xNet+SC: Classifying Places Based on Images by Incorporating Spatial Contexts

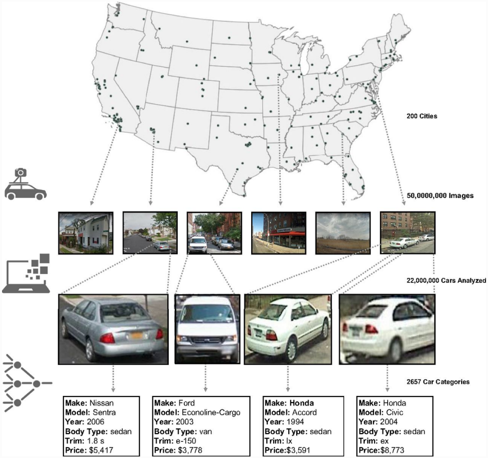
GIScience 2018, August 2018

Bo Yan Krzysztof Janowicz **Gengchen Mai** Rui Zhu

STKO Lab, University of California, Santa Barbara

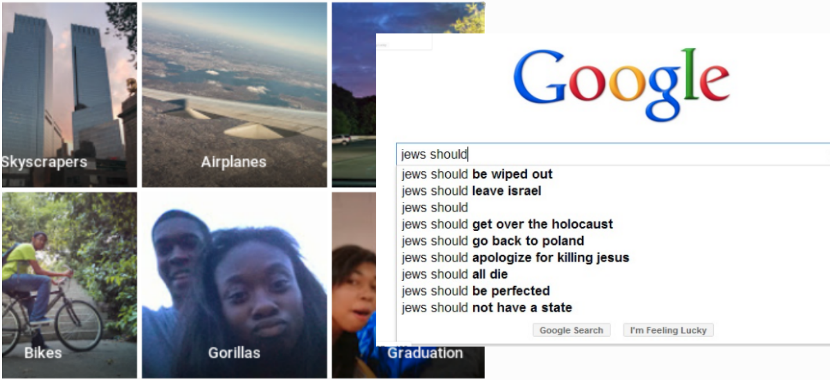
Motivation

- Recent advancements in computer vision such as **deep convolutional neural networks**, have quickly permeated GIScience field.
- E.g., Inferring socioeconomic attributes from cars detected in Google Street View images using deep learning (Timnit Gebru et al. 2017)



Challenges

- Training data can be biased, e.g., representational bias



- Google erroneously labeled photos of black people as *gorillas*, no robust solutions have been established besides simply removing such labels

Challenges

- The visual signal alone may not be sufficient, e.g., when trying to **classify place types based on facades** or interiors
  - Variability of restaurant facades

  - Library or museum?

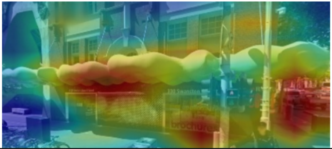
- These cases benefit from more **contextual information**, e.g., by making them **spatially explicit**

Place Facade and Interior Image Classification Task

- Place365-CNN (Bolei Zhou et al. 2017)
  - 10 million images for scene recognition
  - 400+ unique **scene categories** ; many not in line with what we would call place types in GIScience, e.g., *Wave*
  - Different Image classification models
    - AlexNet
    - GoogLeNet
    - VGG16
    - ResNet18
    - ResNet50
    - DenseNet161

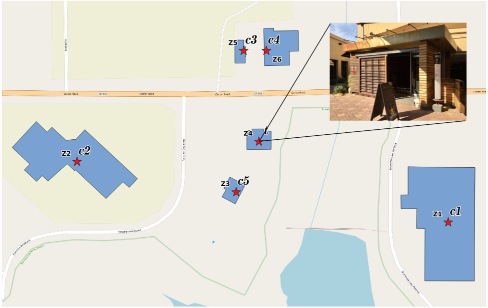
Predictions:

- **Type of environment:** outdoor
- **Scene categories:** diner/outdoor (0.160), restaurant_patio (0.124)
- **Scene attributes:** man-made, no horizon, natural light, open area, cloth, pavement, glass, metal, driving
- **Informative region for predicting the category 'diner/outdoor' is:**



Adding Spatial Context

- Address the outlined challenges by adding spatial context signal, i.e., making learning **spatially explicit**



- Here we use **nearby facades** classified (with potential error) before, e.g., by driving-by, making use of **spatial signatures**

Place Type Alignment

- Class label alignment between Yelp and the Place365 model

Class label	**Places365-CNN category**
Amusement Parks	amusement_park
Bakeries	bakery
Bookstores	bookstore
Churches	church
Cinema	movie_theater
Dance Clubs	discotheque
Drugstores	drugstore, pharmacy
Hospitals	hospital, hospital_room
Hotels	hotel, hotel_room
Jewelry	jewelry_shop
Libraries	library
Museums	museum, natural_history_museum, science_museum
Restaurants	fastfood_restaurant, restaurant, restaurant_kitchen, restaurant_patio
Shoe Stores	shoe_shop
Stadiums & Arenas	stadium

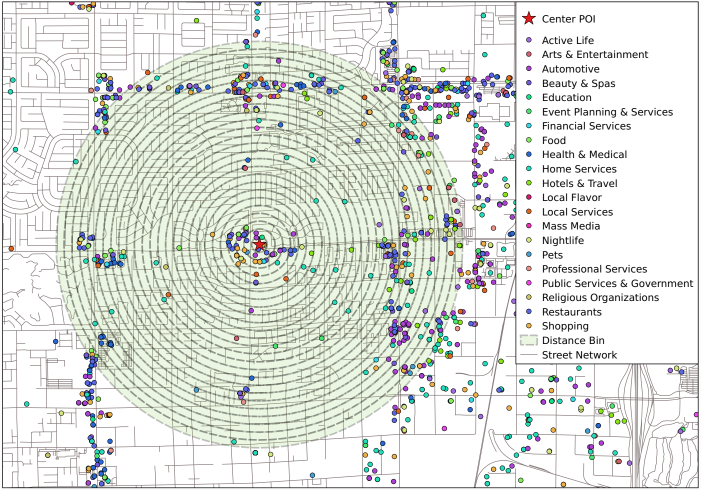
- 15 classes, 50 images/class (Yelp, Google Street View)

Method

- **Image classification models** (baseline models)
  - AlexNet (8 layers)
  - ResNet with 18 layers
  - ResNet with 50 layers
  - DenseNet with 161 layers
- **Spatial context models**
  - Spatial relatedness
  - Spatial co-location
  - Spatial sequence pattern
- **Combination approach**
  - Search re-ranking
  - Bayesian

Spatial Relatedness

- Place2Vec: Learn POI type embeddings using spatial contexts

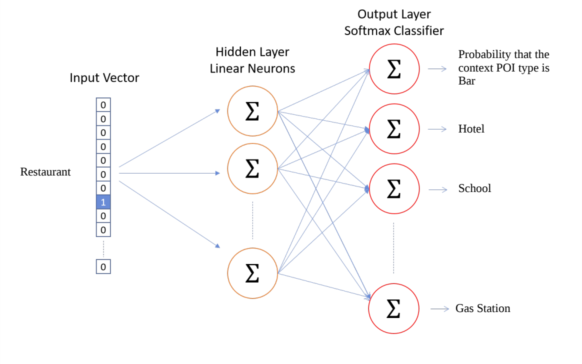


Spatial Relatedness

- Place2Vec (Skip-Gram) (Bo Yan et al. 2017)
  - **Predicts context POI types given center POI types**

$$\hat{y}_{context} = P(t_1, t_2, t_3, \ldots, t_m | t_{center}) \quad (1)$$

$\hat{y}_{context}$ is the *predicted probability* of context POI types, t stands for POI type



Spatial Relatedness

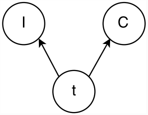
- Place type embeddings (from Place2Vec)
- **Search Re-ranking:** re-rank CNN score of different place types by using spatial relatedness score
  - Calculate spatial context embeddings (averaged place type embeddings)
  - Obtain raw scores for each candidate image label using cosine similarity
  - Normalize raw scores to obtain spatial relatedness scores

$$s_i = \omega^v s_i^v + \omega^r s_i^r \tag{2}$$

where s_i, s_i^v, and s_i^r are the combination score, CNN score, and spatial relatedness score for label i respectively, ω^v and ω^r are the weights for the CNN component and spatial relatedness component, and $\omega^v + \omega^r = 1$.

Spatial Co-location

- **Probabilistic graphical model**



- **Use spatial co-location as Bayesian prior**

$$\begin{aligned} P(t|I,C) &= \frac{P(I,C|t)P(t)}{P(I,C)} = \frac{P(I|t)P(C|t)P(t)}{P(I,C)} \\ &= \frac{P(t|I)P(I)}{P(t)} \frac{P(t|C)P(C)}{P(t)} \frac{P(t)}{P(I,C)} \\ &\propto \frac{P(t|I)}{P(t)} P(t|C) \end{aligned} \tag{3}$$

where I is the image, $P(t|I)$ can be obtained from the CNN model, $P(t|C)$ is the spatial context prior obtained from Eq. 5, and $P(t)$ is the label (type) prior

Spatial Co-location

- Frequency of **co-occurrence** of different labels/POI **types** (restaurant, bar, hotel, school, etc) in space (similar to traditional count-based language models)
- Two assumptions
  - Bag-of-words assumption: distance doesn't matter

$$P(c_i|t) = \frac{count(c_i, t)}{count(t)} \tag{4}$$

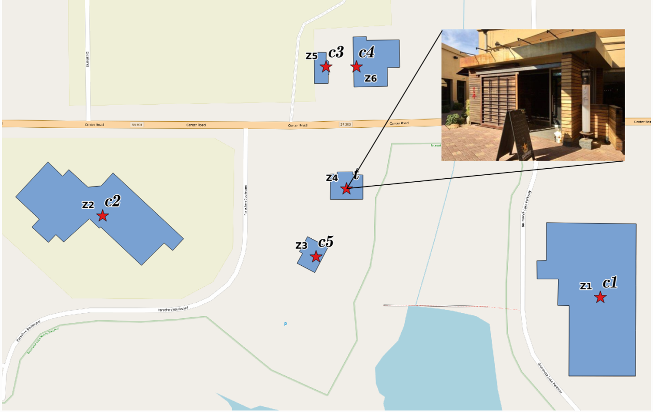
    where c_i is the neighbor type and t is the candidate image label (type)
  - Naive Bayes assumption: no spatial interaction between the context POIs

$$\begin{aligned} P(t|C) = P(t|c_1, c_2, ..., c_n) &= \frac{P(t)\prod_{i=1}^{n} P(c_i|t)}{P(c_1, c_2, c_3, ..., c_n)} \\ &\propto P(t)\prod_{i=1}^{n} P(c_i|t) \end{aligned} \tag{5}$$

    where $C = c_1, c_2, c_3, ..., c_n$ is the spatial context information

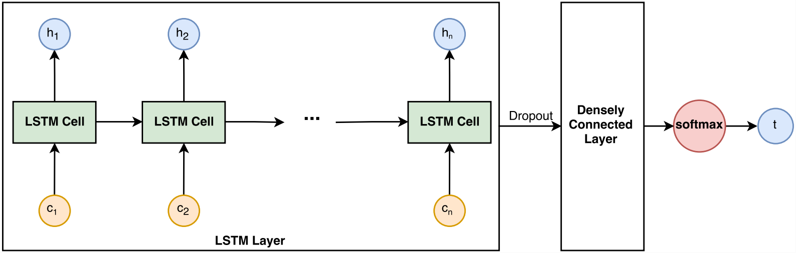
Spatial Sequence Pattern

- Collapse 2D geographic space into 1D sequence
  - **Distance-based**
  - **Morton order-based** (Space-filling curve)



Spatial Sequence Pattern

- Long Short-Term Memory (LSTM)



Use LSTM to obtain spatial context prior
$P(t|C) = P(t|c_1, c_2, c_3, \dots, c_n)$

- Calculate $P(t|I, C)$ using Bayes rule

Classification Result

- 15 classes, 50 images/class (Yelp, Google Street View)
- **Mean Reciprocal Rank**

MRR	AlexNet	ResNet18	ResNet50	DenseNet161
Baseline	0.27	0.28	0.31	0.31
Relatedness	0.27	0.28	0.31	0.32
Co-location	0.30	0.31	0.31	0.32
Sequence Pattern (Random)	0.38	0.40	0.42	0.42
Sequence Pattern (Distance)	**0.41**	**0.42**	**0.44**	**0.44**
Sequence Pattern (Morton order)	0.39	**0.42**	0.43	0.43

- **Accuracy@1**

Accuracy@1	AlexNet	ResNet18	ResNet50	DenseNet161
Baseline	0.07	0.07	0.09	0.09
Relatedness	0.07	0.07	0.09	0.09
Co-location	0.15	0.17	0.17	0.17
Sequence Pattern (Random)	0.18	0.18	0.19	0.20
Sequence Pattern (Distance)	**0.20**	**0.20**	**0.22**	**0.22**
Sequence Pattern (Morton order)	0.19	**0.20**	**0.22**	**0.22**

- **Our model outperform the baseline model by 40% for MRR and double Accuarcy@1.**

Classification Result

- Some intuitive examples

Figure 5 From left to right, images of a restaurant, a hotel, and a museum from Yelp, Google Street View, and Google Maps respectively. The first image is incorrectly classified as library using all 4 CNN models and it is correctly classified as restaurant using the spatial sequence pattern (distance) models. The second image is classified as hospital and library by the original CNN models and is classified as hotel by the spatial sequence pattern (distance) models. For the third image the correct label museum is in the third position in the label rankings of all 4 CNN models while, using the spatial sequence pattern (distance) models, ResNet18 and ResNet50 can correctly label it and in the label rankings of AlexNet and DenseNet161 museum is in the second position.

Conclusion

- **Classifying place types** according to images of **facades** and interiors is hard
- Instead of purely rely on visual signal, combining it with spatial contexts (which is in most cases avaiabale anyway) leads to substantial improvements, e.g., increasing **MRR by over 40%** and **doubling Accuracy@1**
- Complex spatial sequence patterns can be captured using LSTM
- For future work, we can **relax the need of POI datasets** by using the classification results and uncertainty of images of nearby places to improve estimation of the currently seen place (modify our methods to work in a **drive-by-typing** mode)
- Test what we call the **generalized spatial context hypothesis**, namely that we can go beyond facades but apply the same idea of spatial context to tree, cars, and so on.