Name: Divyanshu Srivastava
GitHub Link: Link

Gmail: divyanshuxyz26@gmail.com

Project Documentation: Resume Classification

Dataset:

The dataset used for this project consists of two main components. The first part is obtained from the website Roboflow using the following link: [Roboflow - Resume Dataset]This dataset comprises images related to resumes and is curated for the purpose of training a classifier. The second part of the dataset, representing non-resume images, is collected from Google Images downloaded using the chrome extension. The non-resume dataset includes diverse images such as letters, cover letters, applications, and other relevant content. Size: 1.13GB

Libraries Used:

- NumPy: For numerical operations and array manipulations.
- Matplotlib: For creating visualizations and plotting.
- OpenCV (cv2): For image processing and manipulation.
- TensorFlow: An open-source machine learning framework for building and training models.

Model:

The chosen model architecture for this project is a Convolutional Neural Network (CNN). The decision to use a CNN stems from its ability to capture spatial hierarchies and patterns in images effectively. In comparison to pretrained models, a simple CNN was preferred due to resource constraints—specifically, the absence of a GPU. The model was trained for five epochs on both the training and validation sets. Despite the simplicity of the CNN, it demonstrated robust performance, achieving low training and validation losses.

Metrics:

Various metrics were employed to evaluate the model's performance, as outlined in the accompanying Jupyter Notebook (main.iypnb file). However, due to computational limitations (lack of GPU), the confusion matrix, which provides a detailed breakdown of classification results, was omitted from the metrics presentation. The chosen metrics include precision, recall, binary accuracy, and others, providing a comprehensive assessment of the model's ability to classify resumes and non-resumes accurately.

Conclusion:

Despite the absence of a GPU for model training, the project successfully employed a CNN to classify resumes and non-resumes. The chosen datasets, comprising images from Roboflow and Google Images, provided a diverse range of examples for effective model training. The model demonstrated promising performance, as indicated by the selected metrics. Future work may involve optimizing the model architecture, exploring additional datasets, or leveraging cloud-based GPU resources for more extensive training and evaluation tasks.