IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 10, Issue.12, December 2022 E-ISSN: 2347-2693

A Comprehensive Survey and Comparison on Story Construction Techniques Using Deep Learning for Scene Recognition

Darapu Uma[1*], M. Kamala Kumari[2]

[1]Department of CSE, Pragati Engineering College, Surampalem, A.P, India
[2]Department of CSE, Adikavi Nannaya University, Rajamahendravaram, A.P, India

Corresponding Author: umadarapu03@gmail.com, Tel.: +919951634504

DOI: https://doi.org/10.26438/ijcse/v10i12.1422 | Available online at: www.ijcseonline.org

Received: 29/Oct/2022, Accepted: 01/Dec/2022, Published: 31/Dec/2022

***Abstract*—** Story construction from deep learning is a naive methodology suitable for the digital forensics, smart video surveillance, and intelligent robotics applications. So far deep learning has been utilized for image recognition and classifications. The sequence of those images and classification of them on a temporal basis leads to the development of knowledge on the entire changes in the scenes and finally end up with a story in which the identified scenes are connected with unambiguous changes. This paper retrieves the pros and cons of the research work on recurrent topic transition GAN for Visual Paragraph Generation and relation pair visual paragraph generation. The Existing algorithms proposed for constructing the stories are RTT GAN, RP GAN and BF GAN. These are implemented with respect to different applications human computer interaction, intelligence robotics, digital forensics etc. The survey of the subjected algorithms gives the transparency of their working principles. The current paper presented the visual representation, description, generation of paragraphs with various methodologies along with the comparison.

Keywords—Semantic Region, Attention Module, Discriminator, Scene recognition, Visual features, Generative Adversarial Network

I. Introduction

The present environment of image classification and recognition demand in scene recognition and also a story to be constructed on top of it. Now a day's scene recognition is an essential topic in various applications like digital forensics and intelligent robotics. Digital forensic is one of the branch of forensic science to get the evidence from a crime in a digital manner. For example if any crime happened in a place based on the series of scenes it can detect details of the crime happened and narrate a story behind the scene without any human intervention. Similarly the service of robots in current environment increasing day by day. In this environment robots are recognize objects in the scene and performed specific actions from the input sensors. This paper surveyed on story construction techniques using deep learning. Deep learning is a part of artificial intelligence in which data processing is done through multiple layers in order to extract high level features .In deep learning models like convolution neural networks and recurrent neural networks are used [25].

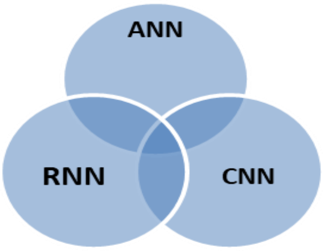

Figure 1.Relation between ANN, CNN and RNN

The above Figure 1 shows that relation among Convolutional neural network (CNN), Artificial neural network (ANN) and Recurrent Neural Networks (RNN) [20].All the neural networks categorized under the neural networks algorithms. Some of the functionalities are same among them. CNN and RNN utilize fixed length input and feed forward recursion. Parameter sharing between RNN and CNN. There is no special relationship between ANN and CNN. Convolution Neural Network (CNN) is a part of Artificial Neural Network (ANN) mainly used for recognition of objects or image [3]. The main aim of CNN is used to reduce the dimensionality of the given input image. Recurrent Neural Network (RNN) is also one of the type neural networks, used for memory usage for short term. RNN is give data back to itself [12]. This paper surveying on paragraph generation models that focus on word RNN, Sentence RNN and Paragraph RNN from [1], [2].

I.I. Definitions

A Scene is a place whereat an individual or objects can act within or navigate. Recognition is a process of identifying an object or a feature. A Story is a series of connected events that written over words either written or spoken, imagery both static and moving objects, body language, performance, music, or any other beget of intercommunication. Story Construction is a narration of story at various scenes in a place.

I.II. Scene Recognition

Scene recognition is an essential description of the picture or image as opposed to list of items with inside the scene [14],[20]. It is used in various applications like Intelligent Robotics, Computer Interaction through human, Autonomous Driving, and Smart Video Surveillance and Digital Forensics. It is beneficial to lessen than the distance among computer systems and people whilst obtaining knowledge for a scene. It is applied to computer vision tasks like description of image and detection of objects. The main aim for scene recognition is to assign the semantic names or marks to the images [20]. Semantic names or marks are characterized with the aid of using humans which include scenes both indoor and outdoor environments and etc. Based on the scene, the above Figure 2 shows the scene recognition of various images.

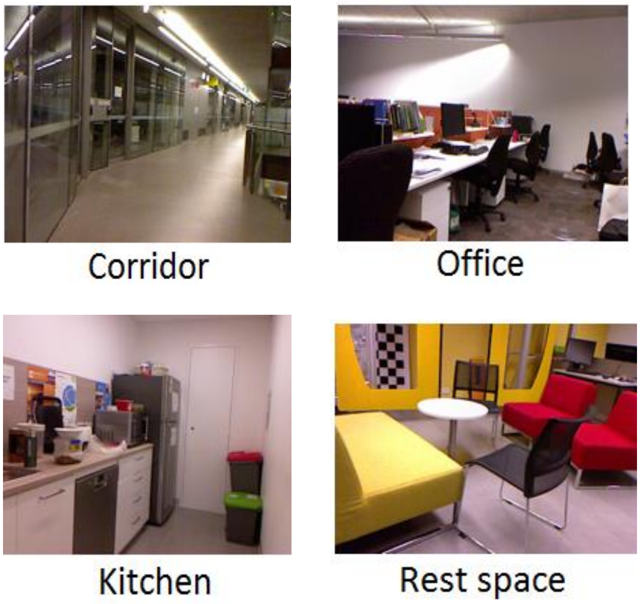

Figure 2.Different Scene Recognition Pictures

Based on the images or objects in a scene it gives the description. In the above figure it differentiates the scene like office, kitchen, rest space and corridor etc. From the given scenes it detects image into various regions with semantic labels and gives description of the image in the form words, short text phrases, sentences and paragraphs.

II. Related Work

This section reviews literature on story construction techniques using deep learning for scene recognition that are related to our work. Now discuss about various paragraph generating techniques.

II.I. Recurrent Topic Transition GAN [1]

This technique produces the coherent paragraph description by applying a method called as Recurrent Topic Transition General Adversarial Network (RTT-GAN)[1].Previous methods failed to cover proper meaning of the generated paragraphs.

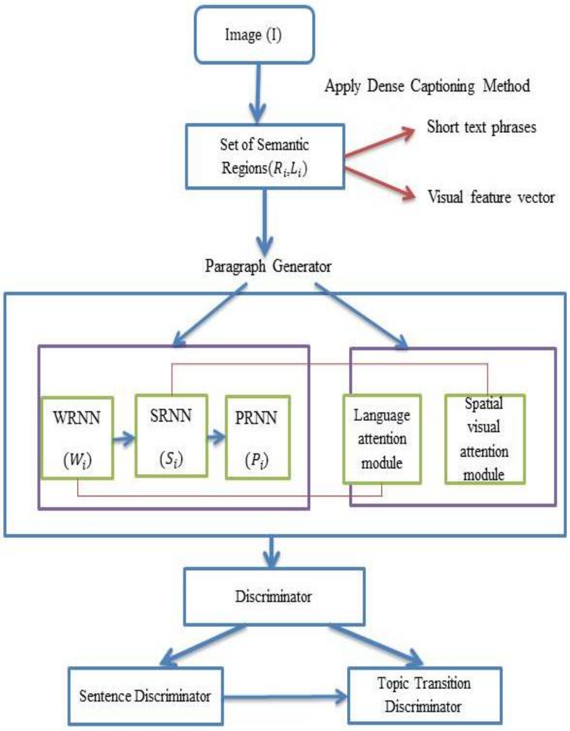
Image (I)
Apply Dense Captioning Method
Set of Semantic Regions(R_i,L_i)
Short text phrases
Visual feature vector
Paragraph Generator
WRNN (W_i)
SRNN (S_i)
PRNN (P_i)
Language attention module
Spatial visual attention module
Discriminator
Sentence Discriminator
Topic Transition Discriminator

Figure 3.Recurrent Topic Transition GAN

From the above Figure 3, recurrent topic transition is a combination of paragraph generator and paragraph discriminator. From the diagram, the input image (I) can be divided into set of meaning full regions ((R_i,L_i) by applying Dense Captioning Method (DCM)[1],[2].Dense Captioning Method(DCM) is used to detect visual features from the image and labelling each feature with a short descriptive phase. Regions are a combination of both visual feature vector and short text phrases. The maximum number of regions used is 50 regions [1]. These labelled regions will key in to the paragraph generator. Paragraph generator frequently generates sentences in a region .Paragraph generator can be sub divided into three level of Recurrent Neural Networks (RNN) groups and two attention modules. The three levels of recurrent neural networks are listed below.

1. Word RNN (WRNN)
2. Sentence RNN (SRNN)
3. Paragraph RNN (PRNN)

And two attention modules are given below

i. Language attention module.
ii. Spatial visual attention module.

Now see the how the recurrent neural networks works

II.I.I. Word RNN (WRNN)

Word RNN merged with a language attention module that describes the local phases of the semantic region and word RNN generates words [1], [2]. In this paper the maximum number of words generated is 30 words is given in Eq (1)

$$W_i=\{W_1, W_2, W_3.....W_{30}\} \quad (1)$$

Where W_i is the number of words generated by the paragraph generator.

Here i=30, maximum number of words is 30.

II.I.II. SENTENCE RNN (SRNN)

Sentence RNN merged with a module called a spatial visual attention module that describes the focus on the semantic region[1].It's responsible for counting number of sentences and additionally that sentences must be within the paragrap h and creating a topic vector of each sentence. Sentence RNN controls the next sentences by selecting the appropriate region of the image. The maximum number of sentences generated is six sentences. Each sentence is a combination of words with in the region

$$S_t = \{W_{t,i}\} \qquad (2)$$

Where S_t is combination of words in the region i

$$S_i = \{(S_1, S_2, \ldots\ldots S_6\} \qquad (3)$$

Where S_i is the number of generated sentences by paragraph generator. Here i=6, maximum number of sentences.

II.I.III. PARAGRAPH RNN (PRNN)

The generated sentences in the previous step will be act as input to the paragraph RNN used in [1],[2].It produces paragraph hidden state as output. Each paragraph contains sentences of respective regions, this can be shown by

From Eq (1), (2) and (3)

$$P_t(S_t|S_{t-1}, R) = \pi P_t(W_{t,i}|W_{t,1:i-1}, S_{1:t-1}, R) \qquad (4)$$

Where P_t is paragraph generated at 't' sentences within region 'R'

$W_{t,i}$ is combination of words

S_t is sentence generator

The generated paragraph will be given to discriminators to check the differences between the generated sentences and making sentences. there are two discriminators used in[1].first discriminator is sentence discriminator(SD) which is used to check the individual sentences and second discriminator is topic transition discriminator (TTD)which is used to collect the individual sentences from sentences discriminator as input and generate coherent paragraph description as output. In order to generate individual sentences Sentence discriminator (SD) used a mechanism known as Long Short term Memory Recurrent Neural Network (LSTM-RNN).In order to get constructed paragraph description topic transition discriminator (TTD) is also used a mechanism called Long Short term Memory Recurrent Neural Network (LSTM-RNN).From [1], the input is Image and output is paragraph description. The merit of this method is Paragraph description with semantic manner and demerit is only first sentences annotations are given for training.

II.II. RELATION PAIR GENERAL ADVERSARIAL NETWORK [2]

Paragraph Generation used for the given input image establish a relation in a subject and object relation ,this produces the coherent paragraph description with subject and object relation by applying a method called as Relation Pair General Adversarial Network (RP-GAN) [2].previous methods, relationship among visual objects employed in an unexpressed manner. From the below diagram can be divided into three parts.

1. Relation pair detection
2. Visual relationships
3. Caption generation

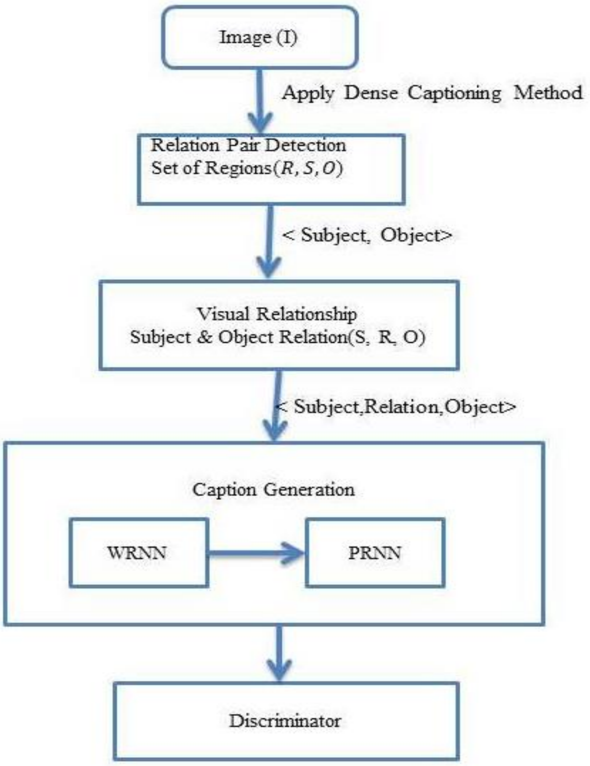

Figure 4. Relation Pair General Adversarial Network (RP-GAN)

II.II.I. RELATION PAIR DETECTION

For the given input image, by applying dense captioning method it divided into regions and also labels with subject and object combination like <subject, object> pair. From the set of regions R, in Relation Pair method classified as subjects (R_{sbj}) and objects (R_{obj}) and these subjects and objects should be less than the regions, R [2].

$$R_{sbj} < R \text{ and } R_{obj} < R \qquad (5)$$

Where R_{sbj} is subjects in the regions

R_{obj} is objects in the region

R is region

II.II.II. VISUAL RELATIONSHIPS

After getting <subject, object> pair from Eq (5) it predict the relation between two regions by applying attention based network mechanism. After this step it generates <subject, relation, object> pair. For every region R it generates visual feature maps and performs Union operation to get minimum number of regions. For region a, b it generates V_a and V_b feature maps, after performing union operation it produces $V_{a,b}$.Here it detects the subject and objects with boxes and perform cross multiplication to get valid relationships. The subject and object boxes pair shown below

(D_{sbj}, D_{obj})

$D_{sbj} \times D_{obj}$ → Valid relationship between subject and object

Where D_{sbj} detecting subject boxes

D_{obj} Detecting object boxes

II.II.III. Caption generation

In this step it converts visual relational features into language by applying long short term memory (LSTM) technique. From [2], it used two types of LSTM techniques. LSTM is a type of recurrent neural network (RNN),is adequate of holding long term storage of data. These two LSTM techniques are listed below.

i. Sentence LSTM

ii. Word LSTM

II.II.III.I. Sentence LSTM

One valid relationship generates one sentence .it checks the number of sentences in generated paragraph.

II.II.III.II. Word LSTM

It generates number of words from the given input vector. Finally the generated paragraph will be given to discriminator. Discriminator check whether the generated paragraph is valid sentences or not. The input is image and output is paragraph. The merits of this technique is Paragraph with proper semantics [7].To got accurate information and demerits is need more robustness to detection of errors in objects because it uses both visual and relational features in order to generate sentences.

II.III. Context Aware Visual Policy Network (CAVPN) and Language Policy Network (LP) [3]

Image Captioning studied used to produce context aware features from the context aware visual policy network (CAVPN) generates gentle paragraph description from the input image [3]. Previous techniques concentrate only on language not for visual features.

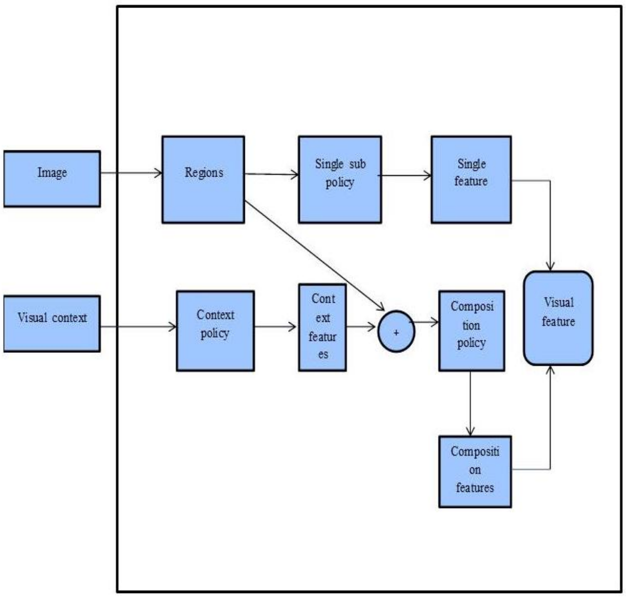

Figure 5. Context aware visual policy network (CAVPN)

From Figure 5, it has two networks one is network of context aware visual policy and network of language policy. Now discuss about each network in detail.

II.III.I. Context aware visual policy network (CAVPN) [3]

From given input image can be divided into regions by applying region policy mechanism. Context aware visual policy network (CAVPN) can be divided into four sub policies[3]. All these sub polices are generating gentle paragraph description Let us see each one in detail.

1. Single sub policy network
2. Context sub policy network
3. Composition sub policy network
4. Output sub policy network

II.III.I.I Single sub policy network

This network extracts features from the regions collected from the given input image. The features came from this its single sub policy network.

The input features at each time step ‘t’ are the detected regions, this can be shown below Eq (6)

$$v_t^{si} = f\,e_t^{si} = SIP^{si}(s_t^{si}, Q_t^{si}) \qquad (6)$$

Where $Q_t^{si} = \{r_1, r_2 \ldots r_n\}$ set of regions

$v_t^{si} = f\,e_t^{si}$ is a sub policy network with single feature at time t

II.III.I.II. Context sub policy network

At each step the time‘t’ every visual context having hidden visual outputs. But every context word may not be suitable at present word generation.so this can be overcome by context sub policy network. This sub policy network may select only the appropriate situation information and integrated with identified region.

From Eq (6), the result of context aware sub policy network is calculated as

$$ct_{t,n} = W_{ct}^{T}[f_t^{ct} ; r_n] \qquad (7)$$

$where\ f_t^{ct}$ is visual features context representation at time t

$ct_{t,n}$ is context aware sub policy network

r_n is nth regions

II.III.I.III. Composition sub policy network

This network takes previous hidden state and means pooled features and produces the composite features. This sub policy is same as to single sub policy network.

$$v_t^c = f\,e_t^o = SIP^c(s_t^c, Q_t^c) \qquad (8)$$

$where\ Q_t^c = \{c_{t,1}, c_{t,2} \ldots . c_{t,k}\}$

v_t^c is composite features

$f\,e_t^o$ is output sub policy features at time t

II.III.I.IV. Output sub policy network

After getting single, composite and context sub policy layers it generates visual output features. It can be shown below

$$f\,e_t^o = SIP^c(s_t^o, Q_t^o) \qquad (9)$$

$where\ Q_t^o = \{v_t^{si}\ \ v_t^c\ ,\ddot{r}\}$ from Eq (7) and (9)

$f\ e_t^o$ is features in output sub policy network at time t

II.III.II. Language Policy Network (LP)

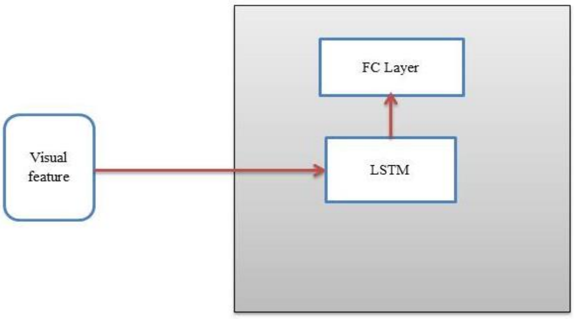

Figure 6. Language policy network

From Figure 6 it is LP network, it produces coherent image description [3]. The visual features came from context aware visual policy network that will act as an input to language policy network. In this network it uses long short term memory (LSTM) [4]. In order to update hidden state and it uses soft max function [3].To get the probability of the each word. LSTM is used also recurrent neural network to hold the data in long term. Soft max is softer version of the max function, and FC layer is a fully connected layer used to calculate output class score.

II.III.III. Hierarchical Representation of CAVPN & LP

The CAVPN and LP networks can also be represented by hierarchical manner [3].These are

i. Sentence level CAVPN and Sentence level LP
ii. Word level CAVPN and Word level LP

II.III.III.I. Sentence level CAVPN and Sentence level LP

Figure 7 describes the visual features of the context aware visual policy network (CAVPN) at sentence level individually given to input to language policy network at sentence level.

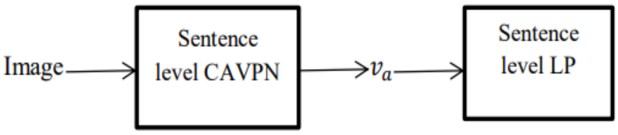

Figure 7. Sentence level CAVPN

Here v_a is the v_aCAVPN(R, { $v_1,\ v_2 \ldots\ldots v_{a-1}$}

Where b=1, 2,….n_s number of sentences and R is the region.

II.III.III.II. Word level CAVPN and Word level LP

Similarly Figure 8 describes the visual features of the context aware visual policy network (CAVPN) at word level individually given to input to language policy network at word level.

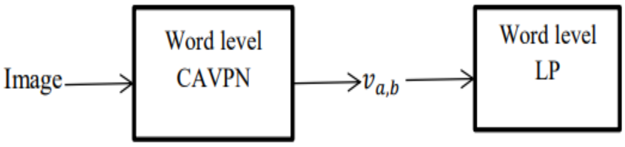

Figure 8. Word level CAVPN

Here $v_{a,b}$ is the $v_{a,b}$= CAVPN(R, { $v_1,\ v_2 \ldots\ldots v_{a-1}$}

Where b=1, 2, …. n_w are number of words.

$v_{a,b}$ is Visual representation of bth word from ath sentences and R is region

The input is image and output is Coherent image description and merit of this method is got best performance by applying image captioning method and demerit is it is applicable to only sequential decision making tasks [3].

II.IV. Backward and Forward Generative Adversarial Networks (BFGAN) [4]

In Sentence Generation is of backward and forward adversarial networks proposed a method Backward and Forward Generative Adversarial Networks (BFGAN) to impose the lexically constrained sentences [4].In previous methods it is very difficult to impose lexically constrained sentences. From the below diagram Figure 9, Backward and Forward Generative Adversarial Networks (BFGAN) worked on these three phases [4].

1. Backward Generator
2. Forward Generator
3. Discriminator

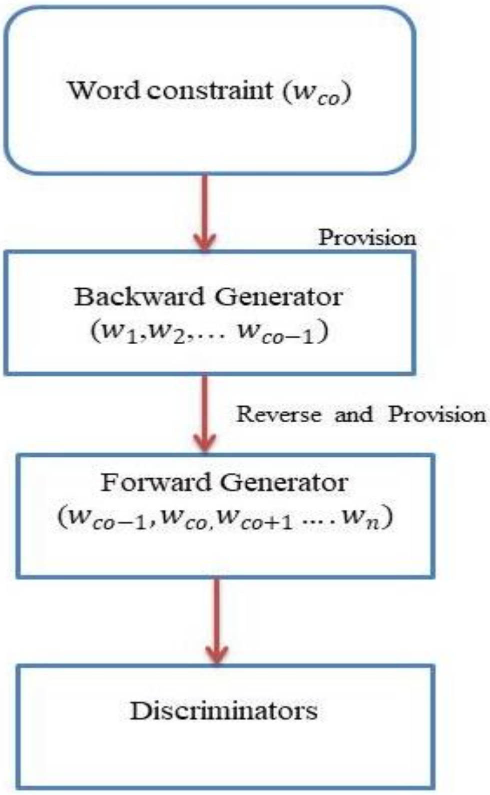
Word constraint (w_{co})
Provision
Backward Generator (w_1,w_2,… w_{co-1})
Reverse and Provision
Forward Generator (w_{co-1},w_{co},w_{co+1} …. w_n)
Discriminators

Figure 9. Backward and Forward Generative Adversarial Networks (BFGAN)

II.IV.I. Backward Generator

Here the input is a word constraint. The word constraint may be a word or phrase or a sentence [4]. That constraint is given to backward generator that takes it as sentences starting point and generated first half sentences backwards. The word constraints Eq (10) is given below

$$(w_1, w_2, \dots w_{co-1}) \qquad (10)$$

1st half sentences backward

$s_{<c} = w_{co-1} \dots\dots w_1$

Where $s_{<c}$ is first half sentences and c is the constraint

II.IV.II. Forward Generator

The sentences generated by backward generator are reverse and give provision to forward generator. It generates the complete sentences with the help of discriminator.

$$w_{co-1}, w_{co}, w_{co+1} \dots . w_n \qquad (11)$$

2nd half sentences forward

$s_{>c} = w_{co+1} \dots\dots w_1$

Where $s_{>c}$ is second half sentence and c is the constraint

By combining Eq (10) and (11), it generates sentences

II.IV.III. Discriminator

Discriminator is used to identify the difference between original sentences and machine generated sentences studied in [1]. The responsibility of discriminator is to combining of two generators by assigning a reward to them [4]. To improve the balanced constrained sentences, long short term memory with dynamic attention mechanism (attRNN-LM) is used as generator [4]. The input of this technique is a word/phrase/sentence and output is Quality of lexically constrained sentence generation. The merit of this is generating lexically constrained sentences and demerit is failed to generate sequences with multiple lexical constraints.

Now compare all the above four methods or techniques with their input and respective outputs. This can be shown in Table 1.This paper surveying on various general adversarial networks based on different language metrics with their performance. This can be shown in Table 2.In this paper it shows five different language metrics METEOR(Metric for Evaluation of Translation with Explicit Ordering), BLEU (Bi Lingual Evaluation Understudy)1, BLEU 2, BLEU 3 and BLEU 4 [3].The language metrics are used for translation of machine to text. These are the metrics used to evaluate text from machine. In this paper surveying sentence or paragraph generation that are applied on the above mentioned language metrics like METEOR,BLEU 1,BLEU 2,BLEU 3 and BLEU 4 are evaluating a generated sentence to a reference sentence. Each method or technique produces paragraph as output by applying various techniques with some attention mechanism or modules.

Table 1 Comparison of four methods with their inputs and outputs

S.No	**Input**	**Technique/Method Used**	**Output**
1	Image	RTT GAN LSTM + attention module	Coherent Paragraph Generation
2	Image	RP GAN-LSTM	Coherent Paragraph Generation
3	Image	CAVPN&LP-LSTM-FC Layer	Coherent Paragraph Generation
4	Word/Phrase	BFGAN	lexically constrained sentence generation

The above Table 1 represents comparison among various paragraph generation techniques [1]. All the four methods consider the input as different types of images with equal dimensions .It describes the inputs of the each method and techniques used in that method and outputs of the methods. In this literature review, all the four methods discussed above are generating coherent paragraphs and sentences. Now this paper also survey the all the four techniques RTT GAN,RP GAN,CAVPN&LP and BF GAN with their accuracy on five different language metrics this can be shown in table 2 [1],[2],[3],[4].

Table 2 Comparison of four methods with their Accuracy on various language metrics

S.No	**Technique/ Method**	**Accuracy on METEOR**	**Accuracy on BLEU 1**	**Accuracy on BLEU 2**	**Accuracy on BLEU 3**	**Accuracy on BLEU 4**
1	[1]	18.39	42.06	25.35	14.92	9.21
2	[2]	17.4	41.94	24.99	15.01	9.38
3	[3]	17.14	42.01	25.86	15.33	9.26
4	[4]	--	--	17.1	10.8	6.9

From Table 2 the high accuracy rate on METEOR language metric is in RTT GAN compared to remaining techniques [2]. Similarly the accuracy on BLEU 1 language metric is in RTT GAN [1]. Similarly for BLEU 2 CAVPN&LP had high accuracy rate [3].For BLEU 3 is CAVPN&LP [3].Finally for BLEU 4 the highest accuracy rate for paragraph generation is on method RP GAN [2].

III. Result Outcomes

Now discuss about the various comparison graphs of the four techniques on the different language metrics like METEOR, BLEU 1, BLEU 2, BLEU 3, BLEU 4 [1],[2],[3]. Table 2 shows different accuracy values and those values had presented in the graphs. All the graphs shown below are two dimensional graphs where the x-axis denotes the four methods listed in the table 2 and y-axis denotes the accuracy values of different language metrics. This paper exposed five graphs, each graph drawn from all the five language metrics. Each method or technique performs well on any one of the metric.

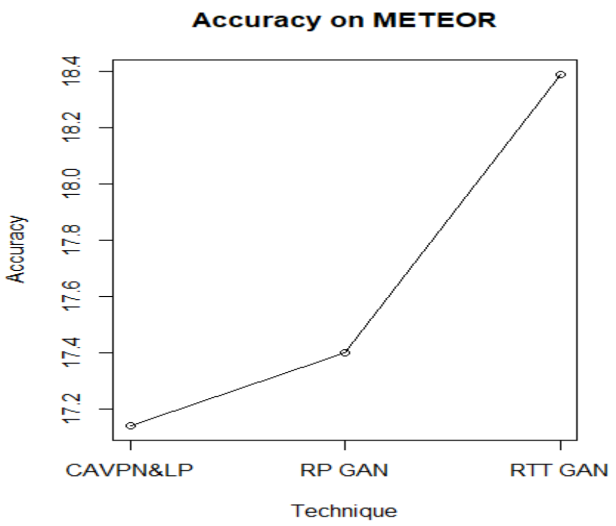

Figure 10. Comparison graph for METEOR language metric

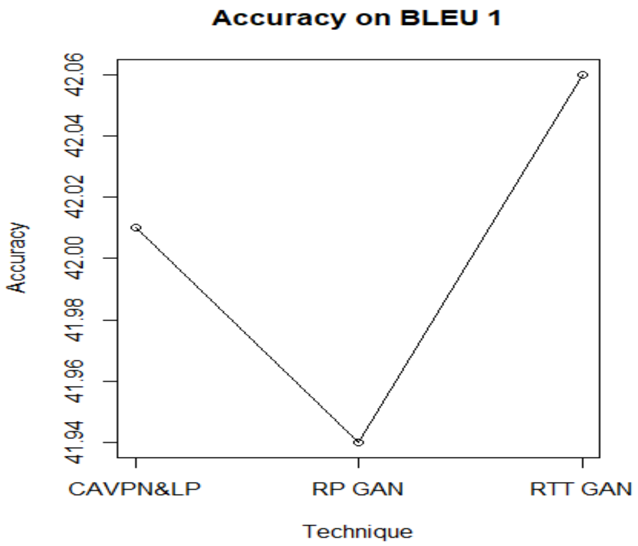

Figure 11. Comparison graph for BLEU 1 language metric

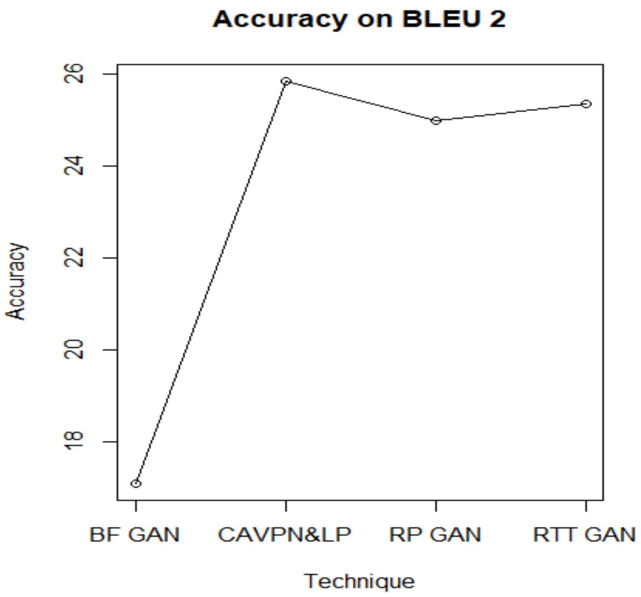

Figure 13. Comparison graph for BLEU 2 language metric

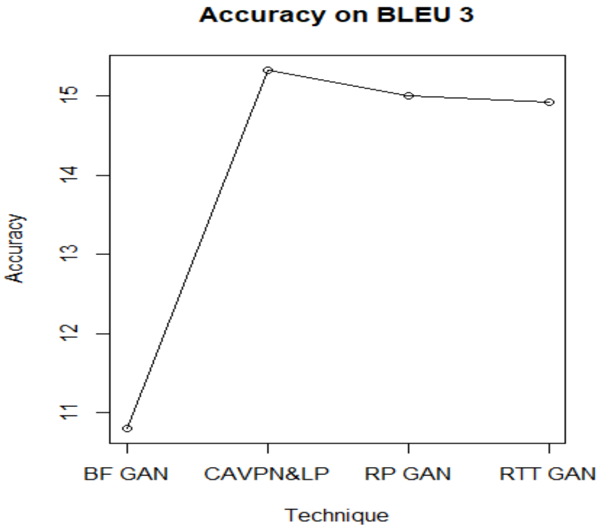

Figure 12.Comparison graph for BLEU 3 language metric

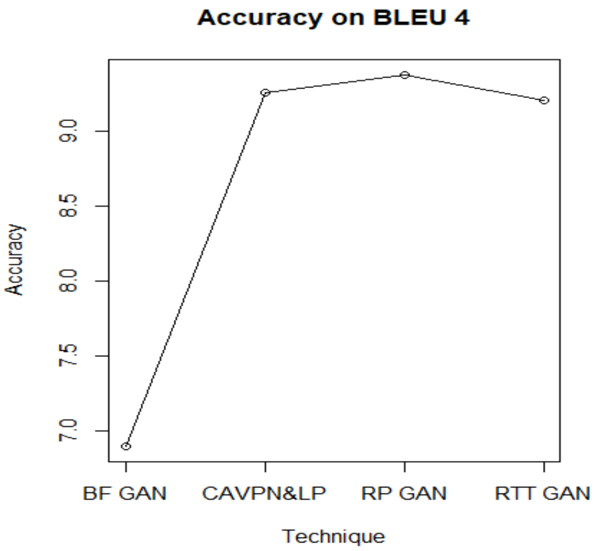

Figure 14. Comparison graph for BLEU 4 language metric

The above figures represent the comparative results of different paragraph generation methods on various language metrics. In this paper different graphs obtained when compared with other generating methods. As per the above diagram Figure 10 shows the comparison of various paragraph generation techniques like RTT GAN,RP GAN,CAVPN&LP and BF GAN on METEOR language metric [1],[2],[3],[4]. From Figure 10, RTT GAN shows highest accuracy rate of 18.39 than the remaining techniques [1]. From Figure 11 shows that the comparison of all the methods like RTT GAN,RP GAN,CAVPN&LP and BF GAN on BLEU 1 language metric.[1],[2],[3],[4]. Graph showed good accuracy on RTT GAN[1].RTT GAN shows highest accuracy rate of 42.06 [1]. From Figure 12 comparison of various paragraph generation methods of RTT GAN,RP GAN,CAVPN&LP and BF GAN on BLEU 2 language metric[1],[2],[3],[4].From the figure 12 CAVPN&LP shows highest accuracy rate of 25.86 [3]. Similarly from Figure 13 shows that the comparison of all the methods on BLEU 3 language metric CAVPN&LP shows highest accuracy rate of 15.33 [3]. And finally from Figure 14 shows that the comparison of various paragraph generation methods on BLEU 4 language metric RP GAN shows highest accuracy rate of 9.38 [2].This paper discussed the above listed four paragraph generated methods of RTT GAN,RP GAN,CAVPN&LP and BF GAN on various language metrics of METEOR, BLEU 1, BLEU 2, BLEU 3, BLEU 4 [1],[2],[3].The above mentioned methods having different accuracy rates on the various language metrics. This paper shows the survey of all four methods and Table 2 denoted all methods generating paragraphs with good accuracy rate based on individual applications. Depending upon the individual applications the techniques produced their results of high accuracy rates.

IV. Conclusion and Future Work

The present paper discussed various algorithms for construction of stories from given set of images. Based on the survey and study on the above algorithms naive methods for story telling can be proposed with respect to application dependent on scene recognitions and stories to be constructed. The above discussed algorithms RTT GAN, RP GAN and CAVPN&LP take images as input and BF GAN takes text as input. A blended mode of input can be considered in future work and create a hybrid algorithm which will take image as well text as input for scene recognitions and storytelling.

References

[1] X. Liang, Z. Hu, H. Zhang, C. Gan and E. P. Xing, *"Recurrent Topic-Transition GAN for Visual Paragraph Generation,"* IEEE International Conference on Computer Vision (**ICCV**), pp.**3382-3391, 2017.**

[2] W. Che, X. Fan, R. Xiong and D. Zhao, *"Visual Relationship Embedding Network for Image Paragraph Generation,"* in IEEE Transactions on Multimedia, Vol.**22**, no.**9**, pp.**2307-2320**, **2020**. doi: **10.1109/TMM.2019.2954750.**

[3] Z. -J. Zha, D. Liu, H. Zhang, Y. Zhang and F. Wu, *"Context-Aware Visual Policy Network for Fine-Grained Image Captioning,"* in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.**44**, no.**2**, pp.**710-722**, **2022.** doi: 10.1109/TPAMI.2019.2909864.

[4] D. Liu, J. Fu, Q. Qu and J. Lv, *"BFGAN: Backward and Forward Generative Adversarial Networks for Lexically Constrained Sentence Generation,"* in IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol **27**, no.**12**, pp.**2350-2361**, **2019**. doi: 10.1109/TASLP.2019.2943018.

[5] K.Kiruba, D. Shiloah Elizabeth, C Sunil Retmin Raj*, "Deep Learning for Human Action Recognition – Survey*," International Journal of Computer Sciences and Engineering, Vol.**6**, Issue.**10**, pp.**323-328**, **2018**.

[6] B. Prasad, U.K. Devi , "*Shape And Texture Based Scene Classification*," International Journal of Computer Sciences and Engineering, Vol.**2**, Issue**.5**, pp.**79-87**, **2014.**

[7] Alejandro López-Cifuentes , Marcos Escudero-Viñolo, JesúsBescós, Álvaro García-Martín *"Semantic-aware scene recognition"* **0031-3203/© 2020** Elsevier.

[8] " Songhao Zhu *, Yuncai Liu*"Automatic scene detection for advanced story retrieval",*Institute of Image Process and Pattern Recognition, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China.

[9] Ramisa, F. Yan, F. Moreno-Noguer and K. Mikolajczyk, *"BreakingNews: Article Annotation by Image and Text Processing,"* in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. **40**, no. **5**, pp. **1072-1085**, **2018.** doi: 10.1109/TPAMI.2017.2721945

[10] C. P. Chaudhari and S. Devane, *"Capturing Semantic Knowledge In Object Localization In Captioning Images,"* 2021 International Conference on Communication information and Computing Technology (ICCICT), pp.**1-4, 2021.** doi: 10.1109/ICCICT50803.2021.9510175.

[11] Stanislav Protasov, Adil Mehmood Khan, Konstantin Sozykin & Muhammad AhmadSignal*,"Using deep features for video scene detection and annotation",* Image and Video Processing, Vol.**12**, pp.**991–999, 2018.**

[12] Y. Choi, S. Kim and J. Lee, *"Recurrent Neural Network for Storytelling,"* **2016** Joint 8th International Conference on Soft Computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems (ISIS), pp. **841-845**, **2016**. doi: 10.1109/SCIS-ISIS.2016.0182.

[13] P. Haritha, S. Vimala and S. Malathi, *"A Systematic Literature Review on Story-Telling for Kids using Image Captioning* - Deep Learning," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp.**1588-1593**, **2022.** doi: 10.1109/ICECA49313.2020.9297457.

[14] H. Zeng, X. Song, G. Chen and S. Jiang, "*Learning Scene Attribute for Scene Recognition,"* in IEEE Transactions on Multimedia, Vol.**22**, no.**6**, pp.**1519-1530**, **2020.** doi: 10.1109/TMM.2019.2944241.

[15] H. Seong, J. Hyun and E. Kim, *"FOSNet: An End-to-End Trainable Deep Neural Network for Scene Recognition,"* in IEEE Access, Vol.**8**, pp.**82066-82077**, **2020.** doi: 10.1109/ACCESS.2020.2989863.

[16] S. Wang, S. Yao, K. Niu, C. Dong, C. Qin and H. Zhuang, *"Intelligent Scene Recognition Based on Deep Learning,"* in IEEE Access, Vol.**9**, pp.**24984-24993**, **2021.** doi: 10.1109/ACCESS.2021.3057075.

[17] J. Guo, X. Nie and Y. Yin, *"Mutual Complementarity: Multi-Modal Enhancement Semantic Learning for Micro-Video Scene Recognition,"* in IEEE Access, Vol.**8**, pp.**29518-29524**, **2020.** doi: 10.1109/ACCESS.2020.2973240.

[18] S. Raghunandan, P. Shivakumara, S. Roy, G. H. Kumar, U. Pal and T. Lu, *"Multi-Script-Oriented Text Detection and Recognition in Video/Scene/Born Digital Images,"* in IEEE Transactions on Circuits and Systems for Video Technology, Vol.**29**, no.**4**, pp.**1145-1162**, **2019**. doi: 10.1109/TCSVT.2018.2817642.

[19] Chen Wanga,b,∗ , Guohua Penga , Bernard De Baets *"Deep feature fusion through adaptive discriminative metric learning for scene recognition"* 1566-2535/© 2020 Elsevier

[20] Lin Xiea,1 , Feifei Leea,1,∗ , Li Liub , Koji Kotani c , QiuChend, *"Scene recognition: A comprehensive survey"* , Elsevier Ltd. **0031-3203**, **2020.**

[21] H. Seong, J. Hyun and E. Kim, *"FOSNet: An End-to-End Trainable Deep Neural Network for SceneRecognition,"* in IEEE Access, Vol.**8**, pp.**82066-82077**, **2020.** doi: 10.1109/ACCESS.2020.2989863.

[22] A. Jalal, A. Ahmed, A. A. Rafique and K. Kim, *"Scene Semantic Recognition Based on Modified Fuzzy C-Mean and Maximum Entropy Using Object-to-Object Relations,"* in IEEE Access, Vol.**9**, pp.**27758-27772**, **2021.** doi: 10.1109/ACCESS.2021.3058986.

[23] G. Chen, X. Song, H. Zeng and S. Jiang, *"Scene Recognition With Prototype-Agnostic Scene Layout,"* in IEEE Transactions on Image Processing, Vol.**29**, pp.**5877-5888**, **2020.** doi: 10.1109/TIP.2020.2986599.

[24] Z. Xiong, Y. Yuan and Q. Wang, *"RGB-D Scene Recognition via Spatial-Related Multi-Modal Feature Learning,"* in IEEE Access, Vol.**7,** pp.**106739-106747**, **2019.** doi: 10.1109/ACCESS.2019.2932080.

[25] S. Wang, S. Yao, K. Niu, C. Dong, C. Qin and H. Zhuang, *"Intelligent Scene Recognition Based on Deep Learning,"* in IEEE Access, Vol.**9**, pp.**24984-24993**, **2021.** doi: 10.1109/ACCESS.2021.30570

AUTHORS PROFILE

Mrs. Darapu Uma persued Bachelor of Technology in Computer Science and Engineering from Pragati Engineering College in 2009 and Master of Technology in Computer Science and Engineering from Pragati Engineering College in 2012.She is currently pursuing Ph.D. in Computer science and Engineering from Adikavi Nannaya University, Rajahmundry, A.P.India and currently working as Assistant Professor in CSE department, Pragati Engineering College (Autonomous) Surampalem, A.P, India.She has published 10 research papers in various reputed national and international conferences and journals. She has secured 3 NPTEL certifications from reputed IITs.Her area of research includes Computer Vision, Machine Learning,Artificial Intelligence.She has a teaching experience of 9 years.

Dr.M.Kamala Kumari persued Bachelor of Technology in Computer Science and Engineering from Acharya Nagarjuna University and Master of Technology from JNTU Kakinada.She received the Ph.D. degree in Computer Science and Engineering from Adikavi Nannaya University, A.P, India, in 2014. She is Associate Professor in CSE department,Adikavi Nannaya University, Rajamahendravaram, A.P, India and currently working as Principal at MSN PG Campus, Kakinada, A.P, India. She has membership in professional bodies like IACSIT**,** CSI and CSTA. She has published numerous papers in various national and international reputed journals and conferences. She has guided more than 50 M.Tech, MCA, B.Tech projects. She has attended over 20 Workshops, FDP and Refresher courses. She has acted as resource person in nearly 30 national and international conferences and webinars. Her area of research interest includes Artificial Intelligence, Machine Learning, Deep Learning, Data Science, Data Analysis,. She has a teaching experience of 20 years and 12 years of research experience.