Interpretable Semantic Photo Geolocation

Jonas Theiner[1], Eric Müller-Budack[2], and Ralph Ewerth[1,2]
[1] L3S Research Center, Leibniz University Hannover, Hannover, Germany [2] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany

Problem Formulation

- **Task**: Predicting the location of a photo at planet-scale without any restrictions
- Humans are far worse than all DNN-based solutions
- DNNs lack of explainability and interpretability

→ How to improve the explainability and interpretability for the gelocation estimation task?

Related Work

- Base Idea for deep geolocation estimation:
  - Divide earth in cells (=*partitioning*)
  - Train a CNN on a classification task (hierarchical)
  - Key for success: Construction of the partitioning

→ Classes are based on the training distribution
→ State-of-the-art: Boundaries are arbitrary shaped (Fig. 1 left two)

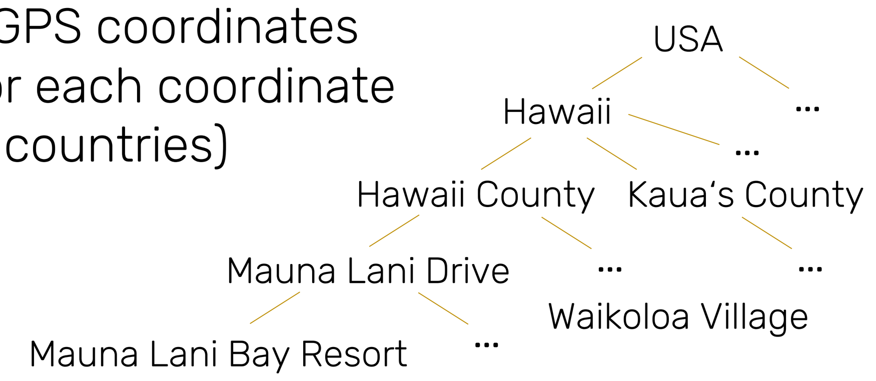
Fig. 1: Partitioning of the earth's survace in cells based on the dataset distribution. Right: Our semantic partitioning (SemP)

Contributions

1. Semantic Partitioning (**SemP**)
   - We incorperate data from Open Street Map (OSM)
   - Making model's output & learning procedure more natural
   - State-of-the art performance on benchmark datasets
2. Investigate the post-hoc interpretability (**concept influence** metric)
   - We measure the influence of semantic visual concepts
   - We provide insights which features contribute to correct and incorrect (and misleading) predictions

Semantic Partitioning (SemP)

- Reverse Geocoding for a set of GPS coordinates
  →Yields an address vector for each coordinate
- Build a multi-tree (i.e., roots are countries)

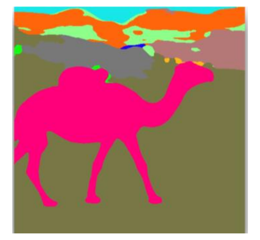


- Partitioning construction for a specific spatial level:
  →Keep finest location in address vector with at least τ_{min} images
  →Each location can be assigned to one parent

- Learning:
  - CNN with multi-head (one head per partitioning)
  - Jointly learns to localize at multiple spatial scales
  - Loss:
    - Standard classification loss (cross-entropy with softmax)
    - Sum per partitioning

Concept Influence (ci)

Given a segmentation and explanation map for an input image,

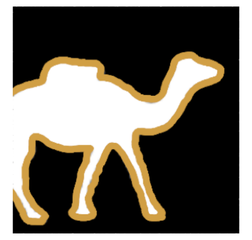


$m(\mathcal{S}, x)$

$e(\hat{c}, x)$ *w.r.t. model's prediction*

we define

$$ci\Big(\underbrace{m(s,x)}_{\text{mask for one concept}}, \overbrace{e_k(\hat{c},x)}^{\text{mask with top-}k\text{ values}}\Big) = \frac{\overbrace{\frac{1}{k}\sum_{i=0}^{w}\sum_{j=0}^{h} e_k(\hat{c},x)_{ij} \wedge m(s,x)_{ij}}^{\text{top-}k\text{ intersection of two binary masks}}}{\underbrace{\frac{1}{wh}\sum_{i=0}^{w}\sum_{j=0}^{h} m(s,x)_{ij}}_{\text{relative concept size}}}$$



Morphological dilation β [px] to cover boundaries: $m^{\beta}(s = \text{"camel"}, x)$

Experimental Results

Concept Influence

Procedure:
- Aggregate concept influence (per concept)
  → for a set of images
- Group by geolocation error
- Order by magnitude

Findings:
- Border between concepts is of interest
  → a skyline, or mountain range
- Identified concepts which give visual cues
  - to a concrete location
  → e.g., tower, building, sky
  - or to rough region
  → e.g., windowpane, tree, flower

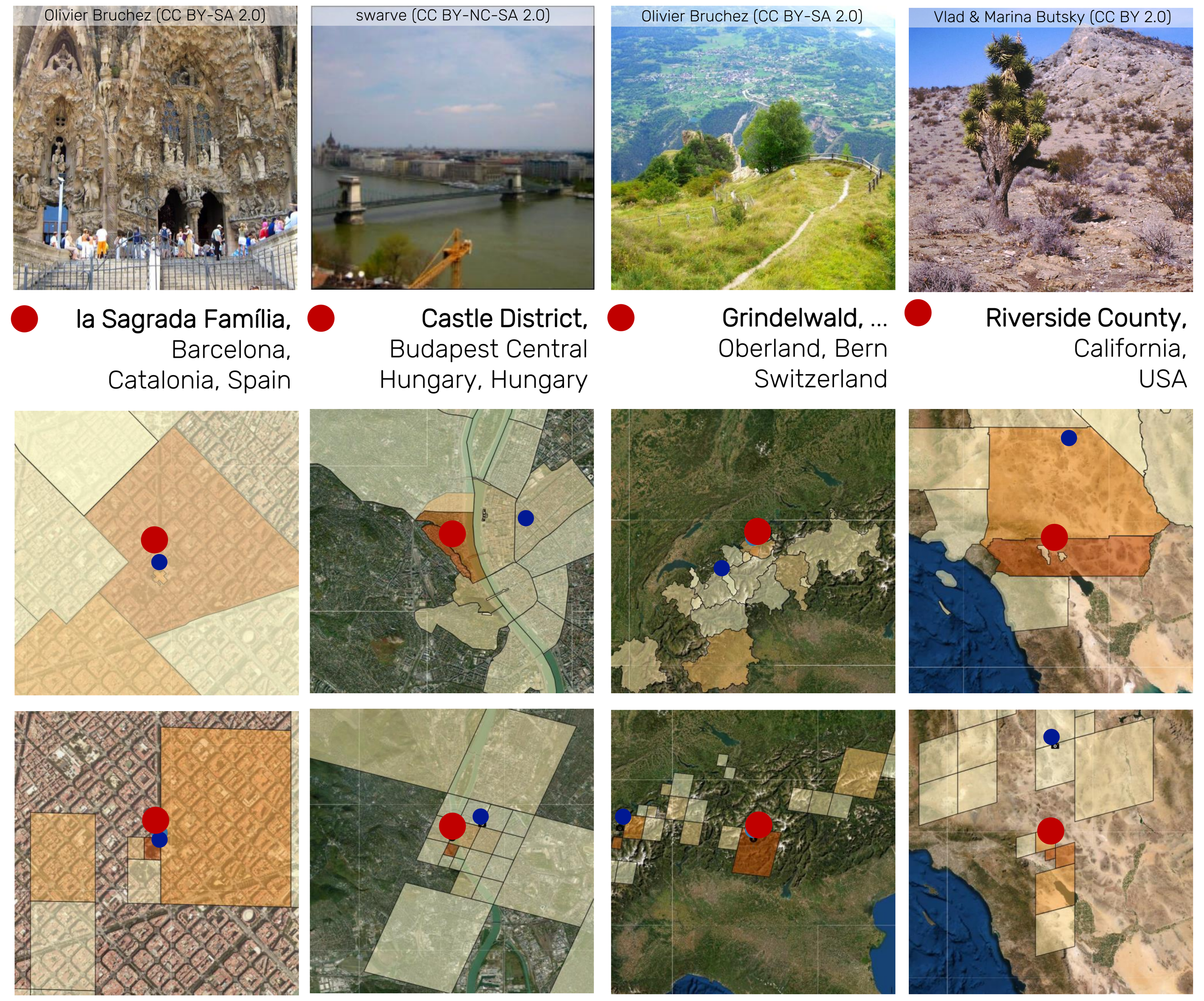


Fig. 2: Qualitative results for **SemP** (top) vs. *s2(M,f*)* (bottom): Visualized top-25 cells; ● predicted location; ● ground truth

Semantic Partitioning

- We compare **SemP** on a fixed setup with the *s2* partitioning
  - Fair comparison (same data, backbone, training, inference)
  - No need to use ensembles or additional retrieval methods
- Directly evaluate a multi-partitioning variant: (*s2(M,f*)* [3]) → leads to state-of-the art results
- Quantitative evaluation on three test sets *(Im2GPS3k* in Tab. 1*)*
- **SemP** intuitively provides better comprehensible output (Fig. 2)

Tab. 1: Results on the *Im2GPS3k* dataset. Percentage of images localized within given radius. Only the partitioning differs.

Method	Street 1 km	City 25 km	Region 200 km	Country 750 km	Continent 2500 km
PlaNet [1] (repr. [2])	8.5 %	24.8 %	34.3 %	48.4 %	64.6 %
CPlaNet [2]	10.2 %	26.5 %	34.6 %	48.6 %	64.6 %
s2(M,f*) (repr.)	11.5 %	30.8 %	41.0 %	55.7 %	70.8 %
SemP({100, 125, 250},f)	12.5 %	**31.4 %**	**42.7 %**	**57.3 %**	**72.0 %**
SemP({50, 75, 100},f)	**13.5 %**	30.8 %	41.2 %	54.7 %	70.2 %

Source code and datasets available at:
https://github.com/jtheiner/semantic_geo_partitioning

Contact

theiner@l3s.de
{eric.mueller, ralph.ewerth}@tib.eu

References

[1] Weyand, T., Kostrikov, I., & Philbin, J. Planet-photo geolocation with convolutional neural networks. ECCV'16 (pp. 37-55). Springer, Amsterdam
[2] Seo, P. H., Weyand, T., Sim, J., & Han, B. Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps. ECCV'18 (pp. 536-551). Springer, Munich
[3] Müller-Budack, E., Pustu-Iren, K., & Ewerth, R. Geolocation estimation of photos using a hierarchical model and scene classification. ECCV'18 (pp. 563-579). Springer, Munich