Assignment 4

Machine Learning

Yuyutsu Saini
2020MT60571
Rishi Jain
2020CS10373

28^{th} November, 2022

COL 774: Machine Learning

Kaggle Team Name: Mission Possimpible

Libraries Used

- numpy
- nltk
- sys
- pandas
- os
- re
- time
- torch
- torchtext
- transformers
- tqdm
- copy
- requests
- random
- sklearn
- csv

Contents

1 Non-Competitive Part

1.1 Convolutional Neural Network

1.1.1 Architecture

We implemented Convolutional Neural Network on the dataset of book covers with the following parameters:

- CONV1: Kernel Size → 5x5, Input Size → 3, Output Size → 32
- POOL1 : Kernel Size → 2x2
- CONV2 : Kernel Size → 5x5, Input Size → 32, Output Size → 64
- POOL2 : Kernel Size → 2x2
- CONV3 : Kernel Size → 5x5, Input Size → 64, Output Size → 128
- POOL3 : Kernel Size → 2x2
- Fully Connected Layer with 128 outputs
- Fully Connected Layer with 30 outputs

We have used ReLU as the activation function for all layers apart from the Pooling layers and used cross entropy loss as loss function.

Batch Size used in Stochastic Descent = 100
Number of epochs = 10
Time taken to train the model = 16m 4s
Accuracy over the Test set is = 12.6 %

1.2 Recurrent Neural Network

1.2.1 Architecture

Note : - Download glove.6B.300d before running the code. Place it in input directory. A bidirectional RNN was implemented with the following architecture using the *Pytorch* library.

- Embedding Layer : Initialized with the vocabulary vectors from the pretrained *GloVe* embedding(`glove_6b_300d`).
- RNN layer : Hidden layer size → 128, bidirectional → True, batchfirst → True
- MLP layers
  - FC1 : Fully Connected Layer with output 128
  - FC2 : Fully Connected Layer with input 128 and output 30 (number of classes)

Batch Size used in Stochastic Descent = 100
Number of epochs = 20
Time taken to train the model = 3m 32s
Accuracy over the Test set is = 45.42 %

Further We have experimented with removing stop words and stemmed the words. The best accuracy that we got with RNN is 53 %.
Number of epochs = 40
Time taken to train the model = 7m 31s
Accuracy over the Test set is = 53.20 %

2 Competitive Part

In this part, we are only using the name of the book to predict the genre of the book. Convolution Neural Network overfits the training data set. We experimented with resnet, vgg, squeeznet, alexnet, densenet and got the accuracy over the CNN was very less and we were able to take the CNN accuracy to slightly less than 29 percent using vgg. So, only RNN was used for training purposes.

We are using BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding by Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. It's a bidirectional transformer pretrained using a combination of masked language modeling objective and next sentence prediction on a large corpus comprising the Toronto Book Corpus for predicting the genre over 30 classes.

First, we removed stopwords using NLTK and then tockenized the book names to get attention mask and special tokens using a BERT pre-trained tokenizer.We have used stochastic gradient descent to optimize the model with batch size 16.

Accuracy over the non competitive test data set = 62.52 %
Time Taken = 35.32 minutes
Epochs =5