsymbol{g}$ with intermediate value.

$$Z^{[i]} = W^{[i]} * A^{[i-1]} + b^{[i]}$$

$$A^{[i]} = g^{[i]}(Z^{[i]})$$

4) Recurrent Neural Network (RNN): CNNs commonly do not do well in a sequential fashion when the input data is interrelated. CNNs have no connection of any kind between previous input and next data. So all of the outputs depend on themselves. Depending on the trained model, CNN takes input and gives output. For doing above task RNN is used. RNN have its memory, so that it is able to remind what happened earlier in data. Earlier means previous inputs. RNN performs best on textual data because text is interrelated (sequential data).Basic formula for RNN is written below.

$$h(t) = f\left(h^{[t-1]}, x(t); \theta\right)$$

$\boldsymbol{f}$ is a function of current hidden state $\boldsymbol{h}$. $\boldsymbol{h^{(t-1)}}$ is a previous hidden state, $\boldsymbol{x(t)}$is current input, and $\square$ is a parameter of function.

5) Long Short Term Memory (LSTM): LSTM is a variant of RNN. It is better than simple RNN because it solves the issues faced by simple RNN. Two major issues faced by simple RNN is *(i)* exploding gradient and vanishing gradient and *(ii)* long term dependency. LSTM uses gates to remember the past and gates are the heart of LSTM. Gates which are available in LSTM are *(i)* input gate *(ii)* forget gate and *(iii)* output gate. They all are sigmoid activation function. Sigmoid means output between 0 and 1, mostly 0 or 1. When output is 0, it means gate is blocking. If output is 1 then pass everything. Below is the equation for above defined gates.

$$i_t = \partial(w_i[h_{t-1}, x_t] + b_i).$$

B. Evaluation Mechanism

Evaluating the trained model is quiet difficult task in image captioning for this purpose various evaluation matrices are created. Most common evaluation mechanisms found in literature are BLEU, ROUGE-L, CIDEr, METEOR, and

SPICE. It is found that BLEU score is most popular method of evaluation used by almost all of the studies. You can verify this from given Fig. 7 and Table VIII.

6) BELU: BLEU stands for bilingual evaluation understudy. It is an evaluation mechanism widely use in text generation. It is a mechanism for comparing the machine generated text with one or more manually written text. So basically it summarizes that how close a generated text is to an expected text. BLEU score is majorly prevalent in automated machine translation but it can be also used in image captioning, text summarization, speech recognition etc. Particularly in image captioning the BLEU score is accuracy that how close a generated caption is to a manual human generated caption of that particular image. The score scale lies between 0.0 to 1.0. Where 1.0 is perfect score and 0.0 is worst score.

We found that almost all studies used bleu as their evaluation matrix and they calculated BLEU-1 to 4 where BLEU-1 is calculating accuracy only on 1 gram, BLEU-2 for 2 grams, BLEU-3 for 3 grams and BLEU-4 for 4 grams.

The BLEU score can be calculated from following formula.

$$BLEU = BP.exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

7) METEOR: METEOR stands for metric for evaluation and translation with explicit ordering. While BLEU takes account of entire text generated overshadowing the score of each and individual sentence generated the METEOR takes care of that. For doing so METEOR enhances the precision and recall functions. Instead of precision and recall the meteor utilizes weighted F-score for mapping unigram and for incorrect word order it uses penalty function.

Formula for weighted function is:

$$F = \frac{PR}{\alpha P + (1-\alpha)R}$$

Where P and R stands for precision and recall calculated as m/c and m/r, where c and r are candidate and reference length and m is number of mapped unigrams among two texts.

Formula for Penalty function is:

$$Penality = \gamma \left(\frac{c}{m}\right)^{\beta}, where\ 0 \le \gamma \le 1$$

Where c is number of matched chunks and m is total number of matches.

Over all meteor score is found by:

$$M = (1 - Penalty) * F$$

8) ROUGE-L: ROUGE stands for recall oriented understudy for gisting evaluation. As clear from its name ROUGE is only based on recall but ROUGE-L is based on its F score which is harmonic mean of its precision and recall values. Following are the formulas for calculating precision, recall and F values

$$P = \frac{LCS(A,B)}{m} \ and \ R = \frac{LCS(A,B)}{n}$$

Here A and B are candidate and reference generated text and m and n are their lengths and LCS stands for longest common sequence since ROUGE-L depends on longest common sequence.

Now for calculating F their harmonic means are calculated.

$$F = \frac{(1 + b^2)RP}{R + b^2 P}$$

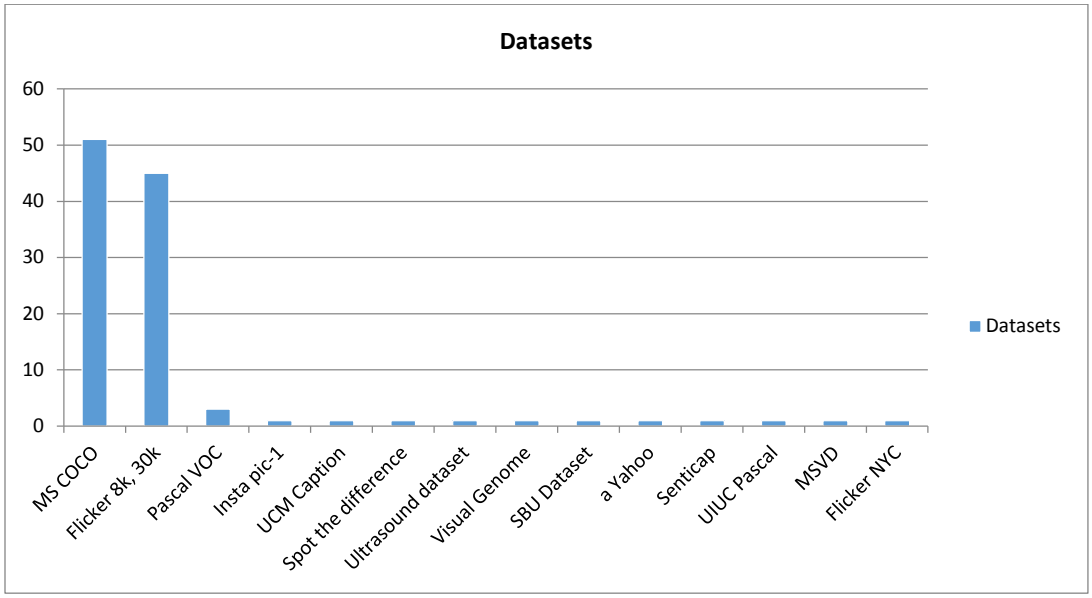

Fig. 2. Datasets used for Image Captioning in Selected Studies.

Fig. 3. MS COCO Dataset Images.

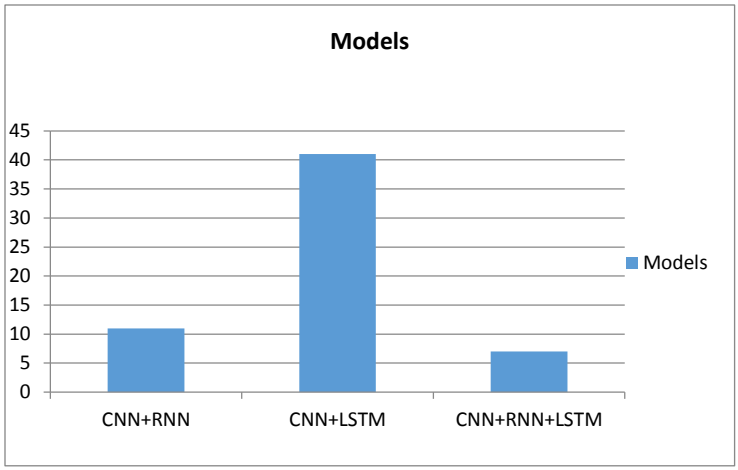

Fig. 4. Deep Learning Model used for Image Captioning in Literature.

TABLE VII. DEEP LEARNING MODEL USED FOR IMAGE CAPTIONING IN LITERATURE

SR#	STUDIES	NETWORKS	
		LSTM	RNN
1	[1],[2],[4],[6],[8],[10],[11],[13],[14],[15],[16],[18],[19],[20],[22],[23],[26],[27],[28],[29],[31],[32],[33],[36],[38],[37],[40],[41],[42],[43],[44],[45],[46],[47],[49],[50],[52],[53],[54],[57],[58]	✓	
2	[5],[7],[9],[12],[17],[30],[34],[35],[51],[56],[21]		✓
3	[18],[24],[25],[39],[48],[59],[60]	✓	✓

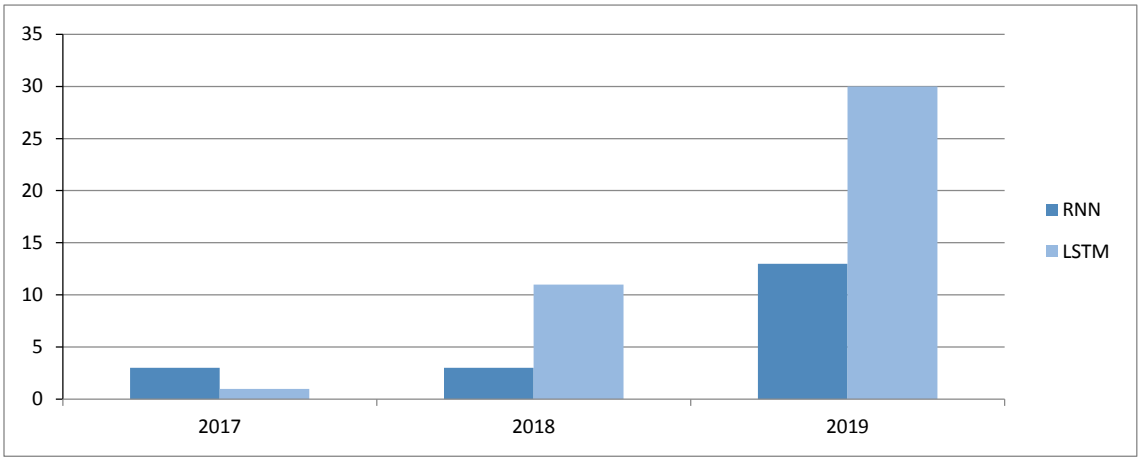

Fig. 5. Use of CNN and RNN Networks through the Years.

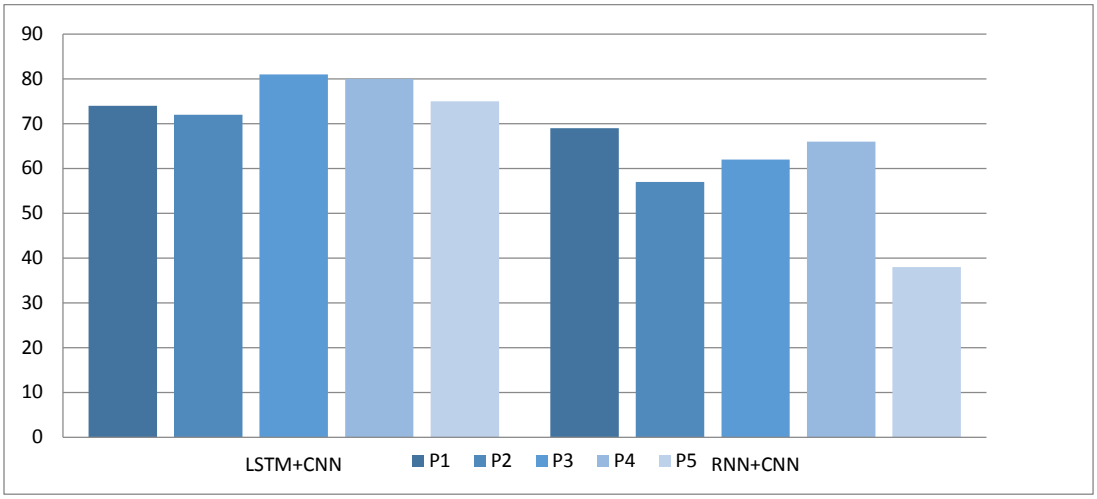

Fig. 6. Comparison of Best Score Achieved by RNN and LSTM (B1 Result Comparison).

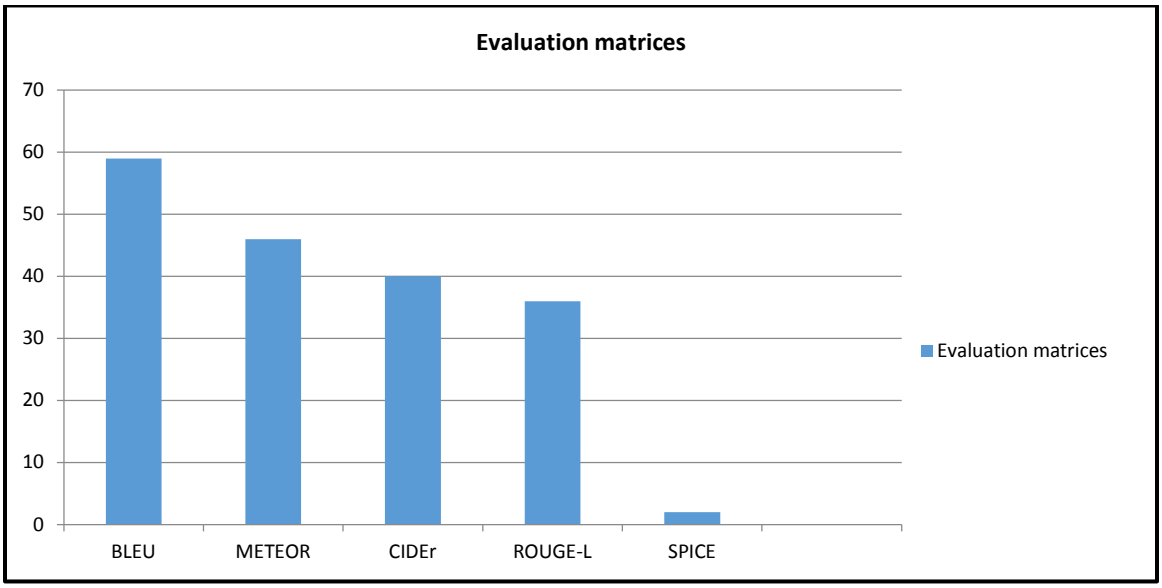

Fig. 7. Evaluation Metrics used in Literature.

TABLE VIII. EVALUATION METRICS USED IN LITERATURE

Sr.No.	STUDIES	EVALUATION METRICS				
		B	M	C	R	S
1	[1], [2], [3], [4] ,[6], [8], [10], [11], [13], [15], [16], [18], [20], [22], [26], [29], [30], [31], [32], [33], [38], [39], [40], [42], [43], [44], [49], [50], [52], [53], [54], [55], [56], [57],	✓	✓	✓	✓	
2	[5], [12], [19], [23], [34], [27], [28], [36], [37], [45], [51], [58], [59]	✓	✓	✓		
3	[7], [35]	✓	✓			
4	[9], [41], [47], [48]	✓				
5	[17]	✓	✓		✓	✓
6	[25]	✓	✓		✓	
7	[34]			✓		
8	[46]	✓		✓	✓	✓
9	[60]	✓	✓	✓	✓	✓

IV. CONCLUSION

This systematic literature review (SLR) presents a detailed analysis of different deep learning models used for image captioning. To perform the study we searched articles from three academic databases, after applying inclusion and exclusion criteria on all article and we selected 61 primary studies to perform a literature review. Using data extraction mechanism we extracted the data and analyzed it deeply. We found various different models and techniques used for image captioning. For image content extraction CNN is the best-suited model and for language generation two frequently used models are RNN and LSTM. It is found that LSTM has outperformed RNN. We also found different studies have used several different mechanisms for scene understanding like encoder-decoder mechanism and attention mechanism. The most suitable dataset for image captioning is MSCOCO because it contains non-iconic images, unlike other datasets.

Throughout our review, we have observed that image captioning is mostly used generally. There are various domains that can take advantage of image captioning to automate their tasks.

1) A model can be trained in medical ultrasound or MRI images or angiographic videos to generate a complete report of a person without any consent from a doctor. Image captioning can be used to generate an automatic report by looking at those medical images of a person.

2) Image captioning can also be used in industries to automate various tasks. A model can be trained on images of a company product manufacturing environment to find out an anomaly in the environment or product automatically. It can also be used also used to detect any mishap in a company like fire or security issues.

3) Image captioning can also be used in agriculture to generate the report of crops for owners by looking at images of crops.

4) Image captioning can also be used in traffic analysis report generation by using CCTV cameras installed on streets and thus guide drivers which is the best suitable path to take and where parking is available.

REFERENCES

[1] Yang, L., & Hu, H. (2019). Adaptive syncretic attention for constrained image captioning. Neural Processing Letters, 50(1), 549-564.

[2] Fu, K., Jin, J., Cui, R., Sha, F., & Zhang, C. (2016). Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. IEEE transactions on pattern analysis and machine intelligence, 39(12), 2321-2334.

[3] Li, J., Yao, P., Guo, L., & Zhang, W. (2019). Boosted Transformer for Image Captioning. Applied Sciences, 9(16), 3260.

[4] Oluwasanmi, A., Aftab, M. U., Alabdulkreem, E., Kumeda, B., Baagyere, E. Y., & Qin, Z. (2019). CaptionNet: Automatic end-to-end siamese difference captioning model with attention. IEEE Access, 7, 106773-106783.

[5] Wu, J., & Hu, H. (2017). Cascade recurrent neural network for image caption generation. Electronics Letters, 53(25), 1642-1643.

[6] Tan, J. H., Chan, C. S., & Chuah, J. H. (2019). COMIC: Toward A Compact Image Captioning Model With Attention. IEEE Transactions on Multimedia, 21(10), 2686-2696.

[7] Huang, G., & Hu, H. (2019). c-RNN: A Fine-Grained Language Model for Image Captioning. Neural Processing Letters, 49(2), 683-691.

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

[9] Dhir, R., Mishra, S. K., Saha, S., & Bhattacharyya, P. (2019). A Deep Attention based Framework for Image Caption Generation in Hindi Language. Computación y Sistemas, 23(3).

[10] Xiao, X., Wang, L., Ding, K., Xiang, S., & Pan, C. (2019). Deep Hierarchical Encoder–Decoder Network for Image Captioning. IEEE Transactions on Multimedia, 21(11), 2942-2956.

[11] Zeng, X., Wen, L., Liu, B., & Qi, X. (2019). Deep learning for ultrasound image caption generation based on object detection. Neurocomputing.

[12] Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).

[13] Xiao, X., Wang, L., Ding, K., Xiang, S., & Pan, C. (2019). Dense semantic embedding network for image captioning. Pattern Recognition, 90, 285-296.

[14] Han, M., Chen, W., & Moges, A. D. (2019). Fast image captioning using LSTM. Cluster Computing, 22(3), 6143-6155.

[15] Dash, S. K., Saha, S., Pakray, P., & Gelbukh, A. (2019). Generating image captions through multimodal embedding. Journal of Intelligent & Fuzzy Systems, 36(5), 4787-4796.

[16] Li, L., Tang, S., Zhang, Y., Deng, L., & Tian, Q. (2017). Gla: Global–local attention for image description. IEEE Transactions on Multimedia, 20(3), 726-737.

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

[18] Su, Y., Li, Y., Xu, N., & Liu, A. A. (2019). Hierarchical deep neural network for image captioning. Neural Processing Letters, 1-11.

[19] Zhang, Z., Wu, Q., Wang, Y., & Chen, F. (2018). High-quality image captioning with fine-grained and semantic-guided visual attention. IEEE Transactions on Multimedia, 21(7), 1681-1693.

[20] Shetty, R., Tavakoli, H. R., & Laaksonen, J. (2018). Image and video captioning with augmented neural architectures. IEEE MultiMedia, 25(2), 34-46.

[21] Dong, J., Li, X., & Snoek, C. G. (2018). Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia, 20(12), 3377-3388.

[22] Ding, S., Qu, S., Xi, Y., Sangaiah, A. K., & Wan, S. (2019). Image caption generation with high-level image features. Pattern Recognition Letters, 123, 89-95.

[23] Wu, Q., Shen, C., Wang, P., Dick, A., & van den Hengel, A. (2017). Image captioning and visual question answering based on attributes and external knowledge. IEEE transactions on pattern analysis and machine intelligence, 40(6), 1367-1381.

[24] Yang, J., Sun, Y., Liang, J., Ren, B., & Lai, S. H. (2019). Image captioning by incorporating affective concepts learned from both visual and textual components. Neurocomputing, 328, 56-68.

[25] Guo, R., Ma, S., & Han, Y. (2019). Image captioning: from structural tetrad to translated sentences. Multimedia Tools and Applications, 78(17), 24321-24346.

[26] Zhang, X., He, S., Song, X., Lau, R. W., Jiao, J., & Ye, Q. (2019). Image captioning via semantic element embedding. Neurocomputing.

[27] Cao, P., Yang, Z., Sun, L., Liang, Y., Yang, M. Q., & Guan, R. (2019). Image Captioning with Bidirectional Semantic Attention-Based Guiding of Long Short-Term Memory. Neural Processing Letters, 50(1), 103-119.

[28] Wang, C., Yang, H., & Meinel, C. (2018). Image captioning with deep bidirectional LSTMs and multi-task learning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(2s), 1-20.

[29] Chen, H., Ding, G., Lin, Z., Guo, Y., Shan, C., & Han, J. (2019). Image captioning with memorized knowledge. Cognitive Computation, 1-14.

[30] He, C., & Hu, H. (2019). Image captioning with text-based visual attention. Neural Processing Letters, 49(1), 177-185.

[31] Zhu, X., Li, L., Liu, J., Li, Z., Peng, H., & Niu, X. (2018). Image captioning with triple-attention and stack parallel LSTM. Neurocomputing, 319, 55-65.

[32] Guo, R., Ma, S., & Han, Y. (2019). Image captioning: from structural tetrad to translated sentences. Multimedia Tools and Applications, 78(17), 24321-24346.

[33] Fu, K., Li, J., Jin, J., & Zhang, C. (2018). Image-text surgery: Efficient concept learning in image captioning by generating pseudopairs. IEEE transactions on neural networks and learning systems, 29(12), 5910-5921.

[34] W.-Y. Lan, X.,-X. Wang, G, Yang, X.-R Li (2019) Improving Chinese Image Captioning by Tag Prediction

[35] Li, X., & Jiang, S. (2019). Know more say less: Image captioning based on scene graphs. IEEE Transactions on Multimedia, 21(8), 2117-2130.

[36] Chen, X., Zhang, M., Wang, Z., Zuo, L., Li, B., & Yang, Y. (2018). Leveraging unpaired out-of-domain data for image captioning. Pattern Recognition Letters.

[37] Fang, F., Wang, H., Chen, Y., & Tang, P. (2018). Looking deeper and transferring attention for image captioning. Multimedia Tools and Applications, 77(23), 31159-31175.

[38] Wang, E. K., Zhang, X., Wang, F., Wu, T. Y., & Chen, C. M. (2019). Multilayer dense attention model for image caption. IEEE Access, 7, 66358-66368.

[39] Xu, N., Zhang, H., Liu, A. A., Nie, W., Su, Y., Nie, J., & Zhang, Y. (2019). Multi-Level Policy and Reward-Based Deep Reinforcement Learning Framework for Image Captioning. IEEE Transactions on Multimedia.

[40] Yu, J., Li, J., Yu, Z., & Huang, Q. (2019). Multimodal transformer with multi-view visual representation for image captioning. IEEE Transactions on Circuits and Systems for Video Technology.

[41] Cui, W., Zhang, D., He, X., Yao, M., Wang, Z., Hao, Y., ... & Huang, J. (2019). Multi-Scale Remote Sensing Semantic Analysis Based on a Global Perspective. ISPRS International Journal of Geo-Information, 8(9), 417.

[42] Yang, M., Zhao, W., Xu, W., Feng, Y., Zhao, Z., Chen, X., & Lei, K. (2018). Multitask learning for cross-domain image captioning. IEEE Transactions on Multimedia, 21(4), 1047-1061.

[43] Ding, G., Chen, M., Zhao, S., Chen, H., Han, J., & Liu, Q. (2019). Neural image caption generation with weighted training and reference. Cognitive Computation, 11(6), 763-777.

[44] Su, J., Tang, J., Lu, Z., Han, X., & Zhang, H. (2019). A neural image captioning model with caption-to-images semantic constructor. Neurocomputing, 367, 144-151.

[45] Poleak, C., & Kwon, J. (2019). Parallel Image Captioning Using 2D Masked Convolution. Applied Sciences, 9(9), 1871.

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

[47] Wu, L., Xu, M., Wang, J., & Perry, S. (2019). Recall What You See Continually Using GridLSTM in Image Captioning. IEEE Transactions on Multimedia.

[48] Guan, J., & Wang, E. (2018). Repeated review based image captioning for image evidence review. Signal Processing: Image Communication, 63, 141-148.

[49] Xu, N., Liu, A. A., Liu, J., Nie, W., & Su, Y. (2019). Scene graph captioner: Image captioning based on structural visual representation. Journal of Visual Communication and Image Representation, 58, 477-485.

[50] Xian, Y., & Tian, Y. (2019). Self-Guiding Multimodal LSTM—When We Do Not Have a Perfect Training Dataset for Image Captioning. IEEE Transactions on Image Processing, 28(11), 5241-5252.

[51] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2016). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence, 39(4), 652-663.

[52] Ding, S., Qu, S., Xi, Y., & Wan, S. (2019). Stimulus-driven and concept-driven analysis for image caption generation. Neurocomputing.

[53] Lu, S., Hu, R., Liu, J., Guo, L., & Zheng, F. (2019). Structure Preserving Convolutional Attention for Image Captioning. Applied Sciences, 9(14), 2888.

[54] Yu, N., Hu, X., Song, B., Yang, J., & Zhang, J. (2018). Topic-oriented image captioning based on order-embedding. IEEE Transactions on Image Processing, 28(6), 2743-2754.

[55] Park, C. C., Kim, B., & Kim, G. (2018). Towards personalized image captioning via multimodal memory networks. IEEE transactions on pattern analysis and machine intelligence, 41(4), 999-1012.

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

[57] Zhang, Z., Zhang, W., Diao, W., Yan, M., Gao, X., & Sun, X. (2019). VAA: Visual Aligning Attention Model for Remote Sensing Image Captioning. IEEE Access, 7, 137355-137364.

[58] He, X., Yang, Y., Shi, B., & Bai, X. (2019). VD-SAN: Visual-densely semantic attention network for image caption generation. Neurocomputing, 328, 48-55.

[59] Li, X., Yuan, A., & Lu, X. (2019). Vision-to-language tasks based on attributes and attention mechanism. IEEE transactions on cybernetics.

[60] Yang, L., & Hu, H. (2019). Visual Skeleton and Reparative Attention for Part-of-Speech image captioning system. Computer Vision and Image Understanding, 189, 102819.

[61] Staniūtė, R., & Šešok, D. (2019). A Systematic Literature Review on Image Captioning. Applied Sciences, 9(10), 2024.