Progress Report on Image Classification Using Decision Trees

Alireza Lorestani, Nasim Fani, Raghav Senwal
Department of Computer Science and Software Engineering (CSSE)
Concordia University
Montreal, Canada

***Abstract*—This report investigates the application of machine learning techniques, specifically Decision Trees and Random Forests, for classifying images in the Places365 dataset. The study explores both supervised and semi-supervised learning approaches to enhance classification accuracy. The results demonstrate the effectiveness of Random Forests in handling image classification tasks and highlight the potential of semi-supervised learning in scenarios with limited labeled data.**

***Index Terms*—Machine Learning, Decision Trees, Random Forest, Image Classification, Semi-Supervised Learning, Supervised Learning**

I. Introduction and Problem Statement

The task at hand involves the application of machine learning techniques, specifically decision trees, to classify images based on the venue from the Places365 dataset. The goal is to explore both supervised and semi-supervised approaches to improve the accuracy of the model. This project aims to develop a robust image classification system capable of correctly identifying venues within the Places365 dataset.

This classification problem has several real-world applications including object detection, automated sorting systems, and scene recognition. The challenge lies in maximizing the accuracy of the model, especially when labeled data is limited.

II. Challenges

A. Limited Labeled Data

A common issue in machine learning is the shortage of labeled data which limits the model's learning ability.

B. High Dimensionality

Images have high-dimensional data making it computationally intensive to process and classify.

C. Model Overfitting

With limited data, there's a risk of the model overfitting to the training set reducing its generalizability.

D. Pseudo-Labeling

Ensuring the quality of pseudo-labels in semi-supervised learning can be challenging.

III. Proposed Methodologies

A. Dataset

The dataset used in this study is the Places365_small dataset. The Places dataset is designed following principles of human visual cognition. In total, Places contains more than 10 million images comprising 400+ unique scene categories. The dataset features 5000 to 30,000 training images per class, consistent with real-world frequencies of occurrence [5]. The dataset can be downloaded from [1]. The images were resized to a consistent size for uniform processing. To increase the accuracy of the models we used some feature extraction methods.

- Chosen classes: Bar, Gymnasium, HospitalRoom, SubwayStation, and Restaurant.
- Number of images/class: 2000 images/class for the supervised and 1200 images/class for the semi-supervised approach.
- Features were extracted using color histograms and Local Binary Pattern (LBP).

B. Model

1) Supervised Learning: A Decision Tree Classifier and Random Forest Classifier were employed for initial supervised learning to establish a baseline model.

2) Semi-Supervised Learning: For semi-supervised learning, a Random Forest Classifier was used to iteratively improve the model by incorporating high-confidence pseudo-labels from the unlabeled dataset. The process involved selecting the predictions with confidence of 70% or higher to augment the labeled data. Other approaches such as choosing a 10% top confidence level were rejected due to providing low accuracy after several tests.

IV. Solving The Problem

A. Supervised Learning

In the supervised learning phase, the data was split into training (80%) and test (20%). Then, the Decision Tree and Random Forest Classifiers were trained using the training data. The trained models were evaluated using the test data. Training curves and evaluation metrics such as accuracy, precision, recall, and F1 score were recorded for both classifiers.

B. Semi-Supervised Learning

Same as the supervised learning approach, data was split into the training (80%) and test (20%). The training data then was split into two parts, unlabeled(80%) and labeled(20%) data. In this phase, a Random Forest Classifier was trained at the beginning using the labeled data. Then, the model was iteratively retrained by adding high-confidence pseudo-labels (predictions with confidence of more than 70%) to the labeled dataset. The accuracy was monitored at each iteration to ensure the quality of the pseudo-labels.

The iterative process involved the following steps:

1) Train the Random Forest Classifier on the labeled data.
2) Predict the labels for the unlabeled data.
3) Select high-confidence predictions and add them to the labeled dataset.
4) Retrain the classifier with the augmented labeled dataset.
5) Repeat the process until no significant improvement in accuracy is observed.

The accuracy progression was monitored at each iteration to ensure the quality of the pseudo-labels.

V. Results

In supervised learning, as Table I indicates, the Random Forest Classifier significantly outperforms the Decision Tree Classifier in all metrics, indicating better overall performance. As shown in Table II, In semi-supervised learning, after many improvements, a significant amount of the unlabeled data was labeled with an accuracy of 42%. The confusion matrices of all models are shown in Fig 1a, Fig 1b, and Fig 2

TABLE I: Supervised Models Performance

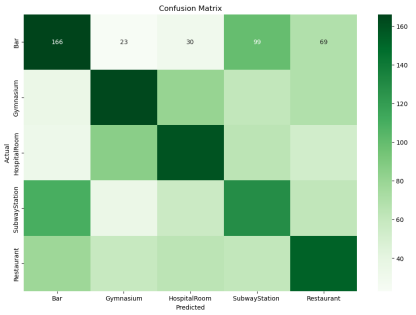
Metric	Decision Tree Classifier	Random Forest Classifier
Accuracy	38.55%	51.90%
Precision	38.75%	51.53%
Recall	38.55%	51.90%
F1 Score	38.60%	51.28%

TABLE II: Semi-Supervised Model Performance

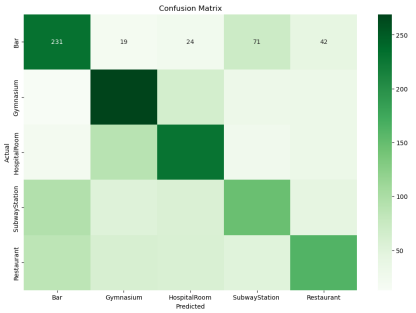
Metric	Random Forest Classifier
Accuracy	42.15%
Precision	40.85%
Recall	42.15%
F1 Score	40.25%

VI. Future Improvements

- Increase Threshold for Pseudo-Labeling: To ensure higher quality of pseudo-labels the confidence threshold can be increased. However, it might result in lowering the number of labeled data.
- Data Augmentation: Applying data augmentation techniques to increase the diversity of the dataset and improve the model's generalizability.
- Validation Set Monitoring: Using a separate validation set to monitor model performance and prevent overfitting.



(a) Supervised Decision Tree Confusion Matrix



(b) Supervised Random Forest Confusion Matrix

Fig. 1: Confusion Matrices for Supervised Learning Models

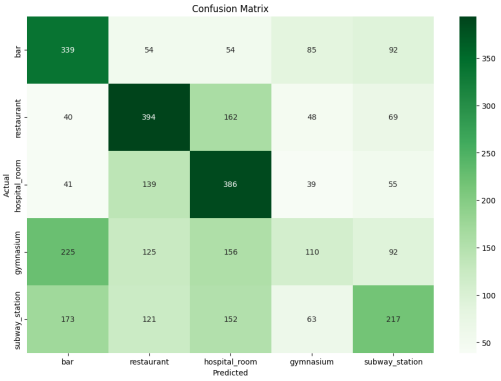


Fig. 2: Semi-Supervised Random Forest Confusion Matrix

- Hyperparameters Grid Search: Exhaustive search over specified parameter values to find the optimal values.

References

[1] Places365 Dataset: http://places2.csail.mit.edu/download-private.html
[2] Scikit-learn Grid Search Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[3] Scikit-learn Decision Tree Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
[4] Scikit-learn Random Forest Documentation: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[5] Zhou, Bolei and Lapedriza, Agata and Khosla, Aditya and Oliva, Aude and Torralba, Antonio, "Places: A 10 million Image Database for Scene Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.