# Zero-Shot Learning for Geolocalization in Restricted Image Domains

CIS 4190/5190

By: Jason Figueroa, Charlie Gottlieb, Tom Holland, Arthur Pogosian
TA: Harshwardhan Yadav

01

# Introduction

# THE PROBLEM

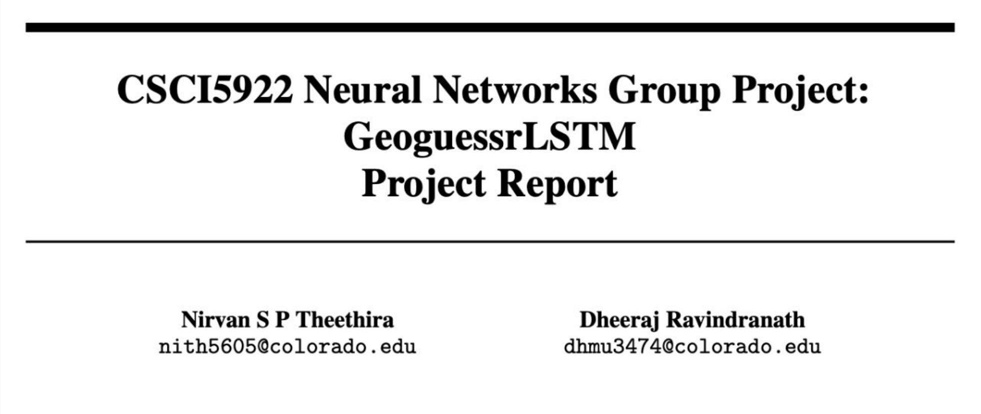
We are addressing the challenge of predicting geographical locations from images.
This would be of aid to law enforcement and investigation, or enhance social media functionalities.

Contributions:
1. Implementation of OpenAI's CLIP
2. Expand Dataset

# Previous Works

- Work 1: “CSCI5922 Neural Networks Group Project: GeoguessrLSTM” by Nirvan S P Theethira and Dheeraj Ravandranath, [https://github.com/Nirvan66/geoguessrLSTM/tree/master](https://github.com/Nirvan66/geoguessrLSTM/tree/master)

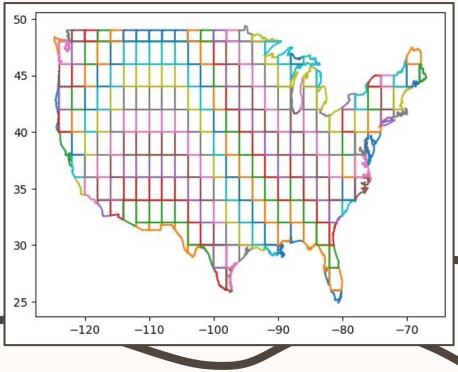
- Work 2: “Learning Generalized Zero-Shot Learners for Open-Domain Image Geolocalization” by Lukas Haas, Silas Alberti, and Michal Skreta, [https://huggingface.co/geolocal/StreetCLIP/tree/main](https://huggingface.co/geolocal/StreetCLIP/tree/main)



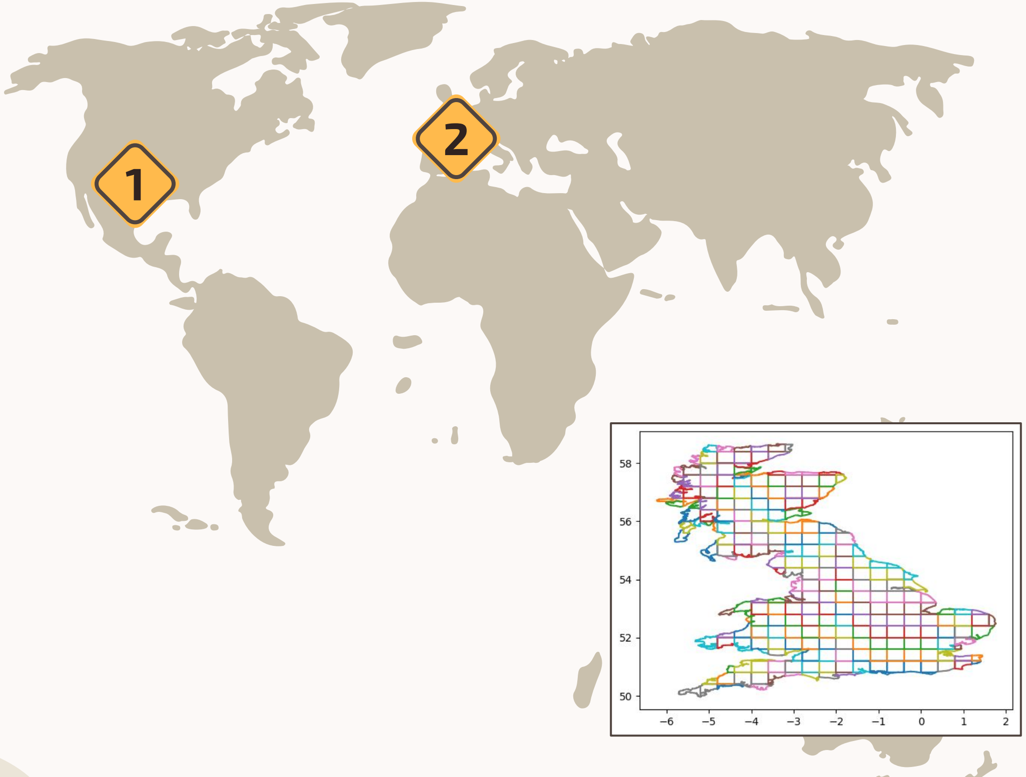
**CSCI5922 Neural Networks Group Project:**
**GeoguessrLSTM**
**Project Report**

**Nirvan S P Theethira**
nith5605@colorado.edu

**Dheeraj Ravindranath**
dhmu3474@colorado.edu



# Contribution #1: Datasets



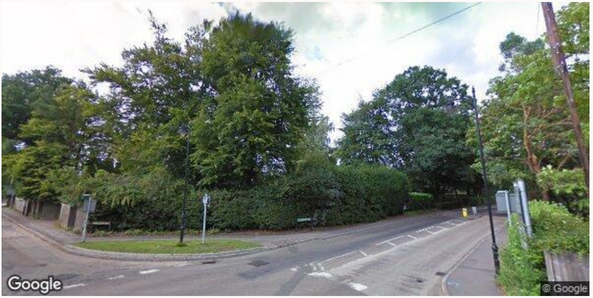
1 US Dataset (Initial):
20580 Images

2 UK Dataset (New)
97260 Images

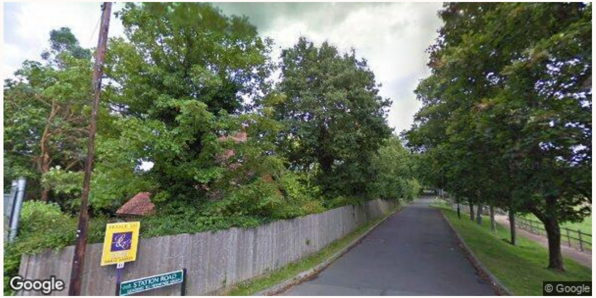
# Example of Images

**City**: Hermitage, Thatcham

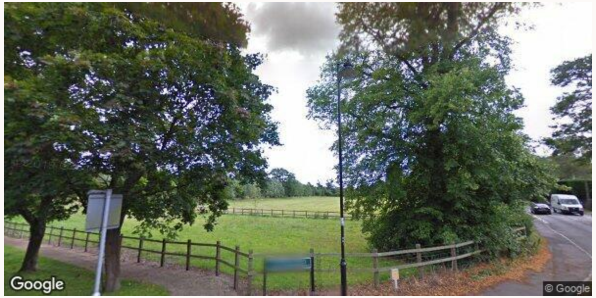
0°



90°

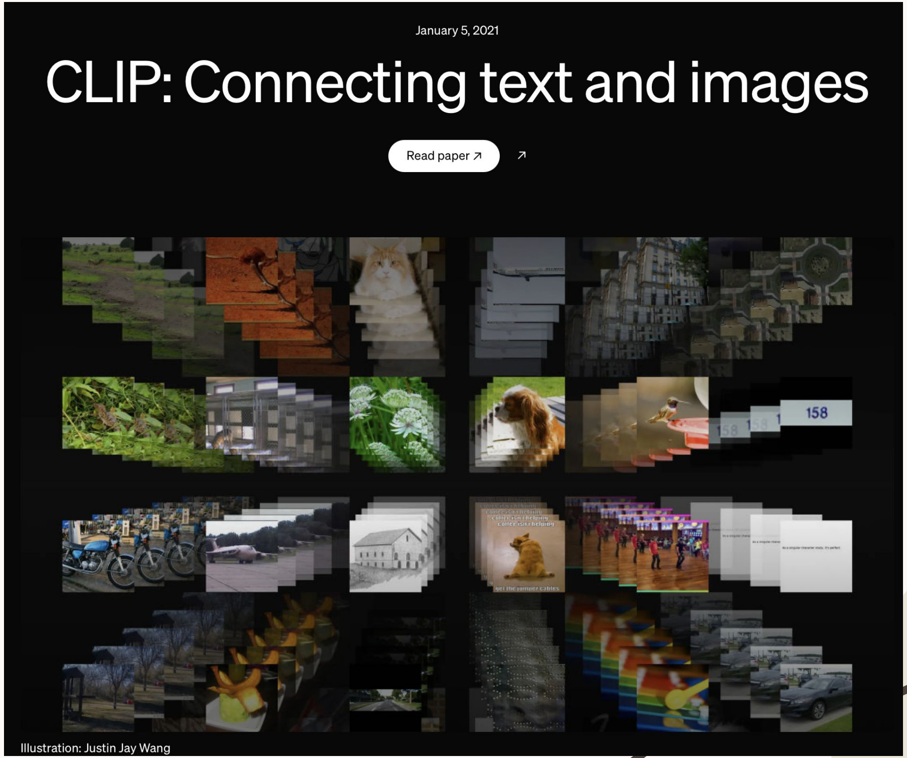


180°



# Contribution #2: Inclusion of OpenAI’s CLIP

- Contrastive Language-Image Pretraining (CLIP) is a model that learns text and images simultaneously
- This allows images to be understood in the context of natural language, linking image information with the vast associative ‘comprehension’ of an LLM


January 5, 2021

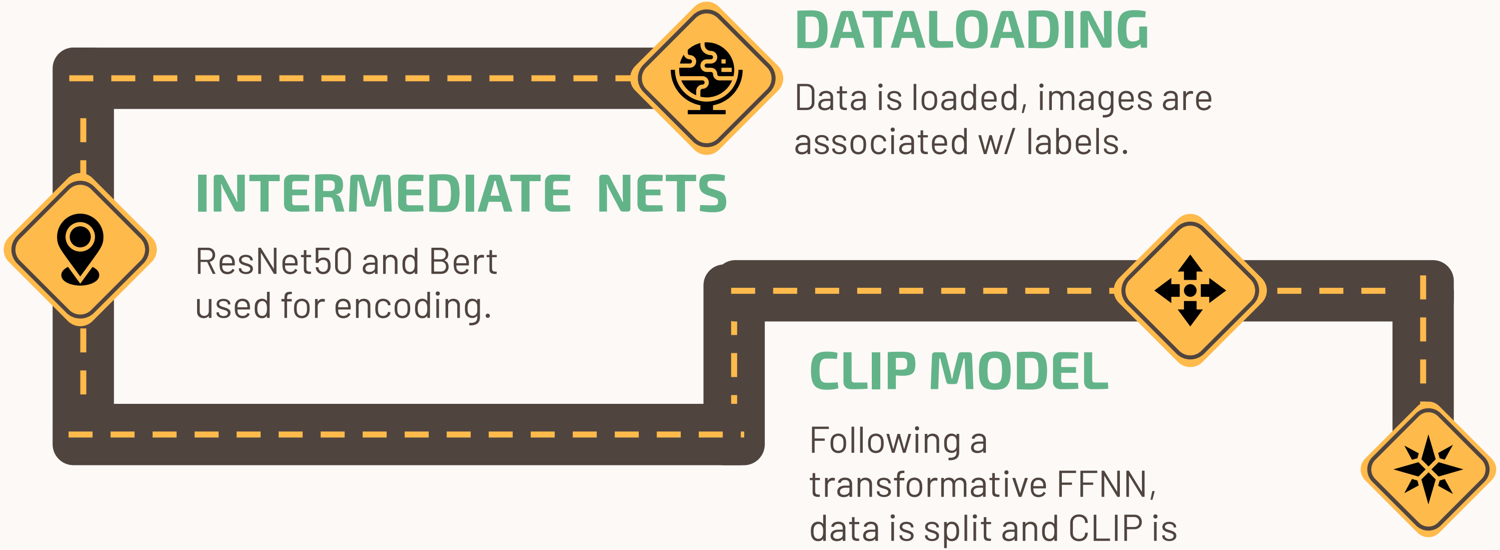
CLIP: Connecting text and images

Illustration: Justin Jay Wang

02

# Methods

# BUILDING OUR NEW MODEL: ARCHITECTURE ROADMAP



## DATALOADING

Data is loaded, images are associated w/ labels.

## INTERMEDIATE NETS

ResNet50 and Bert used for encoding.

## CLIP MODEL

Following a transformative FFNN, data is split and CLIP is tuned.

**TRAINING**!

# COMPLETE ARCHITECTURE

- **Dataset Creation:** Two dataset classes were developed**:**
  - *UKClip*: Grid-based data collection from the UK with individual CSVs associating town names with coordinates.
  - *UKCitiesClip*: Data from the top 150 UK cities, utilizing a single CSV for label association.
- **Data Processing**:
  - Image Encoding: Utilized ResNet-50 CNN to encode images.
  - Text Encoding: Employed DistilBert transformer for text.
  - ProjectionHead NN: Lowered dimensional space and applied transformations using normalization and residual connections.
- **Model Training**:
  - CLIP Architecture: Set up as per OpenAI's guidelines, including a standard cross-entropy loss function.
  - Data Splitting: Implemented train, validation, and test splits.
  - Hyperparameter Tuning: Adjusted weight decays, patience, factor, and learning rates, evaluated through loss metrics on different splits.

03

# Results

# 117,840

Images with accurate location labels gathered across the UK and US

# 3 Datasets for Training

1. UK Grid (~15000 images)
2. Top 150 most populous UK cities (~52500 images)
3. Top 50 most populous UK cities (~30000 images)

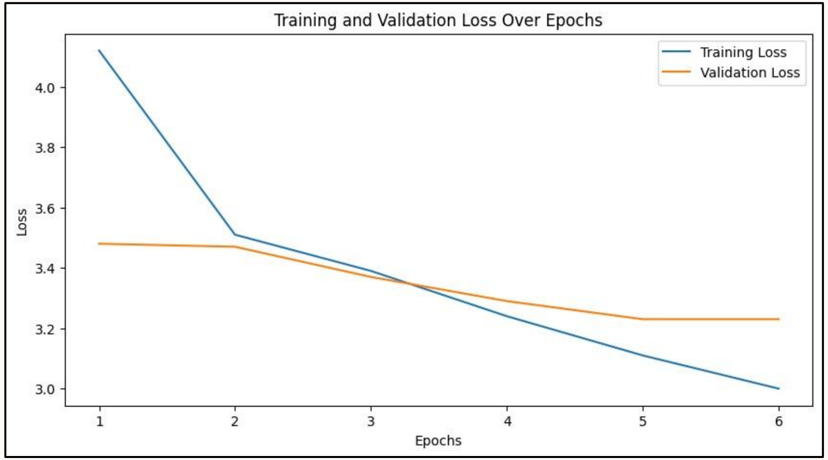
# Model Results (Grid)

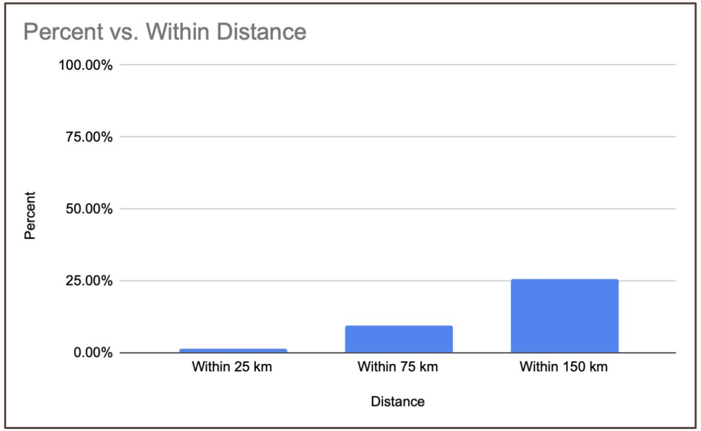
## Test Accuracy: 2.28%

## Avg Distance: 292 km

**Note: Great Britain is at most 500 km from East-West, and 1000 km from North-South**

## Training + Validation





Predicted: Grid 218
True: Grid 108

Predicted: Grid 129
True: Grid 218

Predicted: Grid 15
True: Grid 42

Predicted: Grid 65
True: Grid 43

Predicted: Grid 176
True: Grid 161

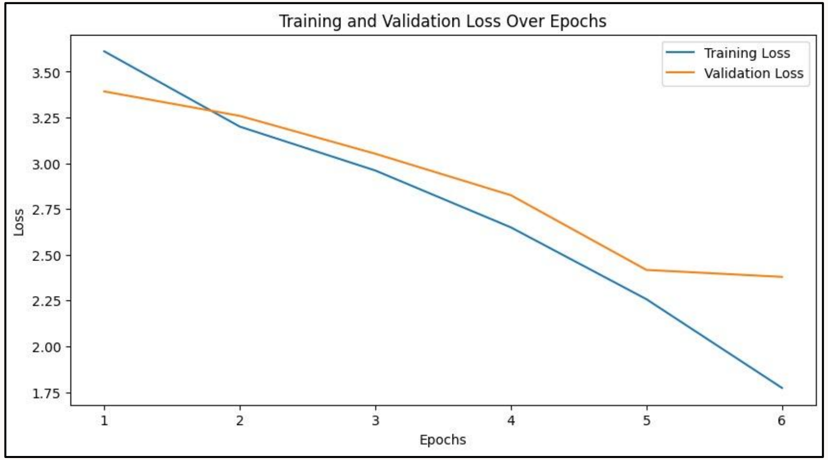
# Model Results (Top 150 cities)

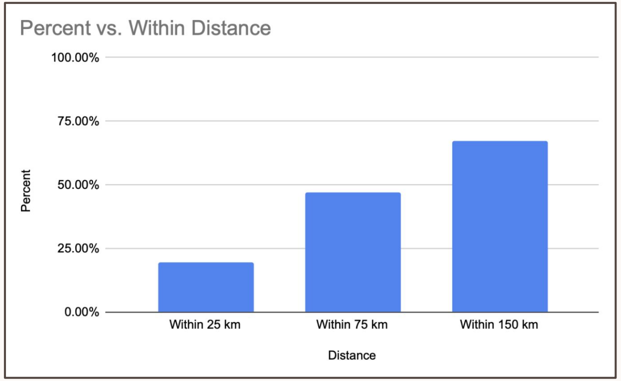
## Test Accuracy: 14.34%

## Avg Distance: 128 km

**Note: Great Britain is at most 500 km from East-West, and 1000 km from North-South**

## Training + Validation





**Note**: Many cities were suburbs of major cities, so we ended up cutting to 102 cities after combining suburbs with the main city.

Predicted: Aberdeen
True: Aberdeen

Predicted: Norwich
True: Norwich

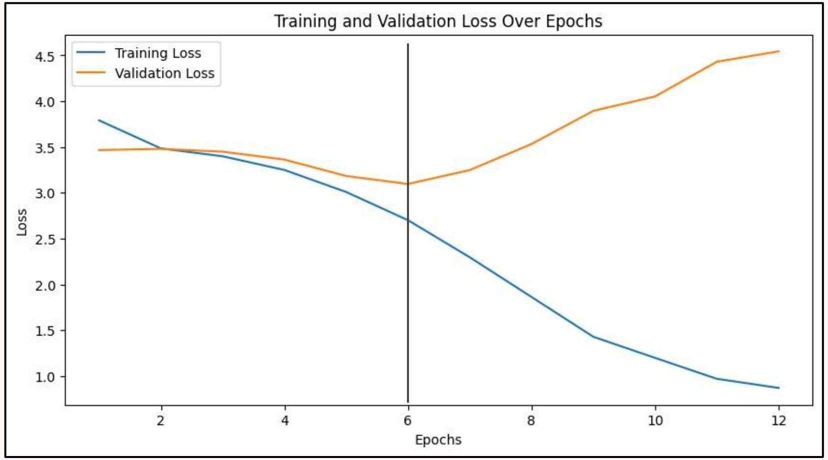
Predicted: South Shields
True: Blackpool

Predicted: Grays
True: Weston-super-Mare

Predicted: Liverpool
True: Liverpool

# Model Results (Top 50 cities)

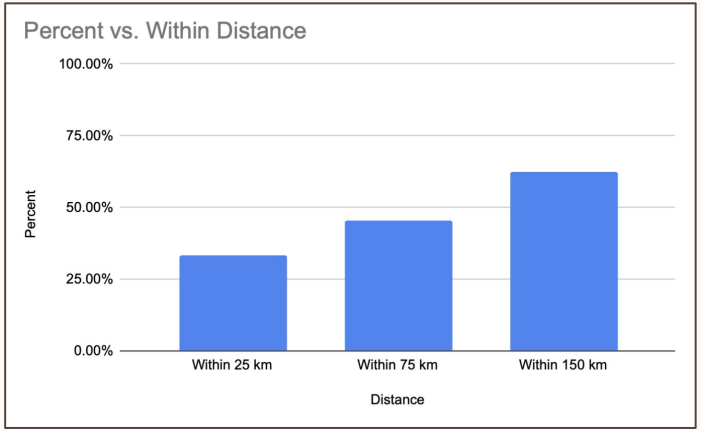
## Training + Validation



## Test Accuracy: 27.08%

## Avg Distance: 131 km

**Note: Great Britain is at most 500 km from East-West, and 1000 km from North-South**



Predicted: Poole
True: Poole

Predicted: York
True: York

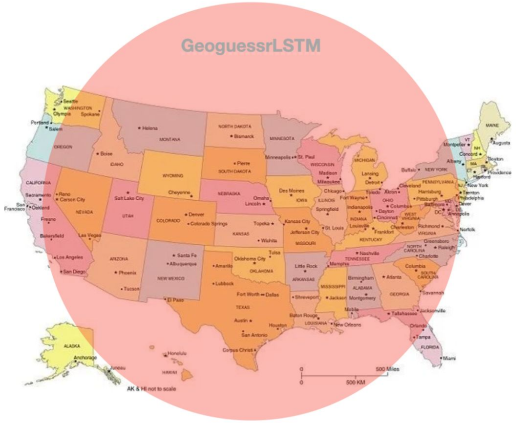
Predicted: Blackpool
True: Blackpool

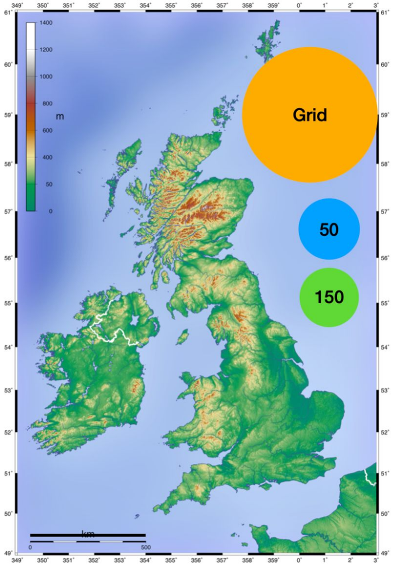
Predicted: Birmingham
True: Liverpool

Predicted: Manchester
True: Liverpool

# Model Comparison

- Visual comparison
- **GeoguessrLSTM:**
  - <u>USA</u>: 4500 km East-West, 1650 km North-South
  - Avg distance 1931 km
- **Our Model:**
  - <u>UK</u>: ~500 km East-West, 1000 km North-South
  - 50 Cities: avg 131 km
  - 150 Cities: avg 128 km
  - Grid: avg 292 km





04

# Conclusions

# Constraints and Future Work

- Ethics and privacy
- Applying model to other countries
- Human Geoguessr Experts
  - Provide basis for how such identification is possible
  - CLIP set basis for initial approach
  - Future models: use experts' strategies

# Thank You!