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

Bo Yan
STKO Lab, University of California, Santa Barbara, USA
boyan@geog.ucsb.edu
https://orcid.org/0000-0002-4248-7203

Krzysztof Janowicz
STKO Lab, University of California, Santa Barbara, USA
jano@geog.ucsb.edu

Gengchen Mai
STKO Lab, University of California, Santa Barbara, USA
gengchen@geog.ucsb.edu

Rui Zhu
STKO Lab, University of California, Santa Barbara, USA
ruizhu@geog.ucsb.edu

Abstract

With recent advancements in deep convolutional neural networks, researchers in geographic information science gained access to powerful models to address challenging problems such as extracting objects from satellite imagery. However, as the underlying techniques are essentially borrowed from other research fields, e.g., computer vision or machine translation, they are often not spatially explicit. In this paper, we demonstrate how utilizing the rich information embedded in spatial contexts (SC) can substantially improve the classification of place types from images of their facades and interiors. By experimenting with different types of spatial contexts, namely spatial relatedness, spatial co-location, and spatial sequence pattern, we improve the accuracy of state-of-the-art models such as ResNet – which are known to outperform humans on the ImageNet dataset – by over 40%. Our study raises awareness for leveraging spatial contexts and domain knowledge in general in advancing deep learning models, thereby also demonstrating that theory-driven and data-driven approaches are mutually beneficial.

2012 ACM Subject Classification Computing methodologies → Computer vision tasks, Computing methodologies → Neural networks, Theory of computation → Bayesian analysis

Keywords and phrases Spatial context, Image classification, Place types, Convolutional neural network, Recurrent neural network

Digital Object Identifier 10.4230/LIPIcs.GIScience.2018.17

1 Introduction

Recent advancements in computer vision models and algorithms have quickly permeated many research domains including GIScience. In remote sensing, computer vision methods facilitate researchers to utilize satellite images to detect geographic features and classify land use [5, 26]. In urban planning, researchers collect Google Street View images and apply computer vision algorithms to study urban change [22]. In cartography, pixel-wise segmentation has been adopted to extract lane boundary from satellite imagery [32] and deep convolutional neural network (CNN) has been utilized to recognize multi-digit house numbers from Google Street View images [10]. These recent breakthroughs in computer

10th International Conference on Geographic Information Science (GIScience 2018).
Editors: Stephan Winter, Amy Griffin, and Monika Sester; Article No. 17; pp. 17:1–17:15
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

vision are achieved, in equal parts, due to advances in deep neural networks as well as the ever-increasing availability of extensive training datasets. For example, the classification error in the latest image classification challenge using the ImageNet dataset is down to about 0.023.[1]

However, such impressive results do not imply that these models have reached a level in which no further improvement is necessary or meaningful. On the contrary, such deep learning models which primarily depend on visual signals are susceptible to error. In fact, studies have shown that deep (convolutional) neural networks suffer from a lack of robustness to adversarial examples and a tendency towards biases [25]. Researchers have discovered that, by incorporating adversarial perturbations of inputs that are indistinguishable by humans, the most advanced deep learning models which have achieved high accuracy on test sets can be easily fooled [6, 11, 28]. In addition, deep learning models are also vulnerable to biased patterns learned from the available data and these biases usually resemble many unpleasant human behaviors in our society. For instance, modern neural information processing systems such as neural network language models and deep convolutional neural networks have been criticized for amplifying racial and gender biases [3, 4, 25, 33]. Such biases, which can be attributed to a discrepancy between the distribution of prototypical examples and the distribution of more complex real world systems [16], have already caused some public debates. To give a provocative example, almost three years after users revealed that Google erroneously labeled photos of black people as “gorillas”, no robust solutions have been established besides simply removing such labels for now.[2]

The above-mentioned drawbacks are being addressed by improvements to the available training data as well as the used methods [23, 3]. In our work, we follow this line of thought to help improve image classification. In our case, these images depict the facades or interiors of different types of places, such as restaurants, hotels, and libraries. Classifying images by place types is a hard problem in that more often than not the training image data is inadequate to provide a full visual representation of different place types. Solely relying on visual signals, as most deep convolutional neural networks do, falls short in modeling the feature space as a result. To give an intuitive example, facades of restaurants may vary substantially based on the type of restaurant, the target customers, and the surrounding. Their facade may be partially occluded by trees or cars, may be photographed from different angles and at different times of the day, and the image may contain parts of other buildings. Put differently, the principle of spatial heterogeneity implies that there is considerable variation between places of the same type.

To address this problem and improve classification accuracy, we propose to go beyond visual stimuli by incorporating spatial contextual information to help offset the visual representational inadequacy. Although data availability is less of an issue nowadays, the biased pattern in the data poses a real challenge, especially as models such as deep convolutional neural networks take a very long time to train. Instead of fine-tuning the parameters (weights) by collecting and labeling more unbiased data, which are very resource-consuming, we take advantage of external information, namely spatial context. There are many different ways one can model such context; in this work, we focus on the types of nearby places. We explore and compare the value of three different kinds of spatial context, namely spatial relatedness, spatial co-location, and spatial sequence pattern.

We combine these context models with state-of-the-art deep convolutional neural network models using search re-ranking algorithms and Bayesian methods. The result shows that,

[1] http://image-net.org/challenges/LSVRC/2017/results#loc
[2] https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

by considering more complex spatial contexts, we can improve the classification accuracy for different place types. In fact, our results demonstrate that a *spatially explicit* model [9], i.e., taking nearby places into account when predicting the place type from an image, improves the accuracy of leading image classification models by at least 40%. Aside from this substantial increase in accuracy, we believe that our work also contributes to the broader and ongoing discussion about the role of and need for theory, i.e., domain knowledge, in machine learning. Finally, and as indicated in the title, our spatial context (*SC*) models, can be added to any of the popular CNN-based computer vision models such as AlexNet, ResNet, and DenseNet – abbreviated to *xNet* here.

The remainder of this paper is organized as follows. Section 2 provides an overview of existing work on spatial context and methods for incorporating spatial information into image classification models. Section 3 presents the image classification tasks and provides information about the convolutional neural network models used in our study. Section 4 explains in detail three different levels of spatial context and ways to combine them in image classification models. Section 5 presents the results. Finally, Section 6 concludes the research and points to future directions.

2 Related Work

There is a large body of work that utilizes spatial context to improve existing methods and provide deeper insights into the rich semantics of contextual information more broadly. For instance, spatial context has been recognized as a complementary source of information in computational linguistics. By training word embeddings for different place types derived from OpenStreetMap (OSM) and Google Places, Cocos and Callison-Burch [7] suggested that spatial context provides useful information about semantic relatedness. In Points of Interest (POI) recommendation, spatial context has been used to provide latent representations of POI, to facilitate the prediction of future visitors [8], and to recommend similar places [34]. By implementing an information theoretic and distance-lagged augmented spatial context, Yan et al. [30] demonstrated that high-dimensional place type embeddings learned using spatial contexts can reproduce human-level similarity judgments with high accuracy. The study showed that such a spatially explicit Place2Vec model substantially outperforms Word2Vec-based models that utilize a linguistic-style of context. Liu et al. [21] used spatial contexts to measure traffic interactions in urban area. In object detection, Heitz and Koller [13] leveraged spatial contexts in a probabilistic model to improve detection result. Likewise, by embracing the idea that spatial context provides valuable extrinsic signals, our work analyzes different kinds of spatial contexts and tests their ability to improve image classification of place types.

Existing work on image classification has realized the importance of including a geographic component. One direction of research focused on enriching images with geospatial data. Baatz et al. [1] took advantage of digital elevation models to help geo-localize images in mountainous terrain. Lin et al. [20] made use of land cover survey data and learned the complex translation relationship between ground level images and overhead imagery to extend the reach of image geo-localization. Instead of estimating a precise geo-tag, Lee et al. [19] trained deep convolutional neural networks to enrich a photo with geographic attributes such as elevation and population density. Another direction of research (which is more similar to our study) focused on utilizing geographic information to facilitate image classification. In order to better understand scenes and improve object region recognition, Yu and Luo [31] exploited information from seasons and location proximity of images using a probabilistic graphical model. Berg et al. [2] combined one-vs-most image classifiers with spatiotemporal class priors to address the problem of distinguishing images of highly similar bird species.

Tang et al. [29] encoded geographic features extracted from GPS information of images into convolutional neural networks to improve classification results.

Our work differs from the existing work in that we explicitly exploit the distributional semantics found in spatial context [30] to improve image classification. Following the linguistic mantra that one *shall know a word by the company it keeps*, we argue that one can know a place type by its neighborhood's types. This raises the interesting question of how such a neighborhood should be defined. We will demonstrate different ways in which spatial contextual signals and visual signals can be combined. We will assess to what extent different kinds of spatial context, namely spatial relatedness, spatial co-location, and spatial sequence pattern, can provide such neighborhood information to benefit image classification.

3 Image Classification

In this section, we first describe the image classification task and the data we use. The task is similar to scene classification but we are specifically interested in classifying different business venues as opposed to natural environment. Then we explain four different deep convolutional neural networks that solely leverages the visual signals of images. These convolutional neural network models are later used as baselines for our experiment.

3.1 Classification Task

Our task is to classify images into one of the several candidate place types. Because we want to utilize the spatial context in which the image was taken, we need to make sure each image has a geographic identifier, e.g. geographic coordinates, so that we are able to determine its neighboring place and their types. In order to classify place types of images, we consider the scene categories provided by Zhou et al. [35] as they also provide pretrained models (Places365-CNN) that we can directly use.[3] Without losing generality, we select 15 place types as our candidate class labels. The full list of class labels and their alignment with the categories in Places365-CNN is shown in Table 1. For each candidate class, we selected 50 images taken in 8 states[4] within the US by using Google Maps, Google Street View, and Yelp. These images include both indoor and outdoor views of each place type. Please note that classifying place types from facade and interior images is a hard problem and even the most sophisticated models only distinguish a relatively small number of place types so far which is nowhere near the approximately 420 types provided by sources such as Foursquare. Places365, for instance, offers 365 classes but many of these are scenes or landscape features, such as waves, and not POI type, such as cinemas, in the classical sense.

3.2 Convolutional Neural Network Models

To establish baselines for our study, we selected several state-of-the-art image classification models, namely deep convolutional neural networks. Unlike traditional image classification pipelines, CNNs extract features from images automatically based on the error messages that are backpropagated through the network, thus fewer heuristics and less manual labor are needed. Contrary to densely connected feedforward neural networks, CNN adopts parameter sharing to extract common patterns which help capture translation invariance and creates sparse connections which result in fewer parameters and being less prone to overfitting.

[3] `https://github.com/CSAILVision/places365/blob/master/categories_places365.txt`

[4] Arizona, Illinois, Nevada, North Carolina, Ohio, Pennsylvania, South Carolina, and Wisconsin

Table 1 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

The architecture of CNNs has been revised numerous times and has become increasingly sophisticated since its first appearance about 30 years ago. These improvements in architecture have made CNN more powerful as can be seen in the ImageNet challenge. Some of the notable architectures include: LeNet [18], AlexNet [17], VGG [24], Inception [27], ResNet [12], and DenseNet [15]. We selected AlexNet, ResNet with 18 layers (ResNet18), ResNet with 50 layers (ResNet50), and DenseNet with 161 layers (DenseNet161). AlexNet is among the first deep neural networks that increased the classification accuracy on ImageNet by a significant amount compared with traditional classification approaches. By using skip connections to create residual blocks in the network, ResNet makes it easy to learn identity functions that help with the vanishing and exploding gradient problems when the network goes deeper. In DenseNet, a dense connectivity pattern is created by connecting every two layers so that the error signal can be directly propagated to earlier layers, parameter and computational efficiency can be increased, and low complexity features can be maintained [15]. These models were trained on 1.8 million images from the Places365-CNN dataset. We used the pretrained weights for these models.

4 Spatial Contextual Information

In this section, we introduce three different kinds of spatial contexts and explore ways in which we can combine them with the CNN models in order to improve image classification. The first type of spatial context is spatial relatedness, which measures the extend to which different place types relate with each other. The second type of spatial context is spatial co-location, which considers what place types tend to co-occur in space and the frequency they cluster with each other. The third type of spatial context is spatial sequence pattern which considers both spatial relatedness and spatial co-location. In addition, spatial sequence pattern considers the interaction between context place types and the inverse relationship between distance and contextual influence. We use POIs provided by Yelp as dataset.[5]

[5] https://www.yelp.com/dataset

4.1 Spatial Relatedness

Since the output of CNN is the probability score for each class label, it is possible to interpret our task as a ranking problem: given an image, rank the candidate class labels based upon the visual signal and spatial context signal. For the visual signal, we can obtain the ranking scores (probability scores) from the CNN architectures mentioned in Section 3. Since the original CNN models has 365 labels, we renormalize the probability scores for each candidate place type by the sum of the 15 candidate ranking scores so that they sum up to 1. This renormalization procedure is also applied to the other two spatial context methods explained in Section 4.2 and Section 4.3. We will refer to the renormalized scores as CNN scores in this study. For the spatial context signal, the ranking scores are calculated using the place type embeddings proposed in [30]. These embeddings capture the semantics of different place types and can be used to measure their similarity and relatedness. In this regard, the task is equivalent to a re-ranking problem, which adjusts the initial ranking provided by the visual signal using auxiliary knowledge, namely the spatial context signal. Intuitively, the extent to which the visual signals from the images match with different place types and the level of relevance of the surrounding place types with respect to candidate place types jointly determine the final result.

Inspired by search re-ranking algorithms in information retrieval, we use a *Linear Bimodal Fusion* (LBF) method (here essentially a 2-component convex combination), which linearly combines the ranking scores provided by the CNN model and the spatial relatedness scores, as shown in Equation 1.

$$s_i = \omega^v s_i^v + \omega^r s_i^r \tag{1}$$

where s_i, s_i^v, and s_i^r are the LBF score, CNN score, and spatial relatedness score for place type i respectively, ω^v and ω^r are the weights for the CNN component and spatial relatedness component, and $\omega^v + \omega^r = 1$. The weights here are decided based on the relative performance of individual components. Specifically, the weight is determined using Equation 2.

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

where acc^v and acc^r are the accuracies for CNN and spatial relatedness measurements for the image classification task. Intuitively, this means that we have higher confidence if the component performs better on its own and want to reflect such confidence using the weight in the LBF score.

In order to calculate the spatial relatedness scores, we use cosine similarity to measure the extend to which each candidate class embedding is related with the spatial context embedding of an image in a high dimensional geospatial semantic feature space. Following the suggestions in [30], we use a concatenated vector of 350 dimensions (i.e., 70D vectors for each of 5 distance bins) as the place type embeddings. The candidate class embeddings can be retrieved directly. Then we search for the nearest n POIs based on the image location, determine the place types of these n POIs, and calculate the average of these place type embeddings as the final spatial context embeddings for images. The cosine similarity score sm_i is calculated between the spatial context embedding of an image and the embedding of each candidate place type class i. Because sm_i ranges from -1 to 1, we use min-max normalization to scale the values to $[0, 1]$. Finally, we apply the same renormalization as for the CNN score to turn the normalized score sm_i' into probability score, i.e. spatial relatedness score s_i^r.

Combining these normalizations together with Equation 1 and Equation 2, we are able to derive that $0 \leq s_i \leq 1$ and $\sum_{i=1}^{N} s_i = 1$ where $N = 15$ in our case. This means that the LBF score s_i can be considered a probability score.

4.2 Spatial Co-location

The spatial relatedness approach follows the assumption that relatedness implies likelihood which is reasonable in cases where similar place types cluster together, such as restaurant, bar, and hotel. However, in cases of high spatial heterogeneity, this assumption will fall short of correctly capturing the true likelihood. An example would be places of dissimilar types that co-occur, e.g., grocery stores and gas stations. Moreover, the LBF method can only capture a linear relationship between the two signals.

Following Berg et al.[2], we also test a Bayesian approach in which we assume there is a complex latent distribution of the data that facilitates our classification task. Intuitively, the CNN score gives us the probability of each candidate class t given the image I, i.e., $P(t|I)$, and the spatial context informs us of the probability of each candidate class given its neighbors $c_1, c_2, c_3, ..., c_n$, denoted as C, around the image location, i.e., $P(t|C)$. We would like to obtain the posterior probability of each candidate class given both the image and its spatial context, i.e., $P(t|I, C)$. Using Bayes' theorem, the posterior probability can be written as:

$$P(t|I, C) = \frac{P(I, C|t)P(t)}{P(I, C)} \tag{3}$$

For variables I, C, and t, we construct their dependencies using a simple probabilistic graphical model, i.e., Bayesian network, which assumes that both the image I and the spatial context C are dependent on the place type t, which intuitively makes sense in that different place types will result in different images and different place types of their neighbors. We know that given information about the image I we are able to update our beliefs, i.e., the probability distributions, about the place type t. In addition, the changes in our beliefs about the place type t can influence the probability distributions of the spatial context C. However, if place type t is observed, the influence cannot flow between I and C, thus we are able to derive the conditional independence of I and C given t. So Equation 3 can be rewritten as:

$$\begin{aligned} P(t|I, C) &= \frac{P(I|t)P(C|t)P(t)}{P(I, C)} \\ &= \frac{P(t|I)P(I)}{P(t)} \frac{P(t|C)P(C)}{P(t)} \frac{P(t)}{P(I, C)} \\ &\propto \frac{P(t|I)}{P(t)} P(t|C) \end{aligned} \tag{4}$$

in which we have dropped all the factors that are not dependent on t as they can be considered as normalizing constants for our probabilities. It follows that the posterior probability $P(t|I, C)$ can be computed using the CNN probability score $P(t|I)$, the spatial context prior $P(t|C)$, and the candidate class prior $P(t)$. Instead of estimating the distribution of spatial context priors, we take advantage of the spatial co-location patterns and calculate the prior probabilities using the Yelp POI data directly. As mentioned earlier, the spatial context C is composed of multiple individual context neighbors $c_1, c_2, c_3, ..., c_n$; hence, we need to calculate $P(t|c_1, c_2, c_3, ..., c_n)$. In order to simplify our calculation, we impose a bag-of-words assumption as well as a Naive Bayes assumption in the spatial co-location patterns. The bag-of-words assumption simplifies the model by assuming that the position (or the order) in which different context POIs occur does not play a role. The Naive Bayes assumption implies that the only relationship is the pair-wise interaction between the candidate place type t and an individual neighbor's place type c_i and there is no interaction between neighboring places wrt. their types, i.e. $(c_i \perp\!\!\!\perp c_j | t)$ for all c_i, c_j. Using spatial co-location, we are able to calculate the conditional probability using place type co-location counts $P(c_i|t) = \frac{count(c_i,t)}{count(t)}$ where $count(c_i, t)$ is the frequency that neighbor type c_i and candidate type t co-locate within a certain distance limit and $count(t)$ is the frequency of candidate type t in the study area. Combining all these components, we can derive:

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

Using Equation 4 and Equation 5, we can derive the final formula for calculating $P(t|I, C)$ shown in Equation 6. For the sake of numerical stability, we calculate the log probability $logP(t|I, C)$ using the natural logarithm. Since the natural logarithm is a monotonically increasing function, it will not affect the final ranking of the classification results.

$$\begin{aligned} logP(t|I, C) &\propto log\bigg(\frac{P(t|I)}{P(t)} P(t|C)\bigg) \\ &= log\bigg(\frac{P(t|I)}{P(c_1, c_2, c_3, ..., c_n)} \frac{\prod_{i=1}^{n} count(c_i, t)}{count(t)^n}\bigg) \\ &\propto logP(t|I) + \sum_{i=1}^{n} log(count(c_i, t)) - nlog(count(t)) \end{aligned} \tag{6}$$

where we also drop $P(c_1, c_2, c_3, ..., c_n)$ as it does not depend on t, so it will not affect the result ranking. The log posterior probability is then used to generate the final ranking of candidate place types and produce the classification results.

4.3 Spatial Sequence Pattern

The spatial co-location approach follows the bag-of-words assumption that the position of spatial context POIs does not matter and the Naive Bayes assumption that the context neighbors are independent of each other. However, in many cases this assumption is too strong. In fact, numerous methods, such as Kriging and multiple-point geostatistics, have been devised to model geospatial proximity patterns and complex spatial interaction patterns. However, incorporating these complex spatial patterns in a multidimensional space would adversely affect the model complexity and make the distribution in Section 4.2 intractable. In order to strike the right balance between the complexity of model and the integrity of spatial context pattern, we propose to capture the spatial sequence pattern in our model by collapsing the 2D geographic space into a 1D sequence.

Specifically, we use the Long Short-Term Memory (LSTM) network model, a variant of recurrent neural network (RNN), in our study. Recurrent neural networks are frequently used models to capture the patterns in sequence or time series data. In theory, the naive recurrent neural networks can capture long term dependencies in the sequence, however, due to the vanishing and exploding gradient problem, they fail to do so in practice. LSTM is explicitly designed to solve the problem by maintaining a cell state and controlling the

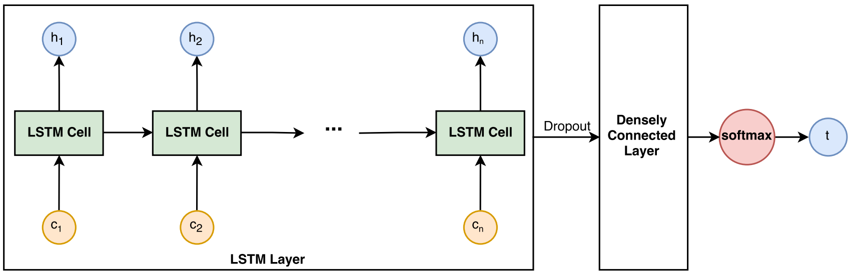

Figure 1 Structure of the LSTM.

input and output flow using forget gate, input gate, and output gate [14]. We use LSTM as a generative model in order to capture the latent distribution of place types using the spatial sequence pattern. In the training stage, the input is a sequence of context place types $c_1, c_2, c_3, ..., c_n$ and the output is the place type t of the POI from which the context is created. The input sequence is ordered in a way so that the previous one is further away from the output than the next one in the collapsed 1D space. Image one would drive around a neighborhood before reaching a destination. For each of the POIs encountered during the route, one would update the beliefs about the neighborhood by considering the current POI and all previously seen POIs. Upon arriving at the destination, one would have a reasonable chance of guessing this final POI's type. The structure of the LSTM model is shown in Figure 1. We apply a dropout after the LSTM layer to avoid overfitting. After training the LSTM model on Yelp's POI dataset, we are able to obtain the spatial context prior $P(t|c_1, c_2, c_3, ..., c_n)$ based on the spatial sequence pattern around the image locations in our test data. We specifically removed the image locations and their context in the training data. Similar to the spatial co-location approach, we use Bayesian inference and log probability to calculate the final result:

$$\begin{aligned} logP(t|I,C) &\propto log\left(\frac{P(t|I)}{P(t)}P(t|C)\right) \\ &= logP(t|I) + logP(t|c_1, c_2, c_3, ..., c_n) - logP(t) \end{aligned} \tag{7}$$

where the candidate class prior $P(t)$ can be computed using the Yelp data. Since we use LSTM as a generative model, in the prediction phase, sampling strategies, such as greedy search, beam search, and random sampling, can be applied based on the distribution provided by the output of the LSTM prediction. However, we only generate the next prediction instead of a sequence, so we do not apply these sampling strategies. Instead, we make use of the hyperparameter *temperature* τ to adjust the probability scores returned by the LSTM model before combining them with the CNN model in a Bayesian manner. Including the hyperparameter τ, the softmax function in the LSTM model can be written as:

$$P(t_i|C) = \frac{exp(\frac{logit_i}{\tau})}{\sum_{j=1}^{N} exp(\frac{logit_j}{\tau})} \tag{8}$$

where $logit_i$ is the logit output provided by LSTM before applying the softmax function and $N = 15$ in our case. Intuitively, when the temperature τ is high, i.e., $\tau \to \infty$, the probability distribution will become diffuse and $P(t_i|C)$ will have almost the same value for different t_i; when τ is low, i.e., $\tau \to 0^+$, the distribution becomes peaky and the largest $logit_i$ stands out to have a probability close to 1. This idea is closely related to the exploration and exploitation trade-off in many machine learning problems. The value of τ will affect the probability scores $P(t_i|C)$ but not the ranking of these probabilities.

In this study, we propose two ways to model the 2D geographic space as a 1D sequence. The first one is a distance-based ordering approach. For any given POI, we search for nearby POIs within a certain distance from it, choose the closest n POIs, and rearrange them by distance with descending order, thereby forming a 1D array. This distance-based method is isotropic in that it does not differentiate between directions while creating the sequence. The second method is a space filling curve-based approach. We utilize *Morton order* here which is also used in geohashing to encode coordinates into an indexing string that can preserve the locality of spatial locations. We use Morton order to encode the geographic locations of every POI and order them in a sequence based upon their encodings, i.e., indexing sequence. After obtaining the sequence, for each POI, we use the previous n POI in the sequence as the context sequence. Other space filling curves could be used in future work.

Because each POI can have multiple place types associated with it, e.g., restaurant and beer garden, the sequence of place types is usually not unique for the same sequence of POIs. As our LSTM input is a sequence of place *types*, we compute the Cartesian product of all POI type sets in the sequence of nearby places:

$$T_{c_1} \times T_{c_2} \times T_{c_3} \times ... \times T_{c_n} = \{(t_{c_1}, t_{c_2}, t_{c_3}, ..., t_{c_n}) | \forall i = 1, 2, 3, ..., n,\ t_{c_i} \in T_{c_i}\} \tag{9}$$

where T_{c_i} is the set of place types associated with POI c_i in the context sequence. In practice, however, we randomly sample a fixed number of place type sequences from each of the Cartesian product for the POI context sequence as the potential combinations grow exponentially with increasing context size.

5 Experiment and Result

In this section, we explain our experimental setup for the models described above, describe the metrics used to compare the model performance for place type image classification, and present the results and findings.

5.1 Implementation Details

For all three types of spatial context, we use 10 as the maximum number of context POIs and a distance limit of 1000m for the context POI search. For the spatial sequence pattern approach, we use a fixed sample size of 50 to sample from the Cartesian product of all POI type sets in the sequence. [6] We use a one-layer LSTM with 64 hidden units. We train our LSTM model using the recommended Root Mean Square Propagation (RMSProp) optimizer with a learning rate of 0.005. A dropout ratio of 0.2 is applied in the LSTM and we run 100 epochs. The same settings are used for all LSTM trainings in our experiment. The total number of POI in the dataset is 115,532, yielding more than 5 million unique training sequences.

For evaluation, we use three different metrics, namely Mean Reciprocal Rank (MRR), Accuracy@1, and Accuracy@5. Another common metric for image classification would also be Mean Average Precision (MAP), but since there is only one true label per type in our task, we use MMR instead.

[6] The median for types per place in Yelp is 3.

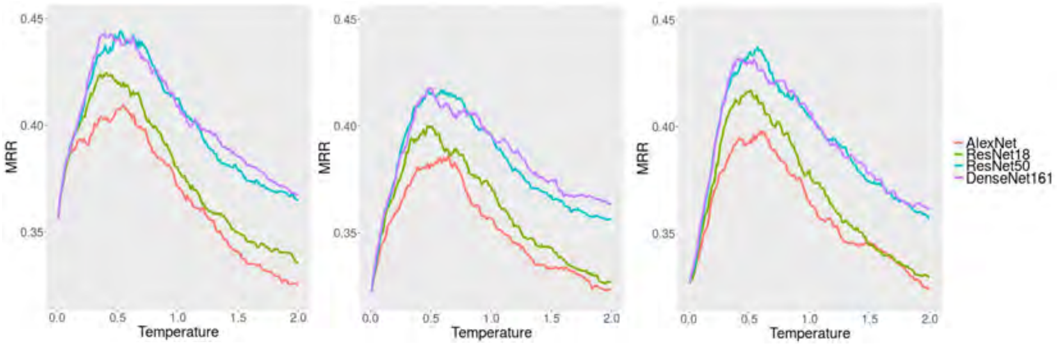

Figure 2 From left to right, MRR result using distance-based sequence, random sequence, and Morton code-based sequence with varying temperatures

5.2 Results

We run the 750 test images we collected, i.e., 50 images per each of 15 types, on the four CNN baseline models (AlexNet, ResNet18, ResNet50, and DenseNet161) as well as the combined models using our three different types of spatial context.[7] In addition to the two methods for converting geographic space into 1D sequences in the spatial sequence pattern approach, we also test one model using random sequences with the same context count and distance limits. We did so to study whether results obtained using the LSTM would benefit from distance-based spatial contexts. A higher result for the spatial sequence based LSTM over the random LSTM would indicate that the network indeed picked up on the distance signal.

The hyperparameter τ can be adjusted; a value of 0.5 has been proposed as a good choice before. In order to test this and find the optimal temperature value, we run the combined model using spatial sequence patterns with three types of sequencing approaches, namely random sequence, distance-based sequence, and Morton order-based sequence.

We test temperature values ranging from 0.01 to 2 with a step of 0.01. We combine the spatial sequence pattern models with all CNN models. The MRR result with respect to temperature are shown in Figure 2. Although there are a slight variations, the MRR curves all reach their peaks around a τ value of 0.5. This confirms the suggestion from the literature. Figure 3 shows selected example predictions. The results for MRR, Accuracy@1, and Accuracy@5 using the baseline models as well as our proposed, spatially explicit models are shown in Table 2, Table 3, and Table 4.[8]

As we can see, by incorporating spatial context in the image classification model, we are able to improve the classification result in general. However, integrating spatial relatedness using the LBF method does not seem to affect the result. This essentially confirms our aforementioned assumption that relatedness does not always imply likelihood. The benefit of incorporating spatial relatedness in cases of spatial homogeneity are likely to be offset by cases of hight spatial heterogeneity in which spatial relatedness may have an negative effect as dissimilar places co-occur.

[7] Transfer learning could be applied to fine tune the CNN models first, but we only have limited images and our hypothesis is that spatial context can be used as a powerful complement or alternative to the visual component for image classification.

[8] The baseline models are not comparable with a random classifier which would yield an expected accuracy of 1/15 in this case, because the baseline CNN models have 365 unique labels and we choose 15 labels in our experiment.

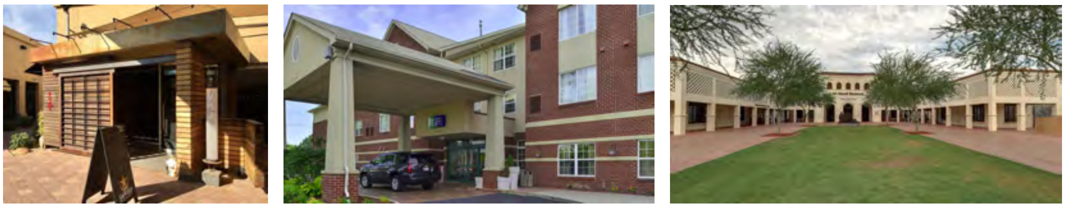

Figure 3 From left to right, images of a restaurant, a hotel, and a museum from Yelp, Google Street View, and Google Maps respectively. The first image is incorrectly classified as library using all 4 CNN models and it is correctly classified as restaurant using the spatial sequence pattern (distance) 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.

Table 2 MRR result using baseline models and proposed combination models using different types of spatial context and sequences

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

The Accuracy@1 measurement is improved by incorporating spatial co-location component in the models. This confirms our previous reasoning that considering the external signal, namely spatial contexts, and assuming a complex latent distribution of the data in a Bayesian manner improve image classification. However, for MRR the improvement is marginal and for Accuracy@5 there even is a decrease after incorporating the spatial co-location component because this type of spatial context falls short of taking into account the intricate *interactions* of different context neighbors. This shortcoming is not clear when only looking at the first few results in the ranking returned by the combined models, but it becomes clearer in later results in the ranking output, thus resulting in a decrease for Accuracy@5 and only a slight increase in the MRR measurement.

The Bayesian combination model using spatial sequence patterns shows better overall results compared with the baseline models, the spatial relatedness model, and the spatial co-location model. This is because the spatial sequence patterns capture spatial interactions between the neighboring POIs that are neglected by the other models. From the result we can see that using a distance-based sequence is better than using a random sequence. To prevent confusion and to understand why the random model still performs relatively well, it is important to remember that this model utilizes spatial context. However, it does not utilize the distance signal within this context but merely the presence of neighboring POI. The results show that a richer spatially explicit context, one that comes with a notion of *distance decay*, indeed improves classification results. Interestingly, the sequence using Morton order, which is widely used in geohashing techniques, does not further improve the result compared to the distance-based sequence. There may be multiple reasons for this. First, we may have reached a ceiling of possible improvements by incorporating spatial contexts. Second, our Morton order implementation takes the 10 places that precede the target place in the index.

Table 3 Accuracy@1 result using baseline models and proposed combination models using different types of spatial context and sequences

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**

Table 4 Accuracy@5 result using baseline models and proposed combination models using different types of spatial context and sequences

Accuracy@5	AlexNet	ResNet18	ResNet50	DenseNet161
Baseline	0.50	0.56	0.59	0.60
Relatedness	0.52	0.56	0.58	0.59
Co-location	0.42	0.44	0.45	0.44
Sequence Pattern (Random)	0.65	0.69	**0.73**	0.73
Sequence Pattern (Distance)	**0.67**	**0.70**	**0.73**	**0.75**
Sequence Pattern (Morton order)	0.65	**0.70**	0.72	0.71

This may result in directional effects. Finally, all space filling curves essentially introduce different ways to preserve local neighborhoods; utilizing another technique such as Hilbert curves may yield different results. Given that the Morton order-based sequence in many cases yield results of equal quality to the distance-based sequences, further work is needed to test the aforementioned ideas.

Summing up, the results demonstrate that incorporating a (distance-based) spatial context improves the MRR of state-of-the-art image classification systems by over **40%**. The results for Accuracy@1 are more than **doubled** which is of particular importance for humans as this measure only considers the first ranked result.

6 Conclusion and Future Work

In this work, we demonstrated that utilizing spatial contexts for classifying places based on images of their facades and interiors leads to substantial improvements, e.g., increasing MRR by over 40% and doubling Accuracy@1, compared to applying state-of-the-art computer vision models such as ResNet50 and DenseNet161 alone. These advances are especially significant as the classification of places based on their images remains a hard problem. One could argue that our proposal requires additional information, namely about the types of nearby places. However, such data are readily available for POI, and only a few nearby places are needed. Secondly, and as a task for future work, one could also modify our methods to work in a *drive-by-typing* mode in which previously seen places are classified, and these classification results together with their associated classification uncertainty are used to improve estimation of the currently seen place, thereby relaxing the need for POI datasets. In the future, we would like to apply transfer learning and experiment with other ways to encode spatial contexts, e.g., by testing different space-filling curves. We plan to develop models to directly capture 2D spatial patterns rather than using a 1D sequence as a proxy and test whether spatial contexts also aid in recognizing objects beyond places and their facades.

References

1 Georges Baatz, Olivier Saurer, Kevin Köser, and Marc Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *Computer Vision–ECCV 2012*, pages 517–530. Springer, 2012.

2 Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2019–2026. IEEE, 2014.

3 Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pages 4349–4357, 2016.

4 Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

5 Marco Castelluccio, Giovanni Poggi, Carlo Sansone, and Luisa Verdoliva. Land use classification in remote sensing images by convolutional neural networks. *arXiv preprint arXiv:1508.00092*, 2015.

6 Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *arXiv preprint arXiv:1707.05373*, 2017.

7 Anne Cocos and Chris Callison-Burch. The language of place: Semantic value from geospatial context. In *15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 99–104, 2017.

8 Shanshan Feng, Gao Cong, Bo An, and Yeow Meng Chee. Poi2vec: Geographical latent representation for predicting future visitors. In *AAAI*, pages 102–108, 2017.

9 Michael F Goodchild and Donald G Janelle. Thinking spatially in the social sciences. *Spatially integrated social science*, pages 3–22, 2004.

10 Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013.

11 Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

12 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

13 Geremy Heitz and Daphne Koller. Learning spatial context: Using stuff to find things. In *European conference on computer vision*, pages 30–43. Springer, 2008.

14 Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

15 Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2261–2269, 2017.

16 Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pages 2280–2288, 2016.

17 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

18 Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

19 Stefan Lee, Haipeng Zhang, and David J Crandall. Predicting geo-informative attributes in large-scale image collections using convolutional neural networks. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 550–557. IEEE, 2015.

20 Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Computer Vision and Pattern Recognition*, pages 891–898. IEEE, 2013.

21 Kang Liu, Song Gao, Peiyuan Qiu, Xiliang Liu, Bo Yan, and Feng Lu. Road2vec: Measuring traffic interactions in urban road system from massive travel routes. *ISPRS International Journal of Geo-Information*, 6(11):321, 2017.

22 Nikhil Naik, Scott Duke Kominers, Ramesh Raskar, Edward L Glaeser, and César A Hidalgo. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences*, 114(29):7571–7576, 2017.

23 Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016.

24 Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

25 Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism. *arXiv:1711.11443*, 2017.

26 Wanxiao Sun, Volker Heidt, Peng Gong, and Gang Xu. Information fusion for rural land-use classification with high-resolution satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 41(4):883–890, 2003.

27 Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 2015.

28 Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

29 Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *Proceedings of the IEEE international conference on computer vision*, pages 1008–1016, 2015.

30 Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. From itdl to place2vec–reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. *Proceedings of SIGSPATIAL*, 17:7–10, 2017.

31 Jie Yu and Jiebo Luo. Leveraging probabilistic season and location context models for scene understanding. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 169–178. ACM, 2008.

32 Andi Zang, Runsheng Xu, Zichen Li, and David Doria. Lane boundary extraction from satellite imagery. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on High-Precision Maps and Intelligent Applications for Autonomous Vehicles*, page 1. ACM, 2017.

33 Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

34 Shenglin Zhao, Tong Zhao, Irwin King, and Michael R Lyu. Geo-teaser: Geo-temporal sequential embedding rank for point-of-interest recommendation. In *Proceedings of the 26th international conference on world wide web companion*, pages 153–162. International World Wide Web Conferences Steering Committee, 2017.

35 Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.