AN INDUCTIVE MAPPING WITH CONVOLUTIONAL REPRESENTATIONS FOR HUMAN SETTLEMENT DETECTION: PRELIMINARY RESULTS

Dalton Lunga, Dilip Patlolla, Lexie Yang, Jeanette Weaver, and Budhendra Bhadhuri

Computational Sciences and Engineering Division
Oak Ridge National Laboratory

ABSTRACT

We test this premise and explore representation spaces from a single deep convolutional network and their visualization to argue for a novel unified feature extraction framework. The objective is to utilize and re-purpose trained feature extractors without the need for network retraining on three remote sensing tasks i.e. superpixel mapping, pixel-level segmentation and semantic based image visualization. By leveraging the same convolutional feature extractors and viewing them as visual information extractors that encode different image representation spaces, we demonstrate a preliminary inductive transfer learning potential on multiscale experiments that incorporate edge-level details up to semantic-level information.

Index Terms— settlement mapping, segmentation, representation learning, convolutional neural networks, inductive transfer learning.

1. INTRODUCTION

Even though the problem of image understanding has a long history in computer vision applications; recent breakthroughs in high performance computing and the availability of large overhead imagery are leading the cause for its surging appeal for disruptive remote sensing (RS) applications. Fueled by the success of methods including deep convolutional neural networks (CNN) in multimedia image classification, similar efforts are being sought for high resolution RS applications that not only achieve human level performance in terrestrial object classification but achieve semantic labeling and building extraction at large scale. However, due to its spatio-temporal nature, RS data offers unique challenges that necessitates the design of new deep neural architectures. For example, by treating building extraction as an object localization challenge a deep learning implementation requires that both spatial and semantic image representations are combined towards training a convolutional classifier for accurate building-level detection[1]. Theoretical frameworks for understanding CNN architectures are yet to be fully explored. However, in the past few years, visualization technologies have emerged to close the gap by providing valuable insights on the information extraction stages of CNNs. Such crucial understanding is attributed to a variety of top performing deep CNNs[2, 3, 4]. Training deep CNNs is computationally expensive and requires elusive knowledge on hyper-parameter tuning. Mild efforts are being directed to study the potential for multiple tasks to leverage on shared image CNN representations[5].

It is with the above motivation that this paper seeks to explore the same representations towards image-level classification, pixel-level segmentation and semantic neighborhood mapping without the need for CNN retraining. An investigation is conducted by visually seeking to understand image representations via probing internal activation maps during the CNN forward pass process. Preliminary results shows that with limited labeling information, a unified representation learning of human settlement structures with overhead imagery has greater potential. Re-purposing of CNN feature extractors from course labels to fine-grained and semantic mapping tasks seeks to inform their wide applicability in RS. Using the visual understanding of CNNs, we highlight the following: (1) inductive transfer learning capabilities with a unified representation feature extractor toward multiple human settlement mapping tasks, (2) use semantic representational space to understand a collection of images, and (3) seeking maximal activation maps to obtain insights on per-class image characteristics to inform unique design of RS driven CNN architectures.

Image-level ground truth acquired from **0.5**-meter overhead imagery is used to train a CNN towards superpixel human settlement mapping. Training data consists of $40,000$ image patches, each of size 144×144 pixels. From 4-large tile scenes we crop out the training patches on a grid including $20,000$ image patches for the validation set. A second set of $20,000$ image patches is created from a different geographic location for generating a second semantic neighborhood mapping result. The CNN architecture consist of **7**-weight layers including **4**-convolutional (**conv**) layers, **2**-fully connected

(**fcn**) layers, and **3**-maxpooling(**pool**) layers. **Pool** layers are configured after each **conv** layer. CNN odel parameters are obtained via a stochastic gradient descent (SGD) technique based on the back-propagation framework. The SGD learning rate is set to 0.00273 via a full hyper-parameter gridsearch, while the activation is performed with ReLU, filter weights initialized from a normal distribution, and the batch size set to 150. The trained CNN feature extractors are re-purposed to assess two different tasks i.e. (1) pixel-level mapping, (2) semantic and topological mapping. We present our early observations and highlight the essentials of probing CNN maps to inform multiscale tasking with a single representational learning learning framework.

2. RELATED WORK

Visual understanding of deep neural network architectures has enabled the capability to extract valuable insights on the internal transformations performed by the CNN filtering process in many computer vision tasks. Visual probing for insight extraction in CNNs can be traced back to the early work in [6], where a direct projection of first layer filters to the image space was performed to assess the learning capacity of the network. A technique to project intermediate and deeper CNN activation maps was demonstrated in [7] via a deconvolution process to yield reconstructed image encodings that were interpreted via the discriminant strong information from the input image pixels. In [3], image patches that maximize selected neuron activation maps were sought by performing a gradient ascent in the image space with the goal of studying their pixel level characteristics. In [8], the authors extend the method towards seeking input images that trigger similar neural stimulation for a given layer. The work in [4] sought to visualize in a more comprehensive manner the representation spaces constructed by all filters of a layer.

We draw lessons from this body of literature to seek insights on the representation spaces constructed by CNNs with **0.5**-meter single band remote sensing imagery. Our main objective being to leverage the representation spaces obtained with image-level ground truth toward assessing fine-grained settlement detection and mapping of images onto a semantic topological geometry.

3. INDUCTIVE MAPPING WITH OVERHEAD IMAGES

Remote sensing overhead image data reside on a spatio-temporal grid and this is in contrast to independent and identically distributed sample-based methodologies widely adopted in traditional machine learning. Human settlement mapping is a typical challenge that requires that an image representation should take the grid nature of the data into account. CNNs have proven to be applicable in leveraging the spatial image grid[6, 9]. Given the range and complex nature of human settlement understanding with overhead imagery, the inductive transfer learning approach is to exploit visualization driven insights and demonstrate a single CNN framework on: (one) superpixel classification of homogeneous regions into single categories, (two) pixel-level classification to seek fine-grained detection of settlement structure boundaries and (three) using image-patches to compute for semantic level neighbourhood mapping. Figure 1 illustrate the generalized conceptual architecture for the envisioned unified representational learning framework with overhead imagery.

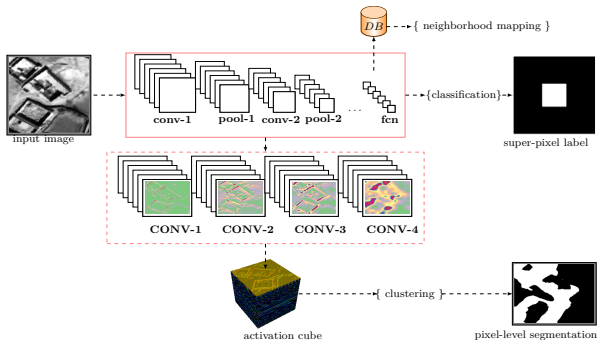


Fig. 1: A unified representational learning with CNN and inductive settlement extraction framework.

3.1. Superpixel based settlement detection

Homogeneous region classification is emerging in popularity with remote sensing imagery. Clear-cut boundaries are sought to distinguish classes (e.g. urban vs forest)[10]. We illustrate the potential applicability of CNN features on a 16×16 superpixel block region for human settlement mapping. We posit that using a large volume of local homogeneous patch level ground truth for training the representation learning framework, could offer greater potential to stimulate automatic learning of coherent local structures that are characteristic and common with overhead imagery containing human settlements. The hierarchical and deeper feature abstraction by CNNs could efficiently generate scale invariant representations that are favourable for seeking homogeneous regions. Figure 2(a) and (b) illustrate the superpixel based mapping generated with a softmax classification on the **fcn** features of the CNN.

3.2. Hypercolumns for settlement detection

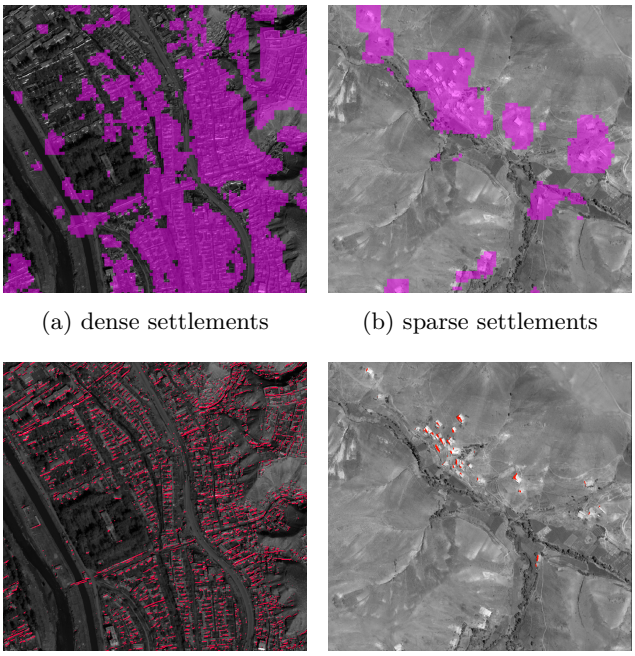
It is inarguable that, at large, recognition algorithms based on CNNs typically use the output of the **fcn** layer as a feature representation. However, as shown by the example in Figure 1, the information from the top **conv** layers and the **fcn** layers appears to be too coarse spatially to enable precise pixel level settlement segmentation. As first demonstrated in [5] and also reflected in Figure 1, lower **conv** layers do retain edge detection information that may be precise in detecting settlement boundaries albeit not capture higher level semantic details to describe the settlement as a whole. Using the approach of [5], we define the hypercolumn at a pixel as the vector of activations of all CNN layers above that pixel. Using the hypercolumns as pixel descriptors, we leverage on image level trained CNN feature extractor to generate the fine-grained mapping shown in Figure 2(c) and (d). The algorithmic implementation utilizes a mini-batch K-means clustering algorithm to generate the pixel-level segmentation results. Figure 2(c) shows the pixel-segmentation results detecting additional settlements that are appear to be omissions in the top-left corner of the superpixel mapping in (a).

3.3. Semantic and topological mapping

Semantic image visualization using very high resolution remote sensing imagery has emerged as another challenging application in the past decade. The most recent attempt at this challenge presented a semi-supervised learning framework employing the notion of superpixel tessellation representations of imagery[10]. The image representation utilizes homogeneous and irregularly shaped regions and relies on hand designed features based on intensity histograms, geometry, corner and superpixel density and scale of tessellation. Of relation, we demonstrate the potential of leveraging top layer **fcn** representations, toward a semantic image representation visualization with thousands of image patches cropped from large scenes. Although the intermediate stages of the CNN could offer representations useful in related mapping tasks, the result in Figure 3 reveals a more homogeneous image content mapping based on top level features. By applying clustering methods in the projected space course level neighborhood segmentation map can potentially be generated.

4. IMAGE INFORMATION EXTRACTION

Each stage of the **conv** filtering process retains class related discriminatory information with the top **conv** and **fcn** retaining object specific semantic information. Figure 3 shows a large scale semantic patch representation obtained with same CNN model for two different geographic locations. The result in (a) is a manifold learning projection[11, 12] of the **fcn** layer features. The projection shows patches close(similar in content) in the **fcn** representation space embedded close in the two-dimensional topological space. The gradual information extraction by CNNs can be visualized to gain useful insights for image representation understanding. As shown in (c), highest activation feature maps can be visualized on each layer to reveal discriminatory and useful information for relevant multiscale analysis. Edge-level and semantic level discriminatory information is consistently extracted across the six example patches in (b) containing *houses, rocks, roads* and *trees.*



(a) dense settlements (b) sparse settlements

(c) dense settlements (d) sparse settlements

Fig. 2: Illustrating superpixel mapping(top row) and pixel-level terrain segmentation(bottom row) on **0.5**-meter 1728×1728 pixel aerial imagery.

5. CONCLUSION

A preliminary investigation is conducted to exploit image-level ground truth towards a single CNN representational model in multiscale human settlement understanding with **0.5**-meter resolution overhead imagery. We visually contrast our early results for superpixel mapping and the hypercolumn induced pixel-level segmentation. The investigation demonstrates a surprising potential to obtain reasonable fine-grained human settlement mapping from image-level ground truth. This outcome offers further grounds to expand our efforts to seek multiple objective functional forms towards RS mapping

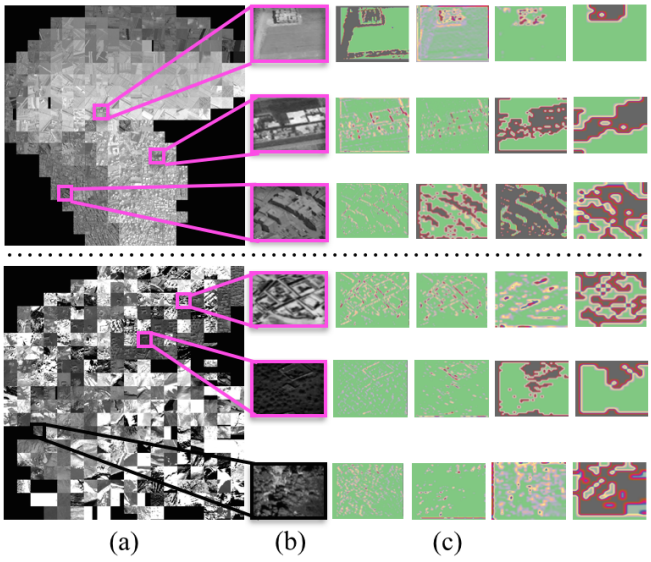


Fig. 3: Topological and semantic mapping of image information for Egypt (top row) and Afghanistan (bottom row). Column (a) shows the semantic neighborhood embedding plane using **fcn** representations. Column (b) are example patches from (a). Column (c) shows reconstructed maximally activated **conv** feature extractors for example patches in (b).

with a unified representation learning framework. In addition, the discriminatory capable **fcn** features when transformed via manifold projections pointed to another direction for seeking image segmentation at semantics level - a very promising avenue for seeking region-based RS semantic neighborhood mapping.

Acknowledgement

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

The authors would like to thank Dr. Jiangye Yuan and Dr Mark Coletti for discussions on related topics.

6. REFERENCES

[1] J. Yuan, "Automatic building extraction in aerial scenes using convolutional networks," *CoRR*, 2016.

[2] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 1–9, IEEE.

[3] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," 2014.

[4] W. Yu, K. Yang, Y. Bai, T. Xiao, H. Yao, and Y. Rui, "Visualizing and comparing alexnet and vgg using deconvolutional layers," in *Proceedings of the 33rd International Conference on Machine Learning Workshop*, 2016.

[5] B. Hariharan, P. Andrés Arbeláez, R. B. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," *CoRR*, vol. abs/1411.5752, 2014.

[6] A. Krizhevsky, S. Ilya, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, Eds., pp. 1097–1105. 2012.

[7] M.D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," *CoRR*, vol. abs/1311.2901, 2013.

[8] M. Aravindh and V. Andrea, "Understanding Deep Image Representations by Inverting Them," 2014.

[9] S. Paisitkriangkrai, J. Sherrah, P. Janney, and A. van den Hengel, "Semantic labeling of aerial and satellite imagery," vol. 9, no. 7, pp. 2868–2881, 2016.

[10] M. Sethi, Y. Yan, A. Rangarajan, R. R. Vatsavai, and S. Ranka, "Scalable machine learning approaches for neighborhood classification using very high resolution remote sensing imagery," in *Proceedings of the 21th International Conference on Knowledge Discovery and Data Mining, Australia, August 10-13, 2015*, 2015, pp. 2069–2078.

[11] D. Lunga, S. Prasad, M. M. Crawford, and O. Ersoy, "Manifold-Learning-Based Feature Extraction for Classification of Hyperspectral Data," *Signal Processing Magazine, IEEE*, , no. 1, pp. 55–66, 2014.

[12] L. van der Maaten and G.E. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9: 2579–2605, 2008.