CSE5DL Assignment Report

This report chronicles answers to questions raised while completing the CSE5DL assignment.

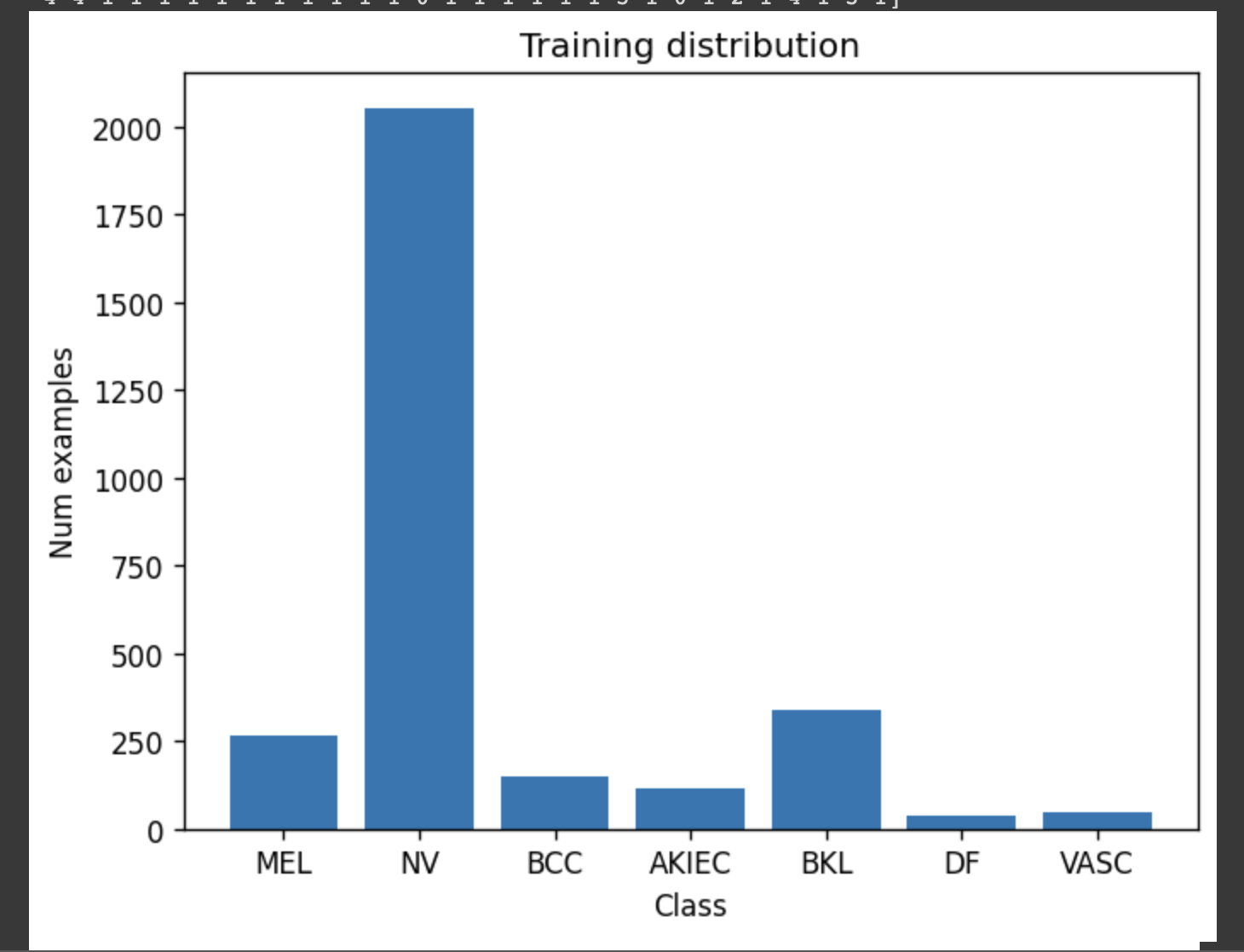
# Task 1

## Task 1a

### Data issues

After plotting the class distribution of the training and validation set, the thing that I noticed is,

Both datasets have a lot of examples for the “NV” class and a lot lesser for other classes. This means that the data is unbalanced and can result in poor performance for underrepresented data.



...

## Task 1b

### Why not use random\_split?

It is vital to make sure that the class distribution is maintained in both the training and validation sets when dealing with data challenges like class imbalance. Without considering the class distribution, employing random\_split alone could produce an unbalanced validation set, which would bias assessment metrics and produce inaccurate performance estimation. It is important to manually split the dataset while making sure that the class distribution is maintained in both the training and validation sets to accommodate for data concerns like class imbalance.

## Task 1c

### Reduce epoch time

When dealing with large training sets and long training times, it can indeed be challenging to debug code due to the extended duration of each epoch. However, there are two ways to significantly reduce the time per epoch for debugging purposes while still using real data and the real training code:

**Use a Subset of Data**: Instead of using the entire training dataset, you can create a subset of the data containing a smaller number of samples.

**Use Smaller Model Architectures**: Another approach is to use smaller model architectures that have fewer parameters and computational requirements. Large models with a high number of parameters tend to have longer training times.

…

### Confusion matrix

…

## Task 1d

### Account for data issues

1. Class Weighting: The class weights were calculated based on the frequency of each class in the training dataset. The class weights were then used in the loss function during training to give higher importance to underrepresented classes. This helps address the issue of class imbalance and ensures that the model pays more attention to minority classes.

The effectiveness of these strategies can be evaluated through several means:

Confusion Matrix: Analyzing the confusion matrix can provide insights into how well the model is able to classify different classes.

…

## Task 1e

### Vertical Flips

In the case of this dataset, it has specific orientation or anatomical features, random vertical flips may introduce unrealistic transformations that are not representative of the real-world data. For example, if the images in the dataset have a specific orientation or symmetry that should not be flipped vertically, applying random vertical flips could introduce inconsistencies and distort the data.

Therefore it is important to be careful with the data augmentation technique to be used.

…

### Effect of Augmentation

...

### [Challenge] 5 crop augmentation

…

## Task 1f

### Experiments

* <See attached excel document>
* <See attached Weights and Bias Report>
* Write a discussion about the key findings from the experimental results.

…

### [Challenge] Batch size

…

# Task 2

## Task 2a

### Data issues

1. **Data Type Issues**: Checked using **train\_df.info()**
   1. This checks for any inconsistencies or unexpected data types in the columns of the **train\_df** DataFrame, such as missing values or incorrect data types.
   2. If there are any issues with data types, it will be reported in the output.
2. **Statistical Summary**: Checked using **train\_df.describe().**

This provides basic statistical information about the columns in the **train\_df** DataFrame, including count, mean, standard deviation, minimum, and maximum values.

...

## Task 2b

### Similar embeddings

According to the visualization Business and SciTech are the two most similar classes.

I can tell because they are overlapping and there is a clear distinction between the other classes.

I wasn’t expecting this because I would expect Business news to be related with stock markets, finance, profits and losses and Science to be related with research and new findings, technology etc.

A picture containing screenshot, colorfulness

Description automatically generated

…

## Task 2c

### Saved model weights

1. **Base Model Architecture**: The DistilBERT model itself consists of a deep neural network with multiple layers. The base architecture of DistilBERT contains multiple transformer layers, which contribute to the overall size of the model.
2. **Pretrained Weights**: When fine-tuning a pretrained DistilBERT model, the initial weights of the model are loaded from a pretrained checkpoint. These pretrained weights are usually large because they contain the knowledge learned from a large corpus of text during the pretraining phase.

...

### What do the longs (int64) represent?

…

### [Challenge] Visualization of fine-tuned DistillBERT model.

When processing text data with the DistilBERT model, the text is first tokenized into individual tokens. Each token is assigned a numerical index that corresponds to its position in the model's vocabulary. These numerical indices are represented as longs (int64) in PyTorch.

…

## Task 2d

### Class distributions and learning rate

...

### Relative performances before and after fixing learning rate

...