PIGEON: Predicting Image Geolocations

Lukas Haas, Michal Skreta, Silas Alberti
{lukhaas, mskreta, salberti}@stanford.edu

Stanford ENGINEERING
Computer Science

CS 330: Deep Multi-task and Meta Learning – Fall 2022

Executive Summary

- **Planet-scale image geolocalization** is considered to be a very **challenging** task, necessitating fine-grained understanding of **visual information** across countries, environments, and time.
- We present **PIGEON**, a **novel deep multi-task** model for planet-scale **Street View** image geolocalization that incorporates semantic **geocell creation** with label smoothing, conducts pretraining of a **CLIP vision transformer** on Street View images, and refines location predictions with **ProtoNets**.
- Motivated by the rising popularity of an online game **Geoguessr** with over 50 million players worldwide, we focus specifically on Street View images and create a **Google Chrome extension** for Geoguessr that deploys PIGEON against humans.
- We build the first AI model which consistently beats human players in Geoguessr, **ranking in the top 0.01% of players**.
- Partnering with the **CTO of Geoguessr** and **Google for Education**, we garner credits to query the **Street View API** and obtain a **novel planet-scale dataset of 400,000 images**.
- Our model achieves impressive results, aided by **positive multi-task transfer** in both an implicit and explicit multi-task setting; we attain **91.96% country accuracy** on our held-out set and **40.36% of our guesses are within 25 km of target**.
- Moreover, applying our pre-trained CLIP model (**StreetViewCLIP**) to **out-of-distribution benchmark datasets** in a **0-shot** setting achieves **near state-of-the-art results**.
- Our results represent an important **novel contribution** towards accurate **planet-scale** image geolocalization.

Relevant Background & Dataset

- The first modern attempt at planet-scale image geolocalization is attributed to **IM2GPS** in 2008 [1], a retrieval-based approach using nearest-neighbor search based on hand-crafted features.
- With the arrival of deep learning to computer vision, Google released a paper called **PlaNet** [2] that first applied **convolutional neural networks** to photo geolocalization.
- More recent work showed that **contextual knowledge** about the image scene can improve predictions [3], and that **vision transformers** and **multi-task** settings [4] contribute to superior performance, further accelerating research in the field.
- Given the **lack of publicly available** planet-scale Street View datasets for image geolocalization, we **sourced** a dataset of 1 million image locations from the Geoguessr game.
- We created a **novel dataset** of **400,000 images** from 100,000 locations **randomly sampled** from a **planet-scale** distribution.
- Finally, we **fundraised education credits from Google**, and wrote code to query the **Street View API**, obtaining four images per location with **randomly initialized compass** directions, with a sample location view presented in *Figure 1*.

Figure 1: Four images comprising a 360-degree panorama in Cuauhtemoc, Mexico in our dataset.

Technical Methodology

- The **technical novelty** of our image geolocalization predictor can be summarized by identifying **six distinct contributions**:

1 Semantic Geocell Creation

- Predicting coordinates directly does not work well, making geocell design **"crucial for performance"** [5].
- We use **planet-scale open-source administrative data** for **semantic** geocell creation and are the **first to employ Voronoi tessellation**.

2 Label Smoothing

- Although we posit a classification problem, our classes remain related to each other via **distance**. To that end, we implement **a smoothing of label distributions over neighboring cells** to **jointly train classes on a single sample**, visualized in *Figure 2*.

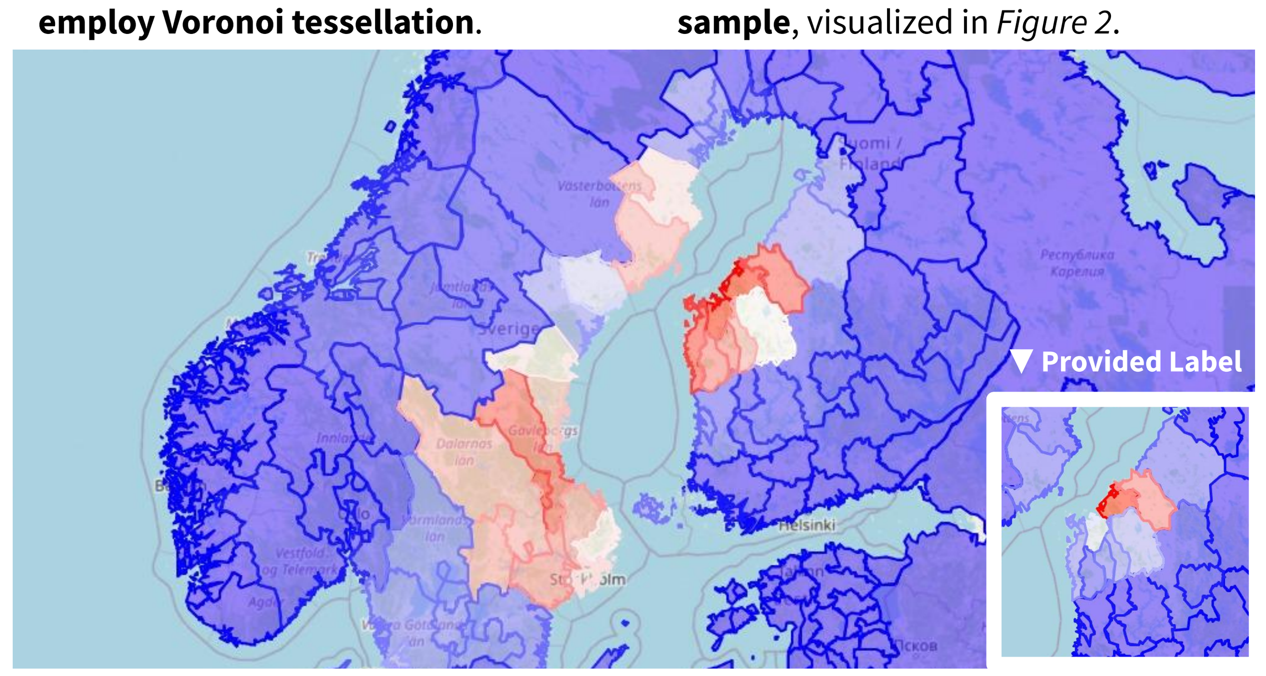

Figure 2: Distribution of probabilities over geocells for a true location in Finland.

3 Vision Transformer (CLIP)

- The **CLIP** vision transformer [6] is **used as a base** for all our models.
- CLIP has demonstrated to be a **great few-shot learner** which is important given the **diversity in our dataset** and the few samples per geocell (~40).

4 Contrastive Pretraining

- We **augment** our dataset with **geographic**, **demographic**, and **geological** auxiliary data.
- We **pretrain CLIP** in an **implicit, contrastive multi-task setting** by using **captions engineered from our auxiliary data** augmentations as visualized in *Figure 3*.

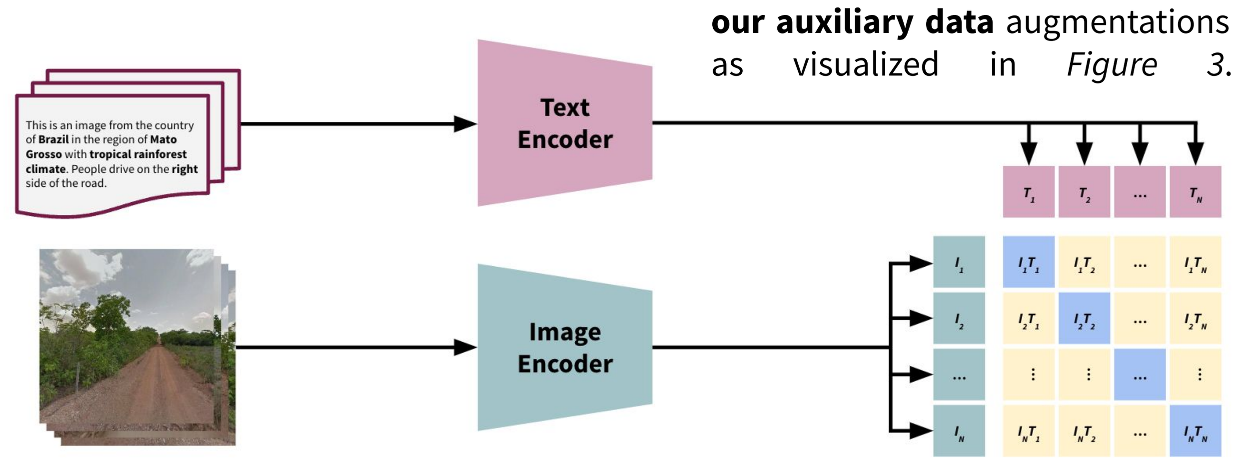

Figure 3: Contrastive pretraining of StreetViewCLIP in an implicit multi-task setting.

5 Multi-task Learning

- Our **multi-task** setup is made **explicit** by creating **task-specific heads** for climate variables, population density, elevation, and the month (season) of the year.
- We **unfreeze the last CLIP layer** to allow for **parameter sharing across tasks**.

6 ProtoNet Refinement

- Once our model selects a geocell, we perform **intra-geocell refinement** using **ProtoNets** [7].
- Tasks are proposed via **OPTICS clustering** in an **unsupervised** manner, visualized in *Figure 4*.
- Our refiner **optimizes** across the **top 5 proposed geocell candidates**.

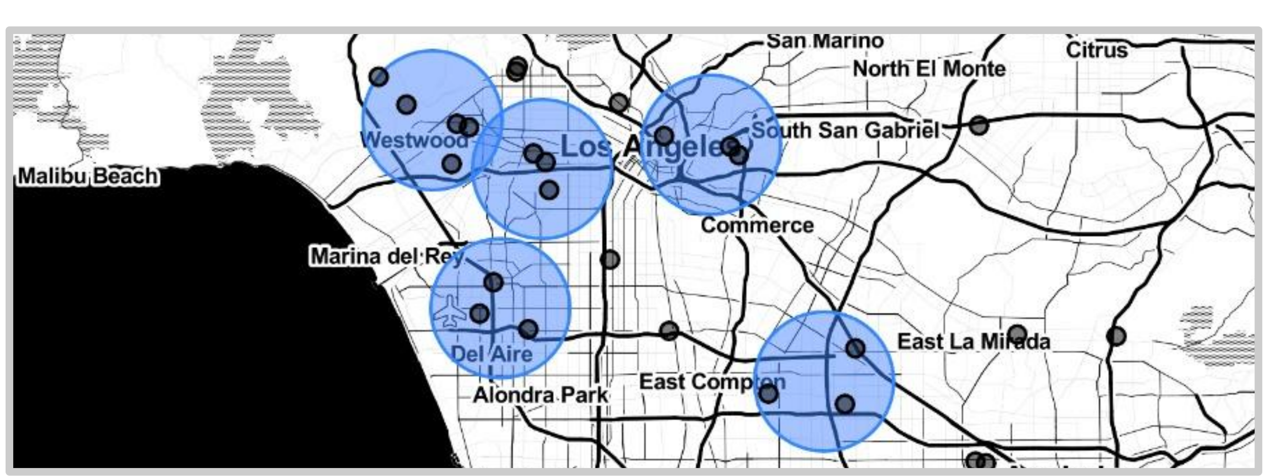

Figure 4: Visualized ProtoNet clusters in the Greater Los Angeles metropolitan area.

Experimental Results

- We perform **detailed ablation studies for our contributions**, with results summarized in *Table 1* which includes standard kilometer-based metrics in line with the literature.

Table 1: Multi-step ablation study on our modeling approach to image geolocalization.

Method	Distance (% @ km)				
	Street 1 km	*City* 25 km	*Region* 200 km	*Country* 750 km	*Continent* 2500 km
CLIP Base	1.28	24.08	55.38	80.20	92.00
+ Label Smoothing	0.92	24.18	59.04	82.84	92.76
+ Four-image Panorama	1.10	32.50	75.32	92.92	98.00
+ Fine-tuning Last CLIP Layer	1.10	32.74	75.14	93.00	97.98
+ Multi-task Parameter Sharing	1.18	33.22	75.42	93.42	98.16
+ Semantic Geocells	1.24	34.54	76.36	93.36	97.94
+ Contrastive CLIP Pretraining	1.32	35.56	**78.86**	**94.54**	98.54
+ ProtoNet Refinement	**5.36**	**40.36**	78.28	94.52	**98.56**

- For the calculation of distance, we exploit the Earth's spherical geometry using the **Haversine formula** in *Equation 1*.

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Equation 1: Haversine formula determining great-circle distance from geographic coordinates.

- **Beyond looking at the share of guesses** within a given distance, we see the results achieved by PIGEON in *Table 2*.

Table 2: Results from the ablation study beyond the standard distance metrics.

Method	Country Accuracy *%*	Mean km Error *km*	Median km Error *km*	Elevation Error *m*	Pop. Density Error *people/km²*	Temp. Error *°C*	Precipitation Error *mm/day*	Season Accuracy *%*	Climate Zone Accuracy *%*	Geoguessr Score *points*
CLIP Base	72.12	990.0	148.0	N/A	N/A	N/A	N/A	N/A	N/A	3,890
+ Label Smoothing	74.74	877.4	131.1	N/A	N/A	N/A	N/A	N/A	N/A	3,986
+ Four-image Panorama	87.64	315.7	60.81	N/A	N/A	N/A	N/A	N/A	N/A	4,442
+ Fine-tuning Last Layer	87.90	312.7	61.81	N/A	N/A	N/A	N/A	N/A	N/A	4,442
+ Multi-task Parameter Sharing	87.96	299.9	60.63	**141.7**	1,094	1.37	**14.48**	**45.74**	74.10	4,454
+ Semantic Geocells	89.36	316.9	55.51	147.1	**1,064**	**1.36**	14.71	**45.74**	**74.66**	**4,464**
+ Contrastive Pretraining	91.14	251.9	50.01	N/A	N/A	N/A	N/A	N/A	N/A	4,522
+ ProtoNet Refinement	**91.96**	**251.6**	**44.35**	N/A	N/A	N/A	N/A	N/A	N/A	4,525

- Notably, our best model achieves a **country accuracy of 91.96%** and a **median error of 44.35 kilometers**.
- Our results further show that geographical, demographic and geological features **can be inferred from Street View images**.
- We additionally test our pretrained **StreetViewCLIP** model on **benchmark image geolocalization datasets** that are **not Street View-specific**. By generating an exhaustive list of country captions, we query StreetViewCLIP to get country-level predictions which we then translate into coordinates. Our results (*Table 3*) are **near SOTA, but in 0-shot**.
- On the latest benchmark IM2GPS3K, StreetViewCLIP achieves an **accuracy of 52.79% for countries not seen during pretraining vs. 41.51% of accuracy for CLIP** for the same countries. Our results highlight that **contrastive pretraining is an effective technique for meta-learning**.

Table 3: Results from zero-shot learning with contrastive pretraining on benchmark datasets.

Benchmark	Method	Distance (% @ km) *Continent* 2500 km
IM2GPS	ViT Base [4]	80.70
	TransLocator [4]	**86.70**
	CLIP (0-shot)	77.22
	Street View CLIP (0-shot)	83.12
IM2GPS3K	ViT Base [4]	70.70
	TransLocator [4]	**80.10**
	CLIP (0-shot)	67.43
	Street View CLIP (0-shot)	76.44

Discussion of Results

- To assess the **interpretability** of our results, we plotted **attention attribution maps** over the images in our dataset, which we visualize in *Figure 5*. Interestingly, the model seems to be **able to learn** "metas" commonly used by players in the Geoguessr game, such as vegetation (left image) and utility poles (right image), **aiding model explainability**.
- Furthermore, we confirmed the accuracy of our results by deploying our model in the Geoguessr game, where our model consistently beats humans, **ranking in the Top 1,000 globally**.
- The results we achieved have vast **social impact** potential. By predicting **climate** based on images, we could be able to assess the risk to the consequences of climate change. Image geolocalization can also be used for attributing location to **archival images**, helping **historical research**, as well as in promoting **geography education** through gamified e-learning.
- Nevertheless, several **limitations** remain. Although PIGEON can successfully identify the vast majority of countries in which photos were taken, it still cannot be used at **extremely precise levels** (street level) that are necessary for detailed geo-tagging.

Figure 5: Attention attribution maps from Canada (left) and New Zealand (right).

Takeaways & Future Work

- Overall, **PIGEON** presents **multiple novel incremental improvements** to **multi-task image geolocalization**, including accurate **semantic geocell creation**, **pretrained vision transformers**, and **ProtoNet intra-geocell refinement**.
- **Going forward**, several extensions can be made to make image geolocalization more precise. Future models can detect **text** included in images to leverage **linguistic information** for predictions. **Road networks** and **compass directions** could further be used for intra-geocell refinement. In the long term, future work could go **beyond Street View**, with the models able to geolocate **any photo taken anywhere in the world**.

Acknowledgements

- We are deeply grateful to **Erland Ranvinge**, the **CTO of Geoguessr**, for responding to our outreach and for providing a sample of locations from the Competitive Duels mode in Geoguessr.
- We are deeply thankful to **Daniel Russell** and **Google** for supporting us with Google Cloud Education credits that allowed us to construct our novel dataset with images from the **Street View API**.
- Additional comments by **Lion Cassens** and **Victor de Fontnouvelle** were particularly helpful in pushing our thinking on strategic project choices.
- We are also grateful to **Prof. Chelsea Finn** and our mentor **Daniel Zeng** for helpful insights throughout the project.

References

[1] James Hays and Alexei A. Efros. im2gps: estimating geographic information from a single image. In Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2008.

[2] Tobias Weyand, Ilya Kostrikov, and James Philbin. PlaNet - Photo Geolocation with Convolutional Neural Networks. In Computer Vision – ECCV 2016, pages 37–55. Springer International Publishing, 2016.

[3] Eric Müller-Budack, Kader Pustu-Iren, and Ralph Ewerth. Geolocation Estimation of Photos using a Hierarchical Model and Scene Classification. In Proceedings of the European Conference on Computer Vision (ECCV), pages 563–579, 2018.

[4] Shraman Pramanick, Ewa M. Nowara, Joshua Gleason, Carlos D. Castillo, and Rama Chellappa. Where in the World is this Image? Transformer-based Geo-localization in the Wild, 2022.

[5] Jonas Theiner, Eric Müller-Budack, and Ralph Ewerth. Interpretable Semantic Photo Geolocation. In 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 1474– 1484, 2022.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. CoRR, abs/2103.00020, 2021.

[7] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning, 2017.