Predicting Drug Treatments from Cellular Images

Boiken Jaho (bj2442)

November 8, 2024

Abstract

This study investigates the application of convolutional neural networks (CNNs) to predict drug perturbations from cellular images. By combining image-derived features with metadata, we demonstrate a complex approach utilizing CNNs and structured data to improve prediction accuracy. Despite challenges such as class imbalance and limited feature differentiation, the results showcase the potential of integrated machine learning techniques in biological research.

1 Introduction

Chemical and genetic perturbations cause distinctive morphological changes in cells, which can be utilized to gain insights into biological processes and drug mechanisms. High-throughput imaging combined with advanced profiling techniques is an effective way to analyze these changes. The dataset presents an expansive resource comprising three million images and morphological profiles of cells treated with various perturbations, which allows the exploration of similarities in their effects. As the manual analysis of such large datasets is inefficient, different methods of analysis can be used. This project employs convolutional neural networks (CNNs) to predict the specific perturbation applied to the cells using their morphological features and metadata. This aims to deepen current understanding of cellular responses to different treatments in a more efficient way.

2 Methods

2.1 Data Collection and Preprocessing

The dataset included images from the *downsampled_data* folder and metadata files. Features were extracted using Cellpose and were merged with perturbation/drug labels from the metadata for each cell image.

2.1.1 Feature Extraction with Cellpose

Cellpose was used to extract structural features from cell images. Each image's extracted properties included:

- Area
- Mean Intensity
- Bounding Box Sum
- Eccentricity
- Solidity

2.1.2 Data Integration

The extracted features were merged with drug metadata using filenames as keys. To handle discrepancies in image naming conventions, preprocessing ensured consistent matching. Cellpose feature segmentation also revealed several instances of each cell type, which required further note.

2.2 Model Architecture

A hybrid CNN model was constructed to process image data alongside metadata:

- Image branch: 2D convolutional layers followed by max pooling and flattening to lower computation complexity and facilitate metadata combination.
- Metadata branch: Dense layers for processing structured data.
- Combined branch: Concatenation of the image and metadata branches, leading to fully connected layers.

The final model was trained using sparse categorical cross-entropy loss with Adam optimizer for dynamic learning rate adjustment. Class weights were applied to mitigate class imbalance, since exploratory analysis showed imbalances of different instances in the set.

2.3 Training and Validation

The model was trained for 20 epochs with early stopping and learning rate reduction, to prevent overfitting and underfitting. Validation was performed on a holdout set to track performance metrics.

2.4 Code Availability

The code is publicly available at `https://github.com/your-repo/project2`.

3 Results

3.1 Training Performance

Figure 1 shows the model's accuracy and loss over training epochs. Initial accuracy was low, but improvements were observed after optimizing hyperparameters and balancing the dataset. Accuracy plateaued at 0.24 despite different adjustments, such as increasing model complexity by adding more layers, adjusting class weights to capture imbalances (DMSO), and resizing images to reduce noise. This could be due to insufficient feature discriminability which prevented differentiation between drugs that could have similar morphological changes. More complex tuning of the model with added layers to capture even subtler differences could be useful however resulted to not be time efficient.

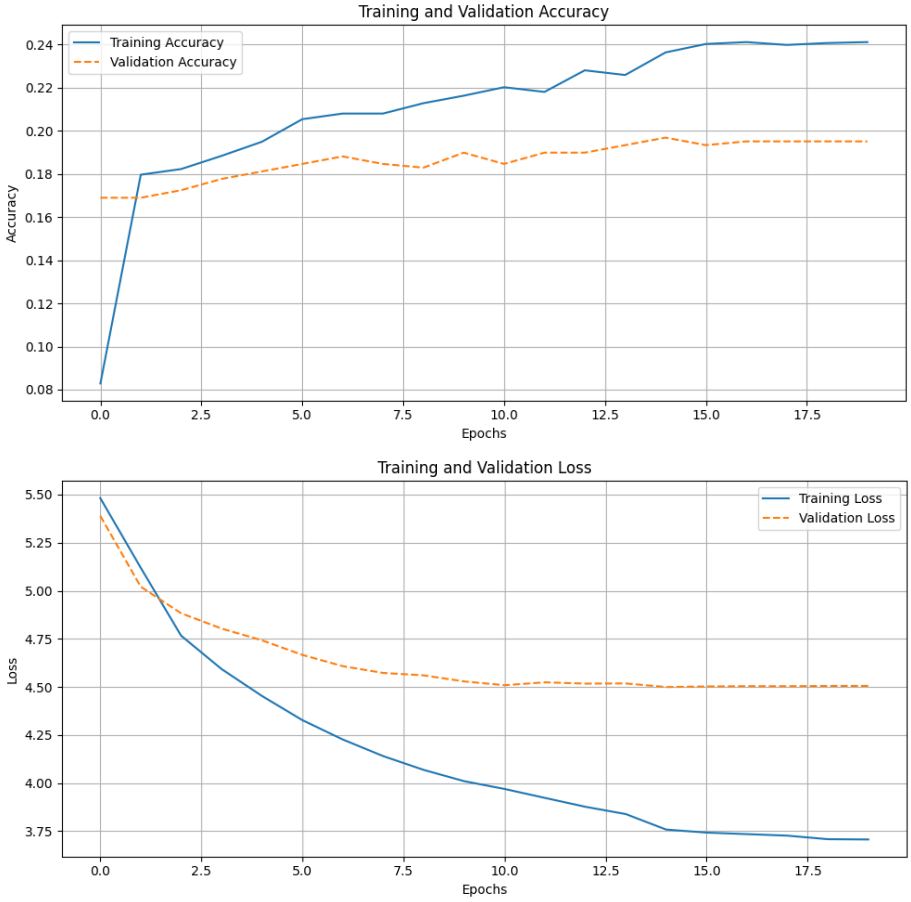

Figure 1: Training and validation accuracy/loss across epochs.

3.2 Drug Prediction Bias

An analysis of drug distribution revealed significant bias towards certain drugs, notably DMSO (Figure 2), in which when tested on unseen cell images the model guessed DMSO more frequently than others, often incorrectly. DMSO is commonly used as a penetrating vehicle for many drugs, which could explain its high representation in this dataset. This likely impacted model performance, as underrepresented drugs were harder to predict accurately.

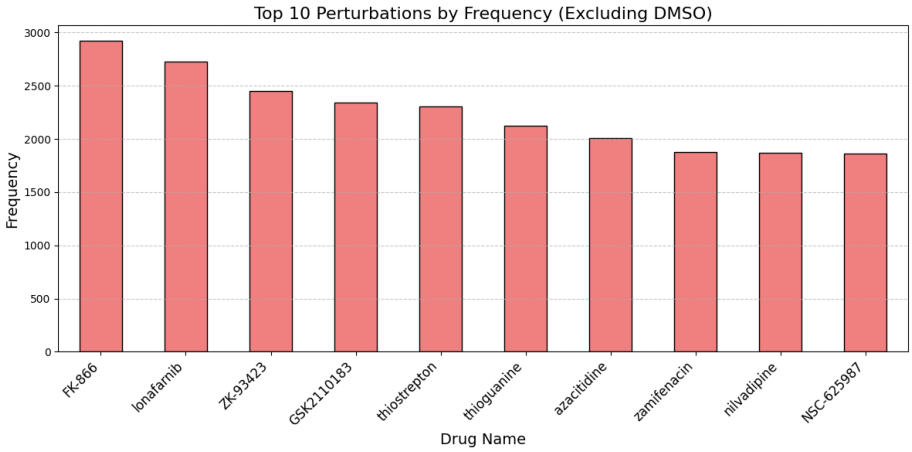

Figure 2: Distribution of top drug perturbations in the dataset.

3.3 Testing on Example Data

When tested on unseen example data, the model often predicted the most common drugs (e.g., DMSO) regardless of actual labels, indicating persistent bias and overfitting to dominant classes. Other than DMSO, FK-866 was commonly guessed, indicating that the model isn't truly learning about the class differences based on the cell segmentation features. This suggests either too subtle differences for a simple CNN model to grasp and learn from.

4 Conclusion

This study highlights the potential of using convolutional neural networks (CNNs) alongside image-derived features and metadata to predict drug perturbations in cells. The hybrid model leveraged structural cell features, such as area, intensity, and eccentricity, combined with metadata to make predictions. Accuracy plateaued due to class imbalance, which impacted the model's ability to generalize. Testing on unseen data revealed the model's bias toward predicting the most frequent perturbations. This suggests that despite the integration of metadata and feature extraction, the subtle morphological differences between

certain drug effects were not fully captured. Future work should focus on exploring more complex architectures, and incorporating additional data preprocessing techniques to enhance model performance and mitigate bias.