Multi-label Classification, Multi-modal Classification and Image Description

A DL LAB Project Report Submitted
in Partial Fulfillment of the Requirements
for the completion of

Deep Learning Lab course

by

BASU VERMA
(142002007)

under the guidance of

Dr Mrinal Kanti Das

IIT PALAKKAD

INDIAN INSTITUTE OF TECHNOLOGY PALAKKAD
PALAKKAD - 678557, KERALA

Acknowledgement

I would like to thank Dr. Mrinal Kanti Das for providing us the opportunity to do a project to get a practical exposure to what we learnt in the Deep Learning course, we thank Ms. Shikha Mallick, Ms. Neha A S and Mr. Rimmon Saloman Bhosale for their valuable tutorials,feedback and support. We thank each other for our mutual understanding and support, it has been a great experience.

May 18, 2021
IIT Palakkad

Basu Verma

Abstract

Now a Days, Image Classification has several application in our daily lives. In this project we did multi-label object classification from the images given and it's description given in 50 lines into 20 predefined labels. Multi-label classification involves predicting zero or more class labels. Unlike normal classification tasks where class labels are mutually exclusive, multi-label classification requires specialized deep learning algorithms that support predicting multiple mutually non-exclusive classes or "labels." Also in this project a model is made and trained on a given data so as to produce textual description when fed an image data. Pretrained ImageNet and VGG16 models as well custom made models are used to achieve the desired results.

Contents

List of Figures

Part. 1

Introduction

1.1 Motivation

We need to build a model for Multilabel image Classification, Multimodal Image Classification and Image Description. Provided data was Pascal50s and abstract50S and we needed to classify the image into given set of multilabels and also to generate a sentence describing about the given image.

1.2 Data description

Data Provided has two files namely Pascal50S.mat and Abstract50S.mat . mat format file is a matlab file format which has struct type data structure same as Dictionary in python . Pascal50s consisted of 1000 images links followed by 50 descriptions for each images.Image link in Pascal correspond to real life images. Whereas Abstract Dataset consisted of 500 images link followed by 50 description lines for each image.And images in Abstract were abstract cartoon images. For use in project images were extracted from links and stored in .npy format and corresponding descriptions were also extracted and stored separately to perform further operations.

1.3 Tools and Technologies

Programming languages	Python
Models Used	Sequential Model, VGG16, MobileNetSSD ResNet50
Other Tools	Jupyter Notebook,Colab Notebook, VS-Code

Table 1.1

Part. 2

Literature Survey

[1] In this paper, they talk about experience in collecting the linguistics data at a relatively low cost and high speed to create corpora of images annotated with multiple one-sentence descriptions.Turker's were divided into participant with qualification test and without qualification test and they were given offers to write about the image in free-form text entry. Using this, they created two corpora totalling more than 40,000 descriptive caption for 9000 images. [2] In this paper, they talk about the five state-of-the-art image description approaches using the new protocol and provide a benchmark for future comparisons.They also talk about the method of generating 50 lines sentences from the original 5 lines sentences of each images using the combination of convolutional neural network and recurrent neural network.

Part. 3

Methodology

In this section we are going to discuss in detail about what testings we have done and information about which model we are using.

3.1 Data Preprocessing

In data preprocessing following steps were done:

- Images were extracted from the links for both the data sets and stored in .npy formats
- Description of all the images were extracted and were stored in lists for corresponding datasets.
- Then for each extracted description Punctuations and stop words were removed .
- From all the above processed data top 10 words with maximum occurrences were taken.
- And these words were matched with given label classes and common words were assigned as labels.
- But this method was not giving very accurate results so a PreTrained Model MobileNet SSD v3 by google was taken which is trained on millions of images .
- So labels were also extracted from images and and these extracted labels were added with labels list extracted from text description.
- Also to increase the accuracy of words matching with labels spacy module was used as it pretrained on thousands of wikipedia articles, journals to understand similarity between words.Such as it gives high similarity between airplane and aircraft . also between car and four wheeler
- This new list is now matched with labels list using spacy module and words with high similarity i.e, above 69% were assigned as labels.

- One special case for labels was also considered that if there is any mention of word tv or monitor in description list after removing punctions and stop words than in the list of labels tv/monitor was also added.It is done as label tv/monitor is combination of two labels and having occurrence of this in label description is very rare like this also in some labels extracted using pretrained model tv or monitor was missing.

3.2 Multilabel and Multimodal Image Classification

In this task, we did multi-label classification and multi-modal classification.

3.2.1 Difference between multilabel and multiclass classification

- **Multiclass Classification:**

  In Simpler term Multiclass classification means that there is more than two mutually exclusive classes for classification of our data sets into and our data can belong to any one of these classes.

- **Multilabel classification:**

  In multi label classification there are many classes in which our dataset can be classified into and our data can belong to more than one class thus having multiple labels thus termed as multilabel classification.For eg. Images given to us can have multiple objects in it like aeroplance, man, car and cat. So it can have labels of corresponding object and thus multiple labels it can belong to.

In multi-label classification, what we basically did was that we made a sequential model using four convolutional layer and each convolutional layer consist of one batchnormalization layer, one maxpooling layer and one dropout model.

Then we trained the model using the images and labels extracted earlier and found the accuracy to be low, approx 21.12%. So, we moved to some pretrained model.

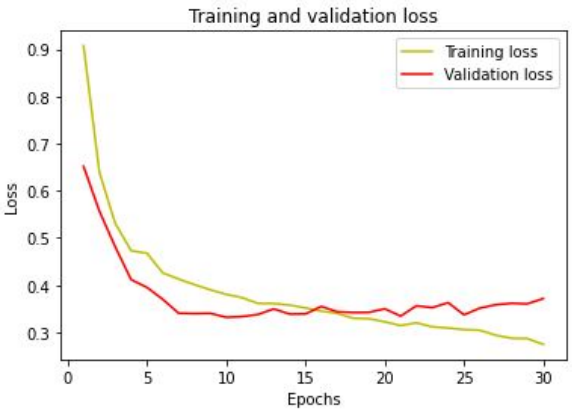


Fig. 3.1 Training and Validation Loss

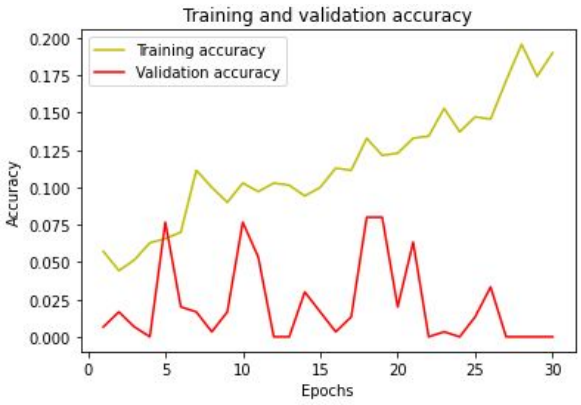


Fig. 3.2 Training and Validation accuracy

Then we used the pretrained model VGG16 for the image classification model and trained on the given input image data and the labels and found the accuracy better than the last model. The accuracy of the model found to be 56.87%.

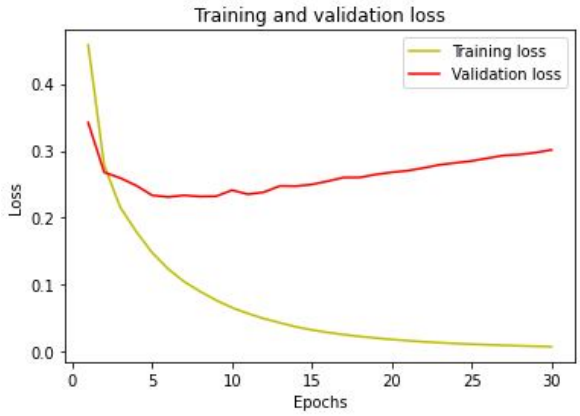


Fig. 3.3 Training and Validation Loss

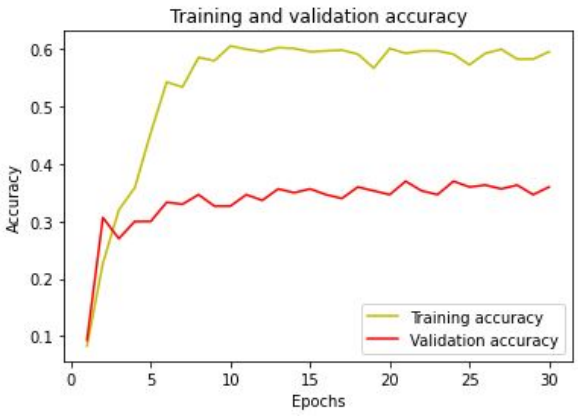


Fig. 3.4 Training and Validation accuracy

Since the accuracy of VGG model is quite bettter than previous model, so we used VGG model for further image classification and predicted on the test image.

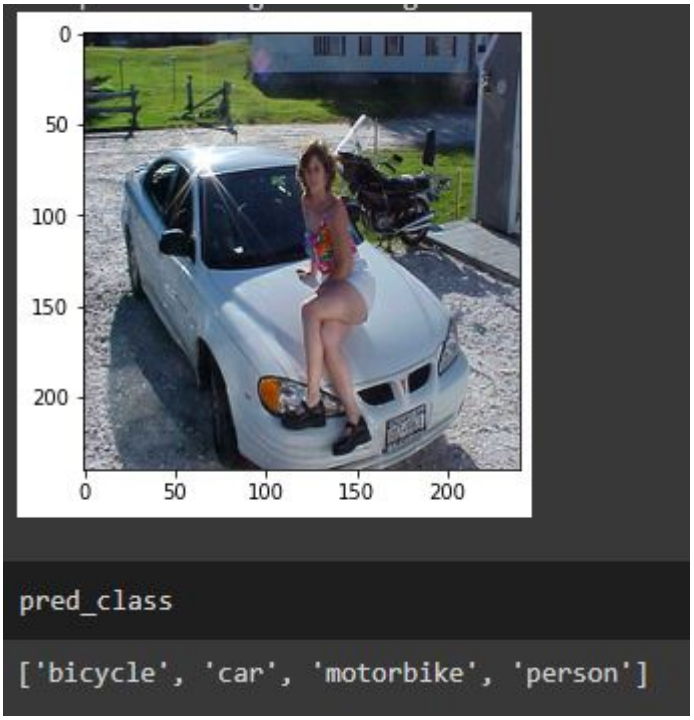


Fig. 3.5 Image Classification with the predicted labels

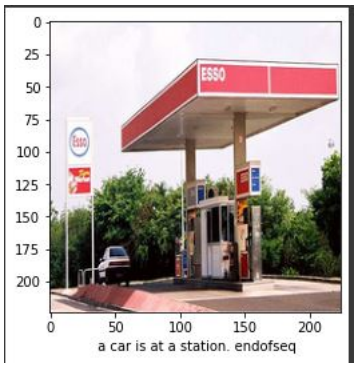
3.3 Image Description

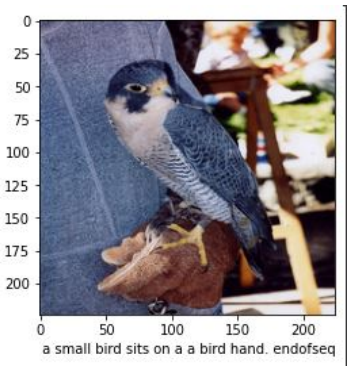
In this task, we had to made a model that can generate a text for the given image.

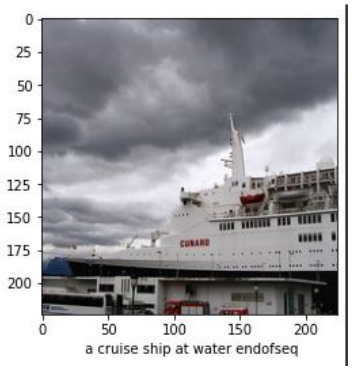
For this, we had made the model of convolutional and recurrent neural network **(CNN + LSTM)**. First we extracted the features from the image using the inbuilt CNN model **ResNet50**. After that we converted the description of each images into list of vectors. For this, first we made the vocabulary of total words available and according to that vocabulary, we converted the description into vectors. Also we added **"startofseq"** and **"endofseq"** at the starting and ending of each sentence, so that the model can understand what is the starting and ending of that sentences.

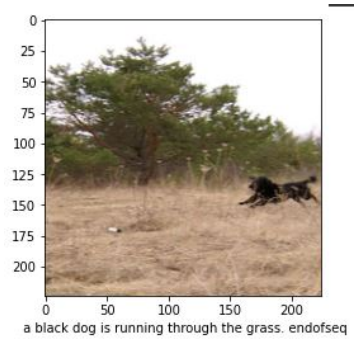
After that we made a generator function which will convert the given image data and the corresponding label description into the input format for LSTM model. It takes the all the input images and corresponding description and gives output as "inseq" and "outseq". "inseq" is the list of a single word and outseq is the list of word just after the inseq word. After that the next word will get added to the inseq list and outseq will consists of word just after the inseq word.This way all the inputs will get transformed into the new inputs for LSTM model. The length of new inputs and outputs is 54,185.

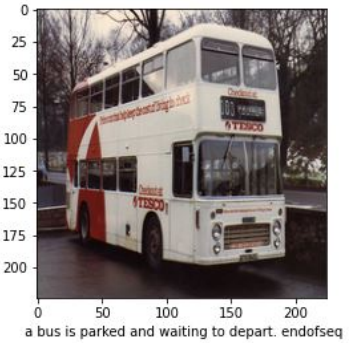
After that we gave that to the combined model of sequential and LSTM and train the model for that inputs. The accuracy comes around 75%. And then we generated the ouput for some images.

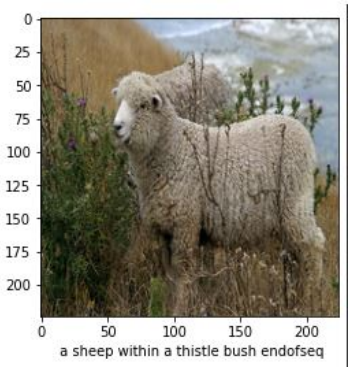












Part. 4

Conclusion

4.1 Conclusion

The techniques which we have learnt in our Deep learning class like convolutional neural network and recurrent neural network are used here in making the model. We have build this model with requirements given with scope for future updations.

4.2 Future-work

1. We want to increase the scope of this project with some user interface
2. We are willing to implement the multilabel classification using some other model to increase the accuracy further.

4.3 My Part

In this project, I did the mult-label and multi-modal classification after the extraction of the labels. And also, the Image Description part.

References

[1] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 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*, pages 139–147, Los Angeles, June 2010. Association for Computational Linguistics.

[2] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015.