COMP6721 – Phase II Report

Abstract

This project investigates image classification of indoor venues—library, museum, and shopping mall—using a Convolutional Neural Network (CNN) implemented from scratch in PyTorch. The dataset was divided into training, validation, and test subsets using an automated split, and data preprocessing was applied to ensure consistency. A grid search was conducted over learning rate and batch size to optimize performance. The final model was selected based on evaluation metrics including accuracy, precision, recall, and F1-score. The best-performing configuration achieved a validation accuracy of 77.97% and a test accuracy of 70%, with balanced classification performance across all classes. The trained model was saved and deployed for real-time prediction on unseen images during the demonstration.

1. Introduction

Indoor scene classification is a challenging and important problem in computer vision with applications in robot navigation, surveillance systems, and context-aware mobile assistance. In this project, we focus on the classification of indoor venues into three categories: library-indoor, museum-indoor, and shopping_mall-indoor. To tackle this problem, we adopt a deep learning approach using a Convolutional Neural Network (CNN) built entirely from scratch in PyTorch, without relying on pre-trained models. The project emphasizes key deep learning tasks such as dataset preparation, model architecture design, hyperparameter tuning, evaluation, and deployment. Through this hands-on implementation, we aim to understand and demonstrate the fundamental mechanics of CNNs for image classification.

2. Dataset Description

The dataset used in this project contains images from three indoor venue categories: library-indoor, museum-indoor, and shopping_mall-indoor. The dataset was structured into two main folders: Training/ and Test/, each containing subdirectories for the respective classes. A validation set was created automatically by splitting the training data into an 80/20 ratio. Before feeding the images into the CNN, all inputs were resized to a fixed resolution of 128×128 pixels, converted to PyTorch tensors, and normalized using the mean and standard deviation values of the training set. This preprocessing ensures consistency across batches and helps accelerate training convergence.

3. CNN Model Architecture

The Convolutional Neural Network (CNN) used in this project was implemented from scratch using the PyTorch framework, without relying on any pre-trained models. The architecture consists of three convolutional layers, each followed by a ReLU activation and max-pooling operation to reduce spatial dimensions and extract hierarchical features. After the convolutional blocks, the feature maps are flattened and passed through two fully connected (dense) layers, with a Dropout layer applied to reduce overfitting. The final output layer returns raw class scores (logits) for the three venue categories. Since PyTorch's CrossEntropyLoss internally applies softmax, no explicit softmax layer is added at the output. The model is optimized using the Adam optimizer.

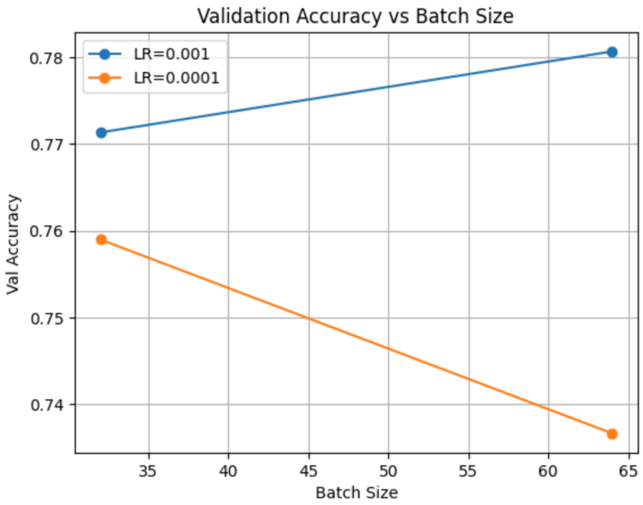
Model Summary:
- Input: 128×128 RGB images
- Conv1: 3 → 32 filters, 3×3 kernel, ReLU, MaxPool(2×2)
- Conv2: 32 → 64 filters, 3×3 kernel, ReLU, MaxPool(2×2)
- Conv3: 64 → 128 filters, 3×3 kernel, ReLU, MaxPool(2×2)
- Flatten, FC1: 128×16×16 → 256, ReLU, Dropout(0.5)
- FC2: 256 → 3 (number of classes)

4. Hyperparameter Tuning

A grid search was conducted over two critical hyperparameters: learning rate and batch size. Each configuration was trained for 15 epochs, and validation accuracy was recorded after each run to identify the best model. The results are summarized in the table below:

Learning Rate	Batch Size	Best Validation Accuracy
0.001	32	77.13%
0.001	64	77.97%
0.0001	32	75.90%
0.0001	64	73.67%

As shown, the configuration with a learning rate of 0.001 and a batch size of 64 outperformed all others, achieving a validation accuracy of 77.97%. This configuration was selected for the final model evaluation on the test dataset.

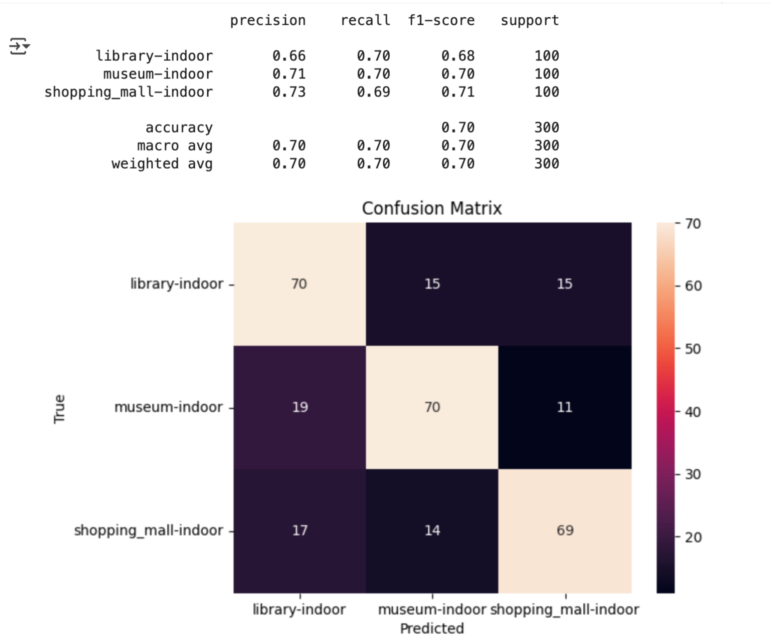


5. Evaluation and Results

The final model was evaluated on the test set. Performance metrics including accuracy, precision, recall, and F1-score were calculated. The following table summarizes the class-wise results:

Class	Precision	Recall	F1-Score
library-indoor	0.66	0.70	0.68
museum-indoor	0.71	0.70	0.70
shopping_mall-indoor	0.73	0.69	0.71

The overall test accuracy was 70%. The confusion matrix further illustrates the classification behavior, with most misclassifications occurring between library and the other two classes.



```
                      precision    recall  f1-score   support

      library-indoor       0.66      0.70      0.68       100
       museum-indoor       0.71      0.70      0.70       100
shopping_mall-indoor       0.73      0.69      0.71       100

            accuracy                           0.70       300
           macro avg       0.70      0.70      0.70       300
        weighted avg       0.70      0.70      0.70       300
```

6. Conclusion

This project successfully demonstrates a scratch-built CNN in PyTorch for indoor scene classification. Through a systematic grid search and careful model evaluation, the system achieved competitive performance with clear potential for further improvements through techniques like data augmentation, transfer learning, or deeper networks.