Journal of Intelligent & Fuzzy Systems xx (20xx) x–xx
DOI:10.3233/JIFS-179351
IOS Press


# Oboyob: A sequential-semantic Bengali image captioning engine

Tonmoay Deb, Mohammad Zariff Ahsham Ali, Sanchita Bhowmik, Adnan Firoze, Syed Shahir Ahmed, Muhammad Abeer Tahmeed, N.S.M. Rezaur Rahman and Rashedur M. Rahman*
*Department of Electrical & Computer Engineering, North South University, Bangladesh*



**Abstract**. Understanding the context with generation of textual description from an input image is an active and challenging research topic in computer vision and natural language processing. However, in the case of Bengali language, the problem is still unexplored. In this paper, we address a standard approach for Bengali image caption generation though subsampling the machine translated dataset. Later, we use several pre-processing techniques with the state-of-the-art CNN-LSTM architecture-based models. The experiment is conducted on standard Flickr-8K dataset, along with several modifications applied to adapt with the Bengali language. The training caption subsampled dataset is computed for both Bengali and English languages for further experiments with 16 distinct models developed in the entire training process. The trained models for both languages are analyzed with respect to several caption evaluation metrics. Further, we establish a baseline performance in Bengali image captioning defining the limitation of current word embedding approaches compared to internal local embedding.

Keywords: Image captioning, CNN, LSTM, natural language processing, computer vision, Bengali image captioning, merge architecture, par-inject architecture, machine translated caption subsampling

## 1. Introduction

An expressive image description is paramount to summarize the contents of an image in a way which tells the story without delving into unimportant details. Text descriptions can aid visually impaired people to draw a mental picture of an image in question. However, there are many ways to express an image while not losing its core meaning. Different descriptions can offer different perspectives on how an image is perceived by its viewer. Taking these things into account, automatically obtaining the sentence level description of an image in different languages has become the challenge for the researchers in computer vision and natural language processing [16]. Though there are substantial research works of representing an image in English, the use of other languages is still an area of exploration. The inclusion of different languages may solve many real-life problems, for instance, early childhood education, image retrieval, and navigation for the blind. These forms of sentence representation of an image are known as image captioning which deals with mainly two challenges. The first challenge is to identify objects in an image in the domain of computer vision, and the second one is to create a correlation among the objects and sentence-level descriptions in the domain of natural language processing [17]. An image may contain various information but extracting the insightful visual information is possible only by emulating the concept of Biological Vision System (BVS). Computer Vision has different approaches involved to mimic the BVS [9, 18], however, one of the major obstacles is to form a machine learning model to merge these two domains for the automatic caption generation in Bengali language.

*Corresponding author. Rashedur M. Rahman, Department of Electrical & Computer Engineering, North South University, Bangladesh. E-mail: rashedur.rahman@northsouth.edu.

ISSN 1064-1246/19/$35.00 

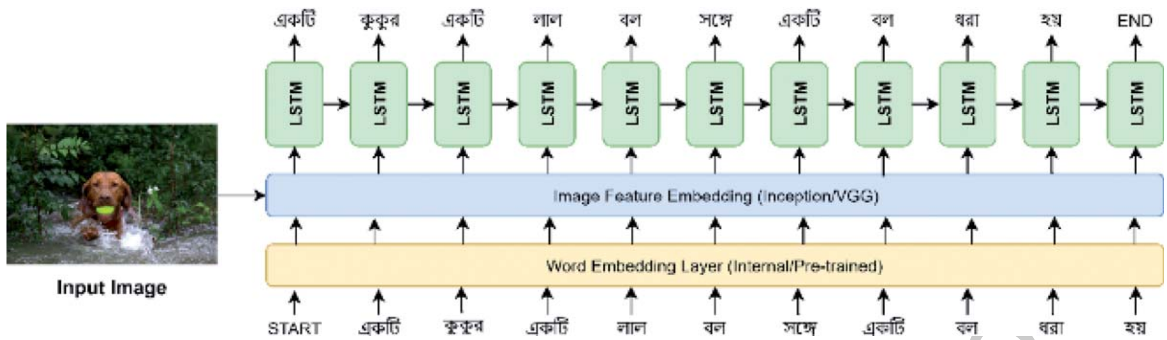


Fig. 1. A standard CNN-RNN involved image captioning process illustrated with Bengali language. Word embeddings and image features are computed through respective CNN models and word embedding techniques.

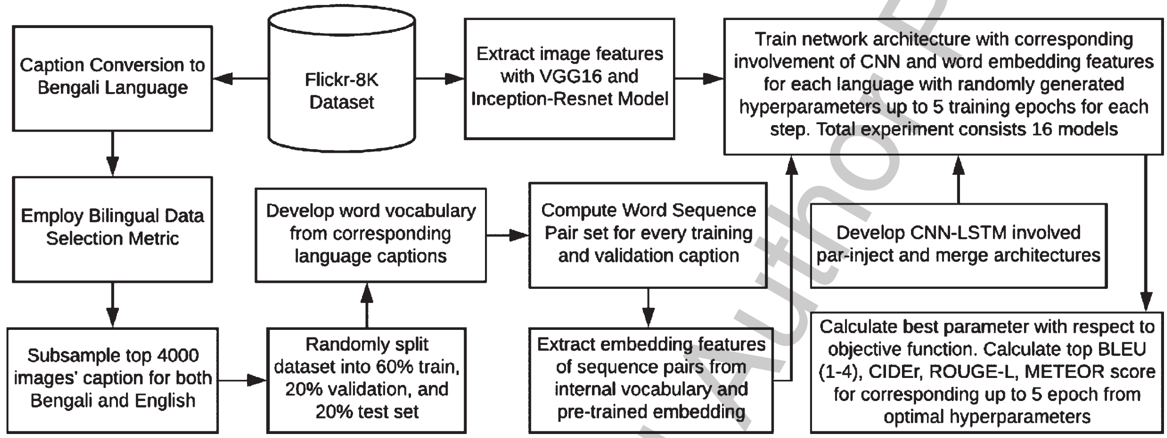


Fig. 2. Experimental workflow summary diagram.



In literature, there are two approaches for automatic captioning, namely bottom-up and top-down approach. According to the first approach, different words are accumulated to correlate with an image and the words form sentences [15]. This approach is easy to implement and able to describe the fine details of an image.

But some features remain ignored at the same time for inconsistent coherency. However, the state-of-art approach is the top-down approach [3, 11] which elects all the visual information through Recurrent Neural Network (RNN). The advantage is that the required parameters for the RNN are obtained from the training dataset [17]. The dramatic revolution in deep learning is incorporated by the Convolutional Neural Network (CNN) with RNN where CNN collects pictorial information from the image and RNN decodes into natural language through incorporating vocabulary-wise embedding from the language model. Figure 1 illustrates sample caption prediction sequence developed through involvement of CNN and LSTM for the context of Bengali language.

Our motivation explicitly aims to address image captioning in Bengali context by addressing a dataset and developing a machine learning model to enrich the amenities in Bengali language. However, considering the scarcity of data and resources, we approach translation of the English captions to Bengali, followed by a manual verification of corresponding subsampled captions by proposed sentence selection metric, and predicting captions for corresponding image as input. In this paper, we propose a baseline CNN-LSTM based Top-Down machine learning model for captioning in the Bengali and compare several captioning techniques for further evaluation of our model performances. Our novelty of the paper lies on the inclusion of a new and completely different language from English language in image captioning model. Figure 2 illustrates the complete experimental flow conducted in this research work. Our major contributions are—

1. Propose Bengali caption dataset through machine translation of Flickr-8K caption set.

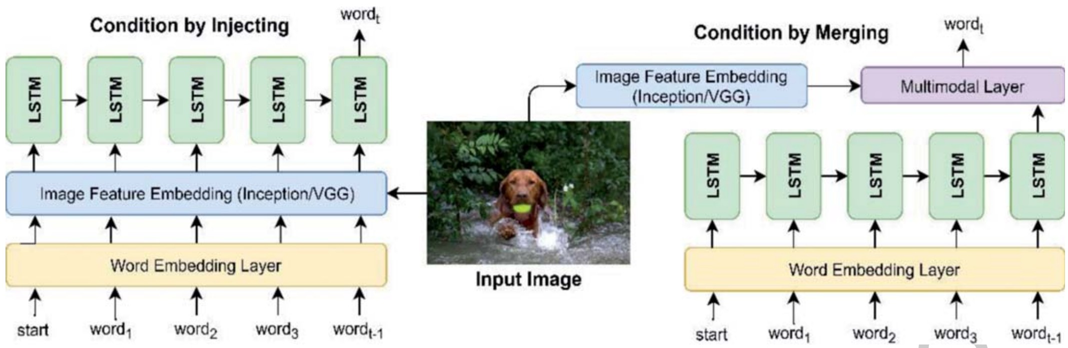


Fig. 3. Illustration of Par-Inject Architecture (left) and Merge Architecture (right).

2. Introduce a novel caption-correlation method to eliminate poorly captioned images.
3. Develop respective CNN-RNN architecture, train with down-sampled dataset and report a comparative performance study.
4. Implement pre-trained word embedding in our context, evaluate and compare experimental results with respect to local embedding.

## 2. Background

In this section, we explore several relevant and significant background works done in this area based on image feature extraction with CNN including novel experiments conducted with non-English languages. We will especially focus on the Bengali language context, mention the limitations and challenges.

### *2.1. Recurrent neural networks involvement*

In line with CNN features [6], RNN is analyzed in many forms in literature. In [11], the authors proposed novel bidirectional mapping between an image and the possible captions. The network model computes visual representation dynamically using RNN and maximum entropy language model, paves the possibility not only to generate captions but also reconstructs an image from captions. In general, RNN has drawback of capturing a long-range mapping. But sequential mapping needs to cover the long distance for image captioning. One of the possible solutions is inclusion of Long Short-Term Memory (LSTM) as a recurrent network [13, 19]. The proposed model handles variable length input-output, connect the visual convolutional model to LSTM, later fed to a convolutional layer to work with spatial-temporal correlation. LSTM model is further investigated with the encoder-decoder architecture. In [3], the author proposed an encoder CNN and a decoder RNN. The decoder RNN was single layer LSTM and a greedy decoding was used at the time of inference. A different approach was taken for an automatic image to captions generator by Ranzato et al. [13] with a sequence level greedy method. In this approach, certain number of words were evaluated for log-likelihood and remaining words were reinforced to optimize arbitrary captioning metrics. In our work, recurrent network involvement is two-fold; first, contributing as a sequence generator for captions, later, to encode the appropriate sequences of word embedding, instead of directly generating them through hand crafted approach.

### *2.2. Recent progress in non-English language*

All previously discussed research works were focused to generate English text whilst research on other languages is still in experimental stage. There have been works related to image caption generation in other languages [14], where the authors developed a Japanese version of the MS-COCO caption dataset [14, 30] as well as an accompanying generative model for text descriptions. Recently, one literature proposed Flickr8k-CN [27, 28], a bilingual extension to Chinese caption generation from image. Increasingly, several experiments conducted on German [36] and Arabic Language [29], where authors developed an English-German dataset to facilitate the captioning process. However, discussed approaches developed a bilingual dataset toward the task. Addressing the limitation, Lan et al. [28] proposed a cross-lingual image captioning including optimized caption fluency through rejection sampling over learning process. Recently an initiative [42] was taken regarding Bengali image captioning through introducing a manually annotated image caption dataset in Bangladeshi context. Though, several researches conducted in Bengali machine translation involving several standard rule-based techniques [12, 20], none are currently state-of-the-art, and out of context to ours. In this research work, we focus to achieve the novelty in automatic image captioning in the Bengali language for minimizing the language barrier with deep learning models. In experiments, state-of-the-art recurrent network model was employed with recent VGG [5] and Inception [21] models. We also report optimum hyperparameters for different models as well competitive performance of models on the subsampled bilingual Bengali dataset.

## 3. Fundamental algorithm overview: Continuous bag-of-words (CBOW) model

This section depicts studies regarding fundamental algorithms applied for the experiments. In this section, we will focus on continuous bag-of-words (CBOW) [26] model, for fixed, lower dimensional, robust word feature representation.

Continuous bag-of-words model [31] was first introduced by Mikolov et al. [26] as a context-based target word prediction weight based fully connected neural network. Briefly, this setup consists of one-hot encoded vector of the word in the input layer, at the same time one-hot encoded context word in the output layer. Fundamentally, between the layers, there exists a lesser-node-based hidden layer, technically define the number of fixed dimensions in which the word should be represented. The complete architecture acts like a bigram model as demonstrated in their work. Grave et al. [31] extended standard CBOW model with position weights, sub-word information. The model represented the words as bag-of-ngrams, rather than prior bag-of-words model [2] with the position-dependent weights. Further, the model was trained on large dataset from Wikipedia and Common Crawl, totaling 157 languages worldwide, later released in FastText [2, 31]. The models[1] consisted a fixed 300-dimensional feature representation per input word. In addition, 5-character *n*-gram, 5-10 negative sampling window size was adopted during model deployment. Our research employed their model for both English and Bengali captions.

[1] https://fasttext.cc/docs/en/crawl-vectors.html

## 4. Comprehensive model architecture overview

In recent advances of CNN-RNN model, several successful experiments employed sequential caption prediction through fusion approach between CNN and RNN. In recent experiments, image features combined with sentence features, resulted in a caption related to the given input image. Tanti et al. [22] generalized the fusion into two sections, inject, and merge architectures.

In inject architecture, image features were involved directly during RNN sequence generation process, where merge architecture compounded to image feature in later stage after word sequence generation. Their work classified inject architecture into three stages, init-inject, per-inject, and par-inject. Init-inject define insertion of visual features as initial hidden state of the recurrent network. In pre-inject, visual feature commit as the first word for sequence model generation, where, in every time step, image feature is concatenated with words in par-inject. According to experimental analysis [22], merge architecture holds visual information intact while learning linguistic features, where, par-inject architecture takes advantage of visual information input for every time step and highly retain visual information than inject-based architectures. Furthermore, merge architectures require less RNN memory size, though achieving competitive performance [24]. Considering this, we have adopted merge architecture according to the works [16, 19]. Additionally, we have used par-inject [11, 23] for having higher performance estimation over other inject models. Further, the selected architectures will be discussed in next sections.

## 5. Experimental procedure

The complete experimental pipeline is divided into several stages. From feature extraction to final model architecture, several techniques are employed, which we will illustrate in the corresponding sections.

### 5.1. Dataset processing

This section illustrates the explanation of dataset used for the experimental work, followed by, translation-based data conversion facilitating Bengali language with further processing, and data elimination approaches conducted for further training of the models.

#### 5.1.1. Dataset description and conversion

For the entire experiment, Flickr-8K dataset [1] is employed, which consists of total 8092 images, taken from Flickr[2]. Corresponding image ids are split into training, validation, and testing; where 6000 are used for training, 1000 for validation and 1000 for testing. Each image consists of 5 human-annotated ground truth captions associated, resulting in 40,460 total sentences. After applying the caption tokenization, followed by word frequency estimation, most frequent words consist of verbs, including “in”, “is”, whilst some of the most frequent nouns include “dog”, “man”. During experiment, token words for corresponding image features are embedded into a vector set and fed. To conduct the experiment in Bengali language, we develop Bengali caption-involved bilingual dataset “Flickr8k-BN” from the existing English captions. To adapt the translation process, Google Translate[3] is employed to convert English sentences to Bengali, resulting a translation set consisting of 40,460 captions. In caption sentences, most frequent Bengali words include “হয়”, “মধ্যে”, “উপর” as verbs, “কুকুর”, “মানুষ” as nouns. Figure 4 illustrates dataset word frequency histogram for both English and translated Bengali sentences, respectively. The machine translation weakness is observed from the figure as Bengali contains higher word frequency along with some English word involvement in translation. Considering above issues, we have applied a novel caption selection metric to reduce frequency rate and machine translation error at an acceptable scale.

#### 5.1.2. Bilingual caption selection metric

Upon translation, through a manual intervention, we observe, machine translation results in ambiguous words, including actual context understanding gap for the target language sentences. This is still an unsolved problem in the natural language processing domain. To overcome the limitation, initially, unique token words are carefully interpreted, which result in observation that, several close-to words are characterized as independent tokens due to some extra characters involved in respective words. To overcome, we have adopted a publicly available Bengali rule-based stemmer, which is not adequate for the task, resulting in irrelevant contextual words, e.g., for input “একটি কালো ছেলে বা লর মধ্যে বসা হয়”, corresponding output is “এক কালো ছাল বালির মধ্য বসা হয়”, which an out-of-the-context and non-grammatical sentence. As per being complex [39], stemming Bengali language leads to another research, where root words should be analyzed manually. The quality of rule-based stemmer [39] could be manually predicted due to having set of conditional statements. From several word pair, output result becomes relatively easy to infer. In our dataset, verification of 40,460 sentences according to the image and English sentence would be a tedious task. To overcome the limitation, we introduce a novel approach for determining top $k$ images containing best captions having correlation between captions per image, and scale of consistent cross-match among the captions. At first stage, each Bengali sentence is represented into fixed dimensional word embedding through FastText [31] pre-trained model.

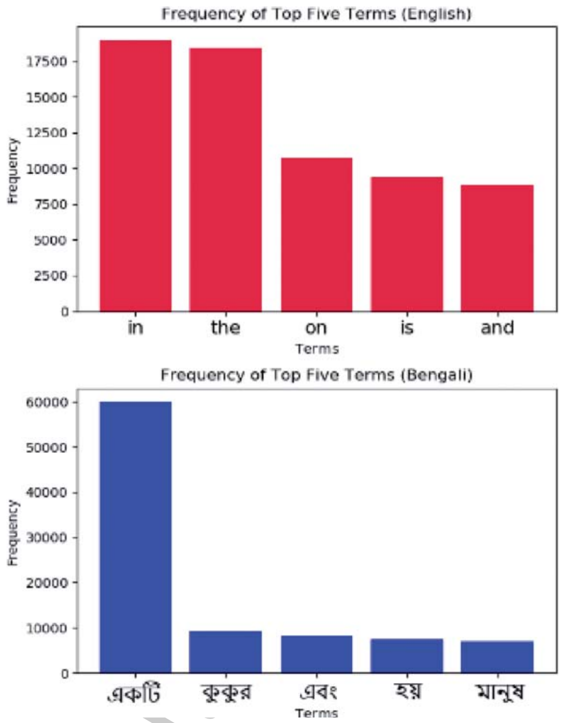


Fig. 4. Word frequency distribution English (top), Bengali (bottom).

Later, for each caption, the average of absolute cross-distance match is computed, followed by a vector summation. The same approach has been applied to estimate match with respect to English language captions. Equation 1 illustrates the equation for computing captioning score for an image. Here, $i$ and $j$ are corresponding indices of $n$ number of caption features, $X$. Both iterations run till $n$, where each would

[2] https://www.flickr.com

[3] https://translate.google.com.bd

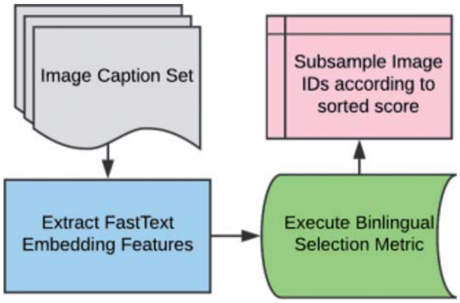


Fig. 5. Sample illustration of bilingual data selection metric from image caption set toward sorted images.

have a one to zero result, and the summation term goes till the last stage of $k$ number of image selection.

$$Score_k = \sum \sum_{i=1}^{n} \sum_{j=1}^{n} \left| X_i - X_j \right| \qquad (1)$$

Later, the computed *Score* for every image are sorted in ascending order, where lower score defines better semantic caption distance. In summary, this lower-score approach illustrates how strong captions are for each image (inter-connected in cross-relationship), a largely important term for our experimental process. A sample workflow illustration has been demonstrated in Fig. 5, where the image captions are put forward into sorted ranking task depending on FastText embedding features. Regarding scoring techniques, Table 1 illustrates a sample caption out of 5 from images having scores highest and lowest, which responds to image having good or bad caption quality for both languages separately. Upon careful observation, we can conclude, for both English, and Bengali, the bad sample caption fails to illustrate the scene properly, due to lack of proper verification. Besides, there is an addition of poor machine translation in Bengali language. Later, the subsampled dataset for both has been employed in later sections.

### 5.2. CNN feature extraction

Currently CNN [6] architecture, with variations used in object recognition tasks with several standard image datasets [35]. The involved trained weights tend to have highly discriminative sampled and optimal features, resulting in competitive accuracy computation [5, 21]. Rather training a distinct new model, we prefer adaptation of selective and high-performing pre-trained models. We have employed Inception-ResNet [21] and VGG-16 [5] models. Both architectures are trained on ImageNet dataset [35], emerged as high-performing object recognition models [7]. Undoubtedly, the prior network has higher depth, few hundred layers, compared to VGG-16. After removal of last classification layer, Inception-ResNet returns 1536 whilst VGG-16 results in 4096-dimensional feature representative vectors. However, in later steps, vectors are compressed to comply with word sequence vector, and hyper parameter optimization-oriented experiments.

### 5.3. Word sequence pair generation

Prior to process image in the caption generation scheme, sentence representation into sequence pair combination is another important and challenging task in our research. In general, an RNN model learns input prefix pair toward prediction of the best possible candidate through probabilistic SoftMax function. However, the training process is kept through fixed input-output RNN sequence pair model, led by the largest number of training sentence length. For example, if we have a sentence with word length 10 whilst max training sentence length can hold 30 words. To accommodate this, the lower sequence is padded with zeros up to highest length. Then, an input-output pair from group of words is computed for training process. However, according to some recent works [3, 4], for training with higher number of examples, single-line padding is not a good choice, rather

Table 1
First caption with scores; English (left), Bengali (right)

| | English | Score | Bengali | Score |
|---|---|---|---|---|
| Good Sample | Black dog in the water with tennis ball in his mouth | 0.14 | একটি বাদামী এবং সাদা কুকুর একটি লাল এবং হলুদ মেরু জাম্পিং হয় | 0.13 |
| Bad Sample | Mountain landscape | 0.45 | দুইজন লোক একটি মুখোমুখি মুখোমুখি | 0.50 |

sequence pair combination of corresponding caption sentences shows better input-output pair representation. As per the suggestion, we have developed a fractioned input-output word sequence pair for training purpose, e.g. a sentence with $n$ number of word would have $n + 1$ sequence pair considering an extra start and end token as identifier during training. Further, zero padding is performed on the sequences followed by pairing with image features depending on architecture mechanisms. Prior discussion is done on the data selection metric, and top 4000 images are selected as experimental set. The known vocabulary has been computed from training set as unique word tokens. Among them, Bengali language consist of 6410 unique tokens whilst English has 4667. However, English has maximum 34 words length training sentence and 21 for Bengali. Further, every training sentence are represented as arrays consisting vocabulary index for corresponding words, followed by zero padding according to the language's sentence length. Embedding of training pairs are performed through network whilst pre-embedded training pair consists of sets of vectors instead of word indices. Table 2 illustrates a sample input-output pair for Bengali sentence where, for image $X$, and sample word pair $Y$, resultant word is $Z$. Further, corresponding word indices are evolved as sequences, and later zero padded according to the maximum length. The table illustrates discussed par-inject model architecture [22]. However, merge model involves same word sequence pairs except image conditioning in each sequence [16, 19]. During generation process, RNN model would result in single word from input condition and follow recursively until the end token prediction.

### 5.4. *Pre-trained word embedding involvement*

The experiment consists two-fold word embedding. Basic embedding structure is derived from the internal vocabulary computed from training set,

Table 2
Sentence sequence model illustration

| X (image feature) | Y (input word) | Z (output word) |
|---|---|---|
| Feature | start | একটি |
| Feature | start, একটি | ছেলে |
| Feature | start, একটি, ছেলে | দাঁড়িয়ে |
| Feature | start, একটি, ছেলে, দাঁড়িয়ে | আছে |
| Feature | start, একটি, ছেলে, দাঁড়িয়ে, আছে | end |

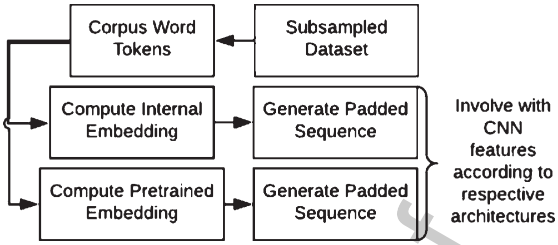


Fig. 6. Sample illustration of word embedding scenario and padding before being fed to network architecture.

represented into vectors prior to being given as input to recurrent network. The embedding models are trained as a part of the experiment. However, there exists a significant token gap in vocabulary, resulting in poor embedding representation through vectors resulting lower number of correlations learning throughout the process. To deal this, we introduce pre-compiled word embedding model to facilitate word representation process. As per discussion in Section 3, CBOW structure is represented according to fundamental vector computation process. For experiment, we have adopted FastText [2, 31] library's models, a robust and widely used word representation model trained on large vocabulary set across several languages and is available for both Bengali and English language. For initiating, the features are computed for each caption, representing a sentence as fixed-dimensional array. This approach is more acceptable where each input representation is robust, as per being trained on billions of tokens whilst being more efficient because no further embedding-related training is required before input to the recurrent network. Figure 6 illustrates embedding workflow that involves visual features for specific architecture structure and caption generation.

### 5.5. *Model architecture development*

This section introduces architecture development process according to prior discussion. It includes preceding illustration about the fundamental feature processing which includes visual and linguistic representation. Initially, we adopt two architectures. Later, a small modification has been performed to facilitate pre-compiled word embedding.

**Par-Inject Architecture** defines recurrent network involvement while conditioning image with word feature during input at every time step. In this stage, at first layer with dropout is introduced followed by an embedding layer word feature pro-

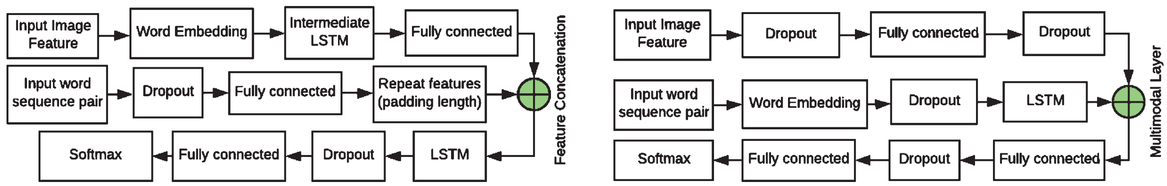


Fig. 7. Par-inject (left) and Merge (right) architecture illustration based experimental workflow.

cessing. An extra fully connected layer is added after image input layer to maintain same dimension of language and word features. Later, image features are repeatedly concatenated in multimodal layer with corresponding word embedding features computed for each time step. Then, processed representation is used as input to the LSTM layer regarding following sequence prediction in output scheme. Further, output feature vector goes through the fully connected layer to match training vocabulary dimension. Finally, best candidate word is selected through probabilistic SoftMax function. Figure 7 visualizes high-level architecture of par-inject.

**Merge Architecture** conditions image by involving corresponding visual feature with final output from recurrent network throughout word prediction process. The fundamental architecture principle demonstrates, image is never used in recurrent neural network whilst the word embedding are directly involved in further sequence generation process, and the output sequence is conditioned with image for next word prediction task, resulting in the visual feature remaining intact compared to par-inject [22], where, visual feature is directly incorporated in recurrent network's sequence prediction process. Embedding layer feature vector discussed in par-inject model are passed to an LSTM network resulting in final prediction. Further, the image feature is down sampled to match word prediction output, followed by a multimodal layer concatenating the output, and a dense layer with identical dimension to vocabulary. Figure 7 illustrates the developed merge model. Here, the green component stands for multimodal layer which concatenates CNN and LSTM features prior to prediction. For both architectures, embedding layer is pre-computed, requiring no further training as per FastText [2] embedding involvement.

### *5.6. Hyperparameter optimization*

**Word Embedding Dimension** is used to represent each word into fixed dimensional vector to facilitate entire learning process of network. In this experiment, an internal word representation technique is involved from fixed vocabulary set of both datasets. However, in case of pre-trained word embedding based training [31], we skip internal embedding, replacing it with a linear activation. Fixed dimension parameters e.g., 100, 200, and 300 are set for local vocabulary experiment for learning word embedding. Later, it is regularized with an extra hidden layer, activation function, and dropout [8].

**Size of Layers** usually impacts networks over feature representation [22] by influencing performance. Variable layer sizes could be used for visual and linguistic feature concatenation, optimization-based tasks e.g., deciding output feature dimension from LSTM. Understanding layer size is essential in addition to other parameters to prevent overfitting the entire network. In experiment, size of layers in feature reduction of CNN and LSTM is (128, 256, 512), and LSTM network hidden parameter and final representation (64, 128, 256), are kept in the three constrained ranges.

**Dropout** [8] essentially prevents a neural network from overfitting. After variable length layer size of fully connected, or LSTM layers, additional dropout layer is assigned. The parameter ranges from 0.3 to 0.5.

**LSTM Projection Dimension** is highly devised in merge architecture [16, 19]. There, two-fold network, followed by a later multimodal layer concatenated two features resulted the next word prediction. As per the architecture, word feature contributes self-conditioned word sequence prediction. Regarding the reason, projection dimension of each LSTM output state was taken care of to evaluate better representative network.

Quantum random method [41] is used to generate hyperparameters with 3% of samples for further experimental analysis according to the library implementation of Talos[4] used in our experiments. The performance for each candidate combination is recorded for 5 epochs. From the best performing

[4] https://github.com/autonomio/talos

model combinations according to objective function is adopted. In addition, our experiment has entirely focused on fixed architecture with variable parameters over the variable architecture that affect main standard, e.g. addition of more layers, or change in visual-linguistic feature concatenation type.

## 6. Result analysis

This section will discuss the experimental results found from our trained model described in preceding sections. We also discuss about optimal hyperparameters and caption quality evaluation with different metrics.

### *6.1. Optimal hyperparameters*

Regarding model development and experimental stage, there already evolved several parameters. The best-performing hyperparameters for Inception-Resnet and VGG-16 visual features are illustrated in Table 3. In general, the merge architecture is simple and lower number of parameters are involved for tuning whilst par-inject architecture requires higher training with larger model size. It is interesting to note that, inception model requires higher CNN-LSTM feature pair compared to VGG and merge architecture that require higher dimension due to multimodal layer ensembled representation. Regarding word vocabulary dimension, inject architecture takes advantage involving CNN with word embedding pair, requiring lower dimensional representation compared to merge architecture. Overall, inception requires lower vocabulary due to having high object recognition. Multimodal layer of the merge architecture remains identical for VGG and varies for Inception-Resnet, having good performance with ELU [34]. For Bengali language, the linguistic representation dimension tends to be higher compared to English, determining a complex language scheme towards more representation attenttion. A higher regularization is required in Bengali pre-trained model. Another interesting point to note that, local vocabulary embedding excluding some merge architectures prefer ELU as activation function, independent of language, defining linguistic models require some non-linearity other than straight linear activation function like ReLU. However, from prior analysis, two decisions can be taken: Inception-Resnet architecture influences LSTM model compared to VGG more efficiently, due to having more

Table 3
Obtained optimal hyperparameters for CNN model and discussed architectures

| CNN Models | Hyperparameter-wise Sections | Par-Inject Architecture Bengali | Par-Inject Architecture English | Par-Inject Architecture Bengali (PTE) | Par-Inject Architecture English (PTE) | Merge Architecture Bengali | Merge Architecture English | Merge Architecture Bengali (PTE) | Merge Architecture English (PTE) |
|---|---|---|---|---|---|---|---|---|---|
| Inception-Resnet Model [29] | Image-LSTM dense | 512 | 256 | 128 | 256 | 256 | 128 | 512 | 512 |
| | Image dropout | 0.3 | 0.3 | 0 | 0.3 | 0 | 0 | 0 | 0.3 |
| | Image activation | ReLU | ReLU | ELU | ReLU | ELU | ELU | ELU | ReLU |
| | Word Vocab. Size | 64 | 128 | – | – | 128 | 64 | – | – |
| | Word LSTM size | 64 | 256 | 128 | 64 | 512 | 512 | 128 | 128 |
| | Word LSTM activ. | ELU | ReLU | ReLU | ReLU | ReLU | ELU | ReLU | ELU |
| | Word LSTM dropout | 0 | 0 | 0.3 | 0.5 | 0.3 | 0.3 | 0.3 | 0 |
| | Inject LSTM size | 256 | 64 | 128 | 128 | – | – | – | – |
| | Inject LSTM dropout | 0 | 0 | 0.3 | 0 | – | – | – | – |
| | Inject LSTM activ. | ELU | ELU | ELU | ReLU | – | – | – | – |
| | Multimodal activ. | – | – | – | – | 256 | 128 | 128 | 128 |
| | Multimodal size | – | – | – | – | ELU | ELU | ELU | ELU |
| VGG-16 Model[6] | Image-LSTM dense | 128 | 128 | 256 | 128 | 256 | 128 | 512 | 256 |
| | Image dropout | 0 | 0.3 | 0 | 0.3 | 0.3 | 0 | 0.5 | 0.3 |
| | Image activation | ELU | ReLU | ReLU | ReLU | ELU | ELU | ReLU | ELU |
| | Word Vocab. Size | 256 | 64 | – | – | 128 | 64 | – | – |
| | Word LSTM size | 128 | 128 | 128 | 128 | 256 | 256 | 128 | 512 |
| | Word LSTM activ. | ReLU | ReLU | ELU | ELU | ReLU | ELU | ELU | ReLU |
| | Word LSTM dropout | 0 | 0.5 | 0.3 | – | 0.5 | 0.3 | 0.5 | 0.3 |
| | Inject LSTM size | 128 | 256 | 128 | 64 | – | – | – | – |
| | Inject LSTM dropout | 0 | 0 | 0 | 0.3 | – | – | – | – |
| | Inject LSTM activ. | ELU | ELU | ELU | ELU | – | – | – | – |
| | Multimodal size | – | – | – | – | 256 | 128 | 256 | 128 |
| | Multimodal activ. | – | – | – | – | ELU | ELU | ELU | ELU |

accurate, lower-dimensional feature representation. Besides, Bengali language model learning process is more complex than English. In addition, a higher regularization is required in Bengali pretrained embedding for obtaining higher dimensional representation including other architectures than English.

The implication is, par-inject requires more memory compared to merge model with more hyperparameter variation and visual-linguistic feature in each time step whilst merge requires least memory through learning visual -linguistic features separately.

### 6.2. Evaluation of caption quality

Optimal hyperparameters found for each architecture is employed in experimental analysis with 5 epochs, from where, highest scored epoch with respect to objective scoring mechanism is selected. In following sections, several highly used scoring approaches that includes Microsoft COCO [30] Evaluation Toolkit is presented. A sample illustration regarding 1 vs. 4 captions scoring from a random test image has been demonstrated in Table 4.

#### 6.2.1. Bilingual evaluation understudy (BLEU)

BLEU [10] involves variable n-gram weighted average to compute difference between actual (reference) and predicted (hypothesis) sentence, resulting in promising scoring compared to human judgment. From illustration given in Table 4, we observe that, BLEU scores are biased toward small sentences for higher scores, including some inefficient estimation for higher values of precision. For Bengali language, BLEU scoring tightly bounds length and vocabularies.

#### 6.2.2. Metric for evaluation of translation with explicit ordering (METEOR)

In the evaluation process, METEOR [32] calculation is performed. This metric is based on unigram matching between reference and hypothesis sentences using the harmonic mean of unigram precision and recall. To evaluate the score over the dataset, we take the aggregation of unigram precision, unigram recall and penalty of harmonic mean, and later combine according as authors' suggestion reported in [32]. In Table 4, METEOR performance is more accurate in cross-match scenario, however, as token-based approach, it is diverse and correct context sentences is underestimated in some cases.

#### 6.2.3. Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

We additionally involve state-of-the-art $ROUGE_L$ [33], a measure based on the Longest Common Subsequence (LCS). The score is computed with an F-measure according to length of the LCS between reference caption and hypothesis caption. Considering this illustration, due to counting subsequence, performance on English captions in Table 4 is higher than Bengali, which is more complex.

Table 4
Sample cross-scoring evaluation result demonstrated for English and Bengali captions from a random test image

| Caption [English] | BLEU | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | METEOR | $ROUGE_l$ | CIDER |
| Young asian woman wearing long shorts and gray collared tshirt is sitting on wooden bench | 0.47 | 0.36 | 0.31 | 0.27 | 0.34 | 0.55 | 0.00 |
| Girl with black purse sitting on wooden bench | 0.87 | 0.79 | 0.68 | 0.50 | 0.41 | 0.69 | 0.00 |
| Woman sits alone on park bench in the sun | 0.44 | 0.00 | 0.00 | 0.00 | 0.20 | 0.44 | 0.00 |
| Woman with handbag is sitting on wooden bench | 0.87 | 0.79 | 0.67 | 0.59 | 0.37 | 0.67 | 0.00 |
| Young woman with black purse sits on wooden bench | 1.00 | 0.79 | 0.56 | 0.00 | 0.45 | 0.71 | 0.00 |
| Caption [Bengali] | | | | | | | |
| একটি এশিয়ান মহিলার ক্যামেরা দিকে পৌঁছেছে যে একটি এশিয়ান শিশুর অধিষ্ঠিত | 0.45 | 0.21 | 0.00 | 0.00 | 0.39 | 0.41 | 0.00 |
| একটি প্রাচ্য মেয়ে তার অস্ত্র একটি শিশুর অধিষ্ঠিত হয় | 0.57 | 0.26 | 0.00 | 0.00 | 0.27 | 0.39 | 0.00 |
| একটি হাসিখুশি এশিয়ান মহিলা তার শিশুকে ধরে রেখেছ | 0.66 | 0.41 | 0.00 | 0.00 | 0.32 | 0.31 | 0.00 |
| একটি হাসিখুশি নারী সাদা পোশাক পরা একটি শিশু যারা তার হাত পৌঁছেছেন | 0.42 | 0.27 | 0.00 | 0.00 | 0.36 | 0.31 | 0.00 |
| মহিলাটি একটি শিশু ধরে রেখেছে এবং একটি ছবির জন্য দাঁড়িয়ে আছে | 0.45 | 0.30 | 0.00 | 0.00 | 0.34 | 0.32 | 0.00 |

Table 5
Evaluation metrices result with several experimental architectures for respective CNN models

| CNN Models | Experimental Model Architecture | BLEU 1 | BLEU 2 | BLEU 3 | BLEU 4 | METEOR | $ROUGE_l$ | CIDER |
|---|---|---|---|---|---|---|---|---|
| Inception | Inject Bengali | 0.55 | 0.38 | 0.27 | 0.15 | 0.32 | 0.51 | 0.22 |
| | Inject English | 0.54 | 0.30 | 0.22 | 0.12 | 0.21 | 0.44 | 0.30 |
| | Inject PTE Bengali | 0.50 | 0.34 | 0.25 | 0.13 | 0.31 | 0.48 | 0.19 |
| | Inject PTE English | 0.53 | 0.32 | 0.21 | 0.13 | 0.22 | 0.46 | 0.40 |
| | Merge Bengali | **0.62** | **0.45** | **0.33** | **0.22** | **0.34** | **0.54** | 0.35 |
| | Merge English | 0.59 | 0.37 | 0.28 | 0.16 | 0.23 | 0.49 | **0.46** |
| | Merge PTE Bengali | 0.61 | 0.44 | 0.32 | 0.18 | 0.33 | 0.53 | **0.37** |
| | Merge PTE English | **0.60** | **0.38** | **0.29** | **0.17** | **0.24** | **0.50** | 0.45 |
| VGG16 | Inject Bengali | 0.55 | 0.37 | 0.26 | 0.13 | 0.33 | 0.49 | 0.20 |
| | Inject English | 0.50 | 0.30 | 0.22 | 0.12 | 0.21 | 0.45 | 0.33 |
| | Inject PTE Bengali | 0.37 | 0.25 | 0.19 | 0.09 | 0.27 | 0.42 | 0.13 |
| | Inject PTE English | 0.56 | 0.33 | 0.24 | 0.13 | 0.21 | 0.46 | 0.31 |
| | Merge Bengali | 0.58 | 0.39 | 0.28 | 0.15 | 0.33 | 0.51 | 0.26 |
| | Merge English | 0.55 | 0.34 | 0.26 | 0.14 | 0.22 | 0.47 | 0.37 |
| | Merge PTE Bengali | 0.56 | 0.39 | 0.27 | 0.13 | 0.32 | 0.50 | 0.23 |
| | Merge PTE English | 0.56 | 0.34 | 0.26 | 0.15 | 0.22 | 0.46 | 0.38 |

#### 6.2.4. Consensus-based image description evaluation (CIDEr)

CIDEr [25] involves Term Frequency Inverse Document Frequency (TF-IDF) weighing for each *n*-gram. The evaluation metric we have used CIDEr-D [38], a modification to CIDEr to prevent scoring in case of poorly judged caption by humans is given a high score by an evaluation metric. Though being a corpus-based metric, for cross-relation illustration in Table 4, CIDEr results in zero score, since it expects a more robust and larger dataset.

### 6.3. Discussion and decision

The metrics in Table 5 illustrate the performance of several architectures involved in previously developed dataset experiments. Considering the optimal hyperparameters with CIDEr as the objective function, corresponding evaluation metrics are computed. From CNN feature extraction scheme, this is clearly observed that Inception-Resnet influence higher performance than VGG16 in all architectures, meaning it is a highly discriminative state-of-the-art feature extraction model. Interesting observation regarding the architecture is, for both Bengali and English language, merge architecture outperform inject regarding all evaluation metrics. However, another observation involves internal or external vocabulary enabled word embedding, where we find mixed observation for the languages. Considering English language, the external pre-trained word embedding representation dominates internal vocabulary due to having diverse vocabulary for robust representation whilst in the metrics, internal vocabulary shows higher result for Bengali language. For Bengali, internal vocabulary conforms to more token patterns compared to pre-trained embedding model, resulting in more robust word representation in Bengali context. Figure 8 demonstrates generated captions from random test images with good performing merge model. This result significantly derives that, current Bengali pre-trained embedding model still requires improvement with diverse set of language tokens. If we consider the intra-language scoring comparison, except CIDEr, metrics for Bengali outperform English language; which indicates higher score estimation for successful selection metric for this language. However, though being a corpus-based metric, CIDEr tends to score English higher comparatively for having lower corpus sentence diversity in Bengali language.


Fig. 8. Caption with merge model for Bengali language test set.

## 7. Conclusion

This research connects Bengali language in image captioning research by introducing a standard experimental analysis and provides a comparative study of recent advancements and techniques currently used in this area. Firstly, we address limitation of resources in Bengali language which has high linguistic complexity and develop a machine translation dataset. We introduce then a novel bilingual sentence selection metric aiming to subsample poorly translated sentence from experiment data set. We further show that unlike English, due to obtaining lower vocabulary-based corpus, Bengali language does not prefer pre-trained word embedding. This paves an open door for further research on modeling Bengali natural language feature extraction with robust representation unveiling improved captions. We establish a baseline experiment scheme for languages other than English toward designing a universal, language-independent image captioning system.

## References


[1] M. Hodosh, P. Young and J. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics, *Journal of Artificial Intelligence Research* **47** (2013), 853–899.

[2] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, Bag of tricks for efficient text classification, arXiv preprint arXiv:1607.01759, 2016.

[3] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, Show and tell: A neural image caption generator, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[4] M. Tanti, A. Gatt and K.P. Camilleri, What is the role of recurrent neural networks (rnns) in an image caption generator? arXiv preprint arXiv:1708.02043, 2017.

[5] K. Simonyan and A. Zisserman, Very deep convolutional networks for largescale image recognition, arXiv preprint arXiv:1409.1556, 2014.

[6] A. Krizhevsky, I. Sutskever and G. E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, *International Journal of Computer Vision* **115**(3) (2015), 211–252.

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *The Journal of Machine Learning Research* **15**(1) (2014) 1929–1958.

[9] T. Deb, A. Arman and A. Firoze, Machine Cognition of Violence in Videos Using Novel Outlier-Resistant VLAD, in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 989–994.

[10] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, Bleu: A method for automatic evaluation of machine translation, in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics*, 2002, pp. 311–318.

[11] X. Chen and C. Lawrence Zitnick, Mind's eye: A recurrent visual representation for image caption generation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2422–2431.

[12] S. Bal, S. Mahanta, L. Mandal and R. Parekh, Bilingual machine translation: English to bengali, in *Proceedings of International Ethical Hacking Conference 2018*, Springer, 2019, pp. 247–259.

[13] M. Ranzato, S. Chopra, M. Auli and W. Zaremba, Sequence level training with recurrent neural networks, arXiv preprint arXiv:1511.06732, 2015.

[14] T. Miyazaki and N. Shimizu, Cross-lingual image caption generation, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12016, pp. 1780–1790.

[15] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg and T.L. Berg, Baby talk: Understanding and generating image descriptions, in *Proceedings of the 24th CVPR, Citeseer*, 2011.

[16] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang and A. Yuille, Deep captioning with multimodal recurrent neural networks (m-rnn), arXiv preprint arXiv:1412.6632, 2014.

[17] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel and Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

[18] P. Bhowmik, M.J.H. Pantho, M. Asadinia and C. Bobda, Design of a reconfigurable 3d pixel-parallel neuromorphic architecture for smart image sensor, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 673–681.

[19] J. Mao, W. Xu, Y. Yang, J. Wang and A.L. Yuille, Explain images with multimodal recurrent neural networks, arXiv preprint arXiv:1410.1090, 2014.

[20] S. Chakraborty, A. Sinha and S. Nath, A bengali-sylheti rule-based dialect translation system: Proposal and preliminary system, in *Proceedings of the International Conference on Computing and Communication Systems*, Springer 2018, pp. 451–460.

[21] C. Szegedy, S. Ioffe, V. Vanhoucke and A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in *AAAI*, 42017, p. 12.

[22] M. Tanti, A. Gatt and K.P. Camilleri, Where to put the image in an image caption generator, *Natural Language Engineering* **24**(3) (2018), 467–489.

[23] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig and M. Mitchell, Language models for image captioning: The quirks and what works, CoRR abs/1505.01809, 2015.

[24] Q. You, H. Jin, Z. Wang, C. Fang and J. Luo, Image captioning with semantic attention, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] R. Vedantam, C.L. Zitnick and D. Parikh, Cider: Consensus-based image description evaluation, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[26] T. Mikolov, K. Chen, G. Corrado and J. Dean, Efficient estimation of word representations in vector space, CoRR abs/1301.3781, 2013.

[27] X. Li, W. Lan, J. Dong and H. Liu, Adding chinese captions to images, in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ICMR '16*, ACM, New York, NY, USA, 2016, pp. 271–275.

[28] W. Lan, X. Li and J. Dong, Fluency–guided cross-lingual image captioning, in *Proceedings of the 25th ACM Inter-*

*national Conference on Multimedia, MM '17*, ACM, New York, NY, USA, 2017, pp. 1549–1557.

[29] V. Jindal, Generating image captions in arabic using root-word based recurrent neural networks and deep neural networks, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 2018.

[30] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár and C.L. Zitnick, Microsoft COCO captions: Data collection and evaluation server, CoRR abs/1504.00325, 2015.

[31] E. Grave, P. Bojanowski, P. Gupta, A. Joulin and T. Mikolov, Learning word vectors for 157 languages, CoRR abs/1802.06893, 2018.

[32] M. Denkowski and A. Lavie, Meteor universal: Language specific translation evaluation for any target language, *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014.

[33] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, Text Summarization Branches Out, 2004.

[34] D. Clevert, T. Unterthiner and S. Hochreiter, Fast and accurate deep network learning by exponential linear units (elus), CoRR abs/1511.07289, 2015.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in *Computer Vision and Pattern Recognition, 2009 CVPR 2009 IEEE Conference on, IEEE*, 2009, pp. 248–255.

[36] D. Elliott, S. Frank, K. Sima'an and L. Specia, Multi30k: Multilingual english-german image descriptions, CoRR abs/1605.00459, 2016.

[37] C. Callison-Burch, M. Osborne and P. Koehn, Re-evaluation the role of bleu in machine translation research, in *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[38] D. Elliott and F. Keller, Image description using visual dependency representations, in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1292–1302.

[39] S. Dasgupta and V. Ng, Unsupervised morphological parsing of bengali, *Language Resources and Evaluation* **40**(3-4) (2006), 311–330.

[40] M.R. Mahmud, M. Afrin, M.A. Razzaque, E. Miller and J. Iwashige, A rule based bengali stemmer, in *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on, IEEE*, 2014, pp. 2750–2756.

[41] M. Herrero-Collantes and J.C. Garcia-Escartin, Quantum random number generators, *Reviews of Modern Physics* **89**(1) (2017), 015004.

[42] M. Rahman, N. Mohammed, N. Mansoor and S. Momen, Chittron: An automatic bangla image captioning system, arXiv preprint arXiv:1809.00339, 2018.