Gaussian Smoothen Semantic Features (GSSF) - Exploring the Linguistic Aspects of Visual Captioning in Indian Languages (Bengali) Using MSCOCO Framework

Chiranjib Sur
Computer & Information Science & Engineering Department, University of Florida.
Email: chiranjibsur@gmail.com

Abstract—In this work, we have introduced Gaussian Smoothen Semantic Features (GSSF) for Better Semantic Selection for Indian regional language-based image captioning and introduced a procedure where we used the existing translation and English crowd-sourced sentences for training. We have shown that this architecture is a promising alternative source, where there is a crunch in resources. Our main contribution of this work is the development of deep learning architectures for the Bengali language (is the fifth widely spoken language in the world) with a completely different grammar and language attributes. We have shown that these are working well for complex applications like language generation from image contexts and can diversify the representation through introducing constraints, more extensive features, and unique feature spaces. We also established that we could achieve absolute precision and diversity when we use smoothened semantic tensor with the traditional LSTM and feature decomposition networks. With better learning architecture, we succeeded in establishing an automated algorithm and assessment procedure that can help in the evaluation of competent applications without the requirement for expertise and human intervention.

Index Terms—image description, language translation, visual captions, semantic learning, Bengali captions.

I. INTRODUCTION

Image captioning [11] applications have gained massive attention from the research community both in academia and industry due to its ability to find a confluence of media and language and create the capability for machines to communicate in different languages with the external world. Image captioning has a wide range of applicability to tag enormous resources in the form of images [48] and video, and vast amounts of new media contents are added daily, which are un-categorized, and no one knows what its content is. This kind of image captioning brings these media in the mainstream web and can enhance the accurate description of the images.

In this work, we have mainly concentrated on the viability of the other language-based image captioning systems. While English is a structural language (syntactic and grammatical), there are unstructured languages (only grammatically) like most of the Indian languages, which are derived from Sanskrit like Bengali (90% Sanskrit vocabulary), Tamil (70% Sanskrit vocabulary), Hindi (50% Sanskrit vocabulary), etc. to name a few. We will be mainly focusing on the Bengali language as a case study and demonstrate how they do for the image captioning counterparts. We have used the Google Translation API (googletrans.Translator) for our work. Though the Bengali language is unstructured, the captioning model performed well for generating the captions. However, this is the first of its kind, never tried research-topic. Due to the absence of any such dataset, we have used a translation application for data generation. It is justified that such a model can be built for a language like Bengali, as it is the fifth-largest widely spoken language in the world and is one of the primitive languages of the world. We have also demonstrated that if we provide a Gaussian smoothening of the semantic layer, we can deliver better captions.
A. Image Caption Problem

The problem of image captioning is rooted in understanding the network, relationship, interaction of the different objects and descriptions. These connections are evident to humans but had to be taught to machines to be able to leverage from these visual features. It is difficult to generalize for different images if the sentences are generated from the object combinations through the determination of the sequentiality and priority of these objects, it is an NP-Hard problem. The models are made to determine the region and the objects in the image heuristically and then the description. It is said to be more than object detection and object subcategory or details interpretation. Image captioning in regional languages require immense exploration with the increasing number of applications expressing with images. Also, a large part of the Indian subcontinent is still dependent on regional language.
as their source of expression. These regional language-based expressions comprise a form of art as these applications integrate well and get constantly improved through human intervention. Image captioning in different Indian languages has its challenges, where each of these languages possesses a different set of language attributes, grammatical properties, and completely different from English and other European languages. These Indian languages are entirely different from their root language Sanskrit. The diversification of these languages was created through different oral transformations, which was evident in ancient India due to the absence of texts and learning centers for everyone. Languages and customs propagated through “Smriti” (or oral propagation) through different people, which later transferred to other parts of the world, including Europe. This fact is evident from the common words like “Matri” became Mother, “Pitri” became Father, “Brahta” became Brother, “Jamity” become Geometry and even “Om” became Amen. However, from the prospect of grammar and way of construction, Indian languages are far different from Latin-originated languages, and in this work, we made an effort to generate sentences in these primitive and ancient culture languages.

The rest of the document is arranged with revisit of the existing works in literature in Section II, the description and statistics of the data in Section III learning network and representation detailed description in Section IV the intricacies of our methodology in Section V experiments, results and analysis in Section VI concluding remarks with future prospects in Section VI.

The main contribution of this work: 1) Bengali (Indian) Language Caption Generators 2) Framework for Translator based training 3) Gaussian Smoothen Semantic Features for Better Semantic Selection 4) Compositional-Decompositional Network Framework for better representation for captions.

II. REFERENCE EXPLOITATION & GOOGLE TRANSLATE

There are around 31 official languages, while the constitution recognizes 22 of these languages. As a result, every such language provides a different prospect of these images, and the development of machine generation in these languages provide better solutions to many applications, including automatic communicator, regional aid provider, and the list is endless. In this work, we made an effort to develop a caption generator and generated a training dataset of captions from the existing English MSCOCO dataset through the use of Google Translation API (googletrans.Translator). While it is true that all the translated captions are not perfect, but it can, at least, provide a good starting point and provide some evaluation standard. Table I provided an estimation of the vocabulary sizes, while Figure 2 provided some instances of the translated captions in different languages for some images. However, we have considered some part of the vocabulary, say top 50%, for likelihood generation, as the whole vocabulary is considerably long, and there are chances of sparsity and negligence of some words due to scarcity of samples in sentences. To find the correct set of vocabulary, we experimented to determine the best set and the cutoff. In Figure 2 being a native speaker of Bengali, we have provided other captions that would be much more realistic than the translated version.

III. LEARNING NETWORK AND REPRESENTATION DESCRIPTION

Learning the representations and correct configuration of the network is essential for the success of language generators. While image feature is helpful, semantic distribution features help in enhancing the captions through direct correlation of the selected objects in images and the likelihood of words in sentences. In this work, we have introduced new architectures utilizing different feature spaces and reported their qualitative and quantitative analysis. While the semantic features are highly sparse, we introduced a Gaussian smoothing for semantic features and shown that the caption gets improved with these kinds of smoothing. Smoothing helps in better capturing of the image region as attention to be decomposed at the hidden level. In this work, we have focused on compositional and decomposition-based architectures as they provide some of the state-of-the-art performance [70] for caption generator without the explicit requirement for large scale image segmentation and region-based feature extraction like [37]. However, they are good ways of expansion of the network and generate better descriptions for images. Our argument is that we can derive better network through decomposition of the causal representation \( h_{x-1} \) than decomposition of both causal \( h_{x-1} \) and previous \( x_{-1} \) contexts. Sequential recurrent neural network has been widely used in generative applications and mathematically, the semantic concept layer is initialized for these networks as the following.

\[
\begin{align*}
  h_0 &= MLP_h(v) \\
  c_0 &= MLP_c(v)
\end{align*}
\]

where \( v \) refers to visual features and \( MLP(.) \) referred to multiple-layer perceptron. Figure 1 provided a generalized architectural overview of the image caption generation system. Later these layer is fused with the Gaussian smoothen semantic features and provided with the perfect context representation for each word.

A. Long Short Term Memory

Long Short Term Memory (LSTM) is widely used sequential recurrent unit because of its scalability and its unmatched generalized solutions for many sequential generation related to languages and other distinct representation learning for classification and most importantly, apart from suppression of the variations or approximation, like any other deep learning architecture, it helps in better generalization. LSTM caption
Mathematically, Long Short Term Memory (LSTM) model, denoted as $f_{L}(\cdot)$, can be described as the followings probability distribution estimation.

$$f_{L}(v) = \prod_{k} \Pr(w_{k} | v, W_{L,1}) \prod \Pr(v | I, W_{1})$$  \hspace{1cm} (12)

using the weights of the LSTM in the architecture is denoted as $W_{L,1}$, $w_{i}$ as words of sentences, $v$ as image features, $Q_{IC}(\cdot)$ and $Q(\cdot)$ are the Image Caption and feature generator function respectively. $Q(\cdot)$ derives $v$ from image $I$.

### B. Gaussian Smoothened Tag Long Short Term Memory

Gaussian Smoothened Tag Long Short Term Memory (GST-LSTM) involves consecutive operations of selection and compositional fusion of different context features, including a semantic layer feature set for better representation. Unlike GSSCN-LSTM, where the input and hidden layer are decomposed through fusion, GST-LSTM only involves decompositional fusion of different context features, including a semantic layer feature set for better representation.

For categorization, $y_{i}$ is evaluated through convergence to the categorization distribution through softmax layer defined as,

$$y_{i} = \sigma(y_{i}) = \frac{\exp(y_{i})}{\sum_{k=1}^{C} \exp((y_{i})_{k})} \hspace{1cm} (10)$$

where we have $\sigma(y_{i}) \in [0, 1]^{C}$ with $C$ as the set of categories. Also we have $x_{t} \in \mathbb{R}^{m}$, $y_{t} \in \mathbb{R}^{C}$, $W_{hy} \in \mathbb{R}^{C \times d}$, $i, f, o, g, c \in \mathbb{R}^{d}$, $h_{t} \in \mathbb{R}^{d}$, $W_{x, s} \in \mathbb{R}^{d \times m}$, $W_{h, s}, W_{c, s} \in \mathbb{R}^{d \times d}$, $b_{s} \in \mathbb{R}^{d}$. The objective minimization function is defined as

$$J(W) = \arg \min_{W} \frac{1}{2T} \sum_{t=1}^{T} \sum_{k=1}^{S} ||y_{t,k} - y_{t,k}||^{2} = \arg \min_{W} \frac{1}{2T} \sum_{k=1}^{S} ||y_{t,k} - y_{t,k}||^{2}$$

and $s$ number of training samples. The parameters are updated with $\alpha \frac{\partial J(W)}{\partial W}$. Value of $\alpha$ determines the learning rate for adaption with the changing topology of the objective function space.

$$x_{t} = (\arg \max(h_{t}, W_{hy})) W_{E} = (\arg \max(y_{t})) W_{E}$$ \hspace{1cm} (11)

where $W_{E} \in \mathbb{R}^{V \times d}$ is the word embedding representation of vocabulary size $V$ and each word is represented with vector of dimension $d$. In most of the architecture, $h_{t} \rightarrow x_{t}$ involves Equation [11] and is a convenience and may not be explicitly mentioned with every model but is implied. Extra set of improvement is
states $h_t$, $c_t$ get fades away with time, and extra attention can revive the image quality of the memory and enhance the performance. The main iteration of the model consists of the following set of equations.

$$\hat{S} = f_s(S)$$  \hspace{1cm} (13)

where we have defined $f_s(.)$ as the Gaussian Smoothing function. $S$ is sparse and operates as a diagonal matrix. In the diagonal matrix, the columns are orthogonal to each other, and the task of each column is to select the feature of interest and discard the others. These columns come directly from the image, and hence we talk about a column tensor and the whole matrix. As we operate different transformations and train the weights, the strength of these individual columns degrades or distributed unevenly. Hence, Gaussian smoothing will help in absorbing better features as it transforms the discrete points to a better structural absorbent. We call $\hat{S}$ as Gaussian Smoothen Semantic Features (GSSF). Due to limitations in GPU memory, we have kept the dimension of $S$ and $S$ same.

$$h_{t-1} = W_{h,m} \hat{S} \odot W_{h,n} h_{t-1}$$  \hspace{1cm} (14)

$$i_t = \sigma(W_x x_{t-1} + W_h h_{t-1} + b_i)$$  \hspace{1cm} (15)

$$f_t = \sigma(W_x x_{t-1} + W_h h_{t-1} + b_f)$$  \hspace{1cm} (16)

$$o_t = \sigma(W_x x_{t-1} + W_h h_{t-1} + b_o)$$  \hspace{1cm} (17)

$$g_t = \text{tanh}(W_x x_{t-1} + W_h h_{t-1} + b_g)$$  \hspace{1cm} (18)

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$  \hspace{1cm} (19)

$$h_t = o_t \odot \text{tanh}(c_t)$$  \hspace{1cm} (20)

Mathematically, Gaussian Smoothened Tag Long Short Term Memory (GST-LSTM), denoted as $f_{GST}(.)$, can be described as the followings probability distribution estimation.

$$f_{GST}(v) = \prod_k \Pr(w_k \mid v, \hat{S} \odot h_{t-1}, W_{L1}) \prod \Pr(S \mid v, W_2)$$

$$\prod \Pr(v \mid I, W_1)$$

$$= \prod_k Q_{IC}(w_k \mid v, \hat{S} \odot h_{t-1}) \prod Q_S(S \mid v) \prod Q(v \mid I)$$  \hspace{1cm} (21)

using the weights of the LSTM in the architecture is denoted as $W_{L1}$, $w_s$ as words of sentences, $v$ as image features, $S$ as semantic features from $Q_S(.)$, $\hat{S}$ as Gaussian Smoothened features, $Q_{IC}(.)$ and $Q(.)$ are the Image Caption and feature generator function respectively. $Q(.)$ derives $v$ from image $I$.

C. Gaussian Smoothened Semantic Long Short Term Memory

Gaussian Smoothened Semantic Long Short Term Memory (GSSCN-LSTM) operated a combination of decomposition of both previous context and hidden layer information and compared to GST-LSTM, it involves further decomposition of features, which are fused to compose representations at a higher level for caption generation, but at the cost of series of weighted transformation weights. The set of equations that describe GSSCN-LSTM is provided below. ([42], [70]) used the decomposition techniques in their architecture.

$$\hat{S} = f_s(S)$$  \hspace{1cm} (22)

where we have defined $f_s(.)$ as the Gaussian Smoothing function.

$$x_{s,t} = W_{x,s;n} \hat{S} \odot W_{x,n} x_{t-1}$$  \hspace{1cm} (23)

$$h_{s,t} = W_{h,s;n} \hat{S} \odot W_{h,n} h_{t-1}$$  \hspace{1cm} (24)

where we have $S$ as the tag distribution and $\hat{S}$ is the Gaussian smoothen tag distribution, $v$ as the image features, $* = i/f/o/g$, $W_{s,n} \in \mathbb{R}^{a \times b}$, $W_{n} \in \mathbb{R}^{d \times s}$, $h_{s,t} \in \mathbb{R}^{a \times b}$ and $x_{s,t} \in \mathbb{R}^{d \times b}$. The decomposed features are processed in the memory network through the following set of equations.

$$i_t = \sigma(W_x x_{t-1} + W_h h_{t-1} + b_i)$$  \hspace{1cm} (25)

$$f_t = \sigma(W_x x_{t-1} + W_h h_{t-1} + b_f)$$  \hspace{1cm} (26)

$$g_t = \sigma(W_x x_{t-1} + W_h h_{t-1} + b_g)$$  \hspace{1cm} (27)

$$o_t = \sigma(W_x x_{t-1} + W_h h_{t-1} + b_o)$$  \hspace{1cm} (28)

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$  \hspace{1cm} (29)

$$h_t = o_t \odot \text{tanh}(c_t)$$  \hspace{1cm} (30)

where $W_{x,s} \in \mathbb{R}^{w \times s}$, $W_{h,s} \in \mathbb{R}^{d \times s}$, where $d$ is the hidden layer dimension, $w_c$ is the word embedding dimension and $s$ is the semantic dimension. Here, we have replaced semantic feature $S$ with Gaussian Smoothen Semantic Features (GSSF) $\hat{S}$ and a gradual smoothen feature helps in better selection for the next phase from the previous states, generated from the model. It must be mentioned that $h_s$ is shadow image representation generated from the image features and continuous fusion of the semantic features and the image features.

Mathematically, Gaussian Smoothened Semantic Long Short Term Memory (GSSCN-LSTM), denoted as $f_{GSSCN}(.)$,
can be described as the following probability distribution estimation.

\[
    f_{\text{GSSCN}}(v) = \prod_k \Pr(w_k | v, \hat{S} \odot h_{t-1}, \hat{S} \odot x_{t-1}, W_L1) \\
    \prod_k \Pr(S | v, W_2) \prod \Pr(v | I, W_1) \\
    = \prod_k Q_{IC}(w_k | v, \hat{S} \odot h_{t-1}, \hat{S} \odot x_{t-1}) \\
    \prod Q_S(S | v) \prod Q(v | I) \\
\]

using the weights of the LSTM in the architecture is denoted as \(W_L1\), \(w_i\) as words of sentences, \(v\) as image features, \(S\) as semantic features through \(Q_S(.)\), \(\hat{S}\) as Gaussian Smoothened features, \(Q_{IC}(.)\) and \(Q(.)\) are the Image Caption and feature generator function respectively. \(Q(.)\) derives \(v\) from image \(I\).

IV. DETAILS OF METHODOLOGY

We have provided some implementation level information, though this is very common for language generator and image captioning research. There are intricacies of the procedures and are considered as the following.

a) Step 1: Data Acquisition: We have used the MSCOCO data, and some of them are shared by [70], which coordinates with the ResNet features and can be easily obtained.

b) Step 2: CNN Transfer Learning Feature: Here, we have extensively used ResNet through transfer learning with 2048 dimension.

c) Step 3: Transfer Learning Translation: The Bengali captions are obtained through Google Translation API (google-trans.Translator) and a plugin provided by them. Since it is a public plugin and protected by DOS attack, the number of times it can translate for you on a certain day is limited, and hence it took me several days to translate the data. Also, since Bengali is unstructured language, the vocabulary dimension is very large and needs to be cleaned and processed with by 5% occurrence rate. The English vocabulary is also needed to be shrink-ed by 5% occurrence rate.

d) Step 4: Transfer Learning Semantic: After the CNN features are extracted, Tag features with 999 dimensions are used, and this is a neural network transformation of the image features.

e) Step 5: Gaussian Smoothing of Semantic: Here, we explain and introduced Gaussian Smoothen Semantic Features (GSSF). Figure 4 provided an instance for an illustration of the effects of the Gaussian Smoothing procedure on semantic features.

f) Step 6: Training Sessions: We trained the LSTM network end-to-end through the use of the all available data, but without the CNN network as it becomes hefty and prone to over-fitting.

g) Step 7: Testing in Original Language Space: We processed the evaluation of the generated Bengali sentences with the ground-truth Bengali sentences.

h) Step 8: Testing in Reference Language Space: For better analysis and clarity, we translated the Bengali sentences and evaluated the performances in the English language to get an overview in comparison to the English language caption generator. Also, it provides some reference of how far we need to go before we can get a perfect captioner. But, unlike English, which is a grammatically structured language, Bengali (with 90% Sanskrit vocabulary) is an unstructured language, and direct comparison of relative performance is difficult.

i) Step 9: Qualitative Evaluation for Semantics Correctness: Also, we have made some semantic relatedness through visual inspection in Qualitative Evaluation section.

V. RESULTS & ANALYSIS

This section provides some details of the experiments. We have used different evaluation techniques to put up different aspects of the generated captions in Bengali language. We also provided instances of the generated caption in Evaluation
of this kind of translation based caption generator analysis is based on two fronts, mainly because no expert can legitimatize the correctness of all these languages, and in the absence of grammar, the semantics of these languages are difficult to define. However, we can offer qualitative evaluations for the working versions of the sentences in these languages. Nevertheless, we have provided both qualitative and quantitative metric based comparisons below.

A. Dataset
We have used the same training-validation-testing split of MSCOCO data that has been used in all the papers, which consists of 123287 train+validation images and 566747 train+validation sentences. Here, each image is associated with at least five sentences from a vocabulary of 8791 words. There are 5000 images (with 25010 sentences) for validation and 5000 images (with 25010 sentences) for testing. We used the same data split, as described in Karpathy’s paper.

For our network, two sets of features are being used: one is ResNet features with 2048 dimension feature vector, and another is the semantic representation with feature vector of 999 dimensions and consisted of the likelihood probability of occurrence for the most appearing set of objects, attributes and interaction criteria as set of combinations of incidents for the dataset. For the translated data, we have restricted the vocabulary of the sentences to 20K and used 32 and 64 batch size based training sessions for each of the models. Also, we found that 1024 dimensions of the hidden layers provided the best performance, while the word embedding is loaded from Stanford GloVe 300 dimension pre-trained model.

B. Evaluation Metrics
Different metrics like CIDEr-D, METEOR, ROUGE_L, BLEU_n, and SPICE are used for our evaluation. However, we think that the BLEU_4 metric is the most sensible way of evaluation as this metric technique takes care of the continuity of object-attribute relationships and at the same time provides organizes the best possible descriptions. Also, BLEU_4 evaluation considers several consecutive lineages based on nouns, adjectives, adverbs from the ground truth, which can never be judged through other criteria. We used Microsoft COCO dataset for our evaluation, and the results are reported based on Karpathy’s division of training, validation, and testing set.

C. Evaluation Procedures
The first evaluation is in Bengal vocabulary space and the performance is dependent on the vocabulary space of the training model. If we consider the function of the translation as $Tr(.)$ and the evaluator function as $f_Q(.)$, then we define the two different evaluation criteria as the followings.

$$E_1 = f_Q(G_L, Tr(R_E)_L)$$
$$E_2 = f_Q(Tr(G_L)_E, R_E)$$

where $Tr(.)_L$ is the translation function to language $L$ and $G_L$ is the generated captions in language $L$ and $R_E$ is the reference caption in language $E$. We define evaluator function $f_Q(.)$ as,

$$f_Q(.) \in \{\text{BLEU}_n, \text{CIDEr}, \text{ROUGE}_L, \text{METEOR}, \text{SPICE}\}$$

Also, The two reference languages frames denoted as $L$ (Any Language other than English) and $E$ (English). The evaluation criteria had varied ranges of metrics, and we evaluated the generated sentences with each of them due to the limitation of grammatical knowledge of different languages and the way the machine generates captions in different languages. Apart from the different evaluation metrics, we have also used Beam Search for generating longer sentences or at least sentences with a higher cumulative probability value.

D. Quantitative Analysis
We made a comparative study of the different architectures and compositional extraction of different representations. Quantitative analysis helps in understanding the capability of
the model in objects based inference in sentences and its distance from the ground truth and also their ability to compose a different form of short sequences from image contexts. Table II provided a comparative overview of evaluation based on the language itself based on the reference of that language. In contrast, Table III provided the same metric evaluation when the generated captions are translated to English language, and a comparison is made based on the ground truth of the reference of the English language. From Table II it is evident that the captioner generator performed very well. Even though the source of data is a translator, which has not reached perfection and is just a prototype with errors. This limitation of the translation models is mainly due to the unstructured nature of the ancient languages like Bengali, which is directly related to Sanskrit and was there from pre-historic times. From Table III we can say that our analysis is limited to constrained vocabulary (20K most appearing Bengali words) and given a better data, this approaches can be improved with precision and as a native Bengali speaker, this model can solve the problem of Bengali language based image captioner applications. We have also used beam search with beam size as 5 and also made sure that the repetition is prevented from appearing as it reduces the performance. Also, we have utilized a dropout rate of 0.5 for the embedding, image features, entry points of the language decoder and the exit points of the language decoders. 0.5 dropout rate provides ample scope for all the variables to be selected or rejected entirely and is very helpful to avoid over-fitting.

E. Discussion

For analysis of the results, we found that our proposed model GST-LSTM model performed the best and also much lower (in the number of weights) in comparison to GSSCN-LSTM and better than LSTM base model in performance. GST-LSTM out-performed GSSCN-LSTM and LSTM by 1% and 4% respectively for BLEU_4 metric and better data will even improve the relative performances. The main reason why the GSSF model performed better has diverse explanations. The addition of the semantic level information and increasing the spread using Gaussian Smoothening (GSSF) has helped considerably. This approach can improve the results as it is providing a better concept layer for the generation of the representation for the captions. GSSF provides a new technique to upgrade the sparse semantic tensor information, mainly when it is acting as a diagonal matrix and its prime work is the selection of the structural information from the media contents. Other factors that helped in better captioning is through the extraction of the usable vocabulary space and also restricting the word embedding. We have used the word embedding of Stanford GloVe through the pre-trained model and gradually smoothened it. Also, for assigning the embedding vector, we have taken the help of a translator model to map the Bengali words with its English counterparts. Table II and Table III provided the numerical results.

F. Qualitative Analysis

It is very difficult to evaluate whether a model is better than the other through averaged numerical. Hence, we have provided some instances of the generated captions in Figure 7 and Figure 8. Whether overall improved captions are generated or not is also difficult to judge from numerical in Table II and Table III. Hence, the following qualitative analysis is adopted and will reflect some of them. Most of the caption generated work used diverse evaluation methods for the generated captions, but there is a requirement of analysis of quality and context correctness of these languages. Also, qualitative metrics are the best way of understanding the acceptability of the generated sentences and hence in this part, we have demonstrated some instances with the original images. Figure 7 and Figure 8 provided some instances of the different caption of different Indian languages. From these generated captions, we can say that the model performed quite well and the sentences represent much better description of the situations.

VI. Conclusion & Future Works

In this work, we have introduced a new memory unit architecture for sequential learning and utilized it for image captioning applications for different unstructured (grammatically) Indian languages like Bengali and is the fifth widely spoken language in the world. The features include a different set of language attributes and grammatical properties and completely different from English. While we have used Bengali language as our reference, other Indian languages can be used in this way as well. The quantitative evaluation criteria for performance can be regarded as a reference, while the qualitative evaluation criteria can be seen as a perfect judge for many situation narrations for images. We devised different architectures (GST-LSTM and GSSCN-LSTM) to provide a comparison with baseline LSTM and also provided a new technique to upgrade the sparse semantic tensor information, mainly when it is acting as a diagonal matrix. Its prime work is a selection of the structural information from the media contents. Since this is the very first work, lots of further work can be done in this paradigm and our performance metric can be used for reference and baseline.

ACKNOWLEDGMENT

The author has used University of Florida HiperGator, equipped with NVIDIA Tesla K80 GPU, extensively for the experiments. The author acknowledges University of Florida Research Computing for providing computational resources and support that have contributed to the research results reported in this publication. URL: http://researchcomputing.ufl.edu

REFERENCES

[1] Sur, C. (2019). Survey of deep learning and architectures for visual captioning-transitioning between media and natural languages. Multimedia Tools and Applications. 1-51.
[2] Lu, D., Whitehead, S., Huang, L., Ji, H., & Chang, S. F. (2018). Entity-aware Image Caption Generation. arXiv preprint arXiv:1804.07889.
[3] Lu, J., Yang, J., Batra, D., & Parikh, D. (2018, March). Neural Baby Talk. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7219-7228).
[4] You, Q., Jin, H., & Luo, J. (2018). Image Captioning at Will: A Versatile Scheme for Effectively Injecting Sentiments into Image Descriptions. arXiv preprint arXiv:1801.10121.
### TABLE II
PERFORMANCE EVALUATION: ORIGINAL GENERATED VS ORIGINAL GROUND TRUTH

| Algorithm  | Language | CIDEr-D | Bleu_4 | Bleu_3 | Bleu_2 | Bleu_1 | ROUGE_L |
|------------|----------|---------|--------|--------|--------|--------|---------|
| LSTM       | Bengali  | 0.93    | 0.31   | 0.42   | 0.55   | 0.70   | 0.53    |
| GST-LSTM   | Bengali  | 0.99    | 0.35   | 0.48   | 0.63   | 0.77   | 0.57    |
| GSSCN-LSTM | Bengali  | 1.01    | 0.34   | 0.44   | 0.38   | 0.73   | 0.55    |

### TABLE III
PERFORMANCE EVALUATION: ENGLISH TRANSLATED VS ENGLISH GROUND TRUTH

| Algorithm  | Language | CIDEr-D | Bleu_4 | Bleu_3 | Bleu_2 | Bleu_1 | ROUGE_L |
|------------|----------|---------|--------|--------|--------|--------|---------|
| LSTM       | Bengali  | 0.62    | 0.24   | 0.34   | 0.46   | 0.58   | 0.39    |
| GST-LSTM   | Bengali  | 0.80    | 0.25   | 0.36   | 0.49   | 0.66   | 0.46    |
| GSSCN-LSTM | Bengali  | 0.78    | 0.25   | 0.35   | 0.49   | 0.66   | 0.46    |

Fig. 7. Qualitative Analysis of Captions. Set 1.

[5] Melnyk, I., Sercu, T., Dognin, P. L., Ross, J., & Mroueh, Y. (2018). Improved Image Captioning with Adversarial Semantic Alignment. arXiv preprint arXiv:1805.00063.

[6] Wu, J., Hu, Z., & Mooney, R. J. (2018). Joint Image Captioning and Question Answering. arXiv preprint arXiv:1805.08389.

[7] Kilickaya, M., Akkus, B. K., Caicici, R., Erdem, A., Erdem, E., & Ikizler-Cinbis, N. (2017). Data-driven image captioning via salient region discovery. IET Computer Vision, 11(6), 398-406.

[8] Chen, F., Ji, R., Su, J., Wu, Y., & Wu, Y. (2017, October). Structcap: Structured semantic embedding for image captioning. In Proceedings of the 2017 ACM on Multimedia Conference (pp. 46-54). ACM.

[9] Jiang, W., Ma, L., Chen, X., Zhang, H., & Liu, W. (2018). Learning to
guide decoding for image captioning. arXiv preprint [arXiv:1804.00887]
[10] Wu, C., Wei, Y., Chu, X., Su, F., & Wang, L. (2018). Modeling visual and word-conditional semantic attention for image captioning. Signal Processing: Image Communication.
[11] Fu, K., Li, J., Jin, J., & Zhang, C. (2018). Image-Text Surgery: Efficient Concept Learning in Image Captioning by Generating Pseudopairs. IEEE Transactions on Neural Networks and Learning Systems, 29(9), 1-12.
[12] Chen, H., Ding, G., Lin, Z., Zhao, S., & Han, J. (2018). Show, Observe and Tell: Attribute-driven Attention Model for Image Captioning. In IJCAI (pp. 606-612).
[13] Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14(2), 48.
[14] Zhao, W., Wang, B., Ye, J., Yang, M., Zhao, Z., Luo, R., & Qiao, Y. (2018). A Multi-task Learning Approach for Image Captioning. In IJCAI (pp. 1205-1211).
[15] Liu, C., Sun, F., Wang, C., Wang, F., & Yuille, A. (2018). Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. arXiv preprint [arXiv:1803.08314].
[16] Chunseong Park, C., Kim, B., & Kim, G. (2017). Attend to you: Personalized image captioning with context sequence memory networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1343-1353).
[17] Tavakoliy, H. R., Shetty, R., Borji, A., & Laaksonen, J. (2017, October). Paying attention to descriptions generated by image captioning models. In Computer Vision (ICCV), 2017 IEEE International Conference on (pp. 2506-2515). IEEE.
Anne Hendricks, Lisa, et al. “Deep compositional captioning: Describing

Karpathy, Andrej, and Li Fei-Fei. “Deep visual-semantic alignments for visual recognition and description.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Gan, Chuan, et al. “Stylenet: Generating attractive visual captions with style.” CVPR, 2017.

Jin, Junqi, et al. “Aligning where to see and what to tell: image caption with region-based attention and scene factorization.” arXiv preprint arXiv:1706.06272 (2017).

Kiros, Ryan, Ruslan Salakhutdinov, and Richard S. Zemel. “Unifying visual-semantic embeddings with multimodal neural language models.” arXiv preprint arXiv:1701.07811 (2017).

Kiros, Ryan, Richard Zemel, and Ruslan R. Salakhutdinov. “A multiplicative model for learning distributed text-based attribute representations.” Advances in neural information processing systems. 2014.

Mao, Junhua, et al. “Deep captioning with multimodal recurrent neural networks (m-rnn).” arXiv preprint arXiv:1412.0322 (2014).

Memisevic, Roland, and Geoffrey Hinton. “Unsupervised learning of image transformations.” Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007.

Pu, Yunchen, et al. “Variational autoencoder for deep learning of images, labels and captions.” Advances in Neural Information Processing Systems. 2016.

Socher, Rich, et al. “Grounded compositional semantics for finding and describing images with sentences.” Transactions of the Association for Computational Linguistics 2 (2014): 207-218.

Sutskever, Ilya, James Martens, and Geoffrey E. Hinton. “Generating text with recurrent neural networks.” Proceedings of the 28th International Conference on Machine Learning (ICML-11). 2011.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. “Sequence to sequence learning with neural networks.” Advances in neural information processing systems. 2014.

L Tran, Du, et al. “Learning spatiotemporal features with 3d convolutional networks.” Proceedings of the IEEE international conference on computer vision. 2015.

T Tran, Kenneth, et al. “Rich image captioning in the wild.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016.

Wu, Qi, et al. “What value do explicit high level concepts have in vision to language problems?.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

Yang, Zhilin, et al. “Review networks for caption generation.” Advances in Neural Information Processing Systems. 2016.

You, Quanzeng, et al. “Image captioning with semantic attention.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

Young, Peter, et al. “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.

Farhadi, Ali, et al. “Every picture tells a story: Generating sentences from images.” European conference on computer vision. Springer, Berlin, Heidelberg, 2010.

Gan, Zhe, et al. “Semantic compositional networks for visual captioning.” arXiv preprint arXiv:1611.08002 (2016).

Girshick, Ross, et al. “Rich feature hierarchies for accurate object detection and semantic segmentation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

Hodosh, Micah, Peter Young, and Julia Hockenmaier. “Frame image description as a ranking task: Data, models and evaluation metrics.” Journal of Artificial Intelligence Research47 (2013): 853-899.

Jia, Xu, et al. “Guiding the long-short term memory model for image caption generation.” Proceedings of the IEEE International Conference on Computer Vision. 2015.

Krisha, Ranjay, et al. “Visual genome: Connecting language and vision using crowdsourced dense image annotations.” International Journal of Computer Vision 123.1 (2017): 33-75.

Kulkarni, Girish, et al. “Babytalk: Understanding and generating simple image descriptions.” IEEE Transactions on Pattern Analysis and Machine Intelligence 35.12 (2013): 2891-2903.

Li, Simon, et al. “Composing simple image descriptions using web-scale n-grams.” Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2011.

Kuznetsova, Polina, et al. “TREETALK: Composition and Compression of Trees for Image Descriptions.” TACL 2.10 (2014): 351-362.
[78] Mao, Junhua, et al. "Learning like a child: Fast novel visual concept learning from sentence descriptions of images." Proceedings of the IEEE International Conference on Computer Vision. 2015.

[79] Mathews, Alexander Patrick, Lexing Xie, and Xuming He. "SentiCap: Generating Image Descriptions with Sentiments." AAAI. 2016.

[80] Mitchell, Margaret, et al. "Midge: Generating image descriptions from computer vision detections." Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2012.

[81] Ordonez, Vicente, Girish Kulkarni, and Tamara L. Berg. "Im2text: Describing images using 1 million captioned photographs." Advances in Neural Information Processing Systems. 2011.

[82] Yang, Yezhou, et al. "Corpus-guided sentence generation of natural images." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.