– Independent Work Report Fall, 2016 –

A Deep Representation for Ground-to-Ground Geolocalization Through Fuzzy Supervised Learning

2017

Daway Chou-Ren

March 3, 2017

1. Note

I'm not sure whether to frame this as an image similarity, image geolocation, or image classification task. I hope to use distance in an image embedding space to geolocate a test image, and then doing image classification is very easy after that since I would just take the most frequent label among its k nearest neighbors by geographic distance.

2. Abstract

Typically, the task of learning fine-grained image similarity relies on obtaining carefully labeled data about image classes, such as whether an image displays a robin or a blue jay, and then training a model on this data to detect both inter-class and intra-class differences. Some recent approaches have begun using free, noisy data from the web to learn similarity on fuzzy classes, yet these works still rely on using a notion of an image class. In this paper, we explore learning a deep representation useful for predicting the visual similarity of iamges without assigning them to a class in a training set. We use deep convolutional networks in a Siamese configuration rather than rely on handcrafted features, and train our model with an extremely weakly supervised process, relying on the geographic distance of images to serve as a proxy for their visual similarity.

3. Introduction

What does it mean for two images to be similar? Are two red-hued images similar, even if they are of different objects? Are an image of a cardinal and an image of a blue jay similar because

they both show birds? Regardless of the semantic interpretation of the word 'similar', any model that determines whether or not two images are similar must first be able to embed them in a visual representation.

Learning visual representations for images has been important for a variety of tasks, including image classification, semantic segmentation, object detection, and even geolocating images. The methods of extracting image representations have changed greatly over the years, moving from manually defined features like histograms of oriented gradients (HOGs) and scale-invariant feature transforms (SIFT)[23][9], to the current state of the art of extracting feature vectors from convolutional neural networks, beginning with the seminal work of Krizhevsky et al. which achieved classification results for the 2010 Imagenet ILSVRC image classification contest far surpassing previous state-of-the-art benchmarks[18]. Since deep learning was proved to be highly effective at image tasks in 2010, the basic two step process for building image understanding models has remained largely the same. First, large quantities of reliably labeled data must be gathered, and then a model, usually a CNN, can trained on this data. Often, Mechanical Turkers are employed to create highly pure labeled datasets.

The popularization of large scale image datasets such as ImageNet, which contains 14,197,122 images belonging to 1000 classes[10], the MIT Places dataset which contains 7 million images for scene classification[41], the SUN scene classification database[36], and the Microsoft COCO dataset of 2.5 million images for common objects in context[22] have allowed researchers to build models highly adept at basic image classification tasks[28]. In a review of deep learning models trained on these massive datasets with many basic image classes (ImageNet contains classes for many animals and plants and items such as 'tennis ball', 'fountain pen', and 'tricycle') Russakovsky et al. concluded that deep learning techniques were able to transfer learning from these dataset classes to other generic classes, such as distinguishing dogs from airplanes.[28] However, more fine-grained image classification, such as between species of flowers, or of dogs of different ages,

required learning different image embeddings.

After the deep learning revolution, a major direction of image understanding research has been to develop more fine-grained datasets for the training of models for highly specific tasks. A quick search for image datasets published in 2016 returns ones for irises, ultrasounds, weather property, tumors, light fields, and food calories.[5][8][7][29][25][26]. Yet, although we have been able to train more and more specific models for finer and finer grained image classification, this research still relies on the gathering of accurately labeled data. It is infeasible to gather large quantities of data for every possible image understanding task.

Noisy Supervised Training

Some more recent approaches have looked into augmenting highly supervised training with weakly supervised web data, thus greatly increasing the amount of data available for these highly specific image understanding tasks. Xu et al.[37] use existing datasets to learn feature representations and part-based object classifiers. They then extract accurate part labels from fuzzy web image data. Kraus et al take this line of work even further and use generic image recognition techniques on noisy web data and exceed state-of-the-art classification accuracies on the CUB-200-2011 dataset, without using any manually labeled data.[17]

Our Approach

This paper follows this recent work in utilizing the large quantity of image data available online through search engines and image hosting sites like Google and Flickr. Rather than use web queries to form fuzzy classes for image classification like Kraus et al, we seek to use geo-tagged images uploaded to Flickr to learn an image embedding useful for image similarity tasks. We do not train on any image pairs manually labeled as similar but rather rely on the physical geographic distance between two images to inform the training of our model. We explore how well geographic distance

can serve as a stand-in for manually labeled similarity data, with a particular focus on exploring heuristics for sampling pairs of similar images for maximal learning efficiency.

We choose to use deep learning representations of images trained through a Siamese network rather than use manually crafted features such as Gabor filters, scale-invariant feature transforms (SIFT), or histograms of oriented gradients (HOG), believing that allowing a model to learn features on its own will be more robust. In a departure from the research of Xu et al and Kraus et al, we also work with a classless representation of our data. We do not assign an image to a fuzzy class, such as belonging to the category 'bridge', but associate it only with its latitude and longitude data. In a sense, this means we treat each image as belonging to a unique geolocation class of size 1.

Using our developed heuristics for sampling pairs of similar images for comparison with dissimilar images, we train a deep Siamese network to learn a low dimensional feature representation, with an objective of learning that pairs of images close in physical distance should be closer in our image embedding space.

The contributions of this paper are [to be completed]

4. Related Work

4.1. Effectiveness of Deep Learning

The field of image similarity relies on learning a useful model for embedding images into a feature space. Recent years have seen some groundbreaking advancements in the application of machine learning for vision tasks, especially in the field of deep learning. Convolutional neural networks[20] are capable of learning low, medium, and high level features, using nonlinear transformations to abstract high level features into more and more basic ones. Krizhevsky et al's momentous performance in the 2012 ILSVRC ImageNet classification competition provided the first demonstration of the effeciveness of deep learning.[18] Recent research has shown that deeper models can perform even

better at a variety of image tasks.[32] Much work on different activation functions has allowed CNNs to become much more sparse, and combined with work exploring deep network depths[30][32] as well as with work allowing models to regulate their own depth[12], deep CNNs have proven extremely effective at learning a variety of useful image representations. Athiwaratkun and Kang show that just the image representation extracted by deep CNNs can be combined with simpler classifiers such as SVMs and random forests to achieve high accuracies for clustering tasks.[2] In recent years, some deep learning models have even achieved classification accuracies surpassing even human performance. In 2015, He et al. achieved a 4.94% top-5 test error on the ImageNet 2012 dataset, surpassing the human error performance of 5.1%[12].

Importantly, deep learning methods do not require the manual crafting of features based on domain-level knowledge. Instead, deep learning models learn to abstract patterns from data automatically. This approach stands in contrast to the majority of work done in the 1990s and 2000s, which made extensive use of manually defined image feature extraction techniques, such as Gabor filters, scale-invariant feature transforms (SIFT), and histograms of oriented gradients (HOG).[13][23][9] In the latter half of the 2000s, hierarchical feature representations such as spatial pyramids, which transform images into segmentations, each of which is locally orderless, proved effective in a variety of image tasks as well[38][11][19]. In the field of content-based image retrieval, spatial envelopes and transformed histograms where used to attempt to capture global scene properties[24][35]. Features extracted using these methods were then used with rigid distance functions such as Euclidean or cosine similarity distances to determine an overall image similarity. Much work has been done on designing better similarity measures for these low-level features. Notably, Jegou et al.[14] adapt the Fisher kernel for use in aggregating local image descriptors into a reduced dimension vector while preserving the bulk of relative distance information.

4.2. Deep Convolutional Neural Networks

Convolutional neural networks have become the de-facto standard in image tasks, as stacked convolutional layers are well suited for learning image descriptors.[15][18][32] Besides convolutional layers, CNNs typically consist of pooling layers, activation layers, fully connected layers, and a loss layer. A convolutional layer consists of a set of kernels K, each of which typically has width and height dimensions smaller than the dimensions of an inputted image, but with a depth matching the depth of the input. During the forward pass of network training, each filter is convolved with a sliding patch across the input's width and height, producing a feature map associated with that kernel. Each feature map codes the activation of the kernel along with the spatial location of that activation. Because the dimensions of K are smaller than the input, convolutional layers only have local connectivity.

Pooling layers, usually in the form of max-pooling, down sample the feature maps produced by the convolutional layers. Max pooling will output the maximum value in each part of a segmentation of a feature map. Pooling layers drastically reduce the computation required to train a network. Activation layers are used to control which nodes in a layer send output to the next layer. The standard activation currently used is the rectified linear activation unit (ReLU), which takes the form

$$f(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{1}$$

Various forms of ReLU have been proposed, notably the parametric ReLU (pReLU), which adds a learnable parameter, α, to control a slope for the negative activation domain and which was used by He et al. to achieve better than human performance for ImageNet classification.[12]

$$f(x) = \begin{cases} x & x > 0 \\ \alpha x & x \leq 0 \end{cases} \tag{2}$$

Since convolutional layers only produce feature maps on local scales, fully connected layers at the end of a CNN allow for high level features to be learned. These layers have full connections to all activated neurons from previous layers, which allow for global mixing of activated feature maps. To complete our discussion of the CNN, the loss layer specifies how a network should penalize incorrect predictions. Typically sigmoid cross-entropy loss is used. Regularization is also used, typically in the form of L1 or L2 weight decay. Dropout layers, which deactivate a random subset of a layer's neurons in each iteration of training, have also proved highly effective for neural network regularization.[31]

The term deep CNN refers to a CNN that has many layers. This definition is highly variable: the popular VGG16 model has 16 layers[30] and versions of ResNet contain 50, 101, and 152 layers[12].

We can more precisely define a CNN as a function f that takes parameters θ and an input image I to produce an image embedding x and a loss L.

4.3. Siamese Neural Networks

The Siamese neural network, first proposed by Bromley et al. in 1994[6] for the purposes of verifying signatures, is a formulation of two copies of a CNN that sharing the same parameters and hyperparameters, as well as their loss layer. Siamese nets are thus well suited for producing image embeddings through pairwise training: the network f takes θ, two images I_1, I_2 as well as an indicator variable p to indicate if these images form a positive (similar) or negative (dissimilar) pair, produces two embeddings x_1, x_2 and one loss L. Siamese networks have been used in a variety of tasks, such as ground-to-aerial geolocalization[21], matching visual similarity for product design[3], comparing image patches[39], and one-shot image classification[16].

5. Distance Metric for Siamese Network

The similarity of two images, $S(A,B)$, can be defined as the Euclidean distance of their feature embedded vectors, $f(A)$ and $f(B)$:

$$S(A,B) = ||f(A) - f(B)||_2^2 \tag{3}$$

Here $f(\cdot)$ might be a feature embedding such as the weight representation of the final convolutional block in a convolutional neural network pretrained on ImageNet.

We use pairwise comparisons of images, grouping sets of three images, p_i, p_i^+, and p_i^-, into two pairs, (p_i, p_i^+) and (p_i, p_i^-). We can compute a hinge loss for these triplets:

$$l(p_i, p_i^+, p_i^-) = \max\{0, g + ||f(p_i) - f(p_i^+)||_2^2 - ||f(p_i) - f(p_i^-)||_2^2\} \tag{4}$$

where g serves as a regulator for the gap between the distance of the image pairs.

5.1. Image Embedding as Used for Image Similarity

With the effectiveness of deep learned image features to tasks like classification, detection, and segmentation, the application of deep learned image features to image similarity is immediate: if a highly accurate ImageNet-trained model predicts two images to belong to the same class, they can be considered more similar than two images belonging to different classes. Distance functions applied to the 1000-class softmax probability outputs of such a model can be used to retrieve a more fine-grained image similarity score than this same-category categorical comparison.

However, the class based approach to image similarity only goes so far as our training data. As Russakowsky et al. discussed[28], pretrained models were only good at segmenting images into basic classes, and demonstrated that they were better tuned for classification of natural classes

such as animals than they were for man-made objects. In recent years, a proliferation of intra-class datasets, such as for hundreds of species of flowers or birds, have allowed deep learning techniques to tackle everything from differentiating species of flowers[1], leaves[27], and birds[4], but the limitations of a class-based formulation of image similarity remain apparent.

5.2. Fuzzy Supervised Training

for the task of image similarity. Krizhevsky et al achieved significantly better sults on the 2010 ILSVRC competition than any previous results.

Following this success, and the widespread use of open-sourced deep learning models for image classification tasks, Russakovsky et al concluded that such approaches were highly accurate at distinguishing basic classes from each other, such as dogs from airplines, but struggled on fine-grained classification.[28] In the years since, a proliferation of intra-class datasets, such as for hundreds of species of flowers or birds, have allowed deep learning techniques to tackle everything from differentiating species of flowers[1], leaves[27], and birds[4].

Similar to classification, much work has been done on object localization or bounding box detection.[40]

The explosion in the effectiveness of deep learning techniques for vision tasks in the past decade has allowed for much progress to be made in the field of fine-grained image similarity. As opposed to image similarity, which is concerned with whether two images display objects belonging to the same class, such as if two images contain robins or if one image is of a blue jay, fine-grained similarity requires a model capable of embedding images into a similarity space that allows for intra-class comparison: of three images of blue jays, which two are more similar?

5.3. Geolocalization

To build PlaNet, a CNN trained to geolocate photos on a global scale, Weyand et al also use a massive geotagged Flickr dataset.[34] However, they look at the geographic density of their photos and divide the earth into variable size latitude longitude bounding boxes which they use as image classes. The existence of classes allows Weyand et al to frame geolocalization as a standard classification problem to which a CNN is easily applied.

6. Related Work

As noted by Wang et al[33], the network structures that are effective at classifying images into object classes are not necessarily well-designed for detecting image similarity, especially when image similar is defined not just as whether two objects are in the same class, but when our desired similar metric must be fine enough to rank similarities within classes. For example, a red book should be judged more similar to a maroon book and should a light green one.

7. Data

Our data consists of all geo-tagged images uploaded to Flickr between 00:00:00 (GMT) January 1, 2006 to 00:00:00 January 1, 2017 with latitude and longitude inside the bounding box [-74.052544, 40.525070, -73.740685, 40.889249]. This bounding box roughly corresponds to the city limits of New York, New York.

8. Network Architecture and Training

9. Testing

10. Results

11. Conclusion

Figure 1: Location of images

References

[1] A. Angelova, S. Zhu, and Y. Lin, "Image segmentation for large-scale subcategory flower recognition," in *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*. IEEE, 2013, pp. 39–45.

[2] B. Athiwaratkun and K. Kang, "Feature representation in convolutional neural networks," *arXiv preprint arXiv:1507.02313*, 2015.

[3] S. Bell and K. Bala, "Learning visual similarity for product design with convolutional neural networks," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, p. 98, 2015.

[4] T. Berg, J. Liu, S. Woo Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2011–2018.

[5] K. W. Bowyer and P. J. Flynn, "The nd-iris-0405 iris image dataset," *arXiv preprint arXiv:1606.04853*, 2016.

[6] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *IJPRAI*, vol. 7, no. 4, pp. 669–688, 1993.

[7] W.-T. Chu, X.-Y. Zheng, and D.-S. Ding, "Image2weather: A large-scale image dataset for weather property estimation," in *Multimedia Big Data (BigMM), 2016 IEEE Second International Conference on*. IEEE, 2016, pp. 137–144.

[8] C. Cortes, L. Kabongo, I. Macia, O. E. Ruiz, and J. Florez, "Ultrasound image dataset for image analysis algorithms evaluation," in *Innovation in Medicine and Healthcare 2015*. Springer, 2016, pp. 447–457.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[13] A. K. Jain, N. K. Ratha, and S. Lakshmanan, "Object detection using gabor filters," *Pattern recognition*, vol. 30, no. 2, pp. 295–309, 1997.

[14] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 9, pp. 1704–1716, 2012.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[16] G. Koch, "Siamese neural networks for one-shot image recognition," Ph.D. dissertation, University of Toronto, 2015.

[17] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei, "The unreasonable effectiveness of noisy data for fine-grained recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 301–320.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[19] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. IEEE, 2006, pp. 2169–2178.

[20] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[21] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5007–5015.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[23] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[24] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.

[25] P. Paudyal, R. Olsson, M. Sjöström, F. Battisti, and M. Carli, "Smart: a light field image quality dataset," in *Proceedings of the 7th International Conference on Multimedia Systems*. ACM, 2016, p. 49.

[26] P. Pouladzadeh, A. Yassine, and S. Shirmohammadi, "Foodd: food detection dataset for calorie measurement using food images," in *International Conference on Image Analysis and Processing*. Springer, 2015, pp. 441–448.

[27] A. Rejeb Sfar, N. Boujemaa, and D. Geman, "Vantage feature frames for fine-grained categorization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 835–842.

[28] O. Russakovsky, J. Deng, Z. Huang, A. C. Berg, and L. Fei-Fei, "Detecting avocados to zucchinis: what have we done, and where are we going?" in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2064–2071.

[29] J. Shi, S. Zhou, X. Liu, Q. Zhang, M. Lu, and T. Wang, "Stacked deep polynomial network based representation learning for tumor classification with small ultrasound image dataset," *Neurocomputing*, vol. 194, pp. 87–94, 2016.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[31] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[33] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[34] T. Weyand, I. Kostrikov, and J. Philbin, "Planet-photo geolocation with convolutional neural networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 37–55.

[35] J. Wu and J. M. Rehg, "Centrist: A visual descriptor for scene categorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 8, pp. 1489–1501, 2011.

[36] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*. IEEE, 2010, pp. 3485–3492.

[37] Z. Xu, S. Huang, Y. Zhang, and D. Tao, "Augmenting strong supervision using web data for fine-grained categorization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2524–2532.

[38] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[39] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4353–4361.

[40] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *European conference on computer vision*. Springer, 2014, pp. 834–849.

[41] 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, pp. 487–495.