17-634: Applied Machine Learning

Assignment #1

Shandon Jude Herft

• Describes your dataset, including formal citations in APA (or similar style) format of the dataset itself, along with any research publications that have used the dataset

I have chosen the “Architectural Styles” dataset, which is publicly available on Kaggle:

Source: DumitruX. (n.d.). Architectural Styles Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/dumitrux/architectural-styles-dataset/data.

This dataset contains 10113 images from 25 architectural styles. It is a mixed between images scraped from Google Images and the dataset from the paper "Architectural Style Classification using Multinomial Latent Logistic Regression" (ECCV2014), made by Zhe Xu.

This folder contains the dataset only with Google Images (g-images) or both datasets joined (architecture-style-dataset).

The dataset made by Zhe Xu can be found on https://www.kaggle.com/wwymak/architecture-dataset.

• Articulate your problem statement, including the “business value” of solving the problem (i.e., what benefit will solving the problem create for end-users or other stakeholders?)

Architectural styles classification is important in the field of heritage conservation, and education in architecture. The goal of this project would be to develop a machine learning model that would be able to classify different building images to fit into one of the 25 architectural styles in the dataset.

This would be highly beneficial for Architects, to help assist in identifying architectural styles of buildings within the urban landscape, and designing appropriate buildings with the right context based on stylistic trends. For Cultural Heritage Conservatives, classifying styles accurately can help with restoration efforts and documentation. For Education, it can help students to understand these different styles. For the field of Real Estate, the style of the buildings can help market value analysis.

• Describes any challenges you anticipate facing, and how you might address them.

1. Imbalance of data: Some styles might have more data available than the other.
   I could try to oversample minority classes, under sampling majority classes, applying adjustments in the training based on the class.
2. Image Quality: Resolution and Quality of images could vary since some of them have been scraped from Google.

Preprocessing these images could be a solution. Adjusting them could help improve the robustness of the model.

3. Overlapping Features: Some features of different architectural styles could be shared. CNN's could be used to identify these patterns from the images. Fine tuning existing models could help transfer learning for use.
4. Overfitting: With small data sometimes, overfitting could occur for some classes. Cross validation can be used to help the model generalize well. Dropout and early stopping could be other techniques.

Citations:

Xu, Z., & Huang, J. (2014). Architectural style classification using multinomial latent logistic regression. European Conference on Computer Vision (ECCV). Springer. Retrieved from https://www.kaggle.com/wwymak/architecture-dataset.

DumitruX. (n.d.). Architectural Styles Dataset. Kaggle. Retrieved from https://www.kaggle.com/datasets/dumitrux/architectural-styles-dataset/data.