Image Caption Generator for Sinhala Using Deep Learning

Damsara Ranasinghe, Randil Pushpananda, Ruvan Weerasinghe

***Abstract*—In this study, for the image caption generation in the Sinhala language, we have implemented a Recurrent Neural Network based model consisting of an InceptionV3 model as an image feature extraction model and a Long Short Term Memory network for the language model by referring to the literature. The different variations of Sinhala versions of the Flickr8K and MS COCO datasets have been constructed and used to train experimental models. Evaluation of the generated captions has been done using both automated and manual approaches. The model trained on the MS COCO dataset with Google translated Sinhala captions has achieved the highest BLEU score of 0.592 and the highest METEOR score of 0.281. After doing the manual caption analysis, it was observed that there could be generated captions which could provide a good idea to the reader while having lower BLEU and METEOR scores.**

***Keywords*—Sinhala image captioning, Caption generation, NLP, Flickr8K, Flickr30K, MS COCO, BLEU, METEOR**

I. Introduction

Sinhalese are the largest ethnic group in Sri Lanka. The native language of the Sinhalese people is Sinhala, and it is also one of the official languages in Sri Lanka. According to the Ethnologue[1], in Sri Lanka, there are 17 million Sinhala language users, with 15 million users who use Sinhala as their first language and about 2 million users who use Sinhala as their second language as of 2019. While having a rich history, the Sinhala language lacks modern Natural Language Processing tools and studies [1].

Humans can get an idea of an image at first glance and interpret the idea in their language . Making a computer have this ability is not an easy task. The image caption generation process needs to identify the image features and express the semantic meaning in natural languages . This is an emerging research area in Natural Language Processing domain . Automated image caption generation models have been proposed and implemented for some languages like English [2, 3, 4, 5, 6, 7, 8], Chinese [9], Japanese [10, 11], Arabic [12], Bengali [13], Hindi [14, 15] etc.

In media and publication domains , image captioning has been used to generate headlines, subtitles, etc. In the medical field, medical image understanding is used to identify the condition of patients by mapping the physiological features and images.

Correspondence: Damsara Ranasinghe (E-mail: damsarar@gmail.com) Received: 08.03.2023 Revised:22.03.2023 Accepted: 19.06.2023

Damsara Ranasinghe, Randil Pushpananda and Dr. A.R. Weerasinghe are from University of Colombo School of Computing, Sri Lanka. (damsarar@gmail.com, rpn@ucsc.cmb.ac.lk, arw@ucsc.cmb.ac.lk)

DOI: https://doi.org/10.4038/icter.v16i2.7266

While a low resource language, Sinhala lacks modern studies and applications in the Natural Language Processing (NLP) area. According to the survey done by de Silva [1], there are several studies have been conducted on corpora, datasets, dictionaries, wordnets, morphological analyzers, Part of Speech (POS) taggers, parsers, named entity recognition tools, semantic tools, phonological tools, Optical Character Recognition (OCR) applications and translators for Sinhala Language but there are no any studies on image caption generation. In this research, Sinhala versions of Flickr8K and MS COCO datasets have been produced and RNN based image captioning model is proposed for the Sinhala language.

The following sections have been organised as follows. Section II discusses the previous studies for this task. Section III presents the experimental setup of the proposed model. Section IV discusses the results obtained from the trained models. Finally, Section V discusses the conclusion and the future work of this study.

II. Literature Review

With the advancement of deep learning, researchers have proposed various deep learning-based techniques for the image caption generation task. These methods are mainly based on encoder-decoder, compositional and attention-based architectures. In Kiros et al. [16], authors proposed two multi-modal log bilinear models named 'Modality-Biased Log-Bilinear Model (MLBL-B)' and 'Factored 3-way Log-Bilinear Model (MLBL-F)' for the sentence generation task. For both IAPR and attributes datasets, the proposed models have performed better than the LBL and n-gram models on the BLEU score. Mao et al. [4], the authors proposed the multimodal Recurrent Neural Network (m-RNN). To prove the validity, they calculated the BLEU scores and perplexity on three benchmark datasets: IAPR TC-12, Flickr8K [17], and Flickr 30K [18]. The proposed m-RNN model could outperform the state-of-the-art techniques at that time. Chen and Zitnickx [19] introduced a model developed using Recurrent Neural Networks for the caption generation task and retrieval task. Their model is based on the model proposed by Mikolov et al.

[1]https://www.ethnologue.com/language/sin

[2]https://artificialintelligence.oodles.io/blogs/ai-powered-image-caption-generator/

[3]https://www.nltk.org/book/ch05.html

This is an open-access document distributed under the terms of Creative Commons Attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

[20], which introduced the external feature layer for language models. Chen and Zitnick [19] integrated the visual features extracted using the CNN into this external feature layer. On PASCAL 1K dataset, their model outperformed the BLEU and METEOR scores of previous Midge [21] and BabyTalk [3] models. Vinyals et al. [5] proposed the 'Neural Image Caption Generator (NIC)', also known as 'Google NIC'. This model is one of the pioneer models which used 'Long Short Term Memory (LSTM)' [22] for the caption generation process and showed remarkable improvement in results. As a part of a study done by Karpathy and Fei-Fei [6], they have introduced a multi-modal Recurrent Neural Network (RNN) based architecture. This model has outperformed the proposed models in Chen and Zitnick [19] and Mao et al. [4], but this has performed worse when compared to the Google NIC [5]. Xu et al. [23] proposed two variants of attention-based models named 'soft attention' based, which uses a determin- istic attention mechanism and 'hard attention' based, which uses a stochastic attention mechanism for the image caption generation task. Their models have outperformed the Google NIC model [5] and Log Bi-linear models [16]. Tanti et al. [7] proposed two variations of RNN models called 'inject' and 'merge' models. The merge model has performed better than the inject model in almost all the experiment cases. This model has outperformed the model proposed in Karpathy and Fei- Fei [6]. However, this performs worse when compared to the model presented in Xu et al. [23]. Huang et al. [24] proposed the 'Attention on Attention (AoA)' module to extend the conventional attention mechanism to identify the relevance of the query and the attention result. This model showed a slight improvement in results compared to the SGAE [25]. Li et al. [8] introduced a transformer-based [26] sequence modelling technique developed only using attention and feedforward layers. Results showed a slight improvement over the AoA [24]. However, ROUGE and CIDER show otherwise. He et al. [27] introduced the 'Image Transformer', a modified encoding transformer [26] with three stacks of spatial graph transformer layer and an explicit decoding transformer with the LSTM network. This model has outperformed the AoA [24] and ETA [8] models, according to the results.

There are several studies conducted on image caption generation for languages other than English. Miyazaki and Shimizu [10] proposed a cross-lingual image caption generation model for Japanese language. They have developed their captain generation model based on the model proposed in Vinyals et al. [5]. Yoshikawa et al. [11] constructed a large Japanese dataset called STAIR Captions. They have trained the image caption generation model proposed by Karpathy and Fei-Fei [6] on the STAIR dataset. This shows an improvement compared to the previously mentioned model [10] results. Peng and Li [9] proposed two models. One takes Chinese words as the input, and the other is to take characters as the input. These models are based on the model proposed in Karpathy and Fei-Fei [6] on their Chinese dataset. Elliott et al. [28] presented the Multi30K dataset, the German version of the Flickr8K dataset. Al-Muzaini et al. [12] presented an image caption generation model for the Arabic language. Their model is based on the 'merge' model presented by Tanti et al. [7]. Deb et al. [13] introduced a sequential semantic image caption engine for the Bengali language. Researchers have conducted several experiments on their dataset by implementing models by adding different extensions to the 'merge' and 'inject' models proposed by Tanti et al. [7]. 'Merge' model architecture has performed well in experiments than others. Mishra et al. [14] proposed a transformer network [26] based image caption generation model for the Hindi Language. Singh et al. [15] presented an encoder-decoder based model for image caption generation in the Hindi language. They have used the Hindi Visual Genome dataset [29] as their dataset.

III. Experimental Setup

A. Constructing the Dataset

While a resource-poor language, there is no image captioning dataset for the Sinhala language. Because of this reason, it is decided to construct an image captioning dataset for Sinhala using the Flickr8K dataset and MS COCO dataset, which are benchmark datasets for this task.

1) Flickr8K Dataset: Flickr8K dataset [17] consists of 8,000 images and five captions for each image resulting in 40,000 total captions. For the dataset construction task, three approaches have been designed. The first approach is to translate all 40,000 captions using the Google Translation API. The resulting dataset of this process consists of 40,000 Sinhala captions for 8,000 images.

After translating the English captions to Sinhala using the 'Google Translation API', it was decided to conduct an experiment to find the effect of the correctness of translated captions on the proposed model. A random sample of 2,000 images has been chosen from the original dataset for this experiment. Human contributors have been used to collect new Sinhala captions for images in the sampled dataset. We asked human contributors to add new captions for the given image as they wished. A small web application has been developed for this task. Some helpful keywords are given to help users build a new caption. These keywords are the set of nouns extracted from the captions in the original dataset using the NLTK Part of Speech (POS) tagger[3].

2) MS COCO Dataset: As the second dataset for the study, 'MS COCO (Common Objects in Context)' [30] large-scale dataset has been used. This dataset is a vital benchmark dataset that has been used in recent studies on image caption generation tasks. This dataset is used not only in image captioning tasks but also in object detection and segmentation tasks. The dataset consists of 1.5 million object instances, 80 object categories, and 91 stuff categories. All the images have five captions, and there are different versions of the MS COCO dataset with a different number of images. The version used in this study has 80,000 images and 400,000 captions. For the MS COCO dataset, only one approach is used to construct the Sinhala version. All the 400,000 captions were translated to Sinhala using Google Translation API. An experiment was performed using this full translated dataset.

B. High-level System Architecture

In this study, InceptionV3 CNN [31] architecture has been used as the image feature extraction model. This architecture has been used as the CNN architecture for the image feature extractor since [12], and [13] also have used InceptionV3 in their models as the visual feature extractor and shown promising results. 'Long Short Term Memory (LSTM)' architecture has been used as the language model in this study. LSTM is an improved version of RNN proposed by the [22]. It addresses the back-propagation error using 'multiplicative gate units' that will learn to open and close access to the constant error flow. As the base model for this task, the 'merge' model proposed by [7] is

used. In the 'merge' model (Figure 1), conditioning is done by merging the image features from the CNN and the output from the LSTM model. Then the output will be inputted into the fully connected layer and then to the Softmax layer to work the probability distributions. Then the caption generation algorithm will use the probability distribution to generate the sequence for the caption.

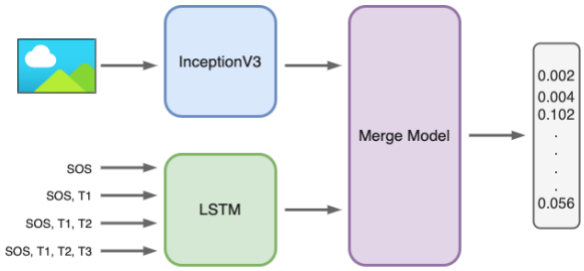

Fig. 1: Merged training model architecture

C. Evaluation

For the evaluation of the generated Sinhala captions, mainly two approaches have been proposed. One strategy is to use automated evaluation scoring metrics. For this, BLEU [32] and METEOR [33] scores have been chosen from the literature. BLEU (Bilingual Evaluation Understudy) score considers the n-grams and assigns weights to the n-grams equally. And measures the similarity between n-grams in reference and generated sentences. The n-gram precision in the BLEU score is calculated by dividing the n-gram matches by the total number of n-grams in the reference sentence. This score value ranges between 0 and 1 and the caption is better when the BLEU score is near 1. METEOR (Metric for Evaluation of Translation with Explicit ORdering) score also considers the n-grams with equal weights assigned and it is an extension of the BLEU score. METEOR score incorporates both precision and recall. The matching process of the METEOR is computationally expensive since it considers both stem and synonym levels when matching while the BLEU score considers only the exact form of the words. The BLEU and METEOR scores have been calculated for each individual image in the testing dataset, andthe average scores are derived from the average individual scores. For the manual caption analysis, three measures have been introduced. Those measures are as follows.

- Context identification - this considers the correctness of identification of the background and the objects in an image
- Grammatical correctness - refers to the correctness of the sentence structure of the generated Sinhala caption
- Overall correctness of the idea - this considers how well the generated caption gives the idea of the image to the reader

IV. Results and Discussion

Several experiments have been performed in this study when developing a model for the image caption generator for the Sinhala language. Two main approaches have been used to evaluate the generated Sinhala captions in the experiments. One is automated caption evaluation using scoring metrics, and the other is doing a manual qualitative evaluation on generated captions.

A. Automated Caption Evaluation

1) Flickr8K Dataset: Using the Flickr8K dataset, several models have been implemented as experiments. The following sections discuss the implementation of these experimental models and their automated evaluation results.

Google Translated Captions with Full Dataset Images: The initial experiment used the Google translated Sinhala captions with the full Flickr8K dataset. It consists of 8,000 images, and each image consists of five Google translated Sinhala captions resulting in 40,000 captions. The 8,000 images have been split into training and testing splits according to 70% and 30%, respectively. This splitting resulted in 5,600 training images and 2,400 testing images. The model achieved average BLEU score of 0.503. At the same time, it has given a METEOR score of 0.187.

Google Translated Captions with Sampled Dataset Images: Then another experiment was carried out on the sampled 2,000 images Flickr8K dataset with the 2,000 Google translated Sinhala captions. The 2,000 images dataset has been split into training and testing images as 1500 and 500 images per split, respectively. This model achieved average BLEU score of 0.118 and average METEOR score of 0.054.

Collected Captions with Sampled Dataset Images: The next experiment was done using the 2,000 captions collected from the human contributors for the sampled 2,000 images. The same 1,500 training and 500 testing image splits have been used in this experiment for the sake of comparison. The model achieved average BLEU score of 0.192 and average METEOR score of 0.096.

By considering the results, it can be identified that there is an improvement in the average BLEU and METEOR scores of the generated captions from the models trained with human collected captions compared to the translated captions. The contextual and grammatical errors in the Google translated captions can be the main reason for this.

2) MS COCO Dataset: After analysing the results obtained from the models trained on the Flickr8K dataset, the model generates good captions for the images with objects related to the Flickr8K dataset. For other images, the captions' quality and accuracy are deficient. To find the reason to overcome this issue, it was decided to train the implemented model on a large-scale dataset. Eighty thousand images version of the MS COCO dataset has been chosen for this task since the MS COCO dataset has images related to 80 different object categories.

For the MS COCO dataset, the model was experimented with 80,000 images with 80,000 translated Sinhala captions. The dataset has been split into training and testing splits according to 70% and 30%, respectively. The resulting training split consists of 56,000 images, and the testing split consists of 24,000 images. After the training, 500 captions were randomly selected from the test images for caption generation and analysis. For the MS COCO dataset with translated captions, the model has given an average BLEU score of 0.592 and an average METEOR score of 0.281.

3) Summary of Automated Caption Evaluation: The BLEU score and the METEOR score have been used for the automated evaluation of captions generated on the different variations of the datasets. Table I summarises the results for the different dataset variations.

By considering the results in Table I, the highest average BLEU score has been given by the model trained using the full MS COCO dataset with the Google-translated Sinhala captions. The model trained using the full Flickr8K dataset with Google-translated Sinhala captions also gave good results. The two

variations of the sampled dataset have given lower results than the large-scale dataset variations. It is observed that the models have shown promising results when they are trained with larger datasets. And another observation is that the model trained with human-collected captions has performed better than the Google-translated captions.

B. Manual Caption Analysis

For this manual evaluation approach, three independent measures have been introduced. Those are identification of the context of the caption, grammatical correctness of the caption, and finally, the overall meaning of the caption concerning the image. This section presents some significant instances observed in the manual evaluation of generated Sinhala captions for Flickr8K and MS COCO datasets.

1) Flickr8K Dataset: For the Flickr8K dataset, captions generated from the model trained on the sampled collected captions have been chosen for the manual evaluation. Figure 2 shows a summary of generated captions of some of the sample images.

In Figure 2a, the model could identify the context very accurately as it could identify snow on the ground and the two dogs. The grammatical correctness of the generated caption is high, and the overall idea given from the caption about the image is also very satisfactory. BLEU and METEOR scores of the generated caption are also high.

The model has generated a grammatically accurate caption for the image shown in Figure 2b after correctly identifying a smiling woman with brown hair. The caption also says that the woman is posing for a photo. This action cannot be correctly identified even by a human easily. BLEU and METEOR scores are given low values even though the caption can give the reader a good understanding of the image.

The model could not identify the context correctly for the image shown in Figure 2c. Even though the generated caption says two dogs are playing in the sand, the image consists of a dog playing with a ball on a grassy lawn. The grammatical correctness of the caption is high, and the BLEU and METEOR scores have given averagely good values for the generated caption.

BLEU and METEOR metrics have given zero as the value for the caption generated for the image shown in Figure 2d, even though the generated caption gives an accurate idea to the reader. This is because the generated caption does not consist of a single word from the provided ground truth reference sentence. The grammatical correctness of the generated caption is also satisfactory.

The image shown in Figure 2e consists of a flock of pigeons lined up on a wall, but the generated caption says a boy is walking by. The context identification in this scenario is very low even though the model has generated a grammatically accurate caption. The BLEU and METEOR metrics have given zero as the score for the generated caption.

2) MS COCO Dataset: Figure 3 shows a summary of generated captions of some of the samples.

For the image shown in Figure 3a, the model has been able to identify the context very accurately. The BLEU score has given the value one, while METEOR has given 0.625 for the generated caption. Grammatical accuracy is also satisfactory, while the generated caption gives an excellent understanding to the reader.

Reference caption:
හිම සහිත බිමක් මතින් දිවි බල්ලන් දෙදෙනෙක්
(Two dogs on a snowy ground)
Generated caption:
හිම සහිත බිමක් සිටින බල්ලන් දෙදෙනෙක්
(Two dogs on a snowy ground)
BLEU score: 0.747 **METEOR score:** 0.701
Context identification: Good **Grammatical correctness:** Average
Overall accuracy: Good

(a) Sample image 1

Reference caption:
මුහුණේ රතු පැහැති පායම් ගාගෙන තොප්පියක් පැලඳ සිනාසෙමින් කතාබහ කරමින් සිටින කාන්තාවක්
(A woman with a red face and a hat talking and smiling and talking)
Generated caption:
සිනාසෙමින් ඡායාරූපයකට පෙනී සිටින දුඹුරු පැහැති හිසකෙස් සහිත කාන්තාවක්
(A woman with brown hair posing for a photo with a smile)
BLEU score: 0.427 **METEOR score:** 0.170
Context identification: Good **Grammatical correctness:** Good
Overall accuracy: Good

(b) Sample image 2

Reference caption:
රතු පැහැති බෝලයක් සෙල්ලම් කරන බල්ලෙක්
(A dog playing with a red ball)
Generated caption:
වැලි සෙල්ලම් කරන බල්ලන් දෙදෙනෙක්
(Two dogs playing in the sand)
BLEU score: 0.513 **METEOR score:** 0.317
Context identification: Bad **Grammatical correctness:** Good
Overall accuracy: Bad

(c) Sample image 3

Reference caption:
කළු ඉහුණු ආවරණ සුදු ඇඳුම් ඇඳගත් දෙදෙනෙක් කඩු හරඹය යෙදෙයි
(Two black-clad white men wield swords)
Generated caption:
ක්‍රීඩාවේ නිරතවන ක්‍රීඩකයින් සිටී
(There are athletes who play sports)
BLEU score: 0 **METEOR score:** 0
Context identification: Good **Grammatical correctness:** Good
Overall accuracy: Good

(d) Sample image 4

Reference caption:
කාප්පයක් එක පෙලට පෙළට සිටින පරාවියන් රැවුවක්
(A flock of pigeons lining a wall)
Generated caption:
පිරිමි ළමයෙක් ගමන් කරයි
(A boy walks by)
BLEU score: 0 **METEOR score:** 0
Context identification: Bad
Grammatical correctness: Good
Overall accuracy: Bad

(e) Sample image 5

Fig. 2: Samples of manual caption analysis on Flickr8K dataset - I

The model has generated a grammatically accurate caption for the image shown in Figure 3b. The model could identify the school bus and the man standing next to the bus very accurately. The overall idea and the accuracy of the generated caption are very satisfactory, and the BLEU and METEOR metrics have given higher values as the score.

TABLE I

SUMMARY OF AUTOMATED CAPTION EVALUATION ON DIFFERENT DATASET VARIATIONS

Dataset variation	Number of train images	Number of train captions	Average BLEU Score	Average METEOR Score
Sampled Flickr8K dataset with translated captions	1,500	1,500	0.188	0.054
Sampled Flickr8K dataset with collected captions	1,500	1,500	0.192	0.096
Full Flickr8K dataset with translated captions	5,600	5,600	0.503	0.187
Full MS COCO dataset with translated captions	56,000	56,000	0.592	0.281

The model has generated a caption for the image shown in Figure 3c, saying that a man is looking at his cell phone. Even though the man is using a mobile tablet, this can be considered a good context identification since even a human cannot differentiate between a tablet from a cell phone sometimes. The BLEU and METEOR metrics have given low scores even though this caption provides a good understanding to the reader. The grammatical correctness of the generated caption is also higher.

The image shown in Figure 3d is a perfect example of poor context and object identification. The model has identified the context as a baseball player lifting his bat in the field while the image actually shows an areal image of a street with people. Even though the model has generated a grammatically correct sentence, BLEU and METEOR have given zero as the score because of the poor context identification.

The model generated an exceptionally accurate caption for the image shown in Figure 3e by the model even though the BLEU and METEOR metrics have given low scores. The model has accurately identified the fire trucks and generated a grammatically accurate caption. The low BLEU and METEOR scores are that there are no fire truck-related words in the reference. The overall idea given in the caption to the reader is very accurate.

Sinhala is a morphologically rich language. As discussed in the above section, it is not good to depend only on the automated evaluation metrics when evaluating the generated captions. The main reason for this is when we consider an image, there can be multiple ways of interpreting the idea. Different people may interpret the same image in different ways. Therefore, it is crucial to do a manual caption analysis.

V. CONCLUSION AND FUTURE WORK

The goal of this study was to introduce a model to generate a syntactically and semantically accurate Sinhala caption for a given image. For this task, there was no dataset available at the beginning of this study. Therefore, we have produced Sinhala versions of two benchmark datasets with Sinhala captions. The Flickr8k dataset consists of 8,000 images with five captions resulting in 40,000 total captions. These 40,000 captions have been translated to Sinhala using the 'Google Translation API'. Then another data collection approach was introduced to collect new Sinhala captions for a sample of the Flickr8K dataset from human contributors. This sample consisted of randomly selected 2,000 images from the original dataset. The MS COCO dataset, which consists of 80,000 images and 400,000 captions, has been chosen as the large scale dataset. These 400,000 English captions have been translated into Sinhala using the Google Translation API. The caption generation model has been implemented based on the RNN merge model proposed by [7] for the English language. The InceptionV3, a CNN, has been integrated to extract image features from an image, and an LSTM network has been employed as the language model. The model is trained on the different variations of the Flickr8K dataset and MS COCO dataset.

When considering the results obtained from models trained on all of these dataset variations, the model trained on the Sinhala version of MS COCO has given the best scores. The model trained on the full Flickr8K dataset with Sinhala captions gave promising results. From the two models trained on the samples Flickr8K dataset, the model trained on the collected captions has outperformed the translated captions version by a short margin. In conclusion, we can say that the model's accuracy will increase if there is more data. When considering the results of the model trained on the sampled dataset, it indicates that the human-entered captions are more accurate than the Google translated captions. Therefore, as a combination of these observations, we can expect a better model with accurate caption generation if we train the model on a large-scale dataset with human-entered captions.

As future works, the following tasks have been identified. Several benchmark datasets are used in the image generation task in the literature that has more images. Since the implemented model gets the opportunity of improving with more data, these datasets can be used to construct Sinhala versions of them to use in their models. Other than using Google-translated Sinhala captions, new captions can be collected from human contributors to train the models. For this caption collection, our proposed approach can be used. Another task is, other than using InceptionV3, different competing architectures like InceptionV4 [34], AlexNet [35], VGG-16, VGG- 19 [36], ResNet50 [37], Xception [38] etc., can be used as the image feature extraction model. And for the language model, alternatives for the LSTM networks can be used. In the later studies, for the image captioning task for the English language, many attention and transformer-based models are presented, and they have also given promising results. These techniques can be used to improve the model for generating Sinhala captions.

Reference caption:
ගොඩනැගිල්ලක පැත්තක ඔරලෝසුවක්
(A clock on the side of a building)
Generated caption:
ගොඩනැගිල්ලක පැත්තක ඔරලෝසුවක්
(A clock on the side of a building)
BLEU score: 1 **METEOR score:** 0.625
Context identification: Good **Grammatical correctness:** Good
Overall accuracy: Good

(a) Sample image 1

Reference caption:
රතු කළු පාසල් බස් රථයකින් පිටත සිටගෙන සිටින මිනිසා
(The man standing outside a red and black school bus)
Generated caption:
පාසල් බස් රථයක් අසල සිටගෙන සිටින මිනිසෙක්
(A man standing next to a school bus)
BLEU score: 0.981 **METEOR score:** 0.566
Context identification: Good **Grammatical correctness:** Good
Overall accuracy: Good

(b) Sample image 2

Reference caption:
ටැබ්ලට් පරිගණකයක් දෙස බලන පිරිමින් කිහිප දෙනෙක්
(A few men looking at a tablet)
Generated caption:
මිනිසෙක් තම ජංගම දුරකථනය භාවිතා කරමින් තම ජංගම දුරකථනය දෙස බලයි
(A man looks at his cell phone while using it)
BLEU score: 0.375 **METEOR score:** 0.170
Context identification: Average **Grammatical correctness:** Good
Overall accuracy: Average

(c) Sample image 3

Reference caption:
ගමන් මලු රැගෙන පාරක ගමන් කරන කිහිප දෙනෙකුගේ ඉහළ දසුනක්
(Top view of several people carrying luggage on the road)
Generated caption:
බේස්බෝල් ක්‍රීඩකයෙක් බේස්බෝල් පිටියක පිත්තක් ඔසවයි
(A baseball player lifts a bat on a baseball field)
BLEU score: 0 **METEOR score:** 0
Context identification: Bad **Grammatical correctness:** Good
Overall accuracy: Bad

(d) Sample image 4

Reference caption:
විශාල ට්‍රක් රථ කිහිපයක් මාර්ගයක පෙළ ගැසී ඇත
(Several large trucks lined the road)
Generated caption:
ගිනි නිවන රථයක් ගිනි නිවන රථයක් අසල නවතා ඇත
(A fire truck is parked next to a fire truck)
BLEU score: 0.214 **METEOR score:** 0.079
Context identification: Good **Grammatical correctness:** Good
Overall accuracy: Good

(e) Sample image 5

Fig. 3: Samples of manual caption analysis on MS COCO dataset – I

REFERENCES

[1] N. de Silva, "Survey on publicly available sinhala natural language processing tools and research," *arXiv preprint arXiv:1906.02358*, 2019.

[2] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *European conference on computer vision*. Springer, 2010, pp. 15–29.

[3] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[4] 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.

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceed- ings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[6] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[7] 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.

[8] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled trans- former for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vi- sion*, 2019, pp. 8928–8937.

[9] H. Peng and N. Li, "Generating chinese captions for flickr30k images," 2016.

[10] T. Miyazaki and N. Shimizu, "Cross-lingual image cap- tion generation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1780–1790.

[11] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "Stair captions: Constructing a large-scale japanese image caption dataset," *arXiv preprint arXiv:1705.00823*, 2017.

[12] H. A. Al-Muzaini, T. N. Al-Yahya, and H. Benhidour, "Automatic arabic image captioning using rnn-lst m- based language model and cnn," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 6, 2018.

[13] T. Deb, M. Z. A. Ali, S. Bhowmik, A. Firoze, S. S. Ahmed, M. A. Tahmeed, N. Rahman, and R. M. Rahman, "Oboyob: A sequential-semantic bengali image caption- ing engine," *Journal of Intelligent & Fuzzy Systems*, vol. 37, no. 6, pp. 7427–7439, 2019.

[14] S. K. Mishra, R. Dhir, S. Saha, P. Bhattacharyya, and A. K. Singh, "Image captioning in hindi language using transformer networks," *Computers & Electrical Engineering*, vol. 92, p. 107114, 2021.

[15] A. Singh, T. D. Singh, and S. Bandyopadhyay, "An encoder-decoder based framework for hindi image cap- tion generation," *Multimedia Tools and Applications*, pp. 1–20, 2021.

[16] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *International conference on machine learning*. PMLR, 2014, pp. 595–603.

[17] C. Rashtchian, P. Young, M. Hodosh, and J. Hocken- maier, "Collecting image annotations using amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 139–147.

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

[19] X. Chen and C. L. Zitnick, "Learning a recurrent vi- sual representation for image caption generation," *arXiv preprint arXiv:1411.5654*, 2014.

[20] T. Mikolov, M. Karafia´t, L. Burget, J. Cernocky`, and S. Khudanpur, "Recurrent neural network based language model." in *Interspeech*, vol. 2, no. 3. Makuhari, 2010, pp. 1045–1048.

[21] Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A., Berg, T. and Daumé III, H., 2012, April. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 747-756).

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] 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*. PMLR, 2015, pp. 2048–2057.

[24] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4634–4643.

[25] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 685–10 694.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[27] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, and N. Pugeault, "Image captioning through image transformer," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[28] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual english-german image descriptions," *arXiv preprint arXiv:1605.00459*, 2016.

[29] S. Parida, O. Bojar, and S. R. Dash, "Hindi Visual Genome: A Dataset for Multimodal English-to-Hindi Machine Translation," *Computación y Sistemas*, vol. 23, no. 4, pp. 1499–1505, 2019, presented at CICLing 2019, La Rochelle, France.

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dolla´r, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[32] 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 of the Associ- ation for Computational Linguistics*, 2002, pp. 311–318.

[33] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[36] K. Simonyan and A. Zisserman, "Very deep convolu- tional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[38] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258