LANDMARK RECOGNITION

Huzeyfe Ayaz huzeyfeayaz23@gmail.com

Yunus Ahmed Stahlschmidt stahlschmidt.yunus@gmail.com

Muhammed Fatih Öztel ffatihoztel@gmail.com

Supervisor: Prof. Dr. Çiğdem Eroğlu Erdem

Introduction

A landmark is a recognizable building or structure that stands out from the crowd. Landmark recognition is an AI product that identifies popular architectures within a given set of images.

Problem

Recognizing over 80.000 different landmarks from a noisy dataset featuring high class imbalance with only very limited hardware capabilities.

Solution

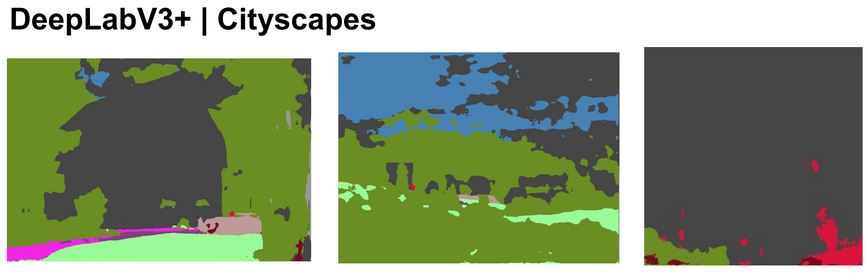
- Extracting features from images to use in a newly proposed data cleaning method for noisy image datasets.
- Creating a model from scratch to predict classes from image embeddings.
- Fine-tuning pre-trained EfficientNet model to predict classes.

Comparing Segmentation Datasets to Extract Features

Original Examples

DeepLabV3+ | ADE20K

DeepLabV3+ | Cityscapes



Cosine Similarities of Images with Different Models

In order to determine which model generates the best embeddings for our purpose. We compared EfficientNetB7 model, which is trained with ImageNet dataset, this generally gives the best results in competitions and various segmentation models that give semantic information for images. We have chosen three images to compare pairwise similarities.

Image 1 (I1)

Image 2 (I2)

Image 3 (I3)

Model Name	Similarity I1-I2	Similarity I1-I3	Similarity I2-I3
EfficientNet B7	0.11	0.51	0.11
Knet S3 UperNet Swin-L	0.78	0.98	0.78
DeepLabV3+ R101	0.44	0.89	0.47
Knet S3 DeepLabV3 R50	0.41	0.88	0.41
Average of top 3 Seg. Emb.	0.64	0.90	0.61

Table 1: Similarity scores of three image pairs

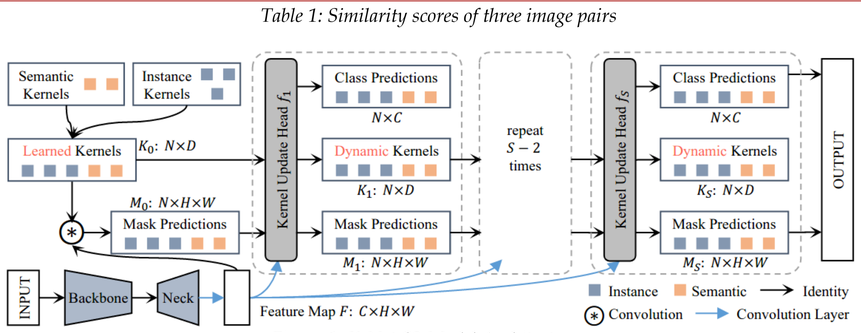


Figure 1: K-Net S3 Model Architecture

Image Processing for Backbone Model

Before feeding images into models, we need to apply some preprocessing steps to images. Segmentation models have inference_segmentor method that converts images into 3, 512, 512 formats in order to gather more information from images like different light conditions.

Dataset

GLDv2c consists of 1.5M samples which include many noisy images. Dataset has 81313 different classes which makes it more challenging to predict class labels accurately. We proposed a new data cleaning method that uses image embeddings to cluster them class-wise with the DBSCAN algorithm. As a result of that, we cleaned nearly 100K images from our dataset.

Image Count per Class	Count
K>20	19693
10<K<=20	16974
5<K<=10	20375
K<=5	24271

Table 2: Class Imbalance

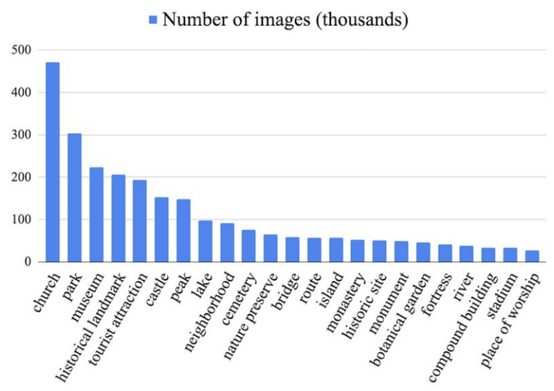


Figure 2: GLDv2 number of images by landmark category

Outlier Detection with Embeddings & DBSCAN

DBSCAN is a clustering algorithm that has many hyper-parameters such as “eps”, “min samples” and “metric”. In order to get the best clustering result, It would be better to apply hyper-parameter tuning for each class separately. But, since each tuning operation takes a significant amount of time, we set “eps=0.33”, “min_samples=2”, and “metric=cosine”. Hence, we obtained relatively good results for outlier detection as shown in the images to the right.

The first row shows the main cluster that covers original images and the second row contains outliers that were removed from the dataset.

Proposed Model Architecture & Fine Tuning

(a) Training efficiency.

	EfficientNet (2019)	ResNet-RS (2021)	DeiT/ViT (2021)	EfficientNetV2 (ours)
Top-1 Acc.	84.3%	84.0%	83.1%	83.9%
Parameters	43M	164M	86M	24M

(b) Parameter efficiency.

Figure 3: Model Comparisons

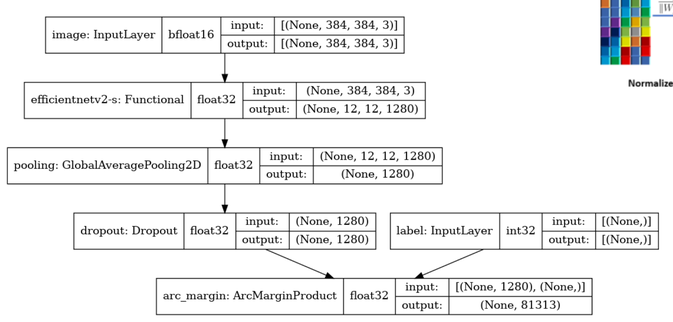


Figure 4: Fine-tuned Model Architecture

EfficientNet is famous for relatively having less number of parameters and high accuracies in the Imagenet dataset. In our project, we picked the latest version of EfficientNet which is EfficientNetv2-S. As a model architecture, we directly give our image inputs to the EfficientNet layers, after taking model features from the model, GlobalAveragePooling is applied to taking the average of the first and second dimensions. Dropout is added to eliminate the overfitting problem in training. Instead of categorical cross-entropy, we used Arcface-loss which gives better results for our problem. Lastly, the model makes predictions among the 81313 classes.

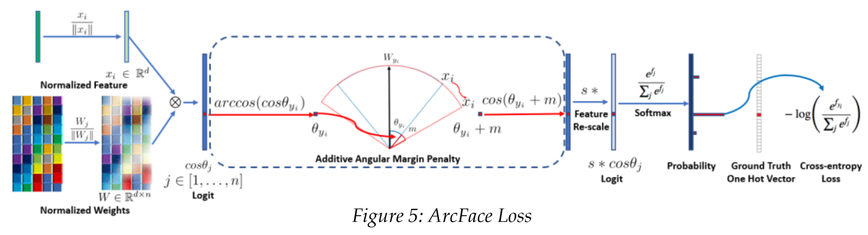


Figure 5: ArcFace Loss

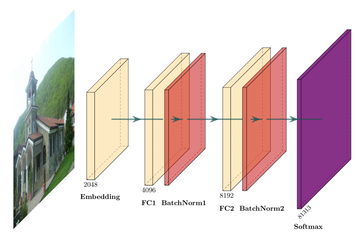


Figure 6: FC Model for Embeddings

Results & Future Work

Top kagglers ensemble numerous models to achieve state-of-the-art results. Our aim was to get similar results while only making use of a single and efficient model which we accomplished as shown in Table 2. For future work, we are planning to put our model to the real test by submitting it to the competition after improving our data cleaning method.

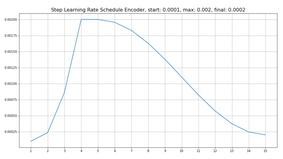


Figure 7: LR by epoch

Epoch #	Acc. at 1	Acc. at 10	Acc. at 100	Acc. at 1k
Epoch 1	0.001	0.001	0.002	0.002
Epoch 5	0.16	0.27	0.44	0.61
Epoch 10	0.34	0.50	0.65	0.78
Epoch 15	0.39	0.56	0.71	0.82

Table 3: Top-K Categorical Accuracy by epoch

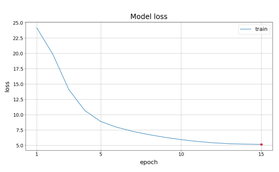


Figure 8: Model Loss History

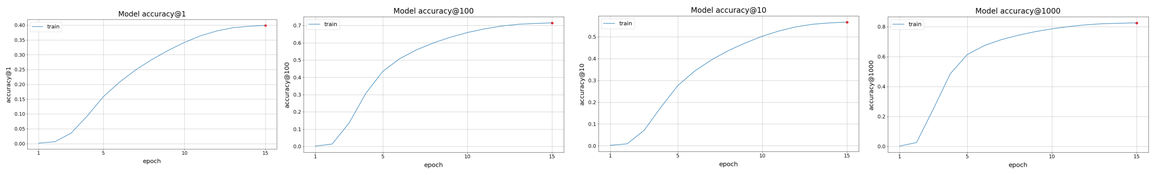


Figure 9: Model Top-K Categorical Accuracy History

References

[1] T. Weyand, A. Araujo, B. Cao, and J. Sim, ‘Google Landmarks Dataset v2 -- A Large-Scale Benchmark for Instance-Level Recognition and Retrieval’, ArXiv Pre-Print Serv., Nov. 2020

[2] Ester, M., H. P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, AAAI Press, pp. 226-231. 1996

[3] Zhang, W., Pang, J., Chen, K. and Loy, C., 2022. K-Net: Towards Unified Image Segmentation. [online] arXiv.org. [Accessed 30 May 2022].

[4] Tan, M. and Le, Q., 2022. EfficientNetV2: Smaller Models and Faster Training. [online] arXiv.org. Available at: <https://arxiv.org/abs/2104.00298> [Accessed 31 May 2022].

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, ‘ArcFace: Additive Angular Margin Loss for Deep Face Recognition’, ArXiv Pre-Print Serv., Feb. 2019, [Online]. Available: https://arxiv.org/abs/1801.07698

Technologies Used