Generating Automated Stories using Visual Data

A possible practical approach to generate automated stories using visual data

Arpit Garg
Faculty of ECMS
The University of Adelaide
Adelaide, Australia
arpit.garg@student.adelaide.edu.au

ABSTRACT

In this paper, we have reviewed several baseline models of storytelling that motivate Neural Machine Translation (NMT) with an attention-based mechanism. This method is used to focus on the parts of the sentence instead of the full sentence. For this purpose, we use the Visual Story Telling Dataset (VIST) provided by the Microsoft and NMT attention-based algorithm by Stanford. This dataset contains both image captions and story captions, and the algorithm used has two approaches global and local. As this algorithm achieved state of the art for the English to German translation, we have tried to implement this algorithm on the mentioned dataset. We have not reached the state of the art through this process, but still, we were able to achieve good results as compared with other algorithms. We have many processes to get the captions of the visual data using images, but the generated captions do not show any human-like personal behaviour. This process tries to move the automated understanding of visual data from captions to anthropomorphic understanding, which would be more subjective and more useful.

KEYWORDS

Neural Machine Translation, Attention-Based Algorithm, Visual Story Telling Dataset, Anthropomorphic.

1. INTRODUCTION

Attention-based Neural Machine Translation successfully achieved state of the art in translation from English to various other languages such as French, German. There are two types of models Global and Local. In global, all the source words used are attached, and in a local-only subset of words used are attached to the data. The algorithm is used with the dataset to generate a narrative, which is more subjective. This approach is very different from giving captions to the images. In this method, a sequence of images is used, and they are connected so they can be in the video format, which depicts the relation between all the images, for example, change in events. It allows the model to create a network to generate literal descriptions of the images and connects them. This generated narration is more conceptual and abstract as the neural network uses previous image narration in memory to generate the current image narration. It gives a clear difference between description and stories generated using visual data.

Figure 1: Example difference between description(grey) and stories narration (white).

For example: “adult male wearing…” is the accurate description of the image whereas “got exhausted by…” is the story narration that contains more interference and is more evaluative and abstract (Figure 1).

There are three types of stories generated using visual data, and this is done to compare the difference between all the three approaches possible to generate narration. For the same dataset, the approaches used are 1. Describing the images in isolation. 2. Describing the images in Sequence. 3. Images in series for stories. All the categories try to find the relationship and create the patterns, For the first approach of describing the images in isolation, considering work by Karpathy et al. [1] gave the features extracted from the recurrent neural network (RNN), convolutional neural network (CNN) as the input and found that the answer to this approach is the same as generating captions for images. They have also achieved state of the art for image caption generating system. In the second approach, i.e. Describing the images in Sequence, Yao et al. [2] had extracted the features from visuals using serial 3D and then used LSTM (Long Short-Term Memory) as the global. They have used the pre-trained CNN model for the extraction of images and used RNN for temporal behaviour. In the latter approach, Liu et al. [3] used the concept of the Bidirectional Attention-based Recurrent Neural Network (BARNN) model, which would help to generate the relationship between the sentences based on images in Sequence. While using this process, Liu et al. [3] also use three different RNN model to encode all the images.

It can be noticed from all the previous observations that there are various approaches available, but none of the approaches is good

enough. Now we need to model and select the most efficient method with some updations in them based on different factors and parameters. We used Microsoft generated automatic evaluation metric for correlation with the human-like thinking and tried to establish an accurate baseline for the automated story-generating task.

2. Motivation

The primary purpose of this process is to generate a more subjective description of the data, which could be the stimulating step in developing the relationship between visuals and language. We are still very far from implementing the natural interactions by artificially intelligent systems, but this could be a critical step towards this. We can see the significant difference between image captioning that is people sitting in a room, and the narration generated that is good time and bonding for the same image in Figure 1. The last portrayal is grounded in the visual sign, yet it brings to hold up under data about social relations and feelings that can also be deduced in the situation. These grounded visual narrations provide more abstract and evaluative language in the vision of language research. If the system successfully recognises the environment, it can make a decision and remarks on the situation indirectly and accordingly.

Storytelling is the essential attribute of the human behaviours as it helps to convey information, protect culture and morals and giving this level of human intelligence and understanding to machine could make a revolutionary change not only in the artificial intelligence society but in the human society also. By implementing this, we are trying to give the power to machines to work together with humans to generate stories; it could be movie-scripts, audibles. In high-level, this approach is beneficial for humanoid robots as these are called humanoid robots. However, they still lack in various attributes when compared to humans, and this is one of that field as they still lack in human-based understanding of visuals, this would take these robots one step closer to the human as these robots would have the same abstract and evaluative understanding of the visuals as the human have. In low-level, this approach is essential as it would help to increase the software quality of the social platforms by reducing the possibility of spreading false and fake news. If anyone can implement state of the art, then this would work as the amplification and ratification system for the various online platforms and could act as an improved version of hearing aids for the indigent category. This research would help to improve the software standard but also improves developers and researcher productivity and reduces human attempts while providing support to humans in various fields according to the implementation level and requirements.

3. Background

For understanding the process of generating stories, we need to understand the few concepts that we are going to use. The displayed text with the images is known as captions or image descriptions, and they are not the same as story generated text. We have two separate datasets one is for Describing the images in isolation, and the other is for Images in series for stories. We are doing the image labelling, but we have the video. So, we are converting the videos to frames. All the videos are the collection of images, and running this together at a fixed rate makes the video. The rate at which they are running per unit time is called the frame rate, and each image is called the frame. We are taking frames from videos and generating the description frame by frame. We are generating stories, so we do not require the audios; we are not considering the audios of the video file. There are various types of video formats possible, which just signify the balancing style of video between quality and size. Every video format has two parts codec and container. Codec is the procedure to encode and decode the video while the container holds the metadata generated while encoding the video. Generating the description of each frame is known as the labelling of the images in this scenario[4].

While considering digital images, every image is made up of picture elements called pixels, and each pixel is the numeric representation of intensities. So for real-life cases, we need to consider the coloured images which are formed by the RGB format that means every intensity of the pixels present in that image can be represented by the mixture of Red, Green, and Blue colour intensity. Before processing any image in the neural network, we need to process the image like changing the size of images and set the parameter of neural networks according to that[5]. All the time, images vectors are considered. Vectors mean it contains images magnitude with directions.

Initially, these all the terms are crucial for our process. Currently, most of all, the background steps of removing audio, converting to frames, were already done by Visual Storytelling Dataset (VIST) to save time and computational power. Hence, we are considering this dataset for our paper, but for the global scenario, we need to do all this background analysis before implementing the algorithm.

4. Research Method

4.1 Research Question

In this paper, we plan to focus on the visuals to generate stories and narrations using artificial intelligence (A.I). In this paper, we are focusing on the sequence model to create automatic stories with the use of image sequence encoder-decoder, which is also based on the previous sentences. The research questions discussed in this paper are:

R.Q. 1: *Is there an approach that can be followed for generating automated stories using visual data?*
R.Q. 2: *If we can achieve a state of the art, in what way it would be beneficial for the society?*
R.Q. 3: *Are there any ethical issues related to this process? If yes, how can we overcome those issues?*

The above mentioned are the three research questions that we have, and we tried to answer each of them in the paper. There was not much work has been done on generating narrations using visuals,

and this approach is still in the developing phase. There might be different answers and approaches possible for each of the research questions, but we have tried to answer that best based on our knowledge.

4.2 Dataset Collection and Construction

The primary source of the data used to answer these questions is Visual Storytelling Dataset (VIST). Training, cross-validation testing of the model is done on this dataset. This dataset comprises of 50,000 unique stories and approximately 20,211 images. The Flickr dataset is the root source for the collection of visuals using its API. Albums included are in the range of 10 to 50 images, and these were pulled at an interval of 48-h. This procedure provides the association between images and stories.

In the root process, Amazon Mechanical Turk employees were requested to select five albums in a non-linear way and drop some lines in the form of a story related to each album. Each image is linked with varies stories, and there are at least five-line stories connected with every image. The primary thing to focus on the dataset is that all names were replaced by gender according to images[6]. This dataset is categorised in two parts description of images in isolation and story of images in sequences. The main problem with the VIST dataset that could be an issue is that all the generated stories were produced by individuals only, so there might be the possibility that there is no connection in the flow or sequence of the stories[7]. There is a high probability that, in some occurrences, the story is not related to the photo's sequence. But we are ignoring this case for this paper. Therefore, we are not anticipating an exact story as an output. We are only worried that our system could be able to generate automated stories in our required language. We only want our model should contain at least some word related to the image sequence. For all these purposes, we have used VIST and Flickr API.

beach (684)	breaking up (350)	easter (259)
amusement park (525)	carnival (331)	church (243)
building a house (415)	visit (321)	graduation ceremony (236)
party (411)	market (311)	office (226)
birthday (399)	outdoor activity (267)	father's day (221)

Table 1: 15 most frequent story album titles with several story phrases

We need to start by extracting the photos. In this process, we would make a list of events of the story. For these events, there must be some possession. Using Flickr and CoreNLP by Stanford, we averaged 5-grams of titles of images, and their description and their pattern dependencies are extracted. Then we used Wordnet 3.0 to collect these phrases, and then these phrases were used to extract the albums from the Flickr API. 2-stage crowdsourcing workflow pattern is used to collect stories that look more abstract with the description aligned to frames. (Figure-2).

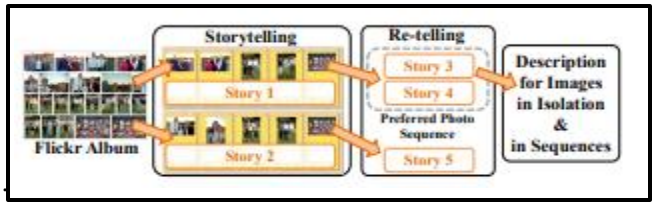

Figure 2: Workflow of Crowdsourcing

Storytelling is the first step in which crowdsource selects the number of images from the album and write the story about it in Sequence. The second step is re-telling in which they are asked to select one photo sequence from the previous step and write the story about it. In two stages, all images in an album are shown according to time, and there are board underframes that contain descriptions[8]. In the initial step, if we click the frame, we would get the board with the story of it. The minimum number of frames required is five to generate stories, and the interface automatically aligns the text with photos. All the users are allowed to skip if we story-making is not possible, and is two users skip the same set of images, then that set would be discarded from all the future events. Re-telling is the same; the only difference is that it shows only two image sequences that are created in the previous stage.

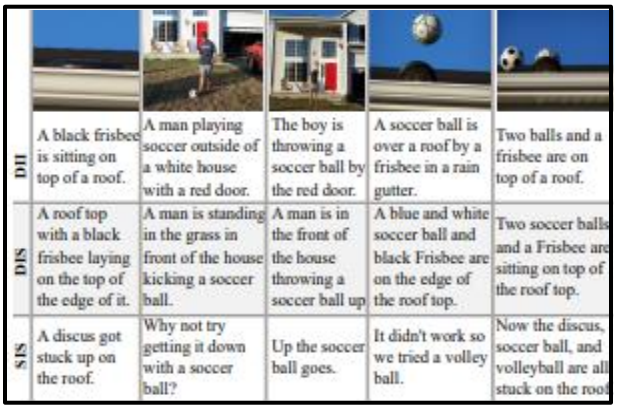

Figure 3: Describing the isolation of images (DII); Describing the image in the required sequence (DIS); and Sequences of the images in the story(SIS).

Crowdsourcing workflow is different for isolation of images and images in sequences. In both the tasks, users are asked to use the image captioning system described by MS COCO, which mainly focuses on describing the main parts of the images [9]. For DII, we use the COCO MS interface of frames and captioning, and DIS, we modified this version with boards of the story appended to it.

Desc.-in-Iso.			Desc.-in-Seq.			Story-in-Seq.		
man	sitting	black	chatting	amount	trunk	went	[female]	see
woman	white	large	gentleman	goers	facing	got	today	saw
standing	two	front	enjoys	sofa	bench	[male]	decided	came
holding	young	group	folks	egg	enjoying	took	really	started
wearing	image		shoreline	female		great	time	

Table 2: Normalised PMI- top-ranked words for each category

At last, after collecting all the data and constructing the data according to the requirements, we perform the data post-processing step. In this, we first need to convert all narrations and descriptions generated to tokens with the use of CoreNLP tokeniser and generalise all the narrations like converting all the names to gender and all the particular places to generalised locations. [10]
Based on the data collection and construction part, we can answer the third research question as there are no ethical issues found for data collection and construction. All the data used are open-source basically for the research work, and all the algorithms used are free to use modified according to the requirement so we have followed the four principles of Ethical issue which are found useful in this research and they are confidentiality, consent, scientific value, and beneficence.

4.3 Data Analysis and Model Creation

We start analysing our data by splitting it into three parts. 80% of the dataset would be divided for training the model and 10% for cross-validation and 10% for testing the model. We have 10,117 albums from Flickr, which contains 210,819 different photos, which means each album has approximately 21 photos. For each album, the average period is 8 hours.

Data Set	#(Txt, Img) Pairs (k)	Vocab Size (k)	Avg. #Tok	%Abs	Frazier	Yngve	Ppl
Brown	52.1	47.7	20.8	15.2%	18.5	77.2	194.0
DII	151.8	13.8	11.0	21.3%	10.3	27.4	147.0
DIS	151.8	5.0	9.8	24.8%	9.2	23.7	146.8
SIS	252.9	18.2	10.2	22.1%	10.5	27.5	116.0

Table 3: Data analysis of the used dataset with the analysis provided[11].

A shallow level of ambiguity is observed between the words in the improved version of describing the images in isolation. While considering single images, we often get repeated image caption words, so we do not consider them much[12]. As we analyse describing the images in the sequence approach, these words are less repeated. In the most advanced approach of Sequence of the story in Sequence, it is more focused on the temporal references and generalised storytelling words, which makes the narration more dynamic and conceptual.
After the extraction, files are saved in the extension of Hierarchical Data Format version 5 (HDF5). After this, each features vector obtained needs to be joined with the vectorised sentence. We are using the data loader to assign every feature obtained with the sentence, respectively. We used the Attention-based NMT (Neural Translation Machine) model to train encoder-decoder.[13] It directly converts original sentences to required sentences by modelling the conditional probability p(y|x) and then divide the conditional probability as:

$$\log p(y|x) = \sum_{j=1}^{m} \log p\left(y_j|y_{<j}, \boldsymbol{s}\right)$$

RNN is used as a decoder in a hidden unit, and CNN would be used for encoding purposes. Similarly, LSTM combined with RNN, which is used as both an encoder and a decoder[14]. We can parameterise the probability by:

$$p\left(y_j|y_{<j}, \boldsymbol{s}\right) = \text{softmax}\left(g\left(\boldsymbol{h}_j\right)\right)$$

where,

$$\boldsymbol{h}_j = f(\boldsymbol{h}_{j-1}, \boldsymbol{s})$$

After using LSTM, our training objective is

$$J_t = \sum_{(x,y)\in\mathbb{D}} -\log p(y|x)$$

The method of observing the previous sentences to generate the current sentence with the use of encoder and decoder is applied. We are following the sequence to sequence approach. In this, our encoder is categorised into two parts. The first part is related to the behaviour of sequences of an image, and another encoder is handling the modelling of the sentence with the previously generated and observed sentence. RNN is capable of learning temporal dependencies, so it is the best choice to use it here for sequence modelling.[15] The encoder that is used for dealing with the behaviour assigns each sentence with the corresponding and associated image. The decoder in the module is again RNN, which tries to reproduce all the information and data encoded. In the training process, we are giving image sequences and previous sequence of sentences to an encoder, and in the decoder, we are sending current sentences. For performing these mentioned steps, we have converted our images from RGB to BGR and cropped them according to the requirements, as we need to give these as input to AlexNet[16].
The words that are used in a generation would be the most common and frequent words that would be appearing in the narrations. Based on the general observations, we have selected that our initial length of every sentence would be approximately equal to 20 words per sentence. Categorical cross-entropy would be used as the loss function, and the learning rate is set to 0.0001 based upon the backtracking improvement. In contrast, ADAM algorithm would be used for the optimisation in training. For measuring the closeness of generated automated translation, we have used the METEOR and BLEU scores. They are considered when we are working on Neural Machine Translation. They measured between 0 to 100 to show how the closeness of the results. After our model training, we would apply cross-validation dataset for updating the required parameters to reduce the loss and then testing data for the final score. By analysing the data in this way, we could easily

comment on the question if the mentioned approach is valid or not[17].

In the attention-based approach, both the global and local approach is used in parallel. It only differs whether we want the attention on all the positions available or the few selected positions.

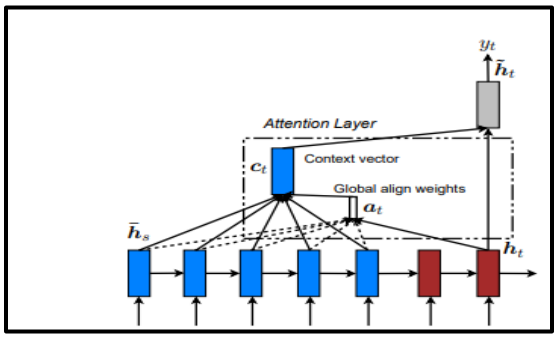

Figure 4: Global attention-based model

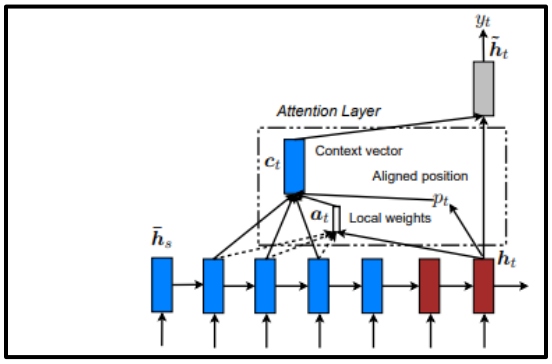

Figure 5: Local attention-based model

In both, the approaches decoding is done in time t, and they take h_t as the first hidden layer at the top of LSTM. The primary purpose is to derive the c_t, which is a context vector that contains all the necessary information to predict the word that is the target y_t. When hidden state h and context vector c is given, we use the simple concatenation layer to merge all the information present to make hidden-attention state given by:

$$\tilde{h}_t = \tanh(W_c[c_t; h_t])$$

Finally, the output of these layers are passed to softmax layer to determine the prediction probability and formula used is:

$$p(y_t|y_{<t}, x) = \text{softmax}(W_s\tilde{h}_t)$$

After training the model on the training dataset, we used 10% of the cross-validation dataset to improve the accuracy of the model by updating the required parameters. After training and cross-validation, our model is ready for testing, and now we can pass real-world videos or the testing videos to generate stories or implement it with some other technology as per the requirements.

5 Findings

We use the automatic evaluation metrics as suggested by Microsoft because this is useful for the benchmark. For the understanding, we find the correlation between predictions and human answers on 3000 stories of SIS set. We tried a normalised approach for the identification of words that are most closely related to each category[18]. We have analysed the correlation between the machine-generated stories and human-like narrations (Table 4).

	METEOR	BLEU	Skip-Thoughts
r	0.22 (2.8e-28)	0.08 (1.0e-06)	0.18 (5.0e-27)
ρ	0.20 (3.0e-31)	0.08 (8.9e-06)	0.16 (6.4e-22)
τ	0.14 (1.0e-33)	0.06 (8.7e-08)	0.11 (7.7e-24)

Table 4: Human narration correlation scores and p-values

For automatic metrics, we used METEOR, BLEU, and Skip-Thoughts to evaluate the relationship between each story of a given sequence. From table 4, we can state that METEOR correlates better as compared to others.

Beam=10	Greedy	-Dups	+Grounded
23.55	19.10	19.21	–

Table 5: Description generated per frame based on METEOR scores

Beam=10	Greedy	-Dups	+Grounded
23.13	27.76	30.11	31.42

Table 6: Baseline of stories based on Meteor scores

Based on all the comparisons, we came to common finding that we can reach the state of the art using Stories in Sequence approach with the METEOR based scores. We can use either a local or global attention-based approach, which is based on the requirements[19]. To show the difference between image captions and narrations, we used DII in isolation and produce captions per image rather than Sequence.

+Viterbi	This is a picture of a family. This is a picture of a cake. This is a picture of a dog. This is a picture of a beach. This is a picture of a beach.
+Greedy	The family gathered together for a meal. The food was delicious. The dog was excited to be there. The dog was enjoying the water. The dog was happy to be in the water.
-Dups	The family gathered together for a meal. The food was delicious. The dog was excited to be there. The kids were playing in the water. The boat was a little too much to drink.
+Grounded	The family got together for a cookout. They had a lot of delicious food. The dog was happy to be there. They had a great time on the beach. They even had a swim in the water.

Figure 6: Example of generated narrations using baseline models

We found that using the size of the beam as 1 that is greedy search improves the quality of stories and results in an increment of

METEOR score by 4.6. In contrast, if we compare the same thing for caption generation, it gives the worst results.

The stories generated by this greedy method are good, but they contain the redundant words, so we tried to include a heuristic approach to prevent this, which is the same phrases that cannot be used more than once in the same narration. This approach improves the METEOR score by 2.3[20]. We defined a collection of grounded visual words that is possible to occur at higher frequency in the caption as compared to the story.

$$\frac{P(w|T_{caption})}{P(w|T_{story})} > 1.0$$

The greedy approach with heuristic implementation contains one more condition that is a word that belongs to grounded visual words; then, it can only be generated by the story in sequence model. This approach results in an increment of 1.3 in the METEOR score[21]. All these approaches are considered only based upon scores. We only tried to implement this model using the baseline conditions. From all these findings, we can answer our research question as it is possible to reach the state of the art of this process, and it would be beneficial for human society. All the ethics are maintained; there are no ethical issues found for this research.

6 Discussion

Findings compared all the possible approaches, and based on the scores and requirements; we can select the best approach to answer the research questions. All the approaches mentioned are good enough to generate the text, but we still can think of modifications in the process to approach fast and which can improve the quality of the research.

We can use the YOLO(You Only Look Once) approach for object detection. The new YOLO v3 approach used the COCO dataset to train the model and classify them, and then we can use the pre-trained model. In our model, also we used COCO, so we simply can detect the objects of the images using YOLO, which would decrease the processing time, and after that, we can pass the outputs to our neural networks where our story would be generated based on the object detection and album classification. This approach would reduce time, saves processing power, and improve the quality of the output.

On 28.05.2020, OpenAI reveals GPT-3, which is an upgraded language model which contains 175 billion parameters. This model contains modified initials parameters, pre-normalisation methods, and reversible tokenisation technique, which shows very high performance on text generation technique and benchmarks in a few-shot, one-shot, and zero-shot settings. It provides more smooth scaling on NLP with more conceptual text generation as compared to present techniques. Initially, the researchers have generated a few news articles using GPT-3, and they found it difficult to differentiate it with human-written articles. It could reduce the processing time, and we need less data for cross-validation, we can use that data for training purposes only. Using this process by just changing the parameters could give the more expected results. So, it can make the revolutionary change in story generation using visuals.

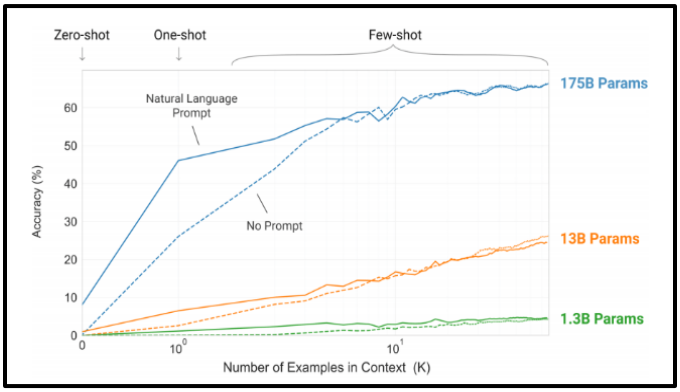

Figure 7: GPT-3 Performance

All our findings from this paper are enough for the usage of these approaches in the real world, but we still can merge our techniques with the above mentioned two techniques to improve the model. As there is still no evidence and evaluation of the above mentioned two techniques on our model, but both these models proved to be the best on the related projects. So, we can try to implement these approaches to our project also. If we successfully implemented these processes with our model, we could increase the scope and usability of the project. As now, it can also be used to detect the authenticity of the news.

7 Threats to Validity

Following the approach, we get the desired result, but there are some limitations that we need to consider while performing the research, and these limitations could make the findings invalid if not maintained.

There are multiple stories generating approaches available but, in this paper, we have selected only three best approaches and compared their results and modified these approaches to get the results regardless of comparing all the other approaches. Our selected approaches are describing the isolation of images, describing the images in Sequence, and the story of the images in Sequence, and we have modified these algorithms only. We have selected these approaches as they give the best results when researched by other researchers. However, if we are changing the parameters of these approaches, there might be a possibility that modifying other approaches could work better than these selected approaches, which could make our research invalid as our primary purpose to find the best story generation approaches possible.

In data collection and extraction, we used the VIST dataset in which Amazon Trunk Workers were requested to select the models and write the narration in some lines about it in a non-linear way, but this is not the standard approach to make the human-like stories. Then we have used this model to train our model, so there is high variance in the model as if we take different albums in the training dataset, we would get a different story, so there are no official

results of this research. We are generating stories, but these stories are highly based on the lines or words used by these workers.

For this research, we are considering largest album dataset that is Flickr which contains more than ten thousand albums and our objective of this research to generate story which is similar to human written stories, so in the real world scenario there is a possibility that frames fails to find the particular album and skip that event. Then it would find the album of next frame; if it does it for two or three frames, then there would be no story generated, and it generates the stories from the related albums, which is just the manipulated story, not the desired story.

One most important advantage of using this dataset is that this is the largest dataset publicly available. However, the main disadvantage with it is there are copyright issues every time with this data, so they keep on adding and deleting the copyrighted material, so everyone gets different results at different times.

In the baseline model, all the narrations generated using the beam size approach as shown in Table 6 and to generate the results beam size is selected as 10 by default because this was found best beam size while image captioning by Devlin et al. [18] and Vinyals et al. [12]. We have tried a similar approach, so we have used this value and got the results, so by changing this value, we could get the undesired result.

For our baseline model after training the dataset, we have used the ADAM algorithm foe the optimisation purposes because it is a combination of Stochastic Gradient Descent with momentum and RMSProp so that it can iteratively update the parameters in less time while considering the batch instead of optimising at every value. It is used here because after applying multiple optimisation techniques on this dataset, this optimiser is giving the best results concerning time, but this is only true for the baseline model and could change as we change the dataset or modify the model.

In data analysis we have used two different attention-based approaches that are global and local attention-based approach in these approaches we are considering the set of words and generating the stories based on that, there might be the possibility that in selecting the set of words we have skipped some important words that are important. We get a different story from what is expected.

We have predefined that the maximum number of words of each sentence would be 20, and there is no standard way that was considered while selecting this. This method was done only based on ordinary observation and could make the results invalid if this is not considered while creating the network.

For the findings and analysis, Meteor scores are considered as they are showing better results for this dataset, but as we have discussed, there is high variance, so by changing the dataset, there might be a possibility that this score fails to give good results. At that time, we again need to consider all the scores and select the score, which gives the best results. For our baseline model, this score is the best, but there is no fixed criterion that we need to use this score only for comparisons, and all the findings are done based on this score. If we change this score, then there could be validity issues with our findings.

Table 5 and Table 6 are generated when we are considering the second half of the story of the images in the sequence model as the validation set. All the scores and future findings are based on that which could fail if we consider the validation set in some other pattern. It is done as per the observation by previous researchers for their research. We contrast the top-1 produced a result and the (changing) number of reference stories for a similar order. We are still not sure about the output when we use a random shuffle. There is no guarantee that this proves to be the best for this research.

To improve the Meteor scores, we have used one heuristic approach in which we have to restrict our model as not to use one word more than once in a narration. By implementing this, our scores were improved. The problem with this heuristic approach is that it makes our stories less abstract and less conceptual there might be a possibility that we need to use a word more than once in our narration to create an impact of the events. If our network fails to do so, we again backtrack towards the more real story, which is like a modified version of image captioning rather than generating stories.

With the heuristic approach, we have used the greedy approach also to select the word from the grounded visual story set. The greedy approach gives the result fast as it would not care about future sentences, but it would not give the optimal results that mean the more abstract and more subjective results that are desired. It is one of the primary purposes of this research, but due to baseline limitations, we have implemented this greedy algorithm, but it could make us deviate from our results.

For all the sentences generated in any story, the word selection is based on pointwise normalised mutual information, which identifies the most closely related word to it from each album. If we pass the video with a smaller number of frames, then the system fails to generate the information, then it tries to include posture verbs wherever possible in large amounts like sitting, standing, going. Because we have put a limit on the occurrence of the word to one, all the posture verbs destroy the meaning of the sentence, and we would not be able to generate the proper narration. We need to pass the video with a large number of frames, so we get the real story; otherwise, for a smaller number of frames, our system fails to produce the story with desired content. Our model is dependent on the number of frames, and a smaller number of frames could make our model invalid.

As from our model, we could see that our model has a high dependency on Flickr, VIST dataset. All the findings are valid until the date on which this paper is written, in the new version, or the updated version, all the findings would change accordingly.

If we considering and commenting on the third question of this research, it is related to the ethical issues. If we are considering all the mentioned cases only for personal use, then all the ethics are maintained, but we are not allowed to use all these for commercial purposes without permission. If we use this for commercial purposes, then there might be some ethical issues generated, and this research would fail for the last question.

8 Related Work

All the previous works and references taken worked on different approaches to this paper. We have tried to combine all the work to answer the given research question. In the early stages, most of the

researchers work on the neural network like CNN and RNN to extract the features of images, which is the crucial step of our research. After that, Stanford researchers worked and published the importance and effectiveness of attention-based learning and the approaches in which the attention-based learning could be most effective. We have used this research with the previous research of neural networks to generate the narrations. On the parallel, for the neural networks, we need a proper and colossal dataset. For training purposes, this dataset is provided by Microsoft researchers, and they have worked on coco-dataset to improve this dataset. The whole paper is based on computer vision and natural language processing, so we need a basic understanding of both the areas which could be possible by reading some of the works done by researchers. We are majorly focused on the language generated by NLP, so we need to consider the important work done on Computational Linguistics using NLP by the previous researchers. These all the works discussed till now can be combined to generate words, but we still need to do more analysis to answer our research question correctly. So, we considered the work done on word alignments of translation to improve the quality of translation, which helps us to generate more abstract and human-like stories. To generate the human-like narrations, we need to focus on the words also, as some words are more emphasised as compared to others, so we need to look for the work done on rare world problems using neural networks. For this paper, we have used the work done by researchers already due to the limitation on computational power and time limits. We have used CoreNLP, which was given by Stanford researchers. For the comparison between captions and stories generated from the images, we have used the predefined model by researchers to generate captions. Our model is dependent on the text generated from the previous image; for this, we have used the research on similarity metrics on previous event descriptions. This model was all based on RNN, so we need to look for the RNN in this field. As all the text-generation is dependent on this field, we have also considered some of the work done in pattern recognition of visuals using computer vision. These mentioned are all the related work that has already been done, but they are done in different fields. In this paper, we have tried to combine all these different researches and used it in our research. For example, attention-based approached till now have only used for translation purposes (translate from one language to another).

We have used this approach to make our narrations better. We also need some metrics for the comparison of the model, and to measure the effectiveness of the approach, we have used the metrics from the research published about n-gram counts and metrics in language models for crawls. After combining and implementing all these different approaches together, we have tried to implement simple methods like some heuristic approaches to improve the quality of the stories. Recently, YOLO and OpenAI also released some models which are related to this field. We have discussed these two approaches in terms of usability and effectiveness in the discussion section of this paper, but there is not enough evidence to state an exact comparison at this stage. There are still some researches and approaches that could be used to improve the efficiency of this model.

9 Conclusion

We have successfully implemented and compared three different approaches to generating automated stories using visuals. We have found that from the Images in Isolation approach to Story of Images in Sequence, we have moved the progress from image captioning to story generation. By comparing, we have cleared the difference between captions and stories. We have argued that generating the more universal and figurative language is the critical element of the automated story generation model as it moves the results more towards human-like understanding. We have implemented two adequate attention-based NMT models that are the global and local approaches to improve the model and select the more abstract words. By using these models, we have discussed baseline models and parameters we have used on the mentioned visuals to train and generate stories and compared them based on METEOR scores as this score is acting as an automatic evaluation metrics and considers some heuristic approaches for the better results with all the threats and limitations discussed. This approach is purely a new field, so there are many limitations, and an amount of future work is required.

10 Future Work

We have discussed the use of GPT-3 and YOLO in the discussion section, it could improve the efficiency of this model, and it also requires much work of this research. There are two attention-based approaches that we have used for the selection of words, and both the global and local approaches are selected directly. Due to the time limitations of this paper, we have used them without any modifications. There are a lot of modified versions of algorithms that are possible, but we cannot randomly select them; we need to study them in-depth before using, which requires much work.

We have considered the automatic scores according to the related work, but this is the most important thing. There is a need for a detailed study of the automatic evaluation of narrations. We are giving all our findings based on this only, so there must be concrete evidence to use this. This approach requires a whole new area of interest and future work. Heuristic, besides with greedy approach, is used to improve the scores, but as already discussed, they are not the correct way of improving the score. We need a standard way to handle the control of the reoccurrence of the words rather than just stopping the model from selecting them if they occurred more than once. This model is purely a separate study, which is time-consuming but much needed.

We could also look for more optimisation techniques. In this, we have used the ADAM optimiser, and we have mentioned the reason above. We have not compared this optimiser with others. We have used it based on theory and related work. We need to work on the optimisation part by comparing it with other optimisers, and this would require lots of detailed study and comparisons based on experimental scores. Our model fails for a smaller number of frames as there are no options available in our model to handle these types of situations. In the future, we need to work on this case

to make sure the story could be generated regardless of the number of frames.

Our model is implemented on the test data, and all findings and results are mentioned are based on that. Although it is ready to implement real-world data with some minor modifications, we have not tried that, and this is out of the scope of this paper. All these mentioned future works are essential to be considered as they played a significant role in the findings of this research. In this research, they are selected based on the related work due to time and resources limitations.

REFERENCES:

[1] Karpathy, A. and Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans Pattern Anal Mach Intell*, 39, 4 (Apr 2017), 664-676.

[2] Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H. and Courville, A. Describing Videos by Exploiting Temporal Structure. *arXiv.org* (2015).

[3] Liu, Y., Fu, J., Mei, T. and Chen, C. W. Storytelling of Photo Stream with Bidirectional Multi-thread Recurrent Neural Network (2016).

[4] Malinowski, M. and Fritz, M. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input (2014).

[5] Polly, W. W. Embers of society: Firelight talk among the Ju/'hoansi Bushmen. *Proceedings of the National Academy of Sciences*, 111, 39 (2014), 14027.

[6] *Energy Minimization Methods in Computer Vision and Pattern Recognition: 9th International Conference, EMMCVPR 2013, Lund, Sweden, August 19-21, 2013. Proceedings*. Springer Berlin Heidelberg, City, 2013.

[7] Thomee, B., Shamma, D., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D. and Li, L.-J. YFCC100M: the new data in multimedia research. *Communications of the ACM*, 59, 2 (2016), 64-73.

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

[9] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. and Dollár, P. Microsoft COCO: Common Objects in Context (2014).

[10] Zhang, Y., Zhang, Y., Bolton, J. and Manning, C. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv.org* (2020).

[11] Ferraro, F., Mostafazadeh, N., Ting-Hao, L., Huang, J., Vanderwende, M., Devlin, M., Galley, M. and Mitchell, M. A Survey of Current Datasets for Vision and Language Research. *arXiv.org* (2015).

[12] Vinyals, O., Toshev, A. and Bengio, S. Show and Tell: A Neural Image Caption Generator. *arXiv.org* (2015).

[13] Ledeneva, Y. and Sidorov, G. Recent advances in computational linguistics. *Informatica*, 34, 1 (2010), 3.

[14] Sutskever, I., Le, Q., Vinyals, O. and Zaremba, W. Addressing the Rare Word Problem in Neural Machine Translation. *arXiv.org* (2015).

[15] Zaremba, W., Sutskever, I. and Vinyals, O. Recurrent Neural Network Regularization (2014).

[16] Oller, D., Glasmachers, T. and Cuccu, G. Analyzing Reinforcement Learning Benchmarks with Random Weight Guessing. *arXiv.org* (2020).

[17] Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D. and Parikh, D. VQA: Visual Question Answering (2015).

[18] Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G. and Mitchell, M. Language Models for Image Captioning: The Quirks and What Works (2015).

[19] Young, P., Lai, A., Hodosh, M. and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2 (2014), 67-78.

[20] Fraser, A. and Marcu, D. Measuring Word Alignment Quality for Statistical Machine Translation. *Computational Linguistics*, 33, 3 (2007), 293-303.

[21] Wolk, K., Wolk, A. and Marasek, K. *Big data language model of contemporary polish*. Polish Information Processing Society (PIPS), City, 2017.