



CNN Plays Geoguessr: Transfer Learning on ResNet50 for Classifying Street View Images

Eric Werner, Jeffrey Heo, Finn Dayton
{ewerner, jeffheo, finndayton}@stanford.edu

Summary of Work

- Created a computer vision (CV) model to play a modified version of Geoguessr, where model would learn to predict the country of an image
- Our model classifies a 224x224 RGB image into 1 of 20 countries
- Modified ResNet-50 (CNN), pre-trained on ImageNet dataset, using feature extraction and fine-tuning
- Built custom loss function to learn geographic similarities between nearby countries, and punish far-away predictions
- Achieved 40.3% accuracy with feature extraction on test set, and **72.5% accuracy** with fine-tuning on test set. Human performance was 7%.

Data

- GeoLocation - Geoguessr Dataset found on Kaggle
- 39,441 images** total labeled by country name for 20 countries
- split into train/validate/test with ratio 90/5/5
- Resized 1536x662 RGB images to 224x224 RGB
- Normalized color channels (mean=[.485, .456, .406] & std=[.229, .224, .225]) and randomly augmented training data via resized cropping, horizontal flipping, and rotation from -5 to +5 degrees



Figure 1: Three sample images left to right: Finland, Singapore, USA

References / Acknowledgements

- Andong Deng, Xingqian Li, Zhibing Li, Di Hu, Chengzhong Xu, Dajiang Dou, "Inadequately Pre-trained Models are Better Feature Extractors" March 2022. <https://arxiv.org/abs/2203.04668>
- Jeffrey Heo, Saharwanj Reddy, Yuhao Ma, Sujan, 2019, Transfer Learning with ResNet-50 for Malaria Cell Image Classification
- Geoguessr, "Geoguessr: Where? World Classification of COVID-19 4-19", <https://www.geoguessr.com/>
- Eric Muller-Budack, Nadir Prati-jren, Ralph Szwed, "Geolocation Estimation of Photos using a Hierarchical Model and Scene Classification", 2018, ECCV
- Geoguessr, "Geoguessr Images (50k)", <https://www.kaggle.com/datasets/subhishaj/geoguessr-images-50k>
- Subha K. 2021, Geolocation - Geoguessr Images (50k)
- Roha Ilyse Bendjilali, Mohammed Beladgham, "Rohad Merit", Abdelmalik, Taieb-Ahmed "Illumination-robust face recognition based on deep convolutional neural networks architectures" May 2020, Indonesian Journal of Electrical Engineering and Computer Science
- Geoguessr, "Geoguessr Images (50k)", <https://www.kaggle.com/datasets/subhishaj/geoguessr-images-50k>

Model

We utilized transfer learning to apply the ResNet50 architecture and weights to our application.

Feature extraction:

- froze weights of first 49 layers before training

Fine-tuning:

- froze weights of first 40 layers before training, leaving last 10 layers to learn features specific to and relevant for Geoguessr screenshots

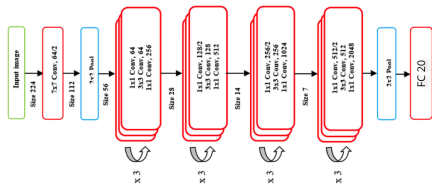


Figure 2: ResNet50 Architecture with last layer FC20 instead of FC1000

$$\text{loss} = L_{\text{CE}} + L_{\text{Haversine}}$$
$$= \sum_{i=1}^n \left(- \sum_{k=1}^K y_k^{(i)} \log \hat{y}_k^{(i)} \right) + \alpha \sum_{i=1}^n \text{dist}_{\text{Haversine}}(\hat{y}^{(i)} - y^{(i)})$$

Figure 3: Custom Loss Function

Our **novel loss function** combined cross-entropy loss (generic classification) with a distance term.

This is not only similar to Geoguessr's scoring scheme of closer = more points, but also allows our model to learn which features nearby countries share, thus learning both country and region.

Discussion

Our project exceeded our expectations, as our initial feature-extraction training with cross-entropy loss barely exceeded random classification.

The confusion matrix captures the results. The diagonal entries are correctly classified, and have overwhelmingly higher frequencies than the off-diagonals. Also, looking closely at high-valued off-diagonals shows that **most false classifications are nearby countries**; the most confused countries were Finland-Sweden-Norway, which are also the only three Nordic countries contained in the dataset. This implies those countries look similar, and also that the model would have scored highly on Geoguessr.

The most exciting result was the **14% test accuracy improvement by using the custom loss function**. We created this loss function with the additional distance penalty to mimic a human in how they 'hone-in' on the region, and then the specific country. In this way, our model learned like a human.

Results

We achieved a **maximum accuracy of 72.5% on the test set**, utilizing fine-tuning and our custom loss function. Purely using cross-entropy loss led to only 58.5% accuracy, and freezing only the last layer with feature extraction resulted in the worst accuracy, with 40.3% of pictures correctly classified.

Model	Train Accuracy	Test Accuracy
Feature Extraction	40.8%	40.3%
Fine-tuning, CE loss	61.9%	58.5%
Fine-tuning, custom loss	76.5%	72.5%

Figure 4: Table of Model Results

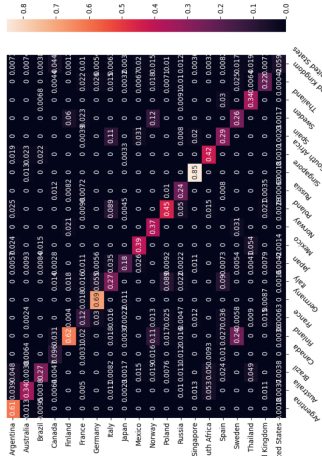


Figure 5: Confusion Matrix of fine-tuned model on test set

Future

- Modify loss to be purely distance
- Increase countries classified
- Within-country specificity
- Try different models: VGG, AlexNet, etc.