Deep Learning Project

Instructor:- Dr. Mrinal Kanti Das

Mentor:- Neha A S

Team 1:- Subhajit Karmakar (111801041) and Vishal Rao (111801046)

Objectives of the Project

- Our project is made up of three tasks.
- First task is based on multi label classification of images in PASCAL 50S dataset in 20 classes.
- Second task is based on multi-modal image classification combining data from images and texts.
- Third task is based on generating description for the images.

TASK 1

Problem Statement

- We are given PASCAL 50S dataset consisting of 1000 images each belonging to one of the 20 classes and having 50 descriptions.
- Our aim is to find the labels of the images by scraping and then use the scraped labels and images to perform multi label image classification.

Data Preprocessing

- We extracted the image links and descriptions from the Pascal50s.mat file
- Obtained images from the links using requests and Image library functions

```
url = dataPascal['train_sent_final'][0][63][0][0]
im = Image.open(requests.get(url, stream=True).raw)
im
```

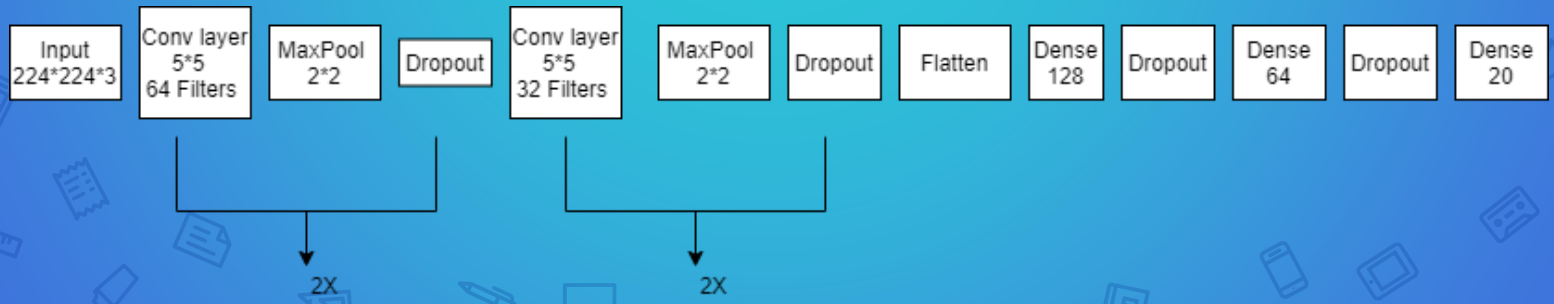
- Resized each dimension of each image to a fixed size

Data Preprocessing

- Built a dictionary with the 20 labels as the keys. Each label has a list of synonyms, obtained from scraping WordNet.
- Converted all image descriptions to lower case, removed punctuations, numbers.
- Lemmatized each word using WordNetLemmatizer function
- Scanned through each word of the description and checked whether that is a synonym of any label, if so we set as 1 at index for that corresponding label in an array of size 20.

Model Architecture

- Our model takes in input of shape 224*224*3.
- Then there are 2 sets of Convolutional Layer of size(5,5),no of filters as 64 and with relu activation function, Max Pool layer of size(2,2) and Dropout layer.
- Then there are 2 sets of Convolutional Layer of size(5,5),no of filters as 32 and with relu activation function, Max Pool layer of size(2,2) and Dropout layer.
- These convolution layers are followed by a flatten layer, a Dense layer of size 128 with relu activation, a drop out layer, a Dense layer of size 64 with relu activation and a dropout layer.
- Finally there is a dense layer of size 20, with sigmoid activation function.



Results

- We splitted our dataset in 8:2 train validation split.
- After training our model on 80% of image data as input and labels as target, we obtain a train accuracy of about 45%
- Precision is 0.52 and recall is 0.22

-

Output

- We used our model to predict the labels for the below image.

- The resultant label obtained from the image.

```
[35] # this is the predicted label

     print("The predicted labels in this image:- ")
     for i in getPredictedLabels(result):
       print(i)

     The predicted labels in this image:-
     person
     dining table

[36] # this is the actual label

     print("The actual labels in this image:- ")
     for i in getLabels(y_test[66]):
       print(i)

     The actual labels in this image:-
     person
     dining table
```

TASK 2

Problem Statement

- We need to use image data and text data and perform multi-modal classification of the images present in PASCAL 50S dataset into the 20 classes.

Data Preprocessing

- Converted all image descriptions to lower case, removed punctuations, numbers.
- Lemmatized each word using WordNetLemmatizer function
- Scanned through few of the descriptions and used Tokenizer() function to convert the texts to encoded integers, used pad_sequences() function to make the encoded array of fixed size.

Text feature extraction

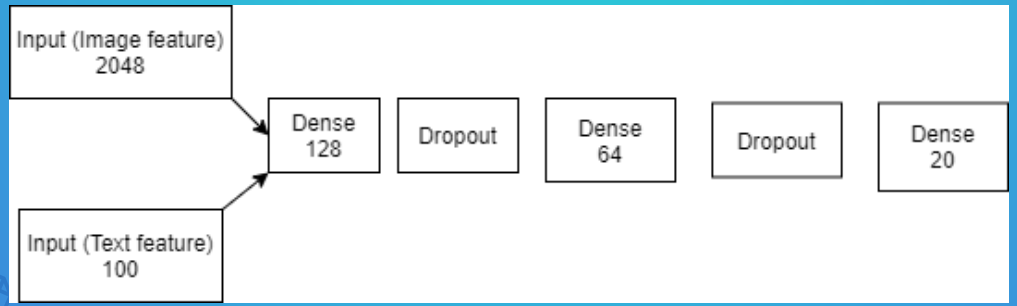
- We trained a simple RNN model on the encoded descriptions and label data.
- We removed the last layer of this model and use this new model to extract features from descriptions.
- This model helps us extract features from text data and gives us an array of features of length 100 for each image

Image Features Extraction

- We used a pretrained model called ResNet .
- We removed the last layer of the ResNet model and passed our image into it, the second last layer returned image features.
- This model helps us extract features from image data and gives us an array of features of length 2048 for each image

Model combining text features and image features

-
  - For each image we combine the text features and image features obtained from the previous models.
  - We passed these combined representation of text and image through a FNN with a final layer having length 20 and a sigmoid function
  - Then according to the output of this sigmoid function we get the predicted labels.
  - The architecture of the model combining image and text features are as follows



Results

- We splitted our dataset in 8:2 train validation split.
- After training our model on 80% of image data as input and labels as target, we obtained a train accuracy of about 47%
- Precision is 0.72 and recall is 0.40

-

Output

- We used our model to predict the labels for the below image.

```
[149] url = dataPascal['train_sent_final'][0][153][0][0]
      im = Image.open(requests.get(url, stream=True).raw)
      im
```

- The resultant label obtained from the image.

```
[146] # the predicted label for the image

      print("The predicted labels in this image:- ")
      for i in getPredictedLabels(result):
        print(i, end=", ")

      The predicted labels in this image:-
      person, car,

[148] # the actual label for the image

      print("The actual labels in this image:- ")
      for i in getLabels(y_test[153]):
        print(i, end=", ")

      The actual labels in this image:-
      person, car,
```

TASK 3

Problem Statement

- We need to build a model for generating image description using combination of RNN and CNN.

Data Preprocessing

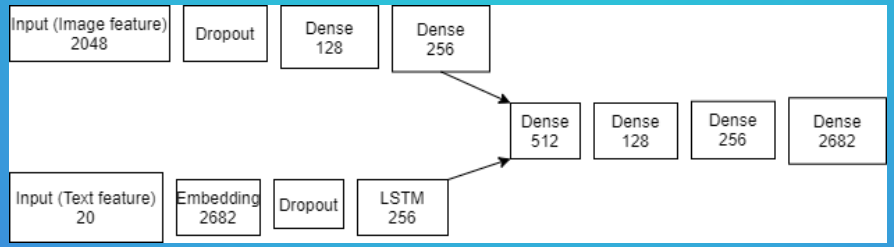
- Converted all image descriptions to lower case, removed punctuation, numbers and added a start word to the beginning of the sentence and added a stop word to the end of the sentence.
- Lemmatized each word using WordNetLemmatizer function
- Scanned through few of the descriptions and used Tokenizer() function to convert the texts to encoded integers, used pad_sequences() function to make the encoded array of fixed size.

Generating Dataset for training

- Took a pretrained ResNet model, removed the last layer and passed each image into the model to obtain image features.
- Scanned through descriptions, converted the words into encodings, splitted each description at different index.
- The image and the part of the description towards the left of the split is treated as input for training, and the next word after the split is treated as output.

Model combining text features and image features

- For each image, we pass the extracted features through a FNN which consists of two dense layers of size 128 and 256.
- For each image, we pass the text encodings corresponding to every image through a RNN, which has embedding layer of size 2682(the no of different words present in all of our descriptions), then it is followed by a dropout layer and a LSTM of size 256.
- Then the resultant output obtained from two models are feed into a FNN having 4 dense layers of size 512, 128, 256 each having relu activation.
- The final layer has 2682 nodes with softmax activation, which gives us a single word out of the 2682 nodes present in our descriptions.
- The architecture of the model is as follows-



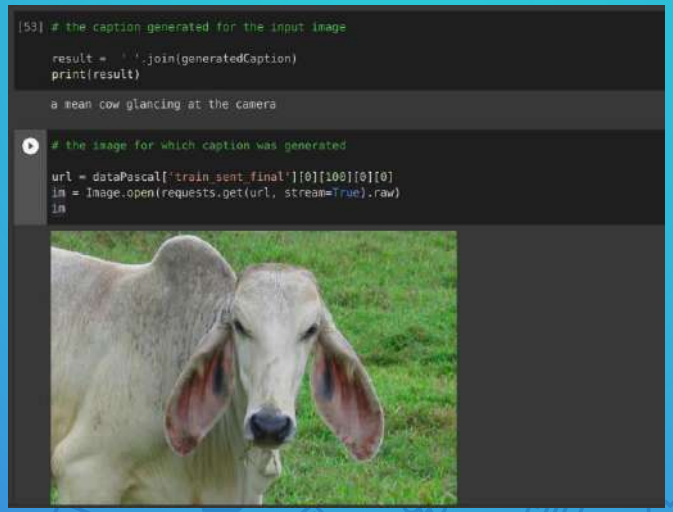
Generating description

- We take a image and use the modified ResNet model to extract its features.
- Then we take an input array with only the start word
- Then we use the image feature and encoded form of input array to predict the next word.
- We append the sampled word to the end of input array and continue it till the length of generated description is 20 or the sampled word is stop word.

-

Output

- We used our model to generate a description for a image. The caption and the image are as follows:-



Learning Outcomes

- We learnt a lot about word scraping, lemmatization and word preprocessing techniques.
- We learnt a lot about designing neural network models and optimizing them by trying various combinations of hyperparamaters.
- We learnt about doing multi-modal image classifications.
- We learnt about converting text data to vector form using various techniques.
- We learnt about manipulating pre trained models like VGG16 and ResNet and using the modified models to extract high level features of the image.
-