Image Captioning from Wikipedia for Multi-Language using Deep Learning Models

Anusha Garlapati
Dept of Electronics and Communication Engineering,
Amrita Vishwa Vidyapeetham,
Amritapuri, India.
garlapatianusha@am.students.amrita.edu

Neeraj Malisetty
Dept of Electronics and Communication Engineering,
Amrita Vishwa Vidyapeetham,
Amritapuri, India.
neerajmalisetty@am.students.amrita.edu
gayathrin@am.amrita.edu

Gayathri Narayanan
Dept of Electronics and Communication Engineering,
Amrita Vishwa Vidyapeetham,
Amritapuri, India.
gayathrin@am.amrita.edu

***Abstract* – With the advancement in deep learning, the consolidation of text classification and image processing has agitated enormous scrutiny in the past years. Captioning images are advanced research mainly in the field of computer vision. Identification of relevant items, their properties, and their connection in images is required for image captioning. This image captioning is a complex and challenging task, several developments are done in this field by researchers. But now, Technology can develop a model that can bring out the closest text or captions explaining an image because of the advancementin deep learning methodologies and the enormous amounts of data that is available. In the beginning, it was contemplated impractical that a computer could characterize an image.**

Even, NVIDIA is developing an app to assist persons with limited or no vision using image captioning technology. The main aim of this project is to develop a model such that it can predict the closest text or caption to that particular image. This analysis can be done by utilizing deep learning methods as these methods are efficient for handling the complexity involved in image captioning. So, this project intends to build a model that predicts captions or labels from images by using deep learning models. The VGG16 architecture is employed in this research for the prediction and generation of captions for the images and this analysis can be compared with other deep learning frameworks such as LSTM or CNN to observe the performance of the model [1].

Keywords – Text-Classification, Image Captioning, Deep Learning, Computer Vision

I. INTRODUCTION

The topic of autonomously producing eloquent words or captions for images has piqued significance in the research area of computer vision along with Natural Language Processing in recent years. Computer vision has made substantial advances in the field of image processing in recent years, such as image categorization and object identification. The development of characterizing the content of an image is referred to as image captioning. Encoder-Decoder architectures are generally used in image captioning methods, with vectors of images as input to the encoder [2]. Image captioning is a key activity that necessitates a semantic comprehension of images as the capacity to generate accurate and precise descriptions for images.

Nowadays, Image captioning is a relatively new and rapidly increasing research area. Several new approaches are being launched on daily basis to attain satisfying outcomes in this research field. Visual Captioning is a crucial endeavor for bettering the human-computer synergy and gaining a better grasp of the principles that underpin human image characterization. Captioning images may be thought of as a whole process. The process is known as Sequence-to-Sequence Process because it translates images, which are considered to be a sequence of pixels, into the sequence of sentences or words [2]. The primary goal of this captioning of images is to create a natural language description for an input image that is given to the model.

II. LITERATURE REVIEW

The authors in this study come up with a hybrid system that utilizes an architecture called Multi-layer Convolutional Neural Network (CNN) to produce vocabulary to characterize an image and utilize other architecture known as Long Short-Term Memory (LSTM) to precisely form substantial sentences utilizing the stemming words [1]. Generally, producing comprehensive and instinctive image characterization has a wide range of applications including captions for news images, characterization of images in the field of medical analysis, text-based image retrieval, Instructions for blind people, and communication between humans and robots.

The concept is established on the detection of objects and actions that must be taken in the image given to the model. Vectors derived from images collected for object detection are utilized in the most effective methods. According to the observation from the survey, it reveals that this analysis was also done by utilizing transformers which extends this object relation transformer technique by decidedly adding content about the dimensional connection among input recognized items by utilizing geometric attention [5]. From the previous papers, it is observed that CNN is utilized to interpret visual information and locate objects in an image, whereas RNN or LSTM is utilized to generate descriptions for images.

III. DATA SET

Each row in data is characterized by a specific individual. Here, this analysis was done by utilizing multiple data sets such as training data, captions for text data, image data, image-pixels data, Resnet-embeddings, and so on. This analysis was done on the large image data set that is released by the Wikimedia Foundation cloud. This data was analyzed for

multiple languages. Here, it has nearly 79 languages. The format of data is shown below:

TABLE I. TRAINING DATA

language	Page-URL
en	https://en.wikipedia.org/wiki/Silver_spoon
Zh-TW	https://zh.wikipedia.org/wiki/%E5%8C%97/kita-sendai-staion

Here, In TABLE1: en stands for the English language.

Zh-TW stands for Chinese (Taiwan)

TABLE II. TRAINING DATA

Page Title	Section Title	Mime-Type
Silver spoon	Historical uses	Image/jpeg
北仙台站	JR 東日本	Image/jpeg

TABLE III. TRAINING DATA

Caption reference	height	width
Two silver-gilt strainer spoons on the table	1194	2139
月台	2112	2816

TABLE IV. TRAINING DATA

Is main page	Context section description
false	Before the place setting became popular around the 18th century.
true	北仙台站是一個位於日本宮城縣仙台市青葉區昭和町，屬於仙山線、仙台市地下鐵南北線

TABLE V. TEST DATA

id	Image URL
0	https://uplaod.wikimedia.org/wikipedia/commons/3/scots_Gaelic_speakers_2011_cesnus.png
1	https://upload.wikimedia.org/wikipedia/commons/e/Thermopylae_Ancient_coastline.jpg

TABLE VI. CAPTIONS DATA

Albert Pike [SEP] Albert Pike
Anna Blount [SEP] Blount and her young daughter,in 1911

Fig. 1. Sample Image Data Set

This is some sample data concerning training, testing, captions, and image data. Before Achieving any results from data, we inhibit pre-processing techniques.

IV. DATA ANALYSIS

One of the dominant steps while interacting with data is exploring data by utilizing data pre-processing techniques. This is one of the most important steps to process the data before giving it to any model for training. There was a challenge in deciding which data to utilize to train the model for getting efficient results as there is more than one data set to accomplish our aim. Here, this pre-processing analysis was done by using different kinds of plots such as tree-maps using sqaurify, bar graphs, and image processing were done by scaling the images and flattening them and text processing was done by using NLP (Natural Language Processing) techniques such as Lowering the text, stop words removal, tokenizing the text, and so on.

Tree-Map: This is used to predict data by nesting rectangles of varied sizes together. The size of Each rectangle is comparable to the quantity of data it performs as a percentage of the total.

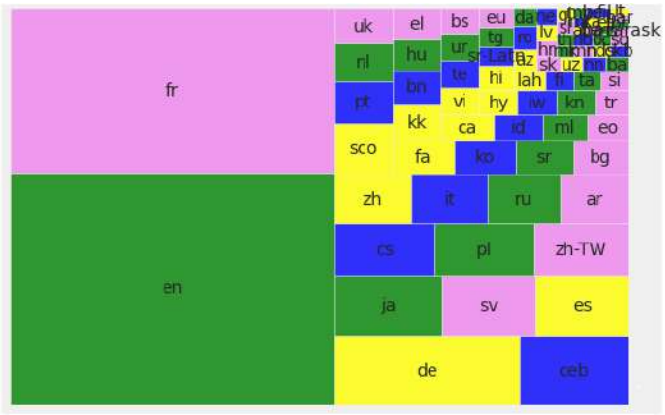

Fig. 2. Tree Map

From Figure2, it interprets that out of 79 languages English and French are most used in data.

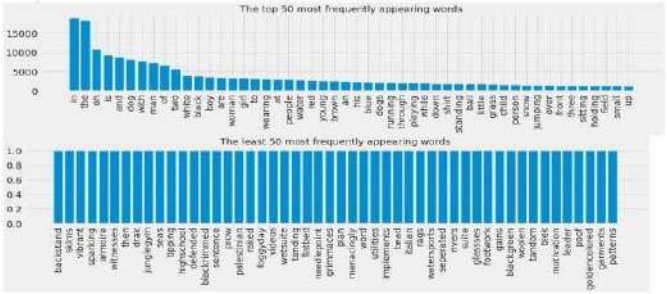

Fig. 3. Top 50 most frequent words

From Figure3, Results interprets that they are the top 50 most frequent words of data after doing text pre-processing techniques on data such as stemming, tokenization.

Tokenization: This can be accomplished in words or sentences. This is the most frequently utilized technique in any NLP Problem statement. It divides sentences into a piece of words.

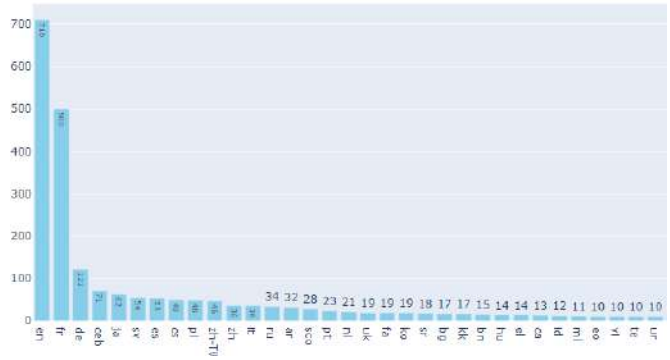

Fig. 4. A bar graph showing Language Distribution

Figure4, shows the language distribution of the data set. And it is observed that English is the most utilized language

Fig. 5. Word Cloud for Page Title

Figure5, it shows the word cloud concerning the page title in data. It shows the most used words in the data.

These are some data analysis techniques concerning text training data and text data. Apart from the text and training data processing, analysis was done even on image data such as re-sizing, flattening the image, and scaling it to pixels (Normalization). These processing techniques are done on images because we cannot just feed an image directly to a model without doing pre-processing on images.

Flattening: This is a method utilized to disciple multi-dimensional array to 1-D array. It's usually utilized in training deep learning tasks while giving input 1-D array for classification models. This is done because multi-dimensional arrays utilize more memory arrays1-D array. To save time complexity while dealing with large data sets, flattening is one of the processing should be performed on images before training the model.

Normalization (Scaling to pixels): This is a technique in image analysis that adjusts the pixel's range. The intention behind this is to adjust an image to its intensity values so that analysis will be efficient.

From Figure6, it is anticipated that results are corresponding to page title after performing image processing techniques on image data.

Fig. 6. Image analysis corresponding to the Page title

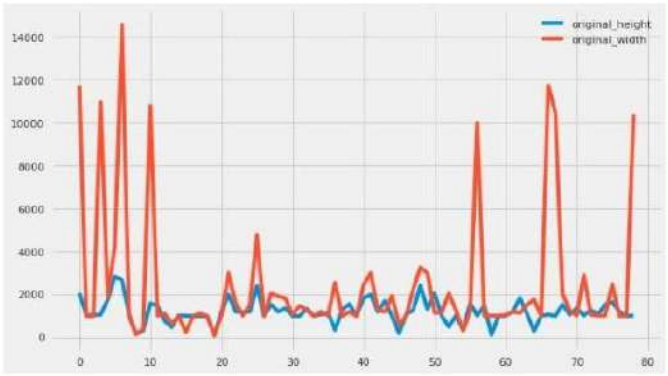

Fig. 7. Graph related to Image dimensions

This interprets the dimensions of images corresponding to height and width.

V. METHODOLOGY

The analysis and perceptions of data were done in two phases. In Phase-I, Deep learning models such as LSTM, VGG16 are utilized for training data and to predict captions for the test data. And in Phase-II, this was developed for further improvements and this project got deployed using Heroku and VS Code. Deep Learning is one of the most quickly evolving and explored fields of analysis that is permeating day-to-day life [7]. It is commonly the development of neural networks by utilizing high-end contemporary frameworks. It enables the advancement of training and the application of considerably bigger neural networks than heretofore.

Researchers have recommended hundreds of different kinds of particular neural networks as improvements or changes to the current models. CNN is most well-known. In this analysis, the model is developed such that text processing is done by utilizing Natural Language Processing and pre-trained models such as Xception are used to extract features from the images,

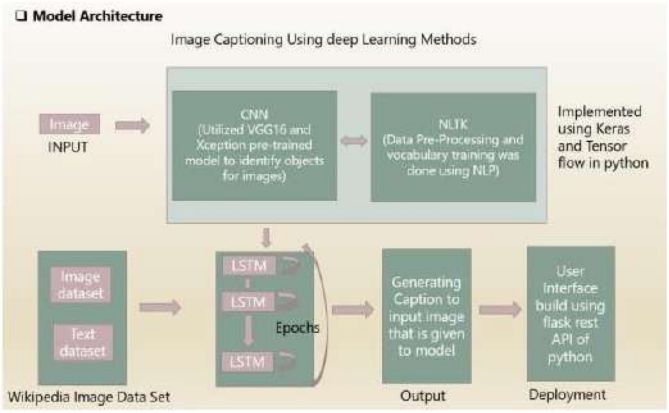

Fig. 8. Model Architecture of our Project

and then they are fed further into LSTM, for generating captions for the test data [4]. By utilizing LSTM, captions for test data are predicted. After all the analysis Results are compared by using the BLEU score, and this project got deployed by utilizing Heroku.

Figure8, it interprets the entire process of our project.

CNN: This stands for Convolution Neural Network where Image data is mapped to a target variable. They have proven to be successful in that they are now the techniques of choice for any form of prediction issue utilizing data as an input to the model. CNN is a multi-layered feed-forward neural network that is built by layering several hidden layers on top of one another in a certain sequence [8]. These layers are frequently outlawed by several layers in CNN, while activation layers are usually enhanced by layers in the convolutional network.

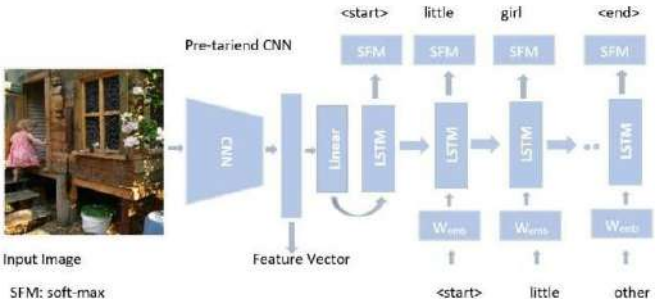

Fig. 9. Overview of Architecture

Figure9, it shows the Architecture of the model how each layer contributes to the model.

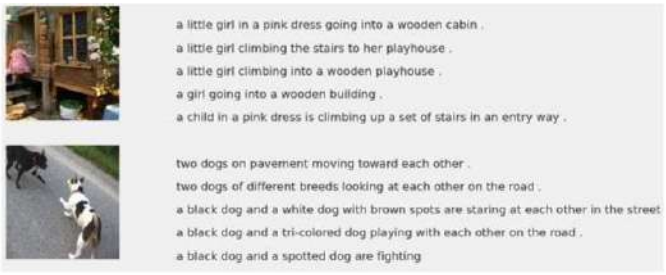

Fig. 10. Training phase results

Prediction of captions for images was analyzed in three phases in this model architecture Feature Extraction for images, Sequence Processor and Decoder.

Feature Extraction: This is done by using pre-trained models such as Xception and VGG16. This is known as transfer learning.

Sequence Processor: This acts as an embedding layer, by interacting with the text data. This contains disciples for extracting text's needed characteristics.

Decoder: This is the concluding phase that utilized advancement techniques to merge image extractor with sequence processor, which is passed to neuron and ultimately for final output phase.

Xception: This is a pre-trained model that has 36 layers that permit it to learn quickly. These are given to the LSTM layer after being developed by a dense layer to provide 2048 vector enhancement of the image [10]. This model is pre-trained on huge data and extracts features from this analysis to utilize them

with the current problem statement analysis. This has trained on image net data that has 1000 various kinds of images for categorization and this model can be loaded directly from applications of Keras. As this model is purely trained on image net data, we have done changes when anticipating this model.

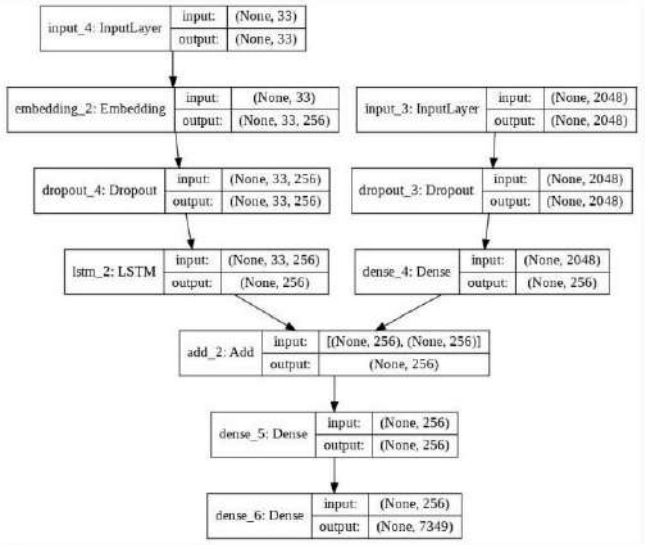

Fig. 11. Model Architecture of Xception Model

Here, From Figure11, input_3 is the input of the pre-trained model.

LSTM:

LSTM stands for Long Short-Term Memory. Here, the CNN model is utilized to get features from images with the help of VGG16-Xception models, which are then input to the architecture of LSTM, which generates the captions for data. This is known as the CNN-LSTM model is primarily developed for the prediction issues involving inputs like images and video captures.

These extracted features from input data by using CNN layers are paired with LSTM for forecasting vectors in further development [4]. These models have a lot of promise and are developing being employed for more complex analysis like text categorization.

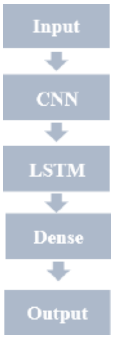

Fig. 12. Model Architecture of CNN-LSTM

Figure12 interprets the overview architecture of the CNN-LSTM model that is utilized for feature extraction and getting captions. The results from the training phase of LSTM are shown below in Figure13.

Model: "model_1"

Layer (type)	Output Shape	Param #	Connected to
input_3 (InputLayer)	[(None, 30)]	0	[]
embedding (Embedding)	(None, 30, 64)	286464	['input_3[0][0]']
input_2 (InputLayer)	[(None, 1000)]	0	[]
CaptionFeature (LSTM)	(None, 256)	328704	['embedding[0][0]']
ImageFeature (Dense)	(None, 256)	256256	['input_2[0][0]']
add (Add)	(None, 256)	0	['CaptionFeature[0][0]', 'ImageFeature[0][0]']
dense (Dense)	(None, 256)	65792	['add[0][0]']
dense_1 (Dense)	(None, 4476)	1150332	['dense[0][0]']

Total params: 2,087,548
Trainable params: 2,087,548
Non-trainable params: 0

Fig. 13. Results from LSTM

Features of images to be anticipated are developed by utilizing this VGG16- Xception Architecture. The weights of VGG16 are frozen when we develop the LSTM model.

Model: "vgg16"

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 224, 224, 3)]	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
fc2 (Dense)	(None, 4096)	16781312
predictions (Dense)	(None, 1000)	4097000

Total params: 138,357,544
Trainable params: 138,357,544
Non-trainable params: 0

Fig. 14. Results from VGG16

With this analysis, our model got trained on many images and text data and was given to predict test data captions.

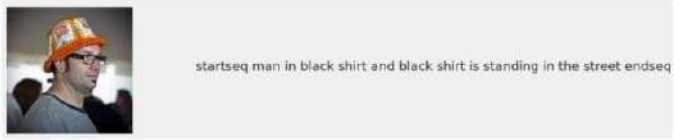

Fig. 15. Testing done on images and predicted caption

So finally, after doing this analysis on the training and testing phase, these are compared using an evaluation metric called BLEU score.

BLUE Score: To figure out a caption, it is correlated to its captions. It stands for Bilingual Evaluation Understudy. This is an algorithm that is utilized for examining the quality of the machine-translated text.

Fig. 16. Figure15 – Image caption predictions compared with BLEU score.

To improve the BLEU score we have developed our model further to get efficient. After development, the test data predictions are shown below from TABLE 7 and TABLE 8.

100% 92366/92366 [44:45<00:00, 34.94it/s]

TABLE VII. TEST DATA PREDICTIONS SAMPLE

image-URL: Scots Gaelic speakers in 2011 census.png[SEP]
closest captions:
Sao Vicente do Penso [SEP] Escudo
Jocs Panamericans de 2011 [SEP] Guadalajara
Bois de pins et chenes de Madren [SEP] carte
Standard time in the United States [SEP] 1913
Hebrides [SEP] Geographic distribution of speakers (2011)

TABLE VIII. TEST DATA PREDICTIONS SAMPLE

image-URL: Thermopylae ancient coastline large.png[SEP]
closest captions:
Oberlin College [SEP] logo
Departmento de Flores [SEP] Escudo
Sekolah Menengh Kebangssan Selandar [sep] cny 4
Academia de Ciencias de Cuba [SEP] Sede.
Geothermal areas in New Zealand [SEP] Geyser Flat

From TABLE7 and TABLE8, these are the inferred results after training and predictions of captions for test data.

TABLE IX. FINAL TEST DATA PREDICTIONS

id	Caption title predicted
0	Sao Vicente do Penso [SEP] Escudo
1	Oberlin College [SEP] logo

TABLE9 infers the final test predictions from data.

```
predict_caption(enc)
```

```
'brown and white dog is running on grass'
```

Fig. 17. Results from final test predictions

```
predict_caption(ec)
```

```
'group of people are walking around street'
```

Fig. 18. Results from final test predictions

VI. DEPLOYMENT

It's a way of bringing our models together in a place where they can be deployed to a web app. Data is anticipated and examined by utilizing VGG16-Xception and CNN-LSTM models. The model is now complete and ready to use. Flask and Tensor flow was utilized in vs code to create the application for testing. Heroku and Git-hub are utilized to deploy the project.

Fig. 19. Web Page of our Project

Fig. 20. Results from Web Page

Fig. 21. Results from Web Page

Fig. 22. Results from Web Page

VII. CONCLUSION

In this Paper, Analysis was done on captions corresponding to text data and image data. This was done in two phases. In Phase-I, the VGG16-Xception model was used to excerpt features from images, and then these are fed into the CNN-LSTM framework for training and to get captions for test data. The model performed well with this data and the results are compared using the BLEU score.

In Phase-II, this model is ready for deployment and this project application was tested using VS Code and deployed in the Heroku app which gives ultimate results for our model.

VIII. FUTURE SCOPE

In practically every sophisticated field of AI (Artificial Intelligence), Captioning of images provides several advancements. This can be further extended by utilizing attention models and by using various Greedy Search Interfaces. This can be even extent for producing captions for live videos as this is a more popular research area nowadays, extracting captions from videos by utilizing text augmentations and can extend to other tasks related to security.

REFERENCES

[1] A. Garlapati, M. Neeraj, G. Narayanan, "Classification of Toxicity in Comments Using NLP and LSTM," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), 2022.

[2] Sanjay S.P., Ezhilarasan N., Anand Kumar M., Soman K.P., "AMRITA-CEN@FIRE2015: Automated story illustration using word embedding (2015), CEUR Workshop Proceedings, 1587, pp. 67 -70

[3] Gautam K.S., Parameswaram L., Thangavel S.K., "A Cascade Color Image Retrieval Framework," (2020) New Trends in Computational Vision and Bio-Inspired Computing Selected Works Presented at the ICCVBIC 2018, pp. 23-36, DOI: 10.1007/978-3-030-41862-5_3.

[4] Viswanathan S., Anand Kumar M., Soman K.P, "A Sequence-Based Machine Comprehension Modeling using LSTM and GRU" (2019), Lecture Notes in Electrical Engineering, 545, pp. 47-55, DOI: 10.1007/978-981-13- 5802-9_5

[5] Wang, Chaoyang, Ziwei Zhou and Liang Xu. "An Integrative Review of Image Captioning Research." Journal of Physics: Conference Series 1748 (2021)

[6] Viktar and Dmitrij Sesok, "Text Augmentation Using BERT for Image Captioning," Applied Sciences 10 (2020).

[7] Ningthoujam C., Chingtham T.S., "Comprehensive Comparative Study on Several Image Captioning Techniques Based on Deep Learning Algorithm (2022) Lecture Notes in Networks and Systems, 281, pp. 229-240, DOI: 10.1007, / 978-981-16-4244-9_18.

[8] Herdade, Simao, Armin Kappeler, Kofi Boakye and Joao Soares. "Image Captioning: Transforming Objects into Words," NeurIPS (2019).

[9] Liu, Shuang & Bai, Liang & Hu, Yanli & Wang, Haoran., "Image Captioning Based on Deep Neural Networks," MATEC Web of Conferences (2018).

[10] Srinivasan, Lakshmi Narasimhan, and Dinesh Sreekanthan, "Image Captioning – A Deep Learning Approach" (2018).

[11] Wang H., Zhang Y., Yu X., "An Overview of Image Caption Generation Methods," 2020 Computational Intelligence and Neuroscience, 2020, art. no. 3062706, DOI: 10.1155/2020/3062706.

[12] Yang Z., Wang P., Chu T., Yang J., "Human-Centric Image Captioning," (2022) Pattern Recognition, 126, art. no. 108545, DOI: 10.1016/j. patcog.2022.108545

[13] Eleison K.C., Hutahaean S.U.I., Tampubolon S.C., Panggabean T.M., Fitriyaningsih I., "An Empirical Evaluation of Phrase-based Statistical machine translation for Indonesia slang-word translator," (2022) Indonesian Journal for Electrical Engineering and Computer Science, 25 (3), pp. 1803-1813, DOI: 10.11591/ijeecs.v25.i3.pp1803-1813.

[14] Lee H., Cho H., Park J., Chae J., Kim J., "Cross Encoder- Decoder Transformer with Global-Local Visual Extractor for Medical Image Captioning," (2022) Sensors, 22 (4), art. no. 1429.

[15] Kumar A., Agarawal A., Ashin Shanly K.S., Das S., Harilal N., "Image Caption Generator using Siamese Graph Convolutional Networks and LSTM," (2022) ACM International Conference Proceeding Series, pp. 306-307.

[16] Mahalakshmi P., Fatima N.S, "Summarization of Text and Image Captioning in Information Retrieval Using Deep Learning Techniques," (2022) IEEE Access, 10, pp. 18289- 18297.