On the Location Dependence of Convolutional Neural Network Features

Scott Workman and Nathan Jacobs
Department of Computer Science
University of Kentucky
{scott, jacobs}@cs.uky.edu

Abstract

As the availability of geotagged imagery has increased, so has the interest in geolocation-related computer vision applications, ranging from wide-area image geolocalization to the extraction of environmental data from social network imagery. Encouraged by the recent success of deep convolutional networks for learning high-level features, we investigate the usefulness of deep learned features for such problems. We compare features extracted from various layers of convolutional neural networks and analyze their discriminative ability with regards to location. Our analysis spans several problem settings, including region identification, visualizing land cover in aerial imagery, and ground-image localization in regions without ground-image reference data (where we achieve state-of-the-art performance on a benchmark dataset). We present results on multiple datasets, including a new dataset we introduce containing hundreds of thousands of ground-level and aerial images in a large region centered around San Francisco.

1. Introduction

The relationship between image appearance and geographic location is complex, fascinating, and well studied. The canonical computer vision task in this domain is image localization. While some images provide strong localization cues and are easily localizable, such as a view of the Statue of Liberty from Ellis Island or the Coliseum in Rome, others only provide weak evidence of their geographic location. For such images, it may only be possible to guess the region in which the image was taken. A wide variety of approaches have been proposed for the former problem, while the latter problem has only received significant attention recently.

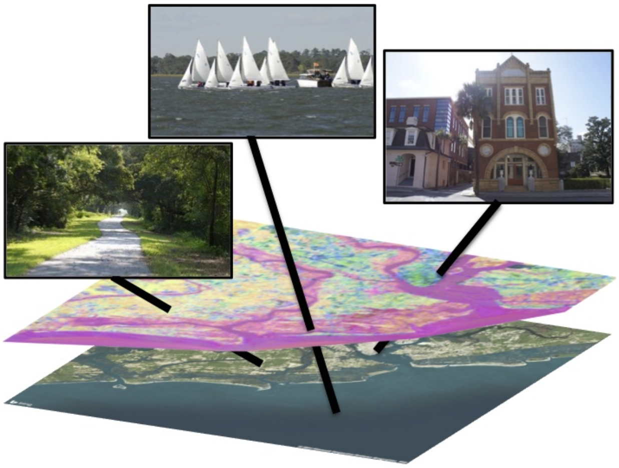


Figure 1: Features learned by a deep convolutional neural network provide strong cues for image geolocalization and geospatial feature extraction on both ground and aerial imagery.

Recent advances in deep convolutional neural networks have lead to major improvements in performance on a wide variety of vision tasks, including: object classification and detection [18, 9, 15], face recognition and verification [26], image super resolution [6], and scene recognition [29]. This work investigates the value of features extracted from a deep convolutional neural network for a variety of image localization tasks (Figure 1).

Main Contributions We make the following contributions:

- We introduce a new dataset that includes ground-level images, with associated aerial images, and an aerial image reference database.
- Using this dataset, and a previously released dataset, we demonstrate that deep learned features have sufficient discriminative power to distinguish between two geographic regions, and that we can extract discriminative, iconic images from the learned models.
- Further, we show that deep features are also useful for understanding and interpreting aerial images and for

image-based search in aerial images, even though they were not explicitly trained for the task.

- We also present *state-of-the-art results* on a benchmark dataset in cross-view image geolocalization. Our approach improves the performance by 1.08% (within the top 1% of ranked candidates) compared to the previous best method, a percentage change of 6.22%.

Together, these results demonstrate the effectiveness of features extracted from convolutional neural networks for finding relationships between aerial and ground-view imagery of the same location.

2. Related Work

Place Recognition and Image Localization Data driven image localization is often reformulated as an image retrieval problem, often called visual place recognition. Standard approaches use machine learning techniques to find visually overlapping images from a reference set with known geographic location. These methods generally fall into two categories, matching using local features [25, 24, 2, 3, 1] or global image features [10, 14]. Recently, Lin et al. [20] introduced the problem of cross-view image localization, learning a mapping between ground and aerial image viewpoints. Many other cues for localization have been explored which take advantage of photometric and geometric properties [28, 12, 13].

Relationship Between Location and Appearance Many attempts have been made to characterize the relationship between location and visual appearance. One common objective is to learn geographically discriminative attributes [22, 11, 7]. Doersch et al. [5] use a discriminative clustering approach to automatically discover the visual style of a region. Zhou et al. [30] present a data-driven attribute analysis for characterizing the identity of a city. Patterson and Hays, and Laffont et al. learn high level scene attributes for scene recognition [21, 19].

Feature Learning with Convolutional Neural Nets Features extracted from convolutional neural networks have proven very powerful for many different problems including image object recognition [18], video classification [17], and a wide variety of other tasks [23]. We take inspiration from Fischer et. al [8] who show that mid-level features compare favorably to SIFT for descriptor matching. We extend this line of research to include problems relating to geospatial image analysis.

3. Deep Features for Image Geolocalization

We explore the application of an existing deep convolutional neural network (CNN) architecture [18] to problems in geospatial image analysis. Instead of training the model ourselves, we take advantage of two publicly available pretrained models. The first is trained on ImageNet [4] for detecting object categories and is available through Caffe [16]. The second is trained on the recently introduced Places Database [29] with the goal of scene recognition. For both, we perform feature extraction using Caffe [16], an open source deep learning framework. We refer to these as ImageNet and Places throughout.

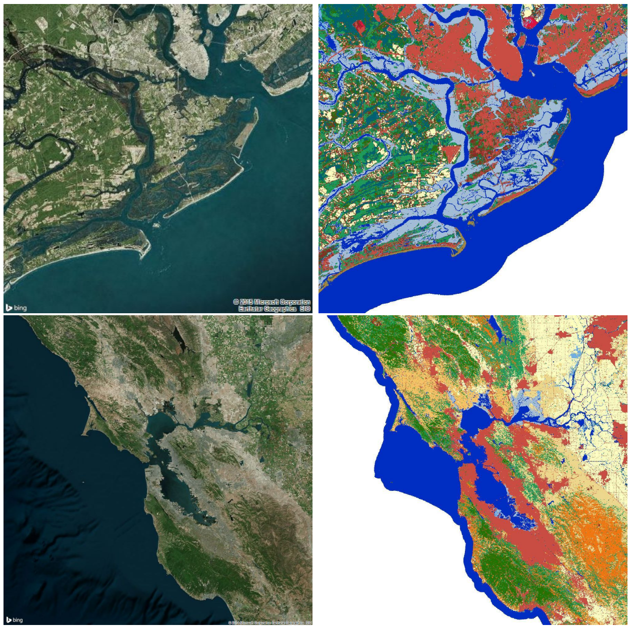


Figure 2: Aerial and land cover maps for the Charleston and San Francisco datasets.

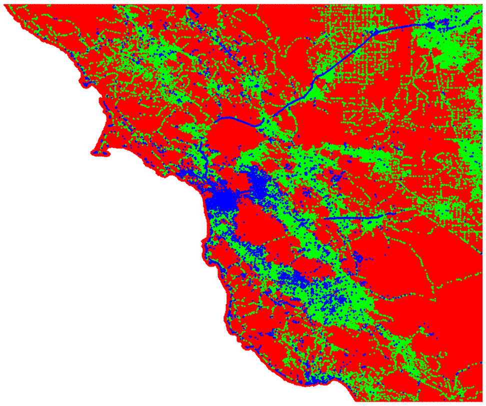


Figure 3: Coverage map of the San Francisco dataset. Red indicates the spatial coverage of aerial imagery, overlaid with Street View (green) and Flickr (blue) image locations.

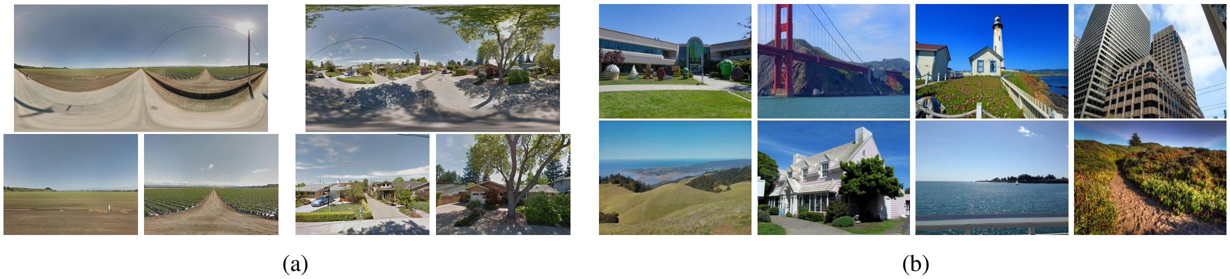


Figure 4: Images from the San Francisco dataset. (a) Example Street View panoramas (top) and their corresponding cutouts (bottom). (b) Example Flickr images after filtering.

In this CNN architecture, features are extracted from images in a layered, feed-forward manner. Initial layers of the architecture consist of convolutions, local response normalization, local pooling, dropout layers, and rectified linear activation units. The top layers of the network are four fully connected layers ‘fc6’, ‘fc7’, ‘fc8’, and the final output layer ‘prob’ that represents a categorical probability distribution. The dimensionality of these top layers in the ImageNet network are 4096, 4096, 1000, and 1000 respectively. In the Places network, the final two feature vectors are 205 dimensional.

3.1. Datasets

We analyze and evaluate the methods we present using two diverse, realistic datasets. An aerial image and a land cover map for these datasets is shown in Figure 2. The first dataset, Charleston, was introduced by Lin et al. [20] and contains 6756 ground-level images with corresponding aerial and land cover images, and a reference map database of aerial and land cover images, without corresponding ground-level images, for a 40km × 40km region around Charleston, SC. Of the ground-level images, 737 are isolated with no other ground-level images nearby.

We introduce a new dataset, San Francisco, containing ground-level and aerial images collected in a 200km × 200km region around San Francisco. We collected aerial imagery for the entire region from Bing Maps, each image of size 256 × 256 and covering a 480m × 480m area. Ground-level images from the region were collected from both Flickr and Google Street View. For Flickr, we queried and downloaded images from 2013 onwards, totaling 114,384 images. We used the pre-trained Places network to filter images that were unlikely to be images of outdoor scenes by manually assigning a label of indoor/outdoor to each of the 205 scene categories. This resulted in a final set of 74,217 images. For Street View, we downloaded 50,000 street level panoramas from which we extracted two side-facing perspective images of size 800 × 600, totaling 100,000 images. Finally, for each ground-level image we downloaded its corresponding aerial image, centered at the same location.

While similar in conception to Charleston, our dataset has several benefits. These include a significantly larger region of interest for localization, many more images, a different region of the country with very different land cover attributes, automatic filtering of non-outdoor images, and a large number of images with very accurate GPS tags (by virtue of Google Street View). In total, the dataset contains 278,561 map images and 174,217 ground-level images and their associated aerial images. As with the Charleston dataset, we identify a set of isolated images for testing, totaling 2,245 ground-view images. Figure 3 visualizes the coverage of our dataset and Figure 4 shows several example images.

4. Experiments

We demonstrate that high-level features extracted from CNNs, specifically the top layers of the ImageNet [18] and Places [29] networks, are highly informative of geographic location. While not always superior, we find that the ‘fc8’ layer of the Places network usually performs the best for any given task, sometimes by a wide margin. The remainder of this section details experiments in three domains: (1) estimating the region of a ground-level image, (2) visualizing land cover differences in aerial imagery and finding similar aerial images, and (3) cross-view localization, in which pairs of aerial and ground-level images are used to localize images in regions without ground-level reference imagery.

4.1. Region Classification

We begin by investigating whether or not such features are useful for classifying which dataset an image is from; in other words, was the picture taken in San Francisco or Charleston? We train an SVM model (with an RBF kernel) to address this problem. For training, we randomly select a set of 10,000 ground-level training images from each dataset, and extract the ‘fc6’, ‘fc7’, and ‘fc8’ features from both networks. From these, we train six separate SVM models, one for each feature level and network. For evaluation, we use the isolated test set images defined in both datasets.

Table 1: Region classification accuracy.

Feature	Accuracy
GIST [27]	81.7 %
ImageNet fc6	82.7 %
ImageNet fc7	82.2 %
ImageNet fc8	80.9 %
Places fc6	85.1 %
Places fc7	85.1 %
Places fc8	84.5 %

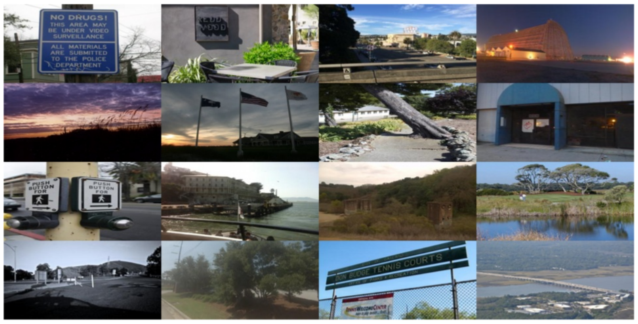


Figure 6: The most ambiguous images based on the SVM score for a region classifier trained on the Places 'fc8' features as described in Section 4.1.

See Table 1 for the accuracy of various CNN features, and the GIST descriptor as the baseline, on this problem. We find that the Places features are clearly superior to the ImageNet features, but that the difference between the various feature levels is negligible.

Figure 5 shows montages of the images with very high and very low SVM scores. Many of the detected images are iconic images of the corresponding region. It also shows images in San Francisco that the classifier determines look most like Charleston, and vice versa. Figure 6 shows a montage of the most ambiguous images, many of which would be very difficult for a person to label correctly.

This experiment demonstrates that CNN features are capturing subtle characteristics of various areas from ground-level imagery. This is, perhaps, the least surprising result in this work, since the CNNs were trained to distinguish object and scene categories in similar ground-level imagery.

4.2. Aerial Imagery Analysis: Land Cover and Image-Based Search

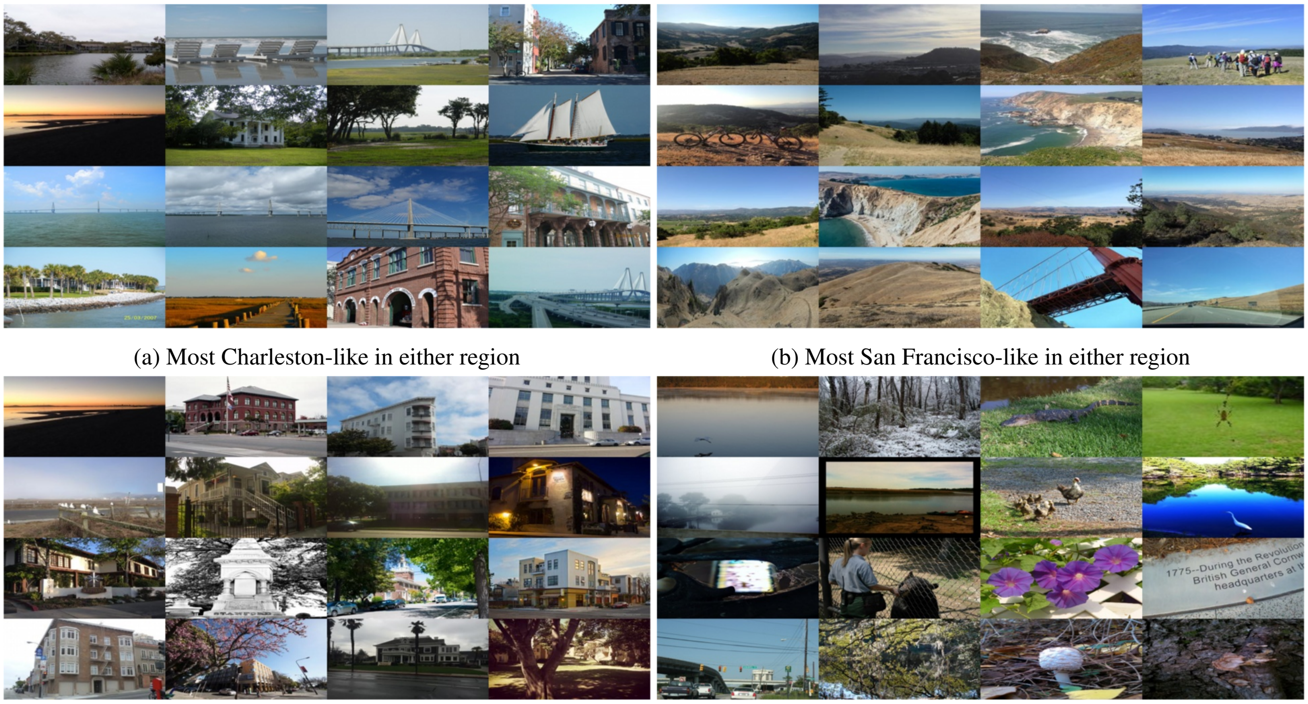
We analyze the effectiveness of ImageNet and Places, neither of which was trained on aerial images, on two problems in aerial imagery analysis. We find that the Places network extracts features that are highly location dependent. To highlight this, we extracted the 'fc8' layer of the Places network on the aerial image reference database from both the Charleston and San Francisco datasets. We then computed the principal component analysis (PCA) decomposition of these features. Figure 7 visualizes a synthetic aerial image for each area using the top three principal components. For each map location we have a 3D PCA coefficient; we use the first, second, and third coefficient as the red, green, and blue color channels of a synthetic aerial image, scaling each color channel to $[0, 1]$ and using natural neighbors interpolation. The result is an image that encodes the dominant feature appearance variations as different colors. We find, upon closer inspection of the aerial and land cover maps (Figure 2), that the top three PCA coefficients of the CNN feature vectors of aerial imagery are closely related to land cover class.

We also investigated a novel user-focused application in image-based search. Consider the following scenario: as a person is browsing a map, they become curious about the ground-level appearance of a particular location. Unfortunately, the location has not been captured by a ground-based panorama service such as Google Street View. We provide a method to search for ground-level images using only the current aerial image and a reference dataset of aerial and ground-level image pairs. Our approach is straightforward: (1) compute CNN features on the current map location, (2) compute the Euclidean distance between this feature and all map images in the reference database, and (3) present the user with the ground-view images that had the most similar aerial images.

The results, shown in Figure 8, demonstrate that we are able to retrieve a diverse and realistic set of ground-level images (as compared to the true ground-level image) by querying on the appearance of the aerial view. While this approach could clearly be generalized, the results are quite compelling. It clearly finds images that would not have been found by matching on the ground-level views. For instance, in Figure 8 (bottom), matching on the aerial view results in ground views that do not contain a building, contrary to what one would expect if the query were the ground view (which contains a building).

4.3. Cross-View Image Matching

Cross-view image localization (i.e. matching ground-level imagery to aerial imagery) has only recently been investigated [20]. The underlying premise is that the mass quantities of dense aerial imagery available today, compared to the relatively sparse coverage of geotagged ground-level images, can be exploited for the task of image localization. When no nearby ground-imagery is available, existing methods which localize via ground-level visual similarity [10] are not applicable. The cross-view localization problem is inherently more difficult than the single-view problem, due to the dramatic differences in viewpoint of the two image sets. This previous work explored several



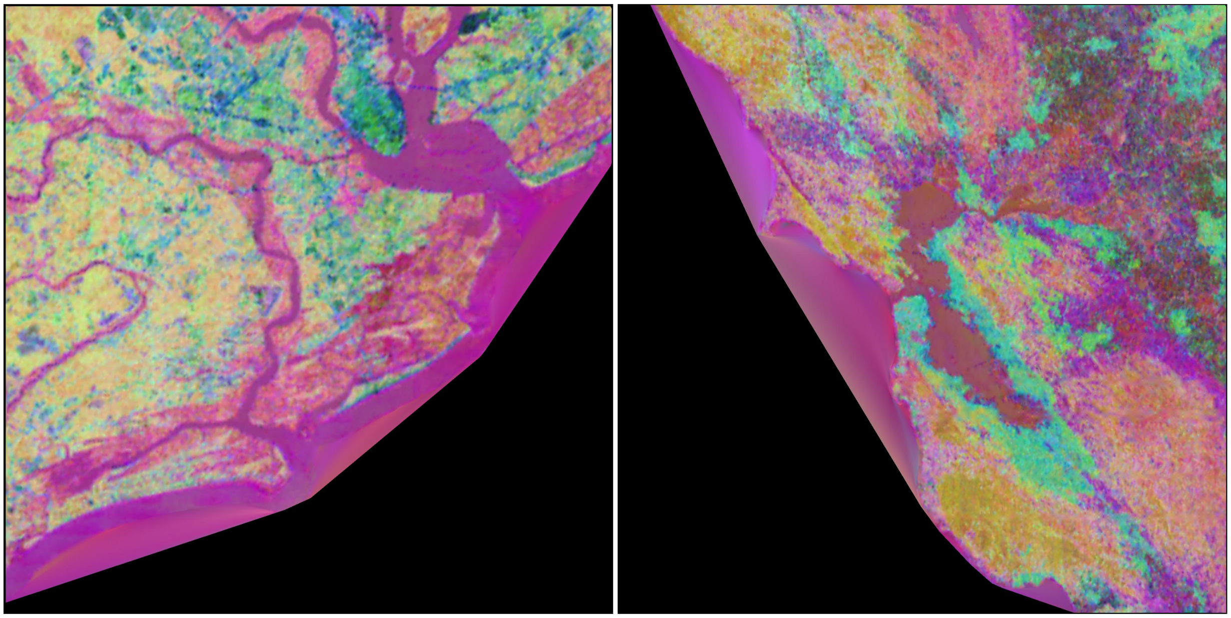
(a) Most Charleston-like in either region

(b) Most San Francisco-like in either region

(c) Most Charleston-like in San Francisco

(d) Most San Francisco-like in Charleston

Figure 5: (a, b) Images with the highest and lowest SVM score for a region classifier trained on the Places ‘fc8’ features as described in Section 4.1. (c, d) Images from the respective regions with the most incorrect SVM scores.



(a) Charleston

(b) San Francisco

Figure 7: Synthetic aerial images, constructed by performing PCA analysis on the Places ‘fc8’ layer output of small aerial images, highlights different types of land cover. For example, regions that are over water (pink), forest (yellow), and urban (green) areas are all clearly visible as unique colors.

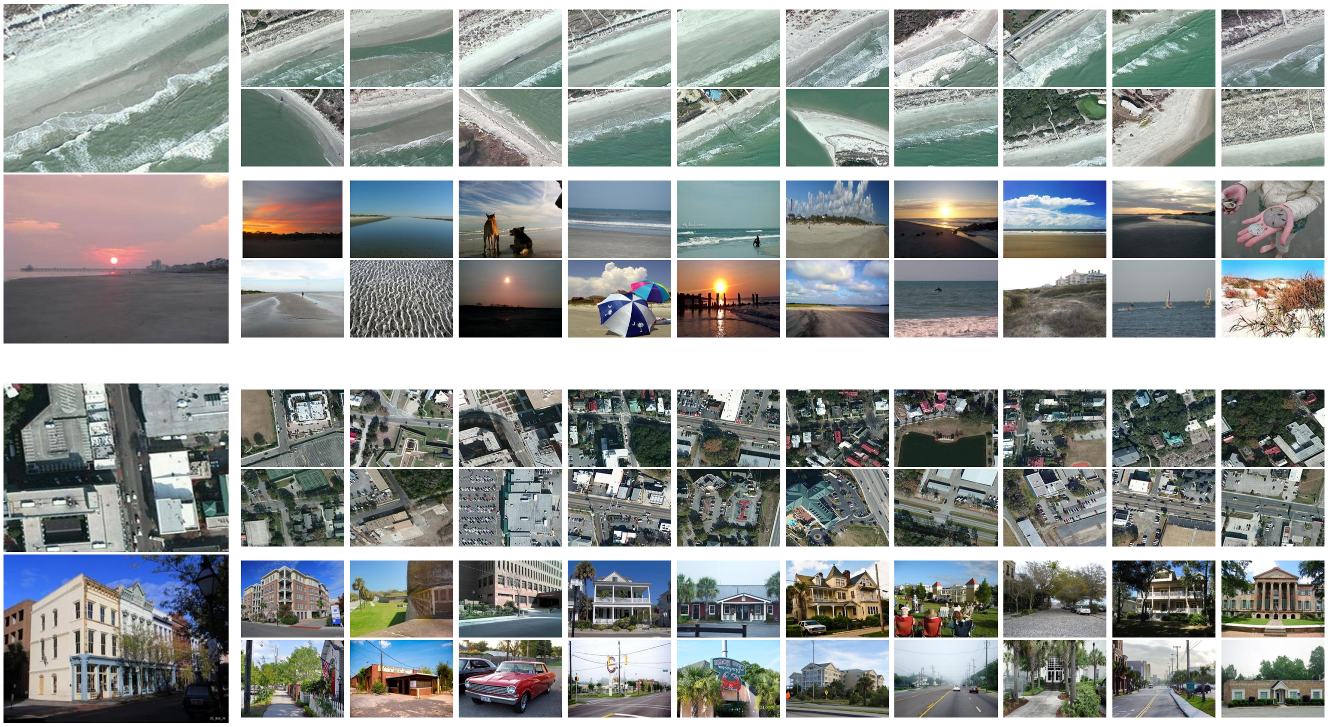


Figure 8: Image retrieval by matching 'fc8' aerial features from the Places network. Given a query aerial image (top, left), we find the most similar aerial images in the reference dataset (top, right), and display their corresponding ground-level images (bottom, right). The results are diverse and realistic as compared to the true ground-level image (bottom, left) at that location.

strategies for characterizing this relationship. Their methods build on global image descriptors, such as color histograms, GIST, and HoG, and combine them with land-cover attributes in an attempt to learn a feature translation between the two viewpoints. Their most successful method combines a feature averaging strategy with a supervised learning technique.

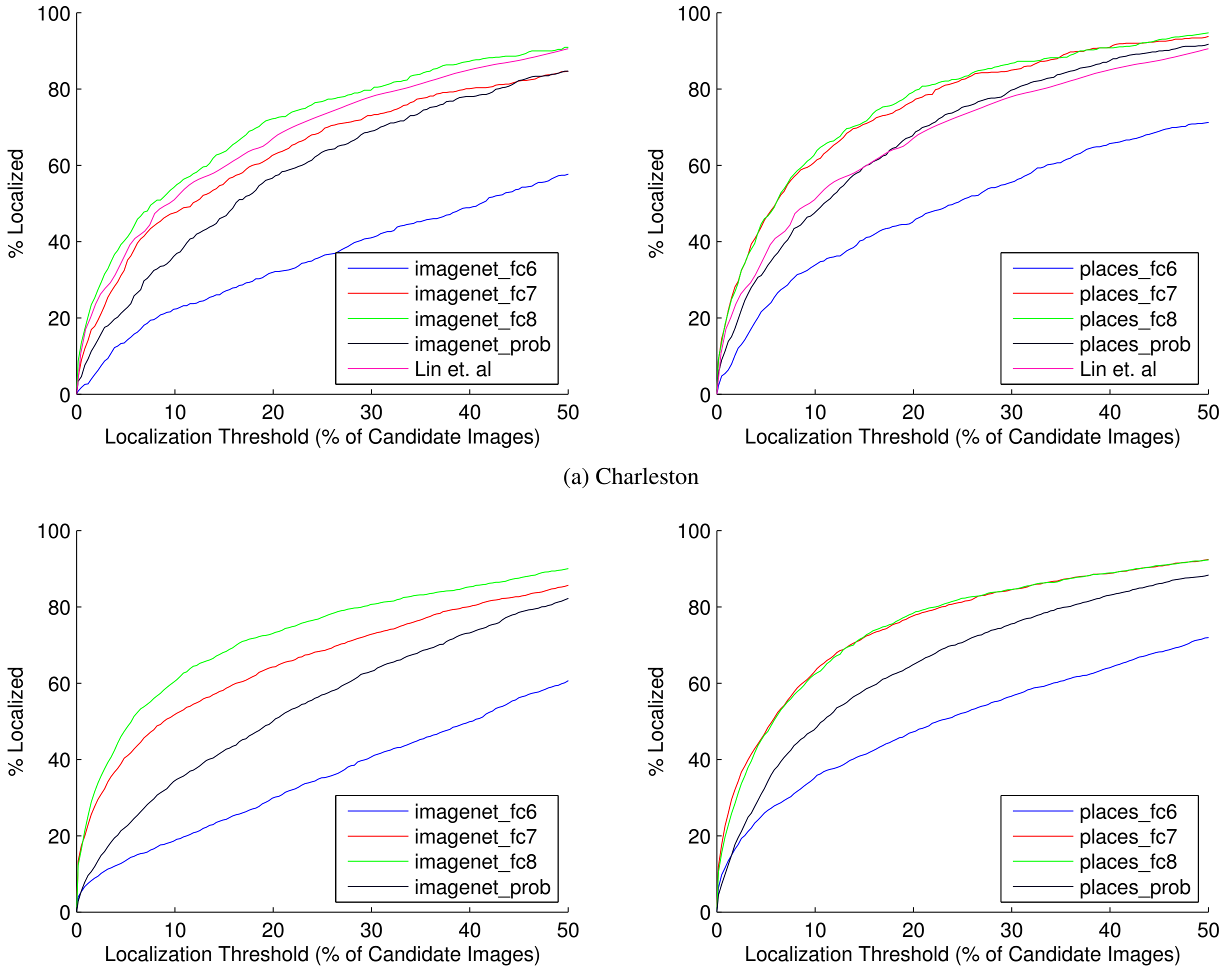
Given our findings that the Places network extracts features that are highly location dependent, even for aerial imagery, we analyze their performance for this task. Our strategy is as follows: given a query image, we first find the closest 30 ground-level images in the training set by comparing their associated 'fc8' feature from Places. For this set of neighbors, we average the 'fc8' features of their corresponding aerial images and use this as our query to search the aerial image reference database. In both cases, we use Euclidean distance as our similarity metric. The result is a score for every map location.

We evaluate this technique on each dataset using the isolated set of images (images for which no nearby ground images exist). The performance metric used is the same as described by Lin et. al [20]. Given the scores for each location, we compute the rank of the ground truth location in the sorted list. Figure 9 visualizes our results as a cumulative distribution function of the fraction of query images correctly localized versus the percentage of candidate images retrieved. In Figure 9 (top), we compare our results on Charleston versus the method of Lin et. al [20]. Our approach is highly effective, outperforming their best method by a large margin without requiring any land cover imagery or more complex methods. In terms of top 1% accuracy, our best result using the Places 'fc8' feature correctly localizes 18.45% of query images versus the 17.37% reported by Lin et. al [20], a 1.08% increase and a percentage change of 6.22%. This trend continues as the localization threshold, i.e. the percentage of candidate images retrieved, is increased.

In Figure 10 we show three example query images from each dataset alongside the localization results computed using the features extracted from the Places network. The localization results are visualized as a heatmap using the similarity scores for each map location. As observed quantitatively in Figure 9, the 'fc8' features outperform the features from other layers qualitatively.

5. Conclusion

In experiments on several geolocation-related computer vision problems, we found that features learned from deep CNNs are easy to compute, discriminative, and give very compelling results on a variety of tasks. The 'fc8' feature



(a) Charleston

(b) San Francisco

Figure 9: Accuracy of localization as a function of retrieved candidate locations. Our method, using Places 'fc8' features, significantly outperforms Lin et. al [20], the previous best method on the Charleston dataset.

from the Places network performs well on all problems we explored, often significantly better than other features. We suppose that this is due to the low dimensionality of the feature (205 vs 4096 or 1000) and the nature of the training data. However, this is an area for further investigation. For classifying the region of ground-images, the CNN features out-performed a commonly used off-the-shelf feature descriptor and also provide a method to identify images that capture the relative appearance of two places. In addition, we found that CNN features give state-of-the-art results on the challenging problem of cross-view image geolocalization. Both the ImageNet and Places CNNs extract strongly location-related features on aerial imagery. This is surprising because they were trained on imagery from a vastly different viewpoint. This points to a promising direction for future research in building deep-learning based models that are directly targeted at problems of localization and location-related feature extraction from ground and aerial imagery.

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7212. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.

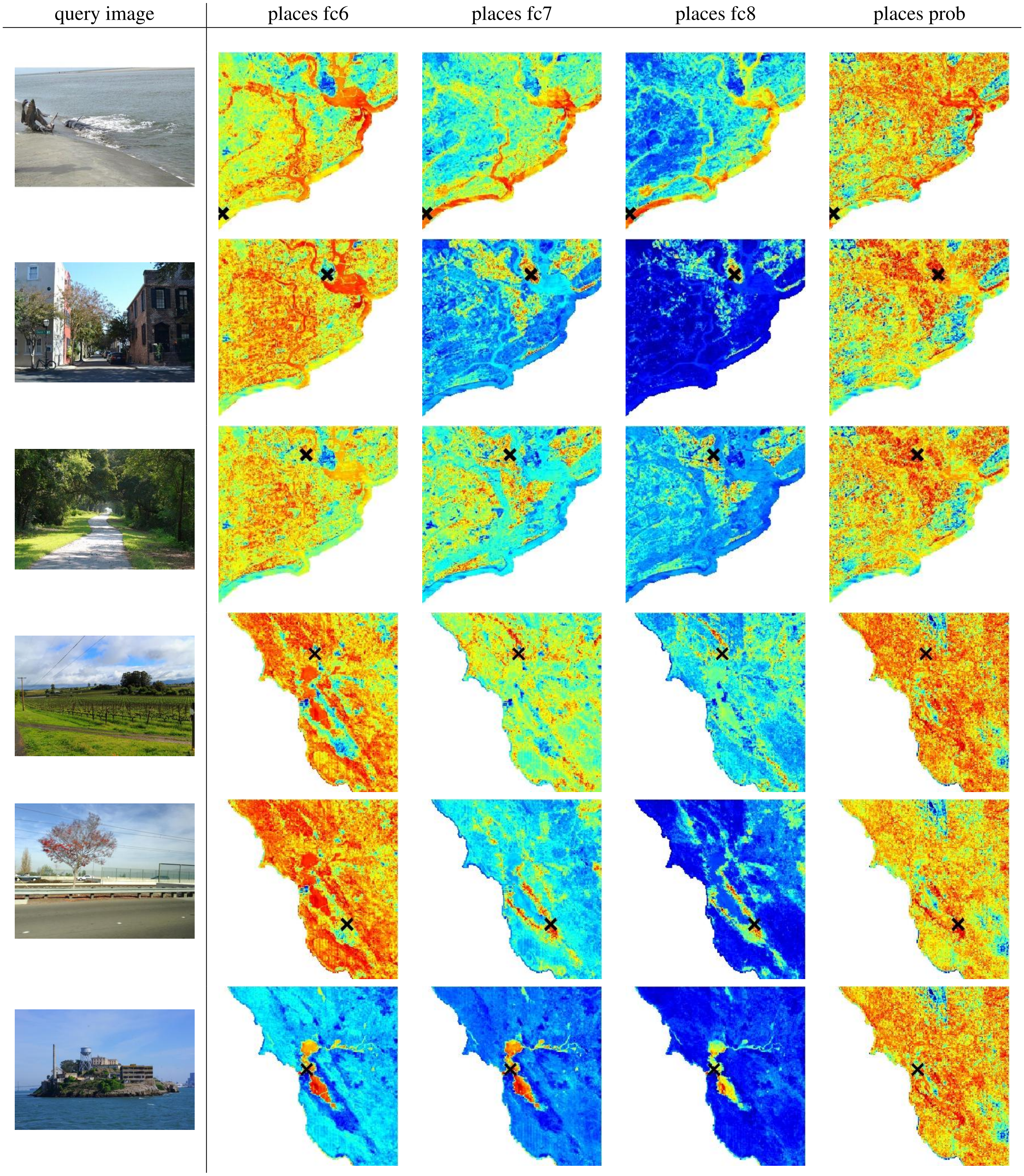


Figure 10: False-color images that represent the likelihood that an image is at a particular location. In each, red represents high likelihood, blue represents low, and the ‘x’ marks the true location. See Section 4.3 for an algorithm description.

References

[1] G. Baatz, O. Saurer, K. Köser, and M. Pollefeys. Large scale visual geo-localization of images in mountainous terrain. In *European Conference on Computer Vision*, 2012.

[2] D. M. Chen, G. Baatz, K. Koser, S. S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, et al. City-scale landmark identification on mobile devices. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[3] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *International World Wide Web Conference*, 2009.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[5] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *ACM Transactions on Graphics (SIGGRAPH)*, 31(4), 2012.

[6] C. Dong, C. C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, 2014.

[7] Q. Fang, J. Sang, and C. Xu. Discovering geo-informative attributes for location recognition and exploration. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(1s):19, 2014.

[8] P. Fischer, A. Dosovitskiy, and T. Brox. Descriptor matching with convolutional neural networks: a comparison to sift. *arXiv preprint arXiv:1405.5769*, 2014.

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[10] J. Hays and A. A. Efros. Im2gps: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[11] M. T. Islam, S. Workman, H. Wu, N. Jacobs, and R. Souvenir. Exploring the geo-dependence of human face appearance. In *IEEE Winter Conference on Applications of Computer Vision*, 2014.

[12] N. Jacobs, K. Miskell, and R. Pless. Webcam Geo-localization using Aggregate Light Levels. In *IEEE Workshop on Applications of Computer Vision*, 2011.

[13] N. Jacobs, N. Roman, and R. Pless. Toward Fully Automatic Geo-Location and Geo-Orientation of Static Outdoor Cameras. In *IEEE Workshop on Applications of Computer Vision*, 2008.

[14] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. Geolocating Static Cameras. In *IEEE International Conference on Computer Vision*, 2007.

[15] M. Jaderberg, A. Vedaldi, and A. Zisserman. Deep features for text spotting. In *European Conference on Computer Vision*, 2014.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[19] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (SIGGRAPH)*, 33(4), 2014.

[20] T.-Y. Lin, S. Belongie, and J. Hays. Cross-view image geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[21] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[22] D. Quercia, N. K. O'Hare, and H. Cramer. Aesthetic capital: what makes london look beautiful, quiet, and happy? In *ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2014.

[23] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014.

[24] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[25] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Transactions on Graphics*, 25(3):835–846, 2006.

[26] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[27] A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[28] S. Workman, R. P. Mihail, and N. Jacobs. A Pot of Gold: Rainbows as a Calibration Cue. In *European Conference on Computer Vision*, 2014.

[29] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. In *Advances in Neural Information Processing Systems*, 2014.

[30] B. Zhou, L. Liu, A. Oliva, and A. Torralba. Recognizing city identity via attribute analysis of geo-tagged images. In *European Conference on Computer Vision*, 2014.