Identification of Plant Diseases using Text Classification

Group 13

Arpit Singh
Deepak Kumar P Honakeri
Milind Choudhary

Introduction

- Aim to perform text classification on the scraped plant disease dataset.
- Compare the performances and accuracies of text-based classifier with an image classifier
- Trained BERT model on multiple settings on the plant disease dataset.

Dataset

- Data scraped from wikipedia through wikipedia API, containing around 2000 types of diseases.
- Dataset includes paragraphs of symptoms and the target label

Data Type	Size
Train	1960
Test	0
Val	0

- Model can’t generalize on the current data. Hence, Data augmentation is required

Data Augmentation

As mentioned earlier, we were successfully in scraping the data from wikipedia but the data had its own problem.

For each unique label we had only 1 unique text. This means that we have 1960 different categories but for each category we have just one data point.

So our aim was to resolve this problem by focusing on two major things: -

1. Extract more data from different websites.
2. Use text augmentation methods to create new data from existing data.

Extracting more data from different websites

- Multiple tools available for extracting data from websites such as Scrappy, Wikipedia API, and Apache Nutch.
- In the case of plant diseases, Wikipedia had the most comprehensive data compared to other websites.
- Other websites had data on some plant diseases, but not all.
- Apache Nutch is the best tool for web scraping, but it couldn't scrape some relevant websites.
- Nutch scraped complete pages without specific information about plant diseases. This means that now you have data but you don’t know that the data is about or which plant disease it refers to.
- Hence, text augmentation methods were used to increase the dataset due to the above reasons.

Using text augmentation methods.

- Text augmentation libraries such as Parrot, Paraphrase, and textattack can generate new data by replacing tokens, adding new ones, translating or paraphrasing.
- However, these libraries work well for single sentences, but not for large datasets.
- To overcome this limitation, the GPT-3 API was used to generate new content by paraphrasing the existing content or creating new content for each label.
- We also tried to split the initial text into multiple data points to create more data for each label.
- This helped in generating more data points for each unique label, which increased the overall size of the dataset.
- By having more data points for each label, the model was able to better learn the patterns and improve its performance.

Results

Model_Name	Dataset	Parameters	Train Accuracy	Val Accuracy	Conclusion
BERT	Wikipedia Scrap	Default	5.1230E-04	0	Lack of Data
BERT (Plant Name)	Wikipedia Scrap	Default	0.2474	0.2425	Hyperparameter Tuning
BERT (Plant Name) + 2 Hidden	Wikipedia Scrap	Default	0.2858	0.2775	Increasing Depth Helps ! (How much?)
BERT (Plant Name) + 3 Hidden	Wikipedia Scrap	Default	0.2229	0.2375	2 Depths was the ideal depth.
BERT (Plant Name) + 2 Hidden	Wikipedia Scrap	learning_rate = 1e-4	0.0051	0	
		learning_rate = 1e-5	0.2858	0.2775	
		learning_rate = 1e-6	0.1325	0.1150	
BERT (Disease)	Augmented (ChatGPT)	Default	0.1574	0.1414	Augmenting Helped.
BERT (Plant Name)	Augmented (ChatGPT)	Default	0.3729	0.3532	Still More Needed!!!

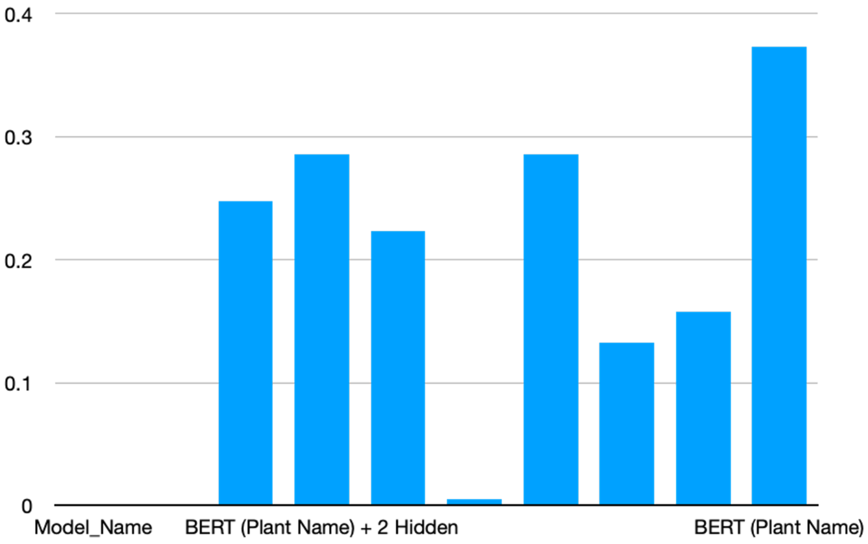

Fig 1. Training Accuracy

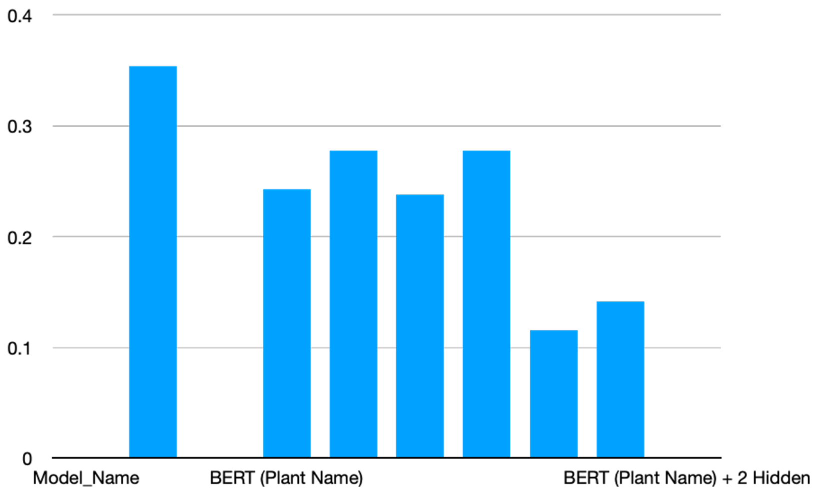

Fig 2. Validation Accuracy

Comparison With CNN + DWT

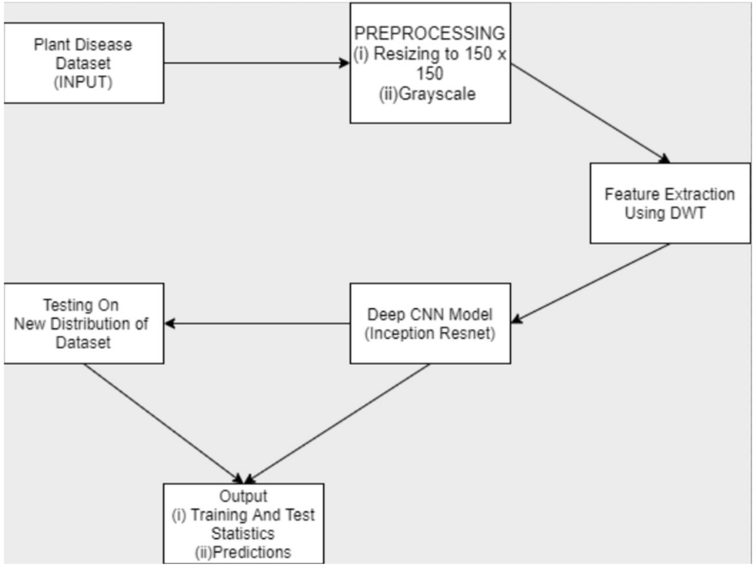

Fig 3. Flowchart of CNN Architecture

Metric	Value
Final Training Accuracy	97.77%
Final Training Loss	0.0724
Final Validation Accuracy	98.57%
Final Validation Loss	0.0339
Test Accuracy	94.05%

Fig 4. Final CNN Metrics

Final Comparative Study

Reference	Dataset	Feature Extractor	Classifier	Accuracy
Mishra et al., 2019	Dataset created from plant village dataset, hosted at Plant Village Disease Classification Challenge	NA	Deep Neural Network (DNN)	88.46%
Prasad et al., 2012	Self-Prepared Dataset	Gabor Wavelet	SVM	89.00%
Mousavi et al., 2016	Public Dataset labelled by Plant Pathologist	Gabor Wavelet	SVM	90.04%
Deshapande et al., 2019	Agricultural fields of Agricultural University, Dharwad	Haar Wavelet	k-NN	85.00%
Zhang et al., 2012	Local Field, ASD Inc., Boulder, Colorado, USA	CWT	MLR	77.00%
Mukherjee et al., 2017	Apple Leaves from Plant Village dataset	NA	Transfer Learning using GoogLeNet	85.04%
Pujari et al., 2014	Department of plant pathology, University of Agricultural Sciences, Dharwad	DWT	Probabilistic Neural Network (PNN)	86.48%
Current Study	Plant Disease Dataset	Bior3.7 Wavelet (DWT)	Transfer Learning using Inception Resnet	94.05%
Current Study	Plant Disease Dataset (Text, Scraping from Wikipedia and Augmentation)		BERT	14.14%

Limitation & Future Works

Lack of data was a major limitation, but it can be overcome by extracting more data and using algorithms to recognize which category the data belongs to. One way to do this is to use the current model to perform classification on new data, which can help in getting more data.

Lack of large processing capabilities was a significant challenge as large BERT models require extensive processing capabilities that were not available on the computers used by our team. Due to this limitation, some potential language models were left unexplored.

Cross-data models were not used in the current approach as the team trained models on both images and text. However, in the future, a multi-data model can be developed that can consider different types of data and produce relevant results. This can help in improving the overall performance and accuracy of the model.

References

1. https://huggingface.co/datasets/glue, Glue Dataset
2. https://github.com/topics/paraphrase-generation?l=python
3. https://nutch.apache.org/
4. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.
5. Gary Ren. Applying NLP Deep Learning Ideas to Image Classification.
6. Yan Y, Kawahara J, Hamarneh G. Melanoma recognition via visual attention. Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26 2019 (pp. 793-804). Springer International Publishing.

Q & A