Image caption generator using LSTM model

Cherku Saikumar
700742475
dept.Computer Science
University of Central Missouri
sxc24750@ucmo.edu

Sai Vinay Reddy Pannala
700740027
dept.Computer Science
University of Central Missouri
sxp00270@ucmo.edu

Appalaswamy Yalamanchily
700747726
dept.Computer Science
University of Central Missouri
axy77260@ucmo.edu

Divya Pothuru
700746292
dept.Computer Science
University of Central Missouri
dxp62920@ucmo.edu

***Abstract*—Globally there are 2.2 billion people who are visually impaired. They need assistance for reading or checking the information.With the help of deep learning we extract the information from images and process and read it out to the people.Deep learning is a technique which processes the information in the images and translates it into useful information.Deep learning is a subset of machine learning and the models contain three or more layers which extract the information.Image captioning projects deal with understanding the information in the images like humans do, producing the text describing what is happening in the image.In this project we propose a neural network approach LSTM for text classification to generate the meaningful captions and VGG16 transfer learning model to process the images and extract features. In the end the performance of the test dataset is evaluated using accuracy and BLEU score to check the accuracy of the predictions. The experimental analysis is conducted on publicly available Flickr8k image dataset.**

[1] ***Index Terms*—BLEU (bilingual evaluation understudy),Deep learning, LSTM(Long ShortTerm Memory), VGG16 and Flickr8k dataset**

I. Introduction

There are various applications of Image caption generator in addition to assisting the visually impaired people this can also be embedded in the web pages in order to automate the few tasks and in the social media post caption generator. Image captioning is fascinating and at the same time a challenging task to perform.Main challenge here is to understand the image like the human brain.CNN is the best technique to understand the images but this can not support the sequence of words in the images.RNN or Recurrent Neural Networks which can provides the memory to understand the sequence of the words and the other version of the RNN is LSTM which is better version of RNN to provide the sequence of the words. In the recent years there are multiple proposals that can predict the captioning of the images but the models suffer with different issues like lack of accuracy and producing meaningful captions for a given image. Though the models achieved good accuracy results they are unsuccessful while producing the meaningful captions for the image.Generative Adversarial Networks overcome the problem with the auto encodes in built.These auto encoders learn in unsupervised learning approach and produce the good results. In this paper image captioning task is performed on the Flickr 8K dataset. In this process the project is divided into two tasks one is detecting the objects using computer vision tasks Regional Object Detector after finding the objects we establish the relationship between the objects using deep learning techniques RNN and LSTM networks.After understanding the semantics of the image it will produce the meaningful captions. In this paper we used the combination of CNN, a simple processing technique and RNN for text generation.To overcome the problem of meaningful caption generation attention network is used.This experiment is performed on the MSCOCO dataset which gave promising results in terms of caption generation and image processing.

The execution of the project is divided into the following tasks: 1.Image feature extraction:

- In this step a deep neural network is used to extract the features in this case we can use transfer learning also like VGG19 structures and DNN densenet.
- 2.Text generation:In this phase a sequence learning model is applied to generate the text LSTM contains input,output and forget gate which remembers the sequence.
- 3.BLEU(BiLingual Evaluation Understudy)) score:To measure the accuracy of the prediction with the original caption we use BLEU score.The score will be in between zero and 1 zero means the predictions are poor and 1 is ideal 100 perent accurate.

In this project we use following machine learning/deep learning libraries:

- Tensorflow
- Keras
- Numpy
- Matplotlib
- Seaborn

[1]https://github.com/Divya-ucm/ImageCaptionGenerator-700746292.git

- scikit-learn

Similar image retrieval is a process of identifying similar images. For any system similar image identification helps to compare the characteristics of the similar items. ICA(Independent Component Analysis) and PCA(Principal Component Analysis) are the best methods to reduce the dimensions and extract the similar features . As a new application image captions are extended to video level. Using Deep learning techniques and Natural Language Processing(NLP). To generate the captions for video sequences it has to look at every event in the video and generate the caption for them. BLEU score and METEOR are used to evaluate the performance of the model. Automated image caption generators with RNN have memory problems. Next word depends on the captions data not on the image. To solve this problem we have used RNN with modified LSTM. which predicts the captions based on the image. To evaluate performance of the predicted captions with ground truth BLEU score metric used.

II. Motivation

Globally there are 2.2 billion people who are visually impaired. They need assistance for reading or checking the information.With the help of deep learning we extract the information from images and process and read it out to the people.Deep learning is a technique which processes the information in the images and translates it into useful information.Deep learning is a subset of machine learning and the models contain three or more layers which extract the information.Image captioning projects deal with understanding the information in the images like humans do, producing the text describing what is happening in the image

III. Objectives

Objectives of this project are:

- Extracting the features of the image using pretrained model VGG16 and predicting the best caption for the image using LSTM model
- Main features of the project are: In the project we are deploying deep learning models to extract the features of the images. To extract the features from the images we use the VGG16 pretrained model. LSTM model to predict the caption for the given image
- Pertain models achieve optimal performance in less time this is because these models have already understood the features of the images.
- These patterns are saved as weights of the model and used to achieve the same performance without training the model again
- Datasets like this size will be trained in less time. LSTM models are used to predict the next sequence of the word with the help of different gates in the model.
- In the results analysis we use BLEU score to measure the performance of the model predictions
- In addition to this we perform cross validation on the data.For each fold accuracy is calculated and if the results are satisfactory training of the model is continued otherwise the model is fine tuned to achieve better accuracy

IV. Data Description

For the experimental analysis we have used the Flickr 8K dataset.Dataset contains two files one is captions file i.e. captions.txt and image files 8056 images. Each image is identified with image id. The captions file contains image ids and their respective captions. The dataset is collected from the Kaggle repository. All the images are in the format of jpg.

man on a bicycle riding on only one wheel .

asian man in orange hat is popping a wheelie on his bike .

a man on a bicycle is on only the back wheel .

a man is doing a wheelie on a mountain bike .

a man does a wheelie on his bicycle on the sidewalk .

Figure 1. 1 sample from dataset

five people are sitting together in the snow .

five children getting ready to sled .

a group of people sit in the snow overlooking a mountain scene .

a group of people sit atop a snowy mountain .

a group is sitting around a snowy crevasse .

Figure 2. 2 sample from dataset

V. Related work

For issues with image captioning, there is a method that has been around for a while. Prior to the widespread usage of transformers, it was possible to utilise convolutional neural networks (CNN) for feature extraction and recurrent neural networks (RNN) for text generation, particularly in Thai. However, these methods need to be improved. The end-to-end image captioning method proposed in this research uses Thai language models called ThaiTC and pretrained vision

Transformers (ViT) and text transformers. With the help of three Thai image captioning datasets, we tested our pre-trained Thai language vision and text transformer experiments. 1) Travel; 2) Food; and 3) Flickr 30k with various challenges. In the past ten years, there has been a substantial advancement in the field of image captioning. This genuinely makes it remarkable that the models created to do so are almost never taught to work on clear photographs and, as a result, they are unable to produce adequate captions on images of rain, especially images of severe rain. Modern breakthroughs in model construction that enable single picture de-raining have been made. However, little has been done to complete the task of captioning images of severe rain as a whole. The focus of this research is on creating an end-to-end architecture that can create appropriate captions for photos from the MS COCO dataset that have substantial rain noise added to them. For this, an end-to-end approach is suggested that employs a GAN-based architecture for deraining images [7] [6].

In today's social media world, practically everyone uses social platforms and actively engages in online communication. Social media users post a lot of images with various captions to their profiles. It takes time to consider the proper caption. A picture's caption must accurately convey the image's meaning and content. In the caption, concise sentences describe the picture. It is possible to create a model for an image caption generator that produces captions for photos of various sorts and resolutions. Using an image captioning model, captions for the input photos are produced in human-understandable language. Using the encoder-decoder principle, CNNs (convolution neural networks) and RNNs (recurrent neural networks) are used [8] [18].

The goal of the novel object captioning job is to create a full caption for any bounding boxes that are not visible in the training images using only those bounding boxes and the context of the image. The image-caption pair synthesis approach is trained to produce more pseudo-label since the training datasets for this task lack description with bounding boxes in the reference caption. We do, however, present an Image-Caption Pair Replacement Algorithm towards semi-supervised novel object captioning (I-CPRA) due to the issue of inaccurate pseudo-label in the current image-caption pair replacement method. Bounding box scaling technique and two-stage semantic graph structure are two of the submodules that make up I-CPRA. To specifically address the issue of the rigorous resolution and aspect ratio replacement condition between novel items in the picture replacement technique [12] [11].

For environmental preservation and land planning, multi-temporal remote sensing (RS) picture analysis of land cover changes is essential. In this research, we examine RS image change captioning (RSICC), a novel challenge designed to produce descriptions of the shifting land cover in multitemporal RS photos in human-like language. We suggest a brand-new RSICC (RSICCformer) model that is Transformer-based. There are three primary parts to it: A caption decoder creates phrases expressing the differences, a dual-branch Transformer encoder (DTE) improves feature discrimination for the alterations, and a CNN-based feature extractor creates high-level features of RS picture pairs. The DTE has a hierarchy of processing phases that it uses to record and identify various changes of interest. Particularly, we employ the bitemporal feature variations as keys [19] [1].

The community has recently become interested in remote sensing (RS) image captioning because it offers more semantic information than more conventional tasks like scene classification. The goal of image captioning is to create an organised, thorough description that captures the essence of an image. The description can be constructed via the encoder-decoder framework or it can be retrieved directly from the ground truth descriptions of related images (retrieval based image captioning). The disadvantage of the former is that it cannot provide new descriptions. The latter could be impacted by incorrect scene or semantic object identification [10] [2].

This study describes gaze-based picture captioning combined with human-centric image retrieval. Advanced picture retrieval has been made possible by the advent of cross-modal embedding techniques, although many systems have only paid attention to the data gleaned from the contents, such as the image and text. Building retrieval methods that specifically reflect human intentions is required to expand picture retrieval. In this research, we focus on the fact that the gaze information gathered from people contains semantic information and suggest a new retrieval strategy via image captioning based on gaze information. We develop a transformer, connect caption, and gaze trace (CGT) model specifically to learn the interaction between images, human-provided captioning, and gaze traces [5] [15].

While researchers are primarily focused on video-related tasks, video captioning is the more heuristic task of the integration of computer vision and natural language processing. Dense video captioning is still regarded as the more difficult assignment because it must take into account every event that takes place in the video and deliver the best captions separately for every event that is presented in the video with a high degree of diversity [4].

We provide a brand-new multi-modal strategy for captioning fashion images that enables users to specify a few semantic characteristics that direct caption production according to individual tastes. We purge, filter, and put together a new fashion picture caption dataset called FACAD170K from the existing FACAD dataset to aid in research and learning. On our constructed FACAD170K dataset, we test the effectiveness of the suggested approach. The results show that our method can perform better than both current models for captioning fashion images and traditional captioning techniques [16] [13].

The text for the image description is produced by conventional image caption algorithms based on the attributes of the image. The projected outcomes, however, frequently omit vital semantic and geographical information. This work suggests a multimodal data-integrated picture caption model to address the aforementioned issues. By include mouse trace and object bounding box as additional information sources,

the model is better able to capture crucial semantic information. The mouse traces are first segmented, converted to bounding boxes, and positioned next to a text word mark. We design a differentiable sampling operator to replace the conventional nondifferentiable sampling operation to index the multilabel classification result and leverage multilabel classification to give previous knowledge. Our method can execute joint training in contrast to earlier two-stage RSI captioning systems, and the joint loss enables the mistake of the generated description to flow into the optimisation of the multi-label classification via backpropagation. In particular, we implement a differentiable sampling operator for the multilabel classification by approximating the Heaviside step function with the steep logistic function [3] [17].

Two-stage RSI captioning systems incorporate an auxiliary remote sensing task to offer prior information, which enables them to produce more accurate descriptions when compared to RSI captioning methods based on the conventional encoder-decoder model. However, the image captioning and the auxiliary remote sensing tasks are performed separately in earlier two-stage RSI captioning systems, which is time-consuming and disregards cross-task interference. We suggest a brand-new joint-training two-stage (JTTS) RSI captioning technique to address this issue. Contrary to the current RS image captioning methods, the second phase uses sequence to sequence neural networks to combine the ground-truth captions of each training image into a single caption, thereby removing duplication from the training set. To combine the standard captions with the summarised captions based on the semantic content of the image, the third phase automatically defines the adaptive weights associated with each RS image. An innovative adaptive weighting approach for LSTM networks is used to accomplish this [9].

The interior characteristics of the items and the external relationships between various objects must be understood in order to accurately describe high spatial resolution remote sensing images. The algorithms used to create image captions currently do not have global representation capabilities, making them unsuitable for summarising complicated situations. In order to do this, we suggest a pure transformer (CapFormer) architecture for captioning remote sensing images. In particular, a scalable vision transformer is used for picture representation, where multi-head self-attention layers can be used to capture the global content. The visual features are gradually translated into complete phrases using a transformer decoder [20].

The process of captioning a picture produces subtitles that are human-like. Automatic picture captioning has emerged as a new area of study and has recently attracted increased interest. Although this is a difficult challenge, a number of ways have been put forth where deep learning techniques have shown to be state-of-the-art in handling these problems for picture captioning. Convolutional neural network (CNN) and recurrent neural network (RNN) variations are used in the majority of techniques [14].

VI. Proposed Framework

The implementation is mainly divided into main two main parts: Image feature extraction and cleaning and Text cleaning feature extraction

A. Image feature extraction:

The VGG16 model is imported from Keras library and imageNet weights are loaded onto the model. In the preprocessing step of images all the images are converted into compatible input shape of 224/224 to VGG16. VGG16 model is applied to extract the features from the image and the extracted image features are flattened and stored as 4096 array format.

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

Total params: 134,260,544
Trainable params: 134,260,544
Non-trainable params: 0

Figure 3. VGG16 model summary

B. Principal Component Analysis:

Principal Component Analysis is used to form the clusters of the similar images. And the PCA is used to reduce the dimensions of the features of the images to 2 dimensions. PCA technique is implemented using scikit-learn library.

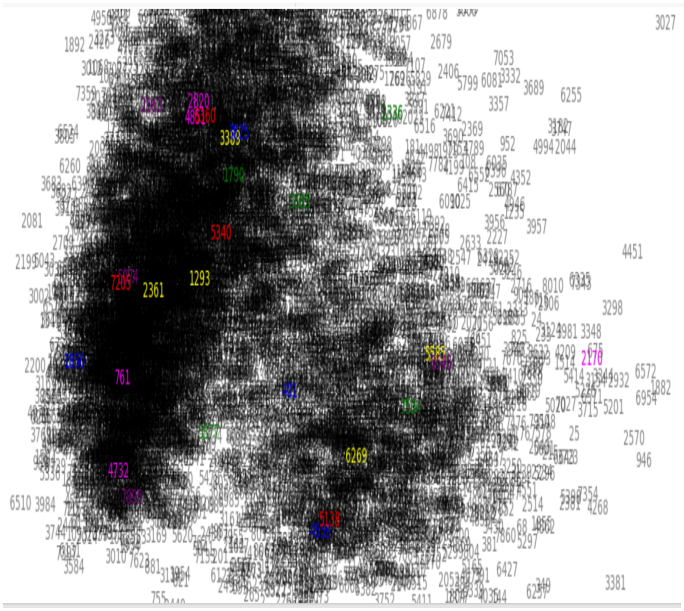

Figure 4. PCA of word embeddings

C. Captions data cleaning and analysis:

Tokenizing the text into vectors.For this task NLTK tokenizer is imported. Since LSTM takes the same length of input captions while training. To make the captions equal length minimum and maximum length of the captions are analyzed and using pad sequences all the captions arrays are padded to make the same length.

D. Data splitting:

After the preprocessing data is split into training and testing.

E. LSTM model training:

The LSTM model is trained with the preprocessed captions. After the training captions are predicted using a test set.'

F. Model evaluation:

Model is evaluated using the BLUE score to check the performance of the predicted score. Higher BLEU score means the predicted caption is similar to the ground truth images.

VII. Methods

A. VGG16

- VGG16 contains 16 layers with weights.
- It accepts input tensor 224/224 with 3 channels
- It contains 13 convolutional layers, 5 max pool layers and 3 dense layers
- Its hyper parameter tuning is less compared to other transfer learning models

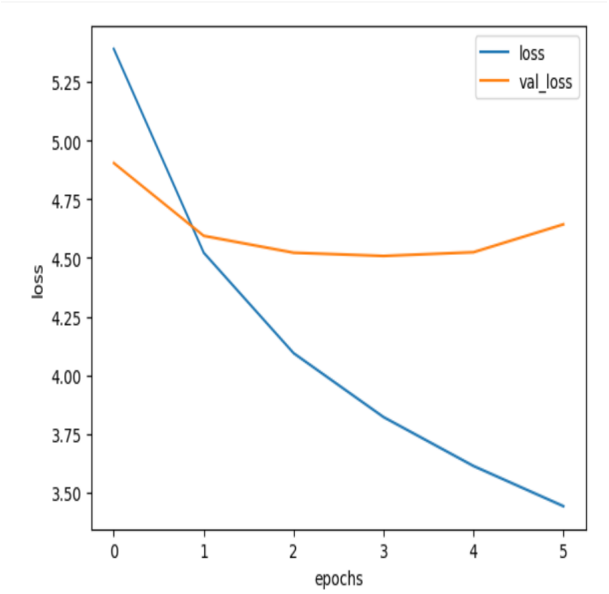

Figure 5. LSTM model accuracy

B. LSTM Model:

- LSTM is a variant of RNN architecture which remembers the previous information.
- LSTM follows two mechanisms: It forgets the information which is not necessary Save information which saves the information of necessary parts of the text
- LSTM mainly consists of 4 parts: Forget gate gets active when it receives unnecessary information Use gate predicts the output from STM and LTM and the other gates are learn gate and remember gate.

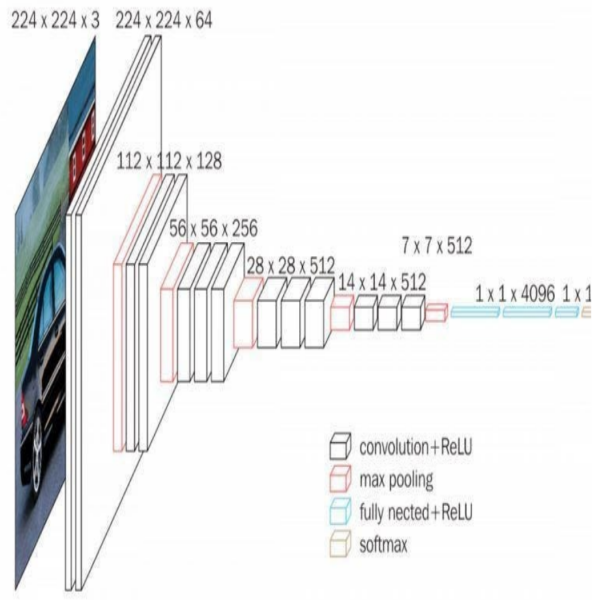

Figure 6. Architecture of VGG16

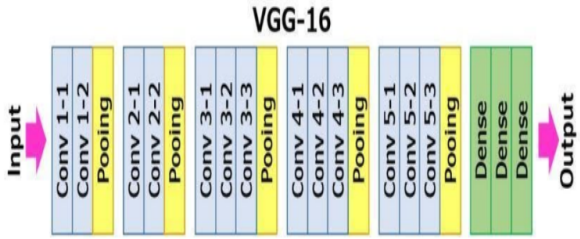

Figure 7. Layers of VGG16

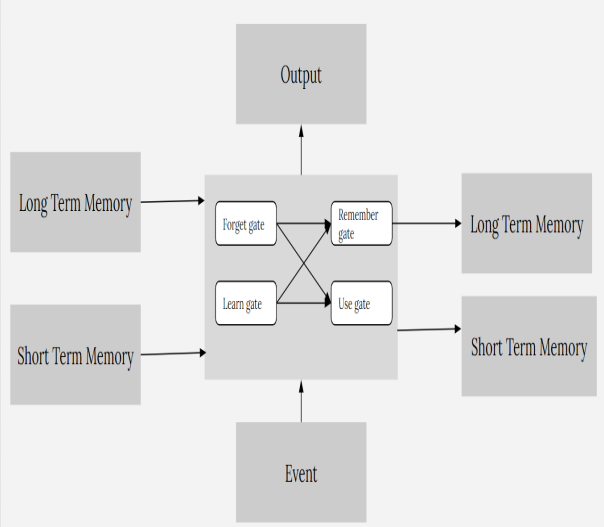

Figure 8. LSTM Block diagram

VIII. Exploratory Data Analysis

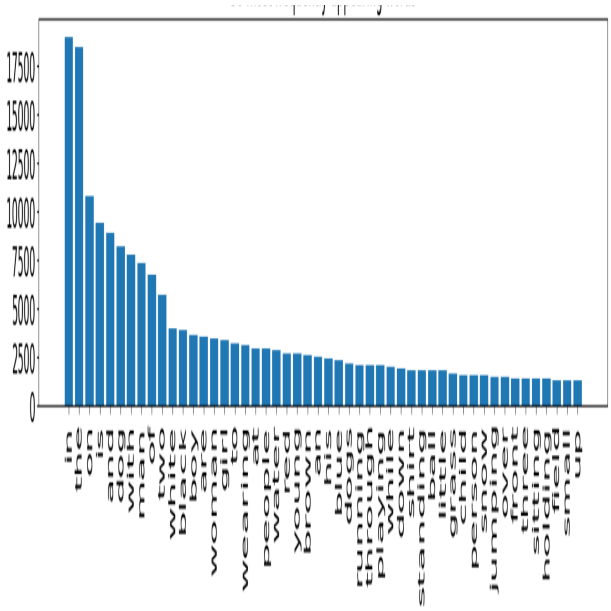

Figure 9. 50 most frequently used words

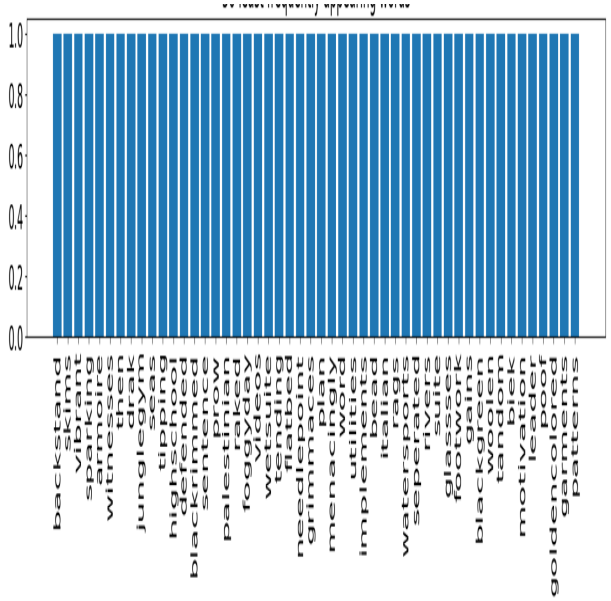

Figure 10. 50 least frequently used words

startseq man in black shirt is standing on the street endseq

Figure 11. Predictions sample2

IX. Results Summary

The model has successfully been taught to produce the expected captions for the images. By fine-tuning the model with various hyperparameters, the caption generation has continuously improved. With a higher BLEU score, the produced captions are more likely to match the real captions on the photos. The validation loss decreases up until the fifth epoch and then increases, while the training loss keeps decreasing. The following were the main findings and observations from the process of training the model and testing it against test data:

Even while the training loss diminishes over time, the validation loss typically increases after the fifth epoch. This suggests that the model is too well-fitted and that training should be stopped. Score doesn't always equate to better captions being produced. If the model overfits on your training data, it will cause the model to examine image details and produce captions that are illogical. The strong and weak captions produced above demonstrate this.

startseq black dog is running in the water endseq

Figure 12. Predictions sample3

startseq man in red shirt is jumping through the snow endseq

Figure 13. Predictions sample 4

true: child in pink dress is climbing up set of stairs in an entry way

pred: boy in red shirt is sitting on the street

BLEU: 7.176794039009363e-232

Figure 14. Predictions sample 5

true: black dog and spotted dog are fighting

pred: black and white dog is running through the snow

BLEU: 1.384292958842266e-231

Figure 15. Predictions sample 6

true: little girl covered in paint sits in front of painted rainbow with her hands in bowl

pred: girl in red shirt is sitting on the street

BLEU: 5.746727420065187e-232

Figure 16. Predictions sample 7

true: man lays on bench while his dog sits by him

pred: black dog is playing in the snow

BLEU: 7.296382734947757e-232

Figure 17. Predictions sample 8

REFERENCES

[1] Mohammad Alsharid, Rasheed El-Bouri, Harshita Sharma, Lior Drukker, Aris T. Papageorghiou, and J. Alison Noble. A course-focused dual curriculum for image captioning. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 716–720, 2021.

[2] Djamila Beddiar, Mourad Oussalah, and Seppänen Tapio. Explainability for medical image captioning. In *2022 Eleventh International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2022.

[3] Chen Cai, Kim-Hui Yap, and Suchen Wang. Attribute conditioned fashion image captioning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1921–1925, 2022.

[4] Yuhu Feng, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. Human-centric image retrieval with gaze-based image captioning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3828–3832, 2022.

[5] Genc Hoxha, Farid Melgani, and Jacopo Slaghenauffi. A new cnn-rnn framework for remote sensing image captioning. In *2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS)*, pages 1–4, 2020.

[6] Jia-Hong Huang, Ting-Wei Wu, Chao-Han Huck Yang, and Marcel Worring. Deep context-encoding network for retinal image captioning. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3762–3766, 2021.

[7] Teetouch Jaknamon and Sanparith Marukatat. Thaitc:thai transformer-based image captioning. In *2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–4, 2022.

[8] Apoorva Krisna, Anil Singh Parihar, Aritra Das, and Arnav Aryan. End-to-end model for heavy rain image captioning. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, pages 1646–1651, 2022.

[9] Nan Lin, Shuo Guo, and Lipeng Xie. Tb-transformer: Integrating mouse trace with object bounding-box for image caption. In *2022 IEEE 24th Int Conf on High Performance Computing Communications; 8th Int Conf on Data Science Systems; 20th Int Conf on Smart City; 8th Int Conf on Dependability in Sensor, Cloud Big Data Systems Application (HPCC/DSS/SmartCity/DependSys)*, pages 810–814, 2022.

[10] Chenyang Liu, Rui Zhao, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022.

[11] Ruchika Malhotra, Tanmay Raj, and Vedika Gupta. Image captioning and identification of dangerous situations using transfer learning. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 909–915, 2022.

[12] Prashant Giridhar Shambharkar, Priyanka Kumari, Pratik Yadav, and Rajat Kumar. Generating caption for image using beam search and analyzation with unsupervised image captioning algorithm. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 857–864, 2021.

[13] Shamima Sukhi, Abu Quwsar Ohi, Md Saifur Rahman, and M.F. Mridha. A survey on bengali image captioning: Architectures, challenges, and directions. In *2021 International Conference on Science Contemporary Technologies (ICSCT)*, pages 1–5, 2021.

[14] Gencer Sumbul, Sonali Nayak, and Begüm Demir. Sd-rsic: Summarization-driven deep remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 59(8):6922–6934, 2021.

[15] Saya Takada, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Generation of viewed image captions from human brain activity via unsupervised text latent space. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2521–2525, 2020.

[16] J Vaishnavi and V Narmatha. Video captioning based on image captioning as subsidiary content. In *2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–6, 2022.

[17] Junjue Wang, Zihang Chen, Ailong Ma, and Yanfei Zhong. Capformer: Pure transformer for remote sensing image caption. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 7996–7999, 2022.

[18] Xinru Wei, Yonggang Qi, Jun Liu, and Fang Liu. Image retrieval by dense caption reasoning. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4, 2017.

[19] Yang Yang. Image-caption pair replacement algorithm towards semi-supervised novel object captioning. In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, pages 266–273, 2022.

[20] Xiutiao Ye, Shuang Wang, Yu Gu, Jihui Wang, Ruixuan Wang, Biao Hou, Fausto Giunchiglia, and Licheng Jiao. A joint-training two-stage method for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.