Project 2: How do cells respond to different drug treatments

Xiaocong Xuan (xx2438)

November 6, 2024

1 Abstract

This study investigates how deep learning can be used to classify drug treatments based on cell morphology, using a MobileNetV2 model trained on a subset of cell images. With careful data preprocessing, the model reached a validation accuracy of 97.82%, showing it can distinguish between treatments with subtle differences in cell structure. The confusion matrix and t-SNE visualization further confirmed that the model effectively captures complex patterns within the data, highlighting its potential for applications in drug discovery and cellular analysis.

2 Introduction

A recent study used data science to analyze an extensive dataset of three million cell images, each showing cells treated with different drugs and modified genes to observe changes in cell structure. These images were captured using high-throughput microscopy, with fluorescent dyes highlighting key cellular components like the nucleus, cytoskeleton, and mitochondria. This setup allowed researchers to investigate how specific genetic and chemical treatments alter cell morphology. The study aimed to reveal patterns linking genetic changes to drug effects by applying machine learning and image processing techniques. This project uses a subset of those images, each subjected to a limited number of treatments, to train a convolutional neural network (CNN). The objective is to develop a model that can accurately predict the drug treatment from any given cell image, showcasing how deep learning can uncover complex biological insights and accelerate drug discovery.

3 Data Processing

In this project, we designed a data processing and analysis workflow to prepare the subset of 2,867 cell images for training a model to predict drug treatments based on cell morphology. The dataset was accompanied by a metadata CSV file containing essential information for each image, including treatment labels. Our first step was to standardize the file names by converting them to lowercase, ensuring consistency when linking each image to its corresponding metadata entry. We then extracted unique identifiers from the file names to accurately match each image with its metadata, which provided details on the specific treatments applied.

After loading the data, we focused on encoding the treatment labels and addressing the class imbalance. The treatment labels, found in the 'Metadata_pert_iname' column (which specifies the name of the chemical compound used as a perturbation), were converted into numerical labels necessary for training the model. A plot of the label distribution (Figure 1) revealed a significant class imbalance, with a small number of classes dominating the dataset. To streamline the classification task and focus on the most representative classes, we retained only the top 50 most frequent classes, ensuring that the model trained on classes with sufficient data. To further address imbalance within this subset, we applied 'RandomOverSampler', which oversampled the minority classes to create a more balanced dataset and reduce the model's tendency to favor the more common classes.

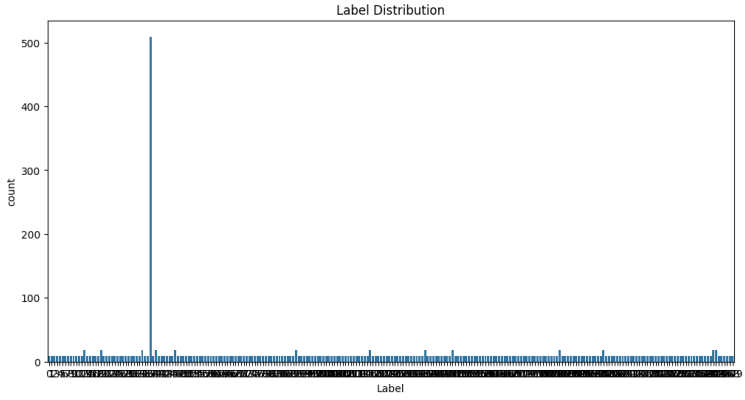

Figure 1: Label Distribution

The dataset was then split into training and validation sets using stratified sampling to preserve class distributions in each subset. A split of 80% for training and 20% for validation was chosen, providing sufficient data for training while retaining a representative sample for validation. Stratified sampling ensured that each class was proportionately represented in both sets, allowing for a balanced evaluation of model performance.

To improve the model's ability to generalize, we applied data augmentation techniques to the training images. These augmentations included random horizontal flips, slight rotations, and adjustments to brightness, contrast, and saturation to create variations that mimic real-world changes in cell morphology. Both the training and validation images were resized to 128x128 pixels and normalized based on ImageNet statistics. This normalization step aligned with the input requirements of pre-trained models and helped with faster model convergence.

We then created a custom 'TiffDataset' class to efficiently handle image loading, label retrieval, and transformations. This class ensured that each image-label pair was returned in a format compatible with PyTorch. Finally, we set up PyTorch 'DataLoader' for both the training and validation sets, with a batch size of 16. The training 'DataLoader' was set to shuffle the data at each epoch, introducing randomness that helps prevent overfitting, while the validation 'DataLoader' maintained a fixed order to ensure consistency during evaluation.

These steps resulted in a well-structured and balanced dataset for training a deep learning model, incorporating data augmentation to enhance generalization and efficient data loading to optimize performance. This setup supports our objective of predicting drug treatments based on cell morphology.

4 Modeling

The data modeling process in this project was carefully designed to maximize the predictive power of a deep learning model for classifying drug treatments from cell images, with a strong focus on efficient feature extraction and dimensionality reduction.

The core of our model architecture is 'mobilenet_v2', a convolutional neural network that is lightweight and computationally efficient, making it well-suited for handling complex image data. Pre-trained on ImageNet, MobileNetV2 brings a robust set of feature extraction layers capable of capturing intricate patterns in visual data. To adapt this model to our specific task, we replaced its final fully connected layer with one that matches the number of classes in our dataset, allowing it to classify drug treatments based on extracted features.

To address the significant class imbalance in the dataset, we calculated class weights for the treatment labels using 'compute_class_weight'. These weights were integrated into the 'CrossEntropyLoss' function, prompting the model to learn effectively from both underrepresented and prevalent classes. By applying a class-weighted loss, we helped the model remain sensitive to all classes, thus minimizing bias towards dominant classes.

In addition to leveraging the neural network's feature extraction capabilities, we employed dimensionality reduction techniques to optimize the model's handling of high-dimensional data. Images were resized to 128x128 pixels, a resolution that strikes a balance between computational efficiency and preserving essential morphological details for classification. This resizing step reduced the dimensionality of each image without sacrificing critical information. Furthermore, the images were normalized according to ImageNet statistics (mean and standard deviation), standardizing the inputs and supporting faster model convergence while enhancing the generalization of learned features.

During training, batches of images and labels were transferred to the GPU to accelerate computation, with mixed-precision training enabled through 'torch.cuda.amp'. This allowed the model to operate in both float16 and float32, lowering memory usage and expediting training without compromising accuracy. Mixed-precision and feature extraction from the pre-trained layers allowed the model to efficiently capture complex morphological patterns while reducing computational demand.

We used the AdamW optimizer, which integrates weight decay to prevent overfitting, along with the 'cosine annealing scheduler' to adjust the learning rate, fostering smoother convergence dynamically. Throughout the training, the model fine-tuned its feature extraction, adjusting weights in later layers to classify treatments accurately while retaining general features from the pre-trained layers. This fine-tuning strategy strikes a balance between

leveraging pre-learned patterns and adapting to the specific nuances of our dataset.

These techniques provided a strong foundation for model training, with feature extraction and dimensionality reduction carefully tuned to balance performance and computational efficiency. Leveraging MobileNetV2's architecture along with class balancing, efficient data loading, and optimized feature extraction, we developed a model well-equipped to accurately predict drug treatments based on subtle morphological features in cell images.

The complete code used for data processing and analysis is available in the following GitHub repository:[https://github.com/xx23438/STATGR5243-Proj2].

5 Results

The validation process was essential in assessing the model's ability to generalize beyond the training data. During validation, the model was set to evaluation mode, disabling layers such as dropout and batch normalization to ensure consistent predictions. Validation images were processed in batches, with each batch passed through the model to generate predictions, which were then compared to the true labels to compute accuracy. Additionally, the validation loss for each batch was calculated using the same loss function as in training, providing a measure of overall error in model predictions.

The model achieved a strong validation accuracy of 97.82%, with a validation loss of just 0.0908. This high accuracy suggests that the model successfully learned to distinguish between drug treatments based on subtle morphological features in the cell images, even when applied to data it had not encountered during training. The drop in validation loss compared to the training loss of 1.6481 further demonstrates the model's ability to generalize, indicating that it effectively captured key patterns relevant to each treatment class. These results confirm that the integration of MobileNetV2's architecture, along with data augmentation, class balancing, and optimized feature extraction, contributed to the model's strong performance in classifying drug treatments, supporting the project's goal of using deep learning for insights into cell morphology.

To further assess the model's performance, we checked at the confusion matrix and t-SNE plot of the feature embeddings. The confusion matrix showed a strong diagonal, indicating that the model accurately predicted most classes, with only a few minor errors. This suggests the model effectively differentiates between drug treatments, capturing the unique morphological traits of each one. The t-SNE plot of the feature embeddings provided additional insight, revealing distinct clusters for many classes. This clustering indicates that the model's feature extraction process effectively grouped similar treatments in a lower-dimensional space. The clear separation between clusters shows that the model learned meaningful, treatment-specific features, with each cluster representing a set of images with similar cell morphology. Together, the confusion matrix and t-SNE plot reinforce that the model not only performs well in terms of classification accuracy but also captures the underlying structure of the data meaningfully.

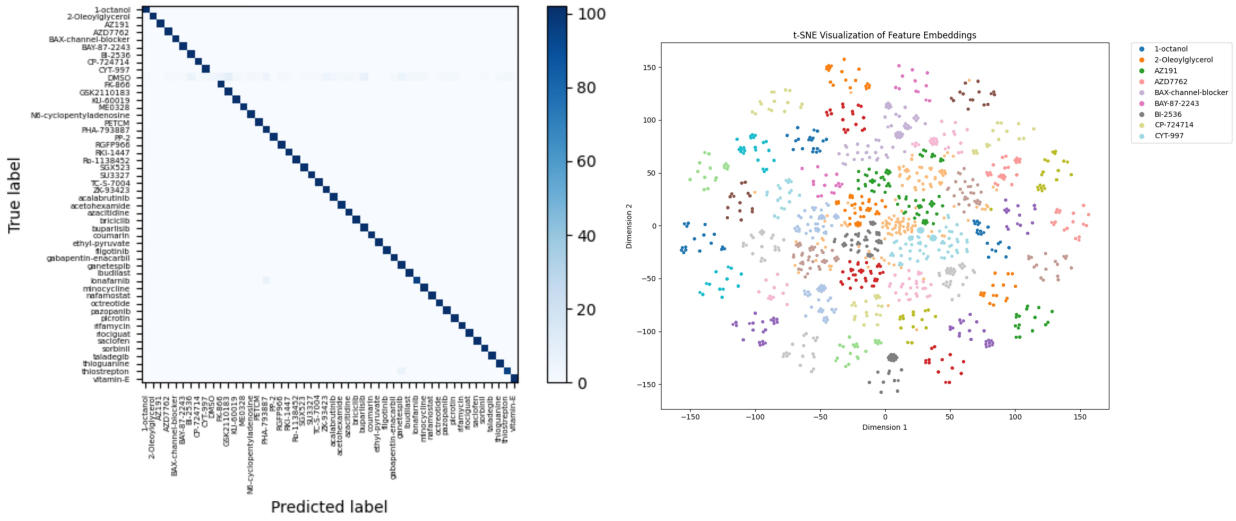

Figure 2: Confusion Matrix & t-SNE Plot

6 Conclusion

In conclusion, this study highlights the potential of deep learning to classify drug treatments based on cell morphology, achieving a high validation accuracy of 97.82% with minimal misclassifications. Using MobileNetV2 and a carefully crafted preprocessing pipeline, we developed a model capable of recognizing subtle morphological differences between treatments. The confusion matrix and t-SNE plot support the model's ability to capture and represent complex patterns in cell images, underscoring its utility in applications like drug discovery and cellular analysis.

However, there were some limitations. The model struggled with certain classes that had limited data, pointing to the ongoing issue of class imbalance. Additionally, the computational power available was insufficient to fully leverage larger models or longer training cycles, potentially restricting the model's performance. The reliance on pre-trained ImageNet features, while effective, could also be improved by using a domain-specific dataset to capture finer cellular details.

Future directions could address these limitations by utilizing more balanced datasets to improve the model's ability to classify rare treatments accurately. Access to higher computational power would enable experimentation with larger models and extended training, potentially enhancing accuracy. Incorporating additional data sources, such as genetic or proteomic information, could also enrich the model's predictions and provide a more holistic view of cellular responses to treatment.