Museum-Indoor vs. Museum-Outdoor Image Classification

Mohammed Yasir
Concordia University
Montreal, QC, Canada
m_m1490@live.concordia.ca

Abdul Majeed
Concordia University
Montreal, QC, Canada
ab_maje@live.concordia.ca

***Abstract*—In today's digital age, automated image classification has become a critical task across various domains, including cultural heritage preservation, tourism, and digital archiving. Museums manage vast collections of images requiring efficient organization. This project focuses on classifying museum images into indoor and outdoor categories using machine learning techniques. We explore supervised learning with Decision Trees, Random Forest, and XGBoost, as well as semi-supervised learning to optimize performance while minimizing labeled data requirements.**

I. Introduction and Problem Statement

This project aims to solve a binary classification problem where images need to be classified into two categories: Museum-Indoor and Museum-Outdoor. Automatic categorization of such images is critical in digital archiving, content management systems, and heritage preservation initiatives. Given the rapid digitalization of cultural artifacts, automating image classification can significantly reduce manual effort and enhance the efficiency of large-scale museum databases.

Applications

- Museum Information Systems: Automatic sorting of museum images into indoor and outdoor categories for better archival organization.
- Tourism Platforms: Improved recommendations based on location types.
- Image Retrieval Systems: Enabling users to search for images based on environment type.

General Strategy

1) The general approach to solving this problem was to train machine learning models using a labeled dataset of museum images. We explored:
2) Supervised Learning using Decision Trees, Random Forest, and XGBoost.
3) Semi-Supervised Learning where only part of the data was labeled initially, and pseudo-labels were iteratively generated for unlabeled data.

Challenges Faced

1) High-Dimensional Features: Each image had a very large feature vector (flattened pixel values), making training slower and prone to overfitting.
2) Limited Labeled Data Simulation: For the semi-supervised approach, only 20% of the training data was labeled. This simulated a real-world scenario where labeling is expensive.
3) Computational Cost: Some models (like XGBoost) were computationally expensive, requiring optimizations to balance performance and speed.
4) Confidence Thresholding in Semi-Supervised Learning: Setting the right confidence threshold for pseudo-labeling required experimentation to avoid introducing too much noise.

II. Proposed Methodologies

A. Dataset

The dataset consists of 10,000 images (5,000 indoor, 5,000 outdoor) sourced from the MIT Places Dataset. Each image was resized to 64×64 pixels and represented as a 1D vector of 12,288 features.

Fig. 1. Sample image of the dataset.

B. Preprocessing

- Resizing: Standardized to 64×64 pixels.
- Flattening: Converted into 1D feature vectors.
- Scaling: Normalized pixel values to $[0, 1]$.

C. Models Used

We experimented with the following machine learning models:

- **Decision Tree:** max_depth=10.
- **Random Forest:** n_estimators=100, max_depth=10.
- **XGBoost:** max_depth=10, n_estimators=100.
- **Semi-Supervised Decision Tree:** max_depth=10, confidence threshold=0.85.

III. Experiments and Results

A. Initial Attempts and Challenges

Direct Supervised Learning

- Initially, Decision Tree and Random Forest were tested directly on the fully labeled training set.
- Decision Trees overfitted on full data, so regularization using max_depth was applied.
- Random Forest showed more robustness due to ensemble learning.

Feature Engineering (Rejected)

- Tried extracting additional features (e.g., color histograms, texture descriptors), but these were not consistently better than raw pixels.
- For simplicity and scalability, we reverted to the flattened pixel vector approach.

B. Successful Approach

The below table shows the comparison metrics of the performance of the models:

Model	Accuracy	Precision	Recall	F1 Score	Time(s)
Decision Tree	0.83	0.83	0.83	0.83	132.53
Random Forest	0.90	0.91	0.87	0.89	107.96
XGBoost	0.90	0.92	0.88	0.90	487.91
Semi-Supervised DT	0.80	0.80	0.80	0.80	216.33

TABLE I
Performance Comparison of Models

C. Supervised Learning Results

- Decision Tree provided a baseline accuracy of 83%. It was fast (75.54 seconds), but lacked robustness.
- Random Forest significantly improved accuracy to 89%, balancing precision and recall. It also trained faster than the Decision Tree.
- XGBoost gave the highest accuracy of 90%, demonstrating superior learning capacity. However, this came at a steep computational cost of 338.13 seconds, indicating that XGBoost may not be the best choice for real-time applications. The comparison graph of results is given below.

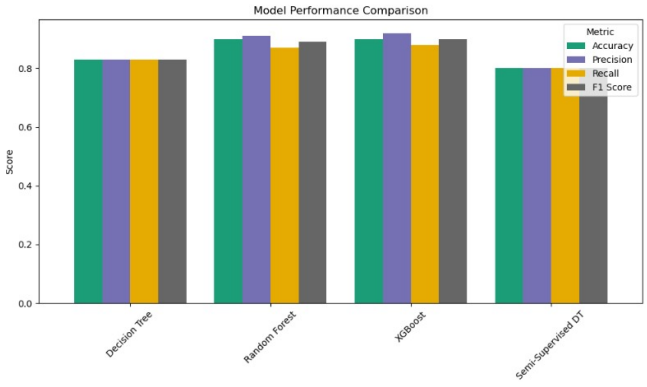

Fig. 2. Comparison graph of results.

D. Semi-Supervised Learning Results

- With only 20% labeled data initially, the Semi-Supervised Decision Tree reached an accuracy of 80
- In each pseudo-labeling iteration, confident samples were added to the training set.
- The process stopped after two iterations when no new confident labels could be added.
- Though slightly less accurate than the fully-supervised models, this method highlights the potential for reducing labeling costs, especially for larger datasets.

- The time taken (158.49 seconds) was longer due to iterative re-training, but still much faster than XGBoost.

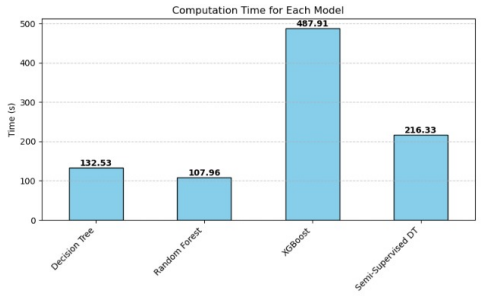

Fig. 3. The computation time of each model is given below.

IV. Conclusion

Machine learning significantly improves museum image classification. While XGBoost offers the best accuracy, Random Forest presents a trade-off between efficiency and performance. Semi-supervised learning reduces dependency on labeled data, making it a viable alternative for large datasets.

References

[1] NeurIPS Proceedings, "Semi-Supervised Learning with Decision Trees," 2022.

[2] CompPhysics, "Machine Learning Summer School: Decision Trees," 2023.

[3] Towards Data Science, "From Decision Trees and Random Forests to Gradient Boosting," 2023.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD), San Francisco, CA, USA, 2016, pp. 785–794.

[5] O. Chapelle, B. Schölkopf, and A. Zien, Semi-Supervised Learning. Cambridge, MA, USA: MIT Press, 2006.

[6] "Decision Tree, Random Forest, and XGBoost: An Exploration into the Heart of Machine Learning."

[7] "Decision Forests for Classification, Regression, Density Estimation, Manifold Learning, and Semi-Supervised Learning."

[8] "Decision Tree, Random Forest, XGBoost: Understand, Choose, and Tune Them Easily!"