International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Neural Story Teller using RNN and Generative Algorithm

Mahipal Singh Rathore[1], Shailendra Patel[2], M Purna Rao[3], and Asha P Sathe[4]

[1,2,3,4]*Department of Computer Engineering, Army Institute of Technology, Pune*

***Abstract* -** When we human being see an image we all come up with different stories about that image based on what we are thinking, feeling and on our individual experiences. This paper aims towards developing an application which give computers the same ability as we human have towards developing a story about an image. So, the objective is to generate short descriptive English story about an image.

In order to generate story, firstly we are extracting the features of image to generate captions and then we try to align the generated image captions (i.e. description of image) with our romance novels (book corpus dataset) to fetch similar sentences to the captions and generate a story out of these fifty most similar sentences.

The features of the image are extracted using the image-net pre-initialised models such as Vgg-16 and that extracted features are then used to create a caption. Now we have to consider a book corpus dataset which has a particular genre (romance) and by using word embedding's such as glove we make every sentence into a vector so that we get the exact similarity between the sentences correct. Now we match the top 50 closest sentences in the novel and align them to form a beautiful story.

I. INTRODUCTION

Storytelling is an art of putting your emotions, sentiments and life experiences into an interesting story that you want to convey to the people. It is very difficult for a computer to write and tell a story. But if computers are learned how the article is configured, it can also write an article like humans. We all know what all machine learning can do. It has a wide range of applications in wide range of fields. One of the application is predicting the emotions and the actions from the image given. The initial steps to it were taken when the MSCOCO data set was open sourced. This data set has the image and a caption related to the image. Now we intend to take this ahead by creating story out of images .This can be done by using the both concepts of image captioning and story generation.

In the past few years combining images or videos has gotten significant attention, partly due to the creation of CoCo, Microsoft's large-scale image captioned dataset. The machine learning algorithms has tackled a diverse set of tasks such as captioning [1], [2], [3], alignment [2], [4] , Q&A, visual model learning from textual descriptions, and semantic visual search with natural multi-sentence queries.

Books provide detailed descriptions about the intentions and mental states of the character. Image captioning provides us Shailendra Patel, M Purna Rao, Prof. Asha P Sathe are with Computer, Army Institute of Technology, Pune, e-mail: shailendra_15174@aitpune.edu.in, e-mail: mrao_15111@aitpune.edu.in, e-mail: asathe@aitpune.edu.in

Mahipal Singh Rathore (e-mails: mahipalrathore_15068@aitpune.edu, in) the relationship between the objects present in the image. Furthermore we are trying to improve the image captioning by using book corpus dataset to match the similar sentences to the caption. So, instead of just caption we are giving a detailed description of the image related to its captions.

Challenges that we come across are, as we are training on MSCOCO data set the caption generated will be like a vague relation between the objects identified in the image which may misguide the whole story. When we find a sentence in the book which has similar meaning it is based on the occurrence of the words in the sentence. So we will be using a large corpus of novels which will give out the wide range of options for the model to make story. And we are also using pre-trained word embedding's such as Glove or Genism which will give out the exact similarity between two sentences not just by considering the occurrence of the words, which may be more relevant.

The neural story teller can have multiple applications. Imagine an app which allows the user to summarize the albums of photo. Also an app can be developed for primary school students which scans the image from the textbook and come up with a descriptive story about that image. Also the neural story teller can have vast number of applications in the entertainment industry.

II. RELATED WORK

Most effort in the domain of vision and language has been devoted to the problem of image captioning. Older work made use of fixed visual representations and translated them into textual descriptions. Recently, several approaches based on RNNs emerged, generating captions via a learned joint image-text embedding [2], [5]. In [6] , the authors go beyond describing what is happening in an image and provide explanations about why something is happening.

For text-to-image alignment, [7] find correspondences be- tween nouns and pronouns in a caption and visual objects using several visual and textual potentials. In [2], the authors use RNN embeddings to find the correspondences. [8] com- bines neural embeddings with soft attention in order to align the words to image regions.

Rohrbach et al. [9] recently released the Movie Description dataset which contains clips from movies, each time-stamped with a sentence from DVS (Descriptive Video Service). The dataset contains clips from over a 100 movies, and provides a great resource for the captioning techniques.

Our effort here is to align images with books in order to obtain longer, richer and more high-level image descriptions. We start by describing our new dataset, and then explain our proposed approach.

III. DATASET USED

A. MS COCO Dataset

The Microsoft Common Objects in Context (MS COCO) is large-scale captioning, object detection and segmentation dataset. The MS COCO dataset contains 91 common object categories with 82 of them having more than 5,000 labelled instances, Fig 1.In total the dataset has 2,500,000 labelled instances in 328,000 images. In contrast to the popular Ima- geNet dataset, COCO has fewer categories but more instances per category. This helps in training advance objects models capable of more effective 2-D localization. As compared to PASCAL VOC and SUN datasets the MS COCO dataset is significantly larger in number of instances per category. MS COCO dataset contains considerably more objects instances per image (7.7) as compared to ImageNet (3.0) and Pascal (2.3).

MS COCO is a new large-scale dataset that addresses all the three main issues in scene understanding: detecting non-iconic views (or non-canonical perspectives) of objects, contextual reasoning between objects and the precise 2D localization of objects.

MS COCO dataset made it possible to use computer vision and NLP (Natural Language Processing) to carry out the task of image captioning successfully. All the previous datasets like ImageNet and Pascal are capable of only image classification but with the inception of MS COCO dataset it is possible to identify the objects in the image and establish relationships between these objects i.e. image captioning.

MS COCO dataset has several features such as: object segmentation, recognition in context, super pixel stuff segmen- tation, 1.5 million objects instances, 80 objects categories, 91 stuff categories, 5 captions per image, and 250,000 people with keypoints.

Later in this paper we have described how MS COCO dataset is employed to generate captions and from captions story.

B. BookCorpus Dataset

In order to train our sentence similarity model we collected a corpus of 11,038 books from the web. These are free books written by yet unpublished authors. We only included books that had more than 20K words in order to filter out perhaps noisier shorter stories. The dataset has books in 16 different genres, e.g., Romance (2,865 books), Fantasy (1,479), Science fiction (786), Teen (430), etc. Table 1 highlights the summary statistics of our book corpus.

The BookCorpus dataset is used to train the sentence simi- larity model. In order to train this model, we first transform the sentences in the bookcorpus dataset into a vector using glove, secondly we also transform the captions generated from the image into the similar vector using glove and finally by using cosine similarity or Euclidean distance we find the next 5-6 sentences similar to the image which will give more enhance description of the image.

IV. GENERATING A STORY FROM AN IMAGE

To generate a story from an image we need to train two models: 1) for image captioning to extract features from the image and 2) for context aware similarity to find similar sentences to the captions of image.

A. *Image Captioning*

It is very hard to describe the content shown in the image for a machine. Image captioning is an application of deep learning concepts such as natural language processing and computer vision. Being able to describe the image involves many sub tasks such as understanding the image and then having the semantics correct to describing it. The deep learning concepts computer vision and natural language help the machine to prominently do both the sub tasks efficiently. The computer vision part extracts the features from the image and pass on through a medium to the natural language part which understands the objects through features extracted and the relation between the objects.

The whole model as a part will maximize the likelihood p(S|I) where target is the sequence of words S = {S1, S2,. . .} and I is the image given as input. This idea of image captioning was intended from the recent successful application in machine learning, Machine translation. The same way as machine translation we have one encoder which encodes the image by extracting the features from it and a recurrent neural network decoder which will produce a sequence of words as caption.

The data set we used to produce the caption of the image is MS COCO. MS COCO dataset has several features such as: object segmentation, recognition in context, super pixel stuff segmentation, 1.5 million objects instances, 80 objects categories, 91 stuff categories, 5 captions per image, and 250,000 people with key points.

1) Model: In this model we use a neural and probabilistic architecture which was inspired from the sequence to sequence model which showed that the caption of the image can be generated correctly if we choose a good encoder and a decoder. Our model will be an end to end model where the input and out will be image and sequence of words respectively. Our encoder will convert the varied dimension into a fixed length which will be given as an input to the decoder. Now the decoder will give out the sequence of words in the form of a vector which is mathematical representation of the words, the mathematical representation is generally the word embedding which were formed out of a large meaning full corpus, pretrained. The purpose will be directly maximize the probability of the image to produce its caption. The mathematical formulae goes as

$$\theta^* = arg\ max_{\theta\ (I,S)} \log p(S\ I; \theta)$$

Where, θ are the parameters of our model, I is an image, and

S its correct transcription. We apply the chain rule to calculate the probability of the whole occurrence of the sequence of words based on just the image, so it will be our decoder for the image captioning.

The main attribute of the LSTM is memory cell which will try to remember the important knowledge of the data where the simple recurrent neural network would be

Fig. 1. Samples of Annotated Images in the MS COCO dataset

TABLE I

SUMMARY STATISTICS OF OUR BOOKCORPUS DATASET

# of books	# sentences of	# of word	# of unique word	Mean # of words per sentence	Median # of words per sentence
11,038	74,004,228	984,846,357	1,316,420	13	11

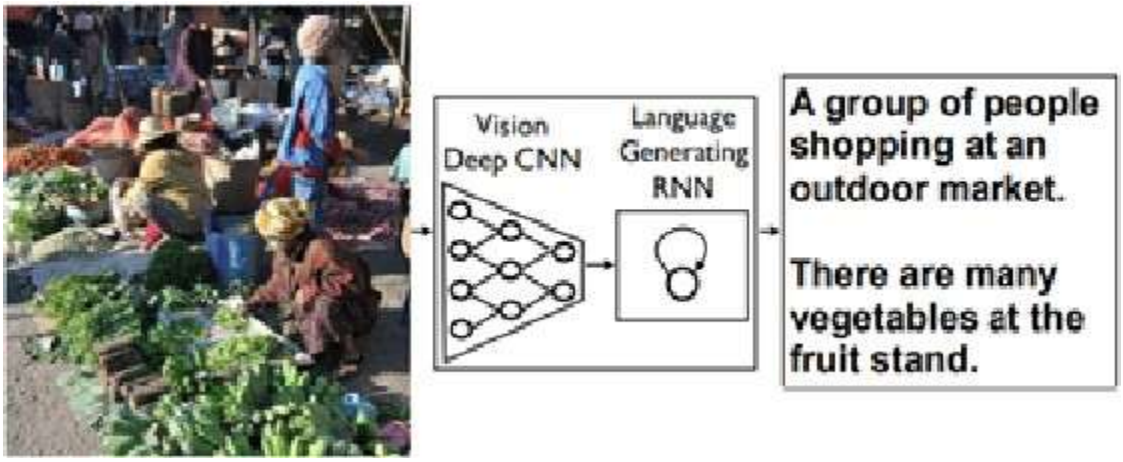

Fig. 2. Example of image captioning.

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t | I, S_0, \ldots, S_{t-1})$$

Now by considering both S and I as the training material

We optimize the log probabilities of whole sequence over whole training set using stochastic gradient descent. It is very common to model the such as recurrent neural network p(St|I, S0, . . . , St 1) where the variable number of words we condition upon t-1 is expressed by a fixed length of hidden states. If we consider in the form of function it would be d

To make RNN more crucial and more efficient we use advanced model of RNN called as LSTM. In encoding part the ImageNet predefined architectures were considered as the best encoders that would extract the features of an image so that the description decoded would be more accurate enough to be accepted.

a) LSTM-based Sentence Generator: The problems such as vanishing and exploding gradients was solved by the advanced model of recurrent neural networks. The selection of this model was inspired from the state of art model of machine translation. The behavior of the cells purely depends upon the gates which will be modified between 0 and 1 whether to keep the information or not. The values in the gates totally depends on the previous hidden states and the current input, the current input is the previous output generated. By seeing the image we can understand the flow of the information from one gate to another gate. The mathematical equations for the gates would be

$$i_t = \sigma(W_{i\,x}\, x_t + W_{im}\, m_{t-1})$$

$$f_t = \sigma(W_{f\,x}\, x_t + W_{fm}\, m_{t-1})$$

$$o_t = \sigma(W_{ox}\, x_t + W_{om}\, m_{t-1})$$

$$c_t = f_t \quad c_t - 1 + i_t \quad h(W_{cx}\, x_t + W_{cm}\, m_t - 1)$$

$m_t = o_t \quad c_t$

$p_{t+1} = Softmax(m_t)$

Where, the various W matrices are trained parameters.

Such multiplicative gates make it possible to train the LSTM robustly as these gates deal well with exploding and vanish- ing gradients. The nonlinearities are sigmoid and hyperbolic tangent h( ). The last equation mt is what is used to feed to a SoftMax, which will produce a probability distribution pt over all words

b) Training: The LSTM model will be trained to predict the word according to the sequence and the probability function such as p(St|I, S0, . . . , St 1) . The probability function will try to maximize the probability of occurrence according to the data that has been provided in the MS COCO. The predefined architecture of Imagenet models will be used as the feature extractors or the encoders in the model which has options of how much we can train. It depends upon the use case, in our case we don't need to train the whole network it will directly give out the specified dimension of vector which can be fed back into the recurrent neural network without any preprocessing. Now the hidden layer of the initial LSTM node would be initialized through the vector from the encoder. The part of the model that gets trained will be the LSTM nodes the output of the first LSTM checked probabilistically p(St|I, S0, , St−x)

$x_{-1} = CNN(I)$

$x_t = W_e S_t, \quad t \in \{0 \ldots\ldots\ldots N-1\}$

$p_{t+1} = LSTM(x_t), t \in \{0 \ldots\ldots\ldots\ldots N-1\}$

We will be using Sp as the stop word, it will be something like <Space> . If the LSTM emitted the stop word means it will be an end to description. The log loss will be calculated and back propagation will be done to alter the gate parameters in order get the maximum probability for the expected sequence. (Refer Figure 4)

The log loss will look like this:

$$L(I,S) = -\sum_{t=0}^{N} \log p_t(S_t)$$

This model uses the LSTM which will not have the bad affects such as vanishing gradient descent and exploding gradient descent. The selection of the pre-defined architecture models such as VGG 16 we also overcome the problem of knowledge through some extent. The training will happen in batches.

B. Context Aware Similarity

We employ sentence similarity methods to compute similar- ities between the captions of the image and each sentence in the book. Meaning of sentences does not only depends on the words present in it but also on the sequence of words present in it. It is collection of words which preserves some meaning. Here in this paper we are more focused on the similarity of the sentences rather than words.(Refer Figure 5)

As we have discussed that meaning of a sentence depends upon two things - Words and there sequence.

Here we are trying 3 models

1. Baselines

In this method we are taking average of word embedding of all words in both sentences and then checking cosine similarity between two.

2. Word Mover's Distance

In this method we are matching two words by finding difference between them, so we will be try to find all the key words from both sentences to match them .So as to find similarity in each key word.

3. Smooth Inverse Frequency

In this method we are removing irrelevant words so that they do not show impact on vectors such as this, that and etc. It does this by removing common component. It computes PCA for result embedding and then it deducts sentence embedding by their first principal component.

All these models have some flaws as first one do not take order of words in account, second one also depends solely on words rather than their order.

So here we tried different approach

1. Shortest path distance from WordNet

It relates parts of sentences or speech, it comprises of 4 things nouns, verbs, adjectives and adverbs. So we do not find similarity in cross dimensions rather in there subnets only. Shortest distance is calculated using hierarchical structure. In order to do so we climb up a hierarchy from both synsets and determine a meeting point. This meeting point or synset is called subsumer and the shortest distance equals the hops from one synset to another .We consider these to find the hierarchical distance or the shortest path.

2. Information content of word

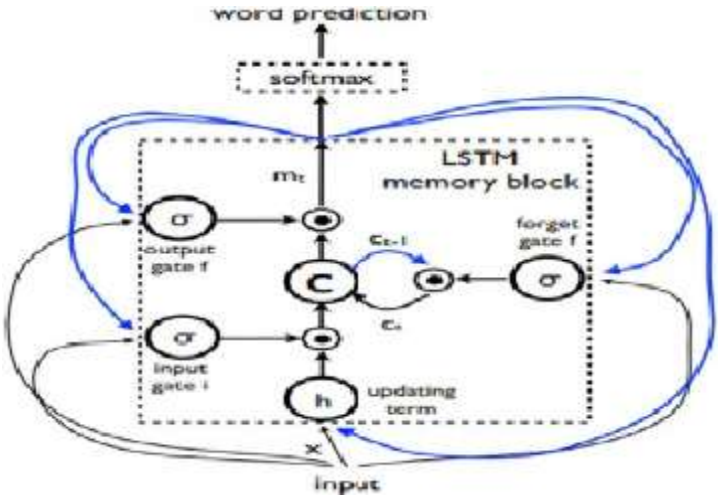

Fig. 3. LSTM : the memory block contain a cell c which is controlled by 3 gates. In blue we show the recurrent connections - the output m at time t-1 is fed back to memory at time t via three gates : the cell value is fed back via forget gate; the predicted word at time t-1 is fed back in addition to the memory output m at time t into the Softmax for word prediction.

We are using statistical information from wordNet for information of word. Using WordNet we are finding frequency of each synset like jewellery - necklaces will have some high similarity value and jewellery- sword will have very lowvalue.

3. Final Step

After finding the synsets, we find the shortest path distances between all the synsets and take the most favourable form a semantic vector. There is also an Intermediate step before making final vector, there are L1 and L2 lists. So here we get S1 and S2 from L1 and L2 respectively. Now cross-comparison is done for all the words from S1 and S2. After this we now determine an index entry for semantic vector. The words with highest similarity value are given more priority while forming vectors V1 and V2.Now we will calculate the dot product or take the cosine similarity of these two vectors to get our result.

Conclusion of Context Aware Similarity

This algorithm uses POS (parts of speech) by tagging each part of sentence so as to ensure that we are comparing rights words and the true meaning of word is clear. Similarity be- tween the words is done using edge based approach. And then semantic vectors are formed to calculate similarity between sentences. Word vectors formed to calculate the impact of the syntactic impact on the sequencer structure of the words in a sentence. The order of word is given less priority or weighted less as in a sentence the meaning of a words is more informative then the sequence of the words.

So overall we are performing two steps first we performing Image captioning using VGG-16 predefined architecture as encoder and LSTM as decoder. We are using MSCOCO dataset for training purpose. We are using predefined weights for encoder and using its output as an input for our decoder. This first stage will take Image as an input and will give caption as an output .We are having a novel dataset containing novels of romance genre, we are segmenting the novel into paragraph and sentences which will decrypt a scenario sp that we can map it to our image .The second stage of our project is of reforming sentence to sentence mapping, here we are finding similarity between the sentences. The output of the first stage will be input to our second stage, second model will take caption as an input and will find similar story by using cosine difference between vectors, so our final result will give story similar to the Image.

V. CONCLUSION

In this paper, we explored a new problem of generating a story about an image which is far more descriptive than the image captions. We proposed an approach that computes the several similarity between the image captions and the sentences in the book. We exploited context aware similarity in order to compute similarities between sentences. This approach can further be improved to train models on specific books for a specific type of application. Using this approach

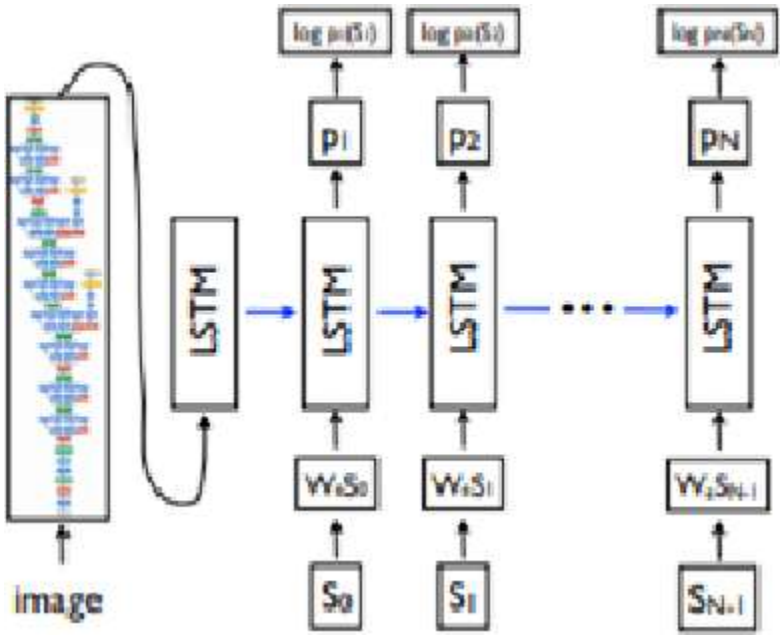

Fig. 4. LSTM model combined with a CNN image embedderand word embeddings. The unrolled connections between the LSTM memories are in blue and they correspond to the recurrent connections in figure 3. All LSTMs sharethe same parameters.

different applications can be developed for the purpose of interactive learning, photo album summarisation and entertainment industry.

ACKNOWLEDGMENT

We would like to take this opportunity to thank our project coordinator Prof. Anup Kadam and Prof. Sagar Rane sir for giving us all the help and guidance we needed. We are grateful to them for their kind support. Their valuable suggestions were very helpful.

We are also grateful to Prof. Dr. Sunil R. Dhore, Head of Computer Engineering Department, AIT, Pune for his indispensable support and suggestions.

REFERENCES

[1] R. Kiros, R. S. Zemel, and R. Salakhutdinov, "Unifying visual-semantic embeddings with multimodal neural language models. CoRR,," pp. 2014– 13.

[2] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for gen- erating image descriptions." in CVPR, 2015.

[3] S. Bengio, D. Erhan, and O. Vinyals, "Show and tell: A neural image caption generator." in arXiv:1411.4555,2014.

[4] C. Kong, D. Lin, and M. Bansal, "What are you talking about? text-to- image coreference." in CVPR,2014.

[5] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books." 2015.

[6] C. Vondrick, H. Pirsiavash, and A. Torralba, ""Inferring the why in images". arXiv.org," 2014.

[7] G. Kulkarni, S. Li, T. Berge, V. Premraj, S. Dhar, Y. Choi, and A. Berge, "Baby talk: Understanding and generating simple image descriptions." in CVPR, 2011.

[8] K. Xu, R. Kiros, R. Zemel, J. Ba, K. Cho, A. Courville, R. Salakhutdinov, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention." in arXiv:1502.03044, 2015.

[9] N. Tandon, B. Schiele, and A. Rohrbach, "A dataset for movie descrip- tion." in CVPR, 2015.