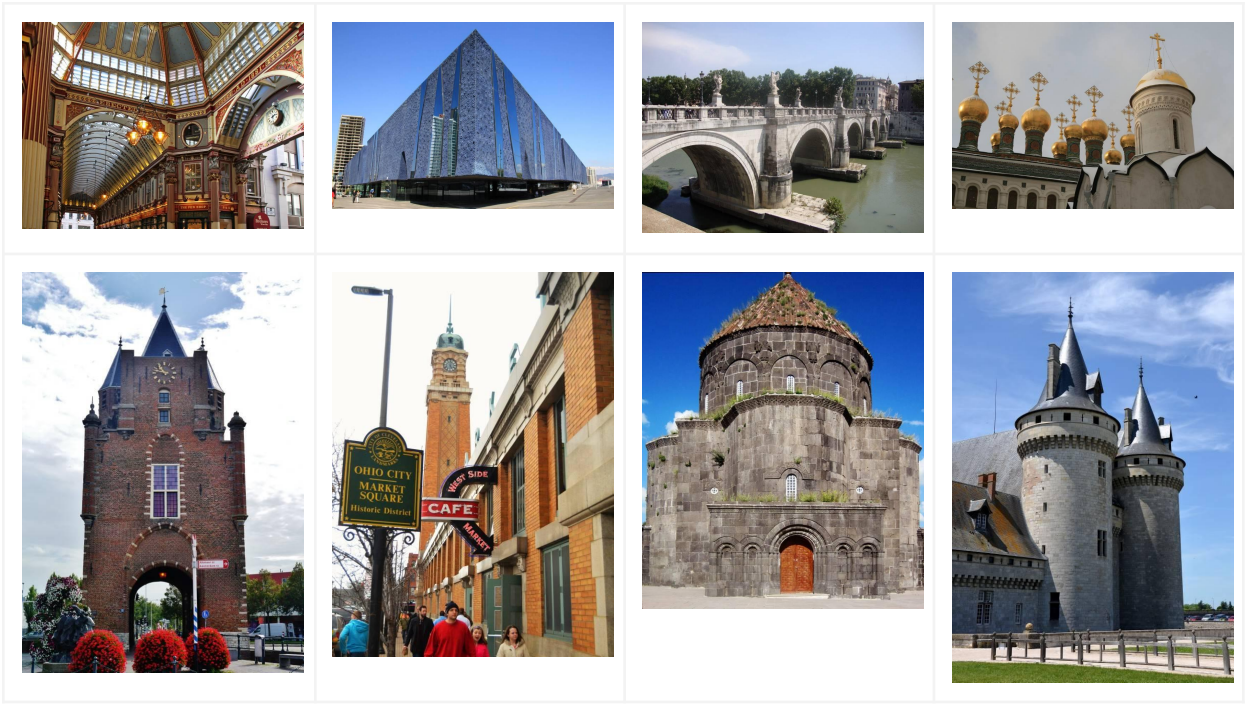
Extreme Image Classification - Landmark Recognition

Vivek Bhatnagar | Varun Tanna | Charlee Stefanski



Landmark recognition technology aims to predict landmark labels directly from image pixels to help people better understand and organize their photo collections. The Task is to take images and recognize which landmarks (if any) are depicted in them.

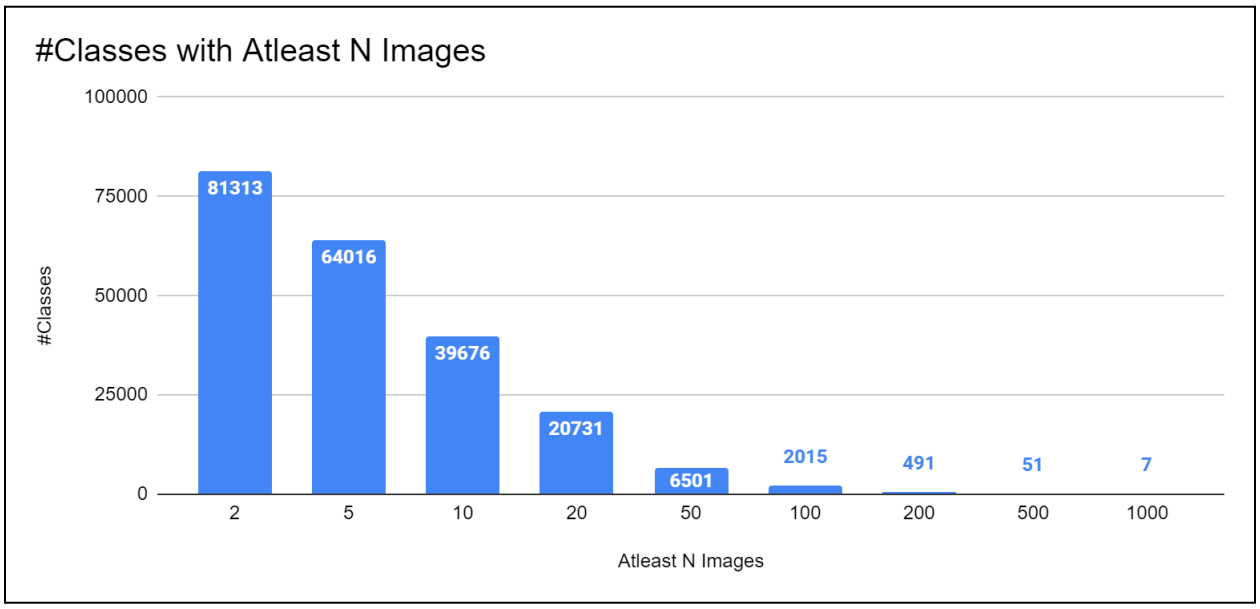
Dataset

https://www.kaggle.com/competitions/landmark-recognition-2021/data

Google Landmarks dataset (GLDv2) (Weyand et al., 2020) is the largest worldwide dataset to foster progress in Landmark recognition and retrieval. GLDv2 training set presents a realistic crowdsourced setting with diverse types of images for each landmark. It contains a large number of classes (there are more than 81K classes in this dataset), and the number of training examples per class may not be very large, as discussed in the next section. The training dataset contains approximately 1.58 million images. There is variability in image sizes, but the most common resolution is 600 * 800, and all images are within 800 * 800 pixels.

Output Categories

The problem/dataset is an example of an extreme classification task—a growing research area in computer vision focusing on multi-class problems involving an extremely large number of labels (ranging from thousands to billions).



- ★ All (81,313) classes have at least two images in the dataset.
- ★ 2,015 classes have at least 100 images in the dataset.
- ★ 51 classes have at least 500 images in the dataset.

We plan to tackle this problem progressively, starting with 51 classes and then generalizing the model to handle 2,015 and 81,313 categories. This will also allow us to scale our project based on the time constraints in which we are to complete the project.

Image Features

Since our project is focused on identifying different landmarks, our project will need to look at the shapes of these places and contrast them with the background in which the pictures were taken. Thus we will need to look at features like edges and corners to both extract the landmark itself and then look at the shape of the building itself. Furthermore, we will need to explore things like ridges and different texture metrics to look at buildings of different shapes and contrast them from one another. For example, two buildings may have a circular dome, but the separating factor could be the pattern of ridges on it.

Model Architectures

Convolutional Neural Networks (CNN) are used to progressively extract higher-and higher-level representations of the image content for the Image Classification/Recognition task. For baseline models, we plan to fine-tune flavors of CNN architecture like YOLOv7 (Wang et al., 2022), ResNet (He et al., 2015), VGGNet (Simonyan & Zisserman, 2014), InceptionNet (Szegedy et al., 2014) and AlexNet (Krizhevsky et al., 2012).

The Vision Transformer (ViT) (Dosovitskiy et al., 2020) has emerged as a better alternative to convolutional neural networks (CNNs) that are currently state-of-the-art in computer vision and, therefore, widely used in different image recognition tasks. ViT models outperform the current state-of-the-art (CNN) by almost 4x in terms of computational efficiency and accuracy. Finally, we plan to explore models that have performed exceptionally well on the GLDv2 dataset—a model with deep orthogonal fusion of local and global features (DOLG) using an EfficientNet backbone and a novel Hybrid-Swin-Transformer (Henkel, 2021).

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., & Zhai, X. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR*. https://arxiv.org/abs/2010.11929

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *CoRR*, *abs/1512.03385*, 1-12. http://arxiv.org/abs/1512.03385

Henkel, C. (2021). Efficient large-scale image retrieval with deep feature orthogonality and Hybrid-Swin-Transformers. *CoRR*, *abs/2110.03786*. https://arxiv.org/abs/2110.03786

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25 (NIPS 2012)*. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Simonyan, K., & Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, *arxiv.1409.1556*. https://arxiv.org/abs/1409.1556

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going Deeper with Convolutions. *CoRR*, *abs/1409.4842*. http://arxiv.org/abs/1409.4842

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. 10.48550/ARXIV.2207.02696

Weyand, T., Araujo, A., Cao, B., & Sim, J. (2020). Google Landmarks Dataset v2 -- A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. *Computer Vision and Pattern Recognition*, *abs/2004.01804*, 1-18. https://arxiv.org/abs/2004.01804