IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

IMPROVED IMAGE CAPTION GENERATION FOR LOW RESOURCE LANGUAGES

Ms.G.Aninthitha
Assistant Professor
SRM Madurai College for Engineering and Technology

Abstract. A descriptive sentence is generated for a given image as part of the artificial intelligence challenge known as image caption creation. Our proposed approach uses two computer vision approaches to comprehend the image's content and a language model from the field of natural language processing to comprehend the image and text into the proper sequence of words. For the purpose of creating captions, the image caption generation domain combines computer vision and natural language processing. The dataset is made up of output images with output images with captions that correspond to the input image. The benchmark dataset utilized here is Flickr 8k[14]. Modern deep learning generative models called diffusion models provide high resolution images. Also diffusion models learn to recover the data by removing noise in the reverse process after first adding noise to the images in the forward process. One such neural network that has been trained on various image and text pairs is called Contrastive Language Image Pre-Training(CLIP). For extracting features from images and text, we use ClipProcessor with CLIPFeatureExtractor and CLIPTokenizer encoders. Transferring captions from one language to another is required for automatic translation. English is a high resource language with a large number of captions available, whereas low resource languages like Tamil and Hindi have less material for conversational AI systems to learn. This paper describes our method for translating generated captions from English to Tamil (Indian Languages) using neural machine translation. Thus CLIP architecture to predict english caption achieves 0.1798 BLEU score and the Neural Machine Translation architecture to translate the english caption to tamil caption achieves 8.7018 accuracy

Keywords: Natural Language Processing, Computer Vision, Diffusion Models, Neural Machine Translation

INTRODUCTION

Image captioning is the process of describing what is seen in an image, such as the entities present and the activities taken, by identifying the items, their properties, and their relationships in the image. Two of the main domains that are involved in this task are Artificial Intelligence and Natural Language Processing. In a typical picture captioning task, a syntactically and semantically accurate and meaningful sentence is expected as an output for an input of an image. The natural language processing field has many uses for image captioning, including social networking use, virtual assistants, editing software recommendations, and more. This helps the blind and visually challenged understand their surroundings, which is tremendously helpful. A promising idea, Natural Language Processing-based image captioning has a number of potential uses. A

number of noteworthy advancements have been made in image captioning in the past, with the exception of a very small number of non-english languages like Chinese [2], Turkish [1], and Arabic [3] are some of the few non-english languages which use deep learning to handle the difficulties of semantically complex languages.The merge model for Tamil language caption generation [4] is used in the experiments with an image captioning model employing a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture.

The evaluation of the research findings produced a Bilingual Evaluation Understudy (BLEU) score that was satisfactory. Based on neural networks and the conditional probability of translated sentences from the source language to the target language, the neural machine translation method works [5].

RELATED WORK

To generate a caption for an image is the main goal of image captioning. The captioning of an image must name the items, events, relationships, and other quiet features that might be absent in the picture. The next stage after feature identification is to create the most relevant, precise description for the image, which must be both syntactically and semantically valid. It employs both natural language processing techniques and computer vision theories to identify and describe objects. Although it's challenging for a machine to mimic human brain function, research in this area has made significant progress. Using Convolutional Neural Network and Long Short Term Memory, deep learning approaches are sufficient to address such issues [6,7]. Recent developments in computer vision have focused on denoising diffusion models, which have produced outstanding achievements in the field of generative modeling. A forward diffusion stage and a reverse diffusion stage are the foundations of a diffusion model, which is a deep generative model. Gaussian noise is added to the input data in numerous phases during the forward diffusion stage to gradually alter it. In the reverse stage, a model must learn to progressively and step-by-step reverse the diffusion process in order to recover the original input data. Despite their recognised computational costs, i.e. poor speeds due to the huge number of sampling steps involved, diffusion models are often recognized for the quality and range of the generated samples. The system must recognise and develop connections between things, people, and animals. This study uses deep learning to find, identify, and produce interesting captions for a given image. To detect, identify, and provide captions, Regional Object Detector (RODe) is used [8]. This approach concentrates on deep learning to enhance the current image caption generation system even further. The proposed solution is tested through experiments utilizing the python programming language on the Flickr 8k dataset [8]. This study examined different deep learning models for creating captions for pictures taken from the Flickr 8k Dataset. Additionally, this work aims to merge a CNN-type encoder for feature extraction from images with a recurrent neural network for caption generation. VGG16 and InceptionV3 are the CNN encoders employed in steps of how the diffusion model performs [9]. A unidirectional or bidirectional LSTM is then given the extracted features to produce captions. The suggested approach generates captions from language using beam search and greedy techniques. Then, using BLEU scores, the generated captions are compared to the real captions. Using the BLEU score, one can assess how similar a sentence is to another [9]. The conversion of text from one language to another is known as machine translation. There are a lot of resources available in English on the internet. This universal language is unfamiliar to many people. It takes a lot of time to manually translate them into native languages like Hindi, Tamil. An effective strategy in these situations is automated machine translation. Here, eight advanced architectures were tested for machine translation and compared for efficiency. The six Indian languages being studied are Hindi, Bengali, Gujarati, Malayalam, Tamil, and Telugu. It has been demonstrated in detail how the Word embedding technique affects the BLEU. Encoder-to-decoder networks work well for short sentences. Attention architecture, on the other hand, is appropriate if the sentence is longer than 20 characters. In these networks, the 4 Layer Bi-directional LSTM is a great option for increasing BLEU. The CFILT, UFAL, and ILCC datasets were taken into account in their work and received a BLEU score of

lower accuracy [10]. The majority of image captions are only available in a select few internet-popular languages. Through automatic translation, the task of machine translation of image captions aims to democratize this information for other low-resource languages. Image information can also be used to improve the quality of translated captions in comparison to standard machine translation. The purpose of the proposed work is to demonstrate a variety of deep learning strategies and methods for translating captions from English to Hindi in the most effective and efficient manner. All metrics show that transformer-based methods perform better than sequence-to-sequence methods, with accuracy scores between 5% and 20% higher. Additionally, transformer-based pre-trained approaches are able to easily resolve ambiguity. Additionally, the results demonstrate that, in situations with such limited resources, text-only methods are sufficient, whereas multimodal methods are unable to improve translation quality. Therefore, the majority of applications for translating image captions from English to Hindi call for text-only pre-trained transformers [11]. To the best of our knowledge, no work has been reported in Assamese regarding the majority of caption generation tasks. 14 million people in the North-East of India speak Assamese, an Indo-European language. The generation of image captions for the Assamese news domain is discussed in this paper. An annotated training corpus is necessary for a good image captioning system. Be that as it may, there is no such standard dataset accessible for this asset obliged language. In this manner, a dataset of 13000 pictures gathered from different web-based Assamese e-papers were constructed. The captions for news images are generated using two distinct architectures. The attention mechanism serves as the foundation for both the first model and the second model. Both qualitative and quantitative evaluations are carried out on these models. Subjective examination of the created subtitles is done as far as familiarity and sufficiency scores in light of a standard rating scale. The BLEU and Consensus-based Image Description Evaluation (CIDEr) evaluation metrics are used to evaluate the quantitative results.They notice that the attention mechanism-based model performs better than the CNN-LSTM-based model [12]. Here German image captioning techniques that use English training data to transfer knowledge. In order to generate German image captions, they investigate four distinct approaches, two of which are foundational and two of which are more advanced and based on transfer learning. For the baseline methods, they train a cutting-edge model using a translated version of the English MS COCO dataset and the smaller German Multi 30K dataset. Both advanced methods are trained on the Multi30K dataset and the translated MS COCO dataset. In one of these methods, an alternative attention mechanism from the literature that performed well in English image captioning is used. They compare how well each method performs for the Multi 30K test set using standard automatic evaluation metrics[13].

Our contribution in the paper are outlined below:

1. In this paper we trained our model using benchmark Flickr 8k dataset thereby improving accuracy of image prediction and image caption generation for Tamil language.
2. We have proposed a mixture of two deep learning architectures and applied it to image caption generation of low resource language by translating the captions in Tamil language.
3. We have used distillbert method for both architectures to compress our large model and thus improvise the performance by predicting the exact low-resource caption to the image.

PROPOSED METHODOLOGY & EXPERIMENTAL ANALYSIS

Our proposed work is a combination of two architectures - CLIP architecture with Neural Machine Translation that are trained on benchmark Flickr 8k dataset. The CLIP architecture uses DistilBERT to optimize the performance of our model thereby predicting the right caption to the whole image. This methodology is divided into main tasks as - 1. Feature extraction from both the images and the text and generating a sequence of possible captions based on the similarity. 2. Projecting the sequence of captions to a 3-stage diffusion model to evaluate the right english caption, 3. The predicted english caption is further

processed to extract the stemming of each token. 4. The tokens are then mapped with its respective tamil tokens and image caption in tamil is generated.

Step 1: Feature Extraction

The image and text is loaded into CLIP architecture. In general the CLIP uses two trained models in parallel - 12 layer text transformer to construct text embeddings and vision transformer to construct image embeddings. CLIPTokenizer is used to encode the text and CLIPFeatureExtractor is used to resize and normalize images for the model. The image is split into 32 patches and flattened after adding noise to the image patch, thereby producing a sequence of low-dimensional linear embeddings. As CLIP learns from an unfiltered noisy data, CLIPs text encoder will provide a linear visual representation of the classifier. Both the CLIPTokenizer and CLIPFeatureExtractor are wrapped into CLIPProcessor along with CLIPModel to simultaneously encode and prepare the images. Similarity between the image and the text is evaluated using CLIPProcessor and CLIPModel. [CLS] is the first token added to every sequence that corresponds to a hidden state used to aggregate sequence during classification tasks. [EOS] is the last token added to every sequence. Finally, this stage generates sequences of captions by predicting how likely the image corresponds to the text using contrastive pre-training. Figure 1 depicts the feature extraction and sequence of captions being generated by CLIP architecture.

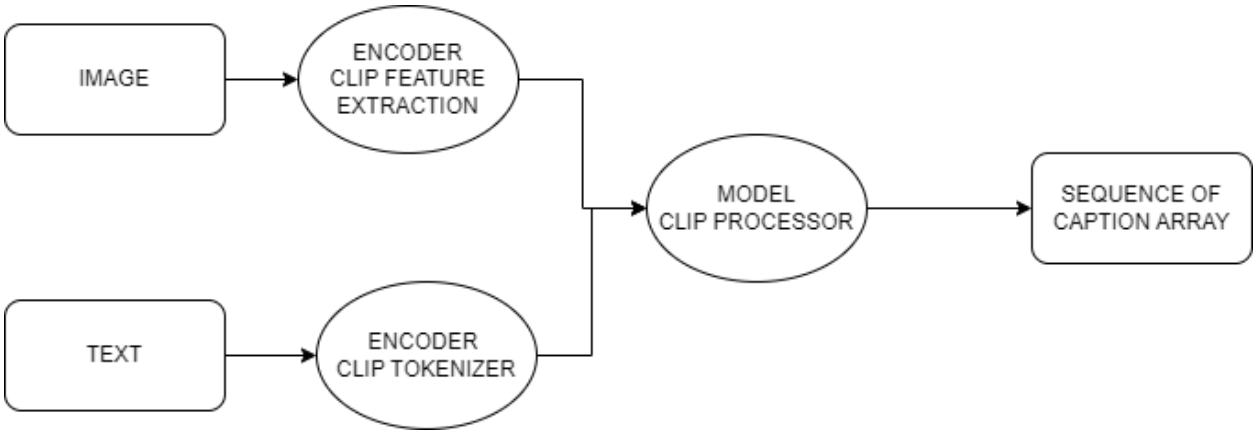

Fig 1: Flow chart for Feature Extraction to Sequence of Caption

Pseudocode_1: Feature Extraction & Caption Sequence Generation

input: (image_path, text)
output: low-dimensional linear text sequences (x_T)

```
image_text = CLIPProcessor(CLIPFeatureExtractor(img), CLIPTokenizer(text))
sequences = CLIPModel(image_text)
for seq in sequences:
        x_T[].append([CLS] + seq + [EOS])
return x_T
```

Step 2: English Caption Prediction using 3-Stage Diffusion Model

The above section provides only a sequence of tokens that are highly similar to image patches. This token sequence is further processed to generate the right caption to the whole image by passing the tokens through a sequence of 3-stage diffusion processes. Each diffusion stage uses DistilBert and fusion methods. Fusion methods include concatenation and addition embeddings. Concatenation embedding, adds a token at the end of the caption sequence. While addition embedding performs a positional embedding. The first stage of diffusion performs denoising of the token sequence using gaussian noise reduction technique thereby

removing all the noised tokens from the sequence. Token sequence x_T is given as input to gaussian noise and the output is the noise-less x_{T-1} sequence. The second stage of diffusion performs positional mask embedding. This positional masking finds the needed word to be included in the caption sequence. The needed word is found by hiding the word in a position and mapping the right word that satisfies the Subject-Verb-Object (SVO) pattern. In general, all english captions are framed as a Subject-Verb-Object pattern. Thus the exact SVO pattern is achieved in the second stage. The third stage places the tokens in the right position that brings a meaningful caption sequence to the image as a whole. Lm-head and softmax is used to extract the ground truth caption of the whole image. The final x_T is taken and the contextual feature is examined and x_0 is projected linearly to a weight matrix lm-head. Softmax is applied to analyze the probability of a predicted token in each position of the caption. The tokens with maximum probability are formed to output the final english text caption. Figure 2 shows the workflow of the 3-stage diffusion model.

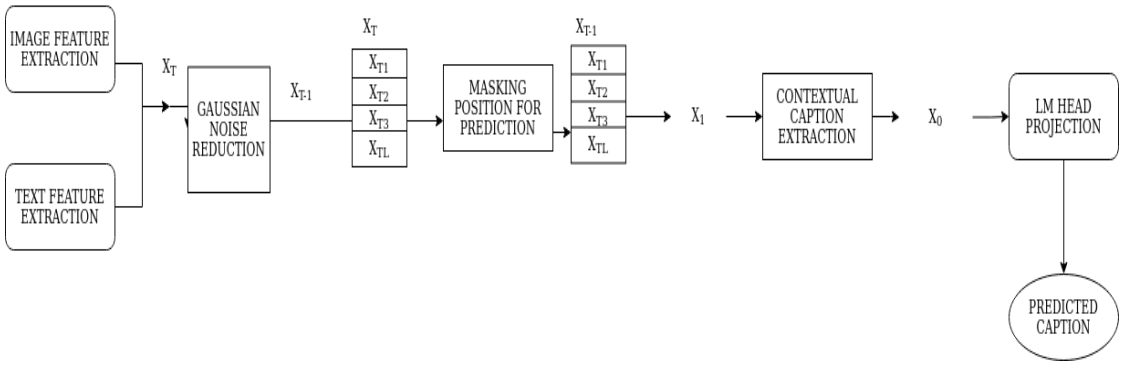

Fig 2:Flowchart for predicting caption from feature extraction

Pseudocode_2: 3-Stage Diffusion Model
input: x_T (sequence of text feature extracted from pseudocode_1) **output: x_0 (predicted caption of whole image)** x_T_1[] ← Remove tokens with noise from x_T[] x_T[L] ← Apply positional masking to x_T_1[] x_0 ← Generate contextual sequence return max_prob(x_0, lm_head, activation=”softmax”)

Step 3: Data Pre-processing for Tamil Caption

The generated english caption is further processed to produce a tamil caption of the whole image. This data pre-processing includes case conversion, tokenization, noise reduction, lemmatization and stemming. The english caption is converted to lower-case and tokenized into words. Punctuations and symbols are removed during noise reduction. Lemmatization here, maps the synonym or generic meaning of the words. Stemming is used to extract the origin word of each token.

Step 4: Neural Machine Translation

We propose another architecture to perform low-resource language to our predicted english caption. The input to the encoder is the pre-processed English caption from the above data pre-processing step, while the output of the decoder is the translated low-resource language. The input sequence is provided as input to the encoder to generate a vector sequence by mapping the English vocabulary corpus. This vector sequence holds semantic information of the words in the input sequence. The decoder translates each vector into low-resource, till the [EOS] of the input sequence is reached. The decoder uses the english-tamil vocab corpus to predict the exact meaningful tamil vector. Finally, a low-resource caption of the whole image is predicted using

both the architecture. Figure 3 shows how the translation of the english caption to tamil caption is processed orderly.

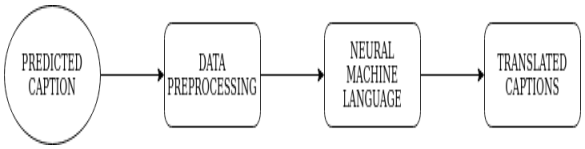

Fig. 3 Flow chart for translation of languages from predicted captions

Pseudocode_3: Neural_Machine_Translation

input: predicted_english_caption
output: predicted_tamil_caption
P ← predicted_english_caption
/*** Data pre-processing ***/
P ← lower_case_conversion(P)
p[] ← tokenization(P)
p[] ← noise_reduction(p[])
p[] ← lemmatization_and_stemming(p[])
/*** Encoder-Decoder ***/
e ← encoder(vector_sequence(p[]))
O ← decoder(e, vocab)

This paper proposes a NLP method for generating captions for low-resource languages. The dataset used is Flickr 8k, which has 8000 images to train the model. Figure 4 depicts the code used to generate english captions for the test_image loaded.

```
model.eval()
with torch.no_grad():
  image = Image.open("./test_image.jpg")
  clip_processor = CLIPProcessor.from_pretrained("./tokenizers/openai/clip-vit-base-patch32-local")
  clip = CLIP.from_pretrained("./models/openai/clip-vit-base-patch32-local")
  plt.imshow(image)
  plt.show()
  inputs = clip_processor(text="", images=image, return_tensors="pt", padding=True)

  image_clip = clip.get_image_features(pixel_values=inputs["pixel_values"]).unsqueeze(0)
  image_clip = image_clip / image_clip.norm(p=2, dim=-1, keepdim=True)

  text_clip = torch.zeros_like(image_clip)
  for _ in range(10):
    restored = torch.randn((1, 18, 768), device=device)
    for i in range(10):

      out, restored = model(restored[:, :MAX_LENGTH, :], image_clip, text_clip, torch.ones((1, MAX_LENGTH)), torch.tensor([[1, 0]]))

      if i % 2 == 0:
        print("inferred: ", dataset.tokenizer.decode(nn.functional.softmax(out, dim=-1).argmax(dim=-1)[0]))
    print()
```

Fig. 4 Implementation for english caption prediction for Flickr 8k

The figure 5 represents how the captions are generated during every stage of our mixed architecture. The model takes an input image, flattens into 32 patches of low-dimensional linear features. The output features obtained for the test image are shown as Sequences obtained in the figure 5. Contextual caption generation represents how meaning captions are formed. Among the three generated meaningful captions, we apply lm-head with softmax, which produces the predicted caption as ‘Clouds on the sky’. Even though the three captions give meaningful captions. The one with high probability is fixed as the predicted caption. Further we used Neural Machine translation to translate the generated english caption to low-resource (tamil) caption. The neural machine language performs lemmatization and stemming on the english caption, by extracting the ‘cloud’ from ‘clouds’. And finally translated the english caption with its appropriate tamil vectors with the help of vocab.

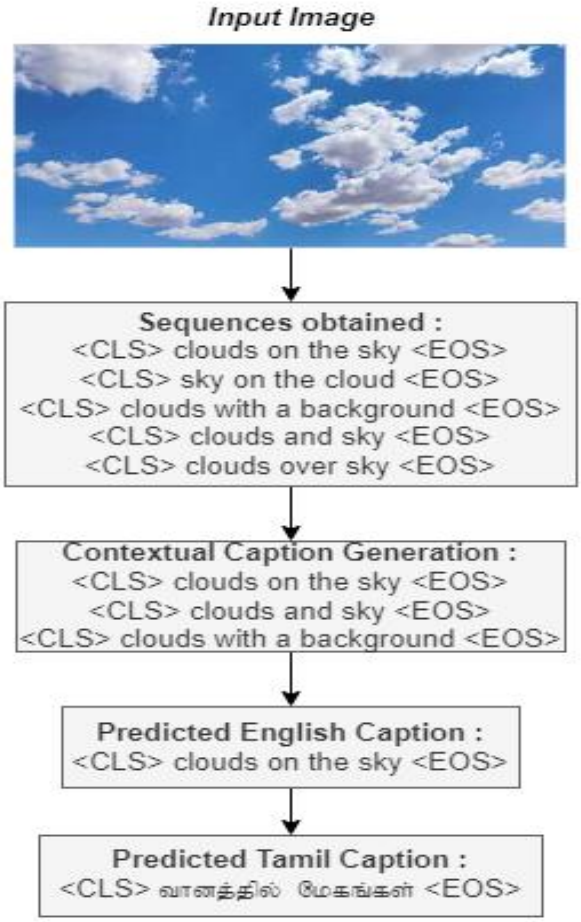

Fig. 5 Screenshot of low resource language image of predicted english caption translated to tamil caption

CONCLUSION AND FUTURE WORK

In this paper, we proposed a combination of two architectures - CLIP architecture with Neural Machine Translation, that are trained on benchmark Flickr 8k dataset. We used Vision Transformer (ViT) with DistilBERT and Neural Machine Language to generate low-resource captions of the image. Our proposed model is trained using DistilBERT thereby using only 40% of the parameters to train the model. Thus the training takes less time compared to BERT models. The neural machine language also translated the english caption with its appropriate tamil semantic caption. Ground truth caption for the whole image is generated and translated. The CLIP architecture to predict english caption achieves 0.1798 BLEU score and the Neural Machine Translation architecture to translate the english caption to tamil caption achieves 8.7018 accuracy. Even though color is extracted and complex images that include exact scenarios in an image are not captioned. Our model generates only a generic caption. This work can be extended to the Flickr30k dataset. Also generate other low-resource captions like french, korean,hindi.

REFERENCES

[1] Yılmaz, B. D., Demir, A. E., Sönmez, E. B., & Yıldız, T. (2019). Image Captioning in Turkish Language. In *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)* (pp. 1-5). IEEE.

[2]Zhang, C., Dai, Y., Cheng, Y., Jia, Z., & Hirota, K. (2018, December). Recurrent attention LSTM model for image Chinese caption generation. In *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)* (pp. 808-813). IEEE.

[3]Al-Muzaini, H. A., Al-Yahya, T. N., & Benhidour, H. (2018). Automatic Arabic image captioning using RNN-LSTM-based language model and CNN. *International Journal of Advanced Computer Science and Applications*, *9*(6).

[4] Rajalingam, G., & Wickramaarachchi, W. U. (2021). Image Captioning in Tamil Language with Merge Architecture.

[5] Choudhary, H., Pathak, A. K., Saha, R. R., & Kumaraguru, P. (2018, October). Neural machine translation for English-Tamil. In *Proceedings of the third conference on machine translation: shared task papers* (pp. 770-775).

[6] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj and R. K. Mishra, "Image Captioning: A Comprehensive Survey," 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), 2020, pp. 325-328, doi: 10.1109/PARC49193.2020.236619.

[7]Croitoru, F. A., Hondru, V., Ionescu, R. T., & Shah, M. (2022). Diffusion models in vision: A survey. *arXiv preprint arXiv:2209.04747*.

[8] Kumar, N. K., Vigneswari, D., Mohan, A., Laxman, K., & Yuvaraj, J. (2019, March). Detection and recognition of objects in image caption generator system: A deep learning approach. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)* (pp. 107-109). IEEE.

[9] Takkar, S., Jain, A., & Adlakha, P. (2021, April). Comparative study of different image captioning models. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1366-1371). IEEE.

[10] Gogineni, S., Suryanarayana, G., & Surendran, S. K. (2020, September). An Effective Neural Machine Translation for English to Hindi Language. In *2020 International Conference on Smart Electronics and Communication (ICOSEC)* (pp. 209-214). IEEE.

[11] Bisht, P., & Solanki, A. (2022). Exploring Practical Deep Learning Approaches for English-to-Hindi Image Caption Translation Using Transformers and Object Detectors. In *Applications of Artificial Intelligence and Machine Learning* (pp. 47-60). Springer, Singapore.

[12] Das, R., & Singh, T. D. (2022). Assamese news image caption generation using attention mechanism. *Multimedia Tools and Applications*, *81*(7), 10051-10069.

[13] Biswas, R., Barz, M., Hartmann, M., & Sonntag, D. (2021, November). Improving German Image Captions Using Machine Translation and Transfer Learning. In *International Conference on Statistical Language and Speech Processing* (pp. 3-14). Springer, Cham.

[14] Datasets Hodosh, Micah, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics." Journal of Artificial Intelligence Research 47 (2013): 853-899.