Transfer Learning for Urban Feature Extraction in High Resolution Satellite Data

Anukriti Pathak (1), Richa Joshi (1), Poonam S. Tiwari (2), Hina Pande (2), Shefali Agrawal (2)

[1]Banasthali Vidyapeeth, Rajasthan, India
[2]Indian Institute of Remote Sensing, ISRO, Dehradun
E-mail : poonam@iirs.gov.in

Abstract : With the increasing population and hence urbanization, there is more need than ever to work upon and improve automatic urban feature extraction methods. Although many methods have been proposed over decades, urban feature extraction is still a challenging problem due to complexity of the scene characteristics. Urban feature extraction in high spatial resolution remote sensing images plays an important role in city planning, navigation, population estimation and many other applications.

Deep convolutional neural network (CNN) transfer has recently shown strong performance in scene classification of high-resolution remote-sensing images. Recently developed extensions of the CNN frameworks made it possible to perform dense pixel-wise classification of input images. CNN is rarely used from scratch due to limitations in availability and huge effort required for the generation of training data. Transfer learning is a methodology where pre-trained or pre-learned model on one sort of information can be utilized on another comparative sort of information without training the model from scratch. The objective of this study is to explore the possibility of using convolutional neural networks for efficient extraction and improved performance when using high resolution satellite images.

In this study, ResNet- 50 is used as backbone for UNet architecture. ResNet-50 model comprises of large number of layers but with reduced complexity and also skip associations. Skip connections enable the model to learn identity function that helps solve gradient vanishing problem. The architecture is used on a multispectral high resolution satellite image of a part Gandhinagar, Gujarat. The pre-trained deep CNN models were fine-tuned on the dataset. Training results obtained from the minimal dataset yielded an accuracy of 79.29% and an Intersection over union (IoU) score of 0.6152. The results from larger dataset displayed an increase in accuracy to 83.90% and an IoU score of 0.7022. The accuracy of extraction was evaluated based on, four indicators - False Alarm (FA), Missed Alarm (MA), Overall Alarm Rate (OA), and Kappa. The results on real dataset indicate that the selected model is able to extract 83% of feature of interest correctly. The improvement in accuracy indicates that the model is effectively transferring the learning to the next level.

It can be concluded that, as opposed to the minimal dataset, splitting the image into large dataset can be used for improving the accuracy of prediction. Not only does it increase the training accuracy greatly, but it also affects the ground truth vs. predictions and the IoU score. Thus the study successfully addresses the problem of urban feature extraction using multispectral satellite data based on transfer learning techniques.

Key words: transfer learning, deep CNN, urban features, multispectral satellite data

1. **Introduction:** An image is a visual two-dimensional representation of objects in a real scene. Remote sensing image are representations of parts of the earth surface as seen from space. Remote sensing is the process of acquiring information or data about the earth's surface without being in contact with it by detecting and monitoring the physical characteristics of an area. Remote sensing data provides essential and significant information that helps in monitoring various applications. Since high resolution remote sensing imagery is becoming more accessible and affordable, urban feature extraction has been of great practical interest [1]. With the advancement of remote sensing technology, the spatial resolution of remote sensing images has been continuously improved. Urban feature extraction in high spatial resolution remote sensing images plays an important role in city planning, navigation, population estimation and many other applications. Spatial resolution refers to the size of the smallest objects that can be distinguished in an image. In digital imagery, spatial resolution is limited by the pixel size of the imagery. A pixel is the smallest two-dimensional area sensed by the remote sensing device.

   Although many methods have been proposed over decades, urban feature extraction is still a challenging problem due to complex scenes. Our current approach has been applied to multispectral satellite images. The purpose of this paper is to perform an extensive research on the possibility of using convolutional neural networks for efficient accuracy and better results on high resolution satellite images. Multispectral imaging is a critical tool for better understanding of image formation and reflectance phenomena. The relevant data of the Earth remote sensing is provided in the form of multispectral images. Multispectral imaging is done by capturing image data within specific wavelength ranges across the electromagnetic spectrum. A multispectral image consists of several bands of differing continuous wavelengths. Multispectral image brings a significant advance in remote sensing in that it is used for variety of applications.

   With the increasing population and hence urbanization, there is more need than ever to work upon and improve automatic urban feature extraction methods. Deep learning, which is a subfield of machine leaning, is based on learning levels of representations [2]. The concept of deep learning comes from the study of Artificial Neural Network. Deep learning algorithms have seen a massive rise in usage and popularity for remote sensing analysis over the past few past years. Neural networks, the basis of deep learning algorithms, have been intensively used in the remote sensing sphere for generic, robust and reliable result. Deep learning is a class of machine learning that uses multiple layers to progressively extract higher layers features from the raw input. There have been many proposed techniques before the introduction of machine learning in remote sensing. Machine learning is an algorithm that learns patterns in data using pre-trained models, and then predicts similar patterns in new data. The traditional algorithms differ from new school algorithms in that the previously used algorithms are programmed to perform a task and not to learn to perform a task. Thresholding, K-means clustering, Histogram-based image segmentation, Edge detection comes under this category. The major barrier to achieve scalable solution with these algorithms is that prior information and assumptions are not generalizable over extended areas [3].

   Therefore, there have been extensive research efforts in applying Convolutional Neural Networks (CNN) [4], [5], [6]. Deep learning is the fastest growing trend in remote sensing technology. Recently developed extensions of the CNN frameworks made it possible to perform dense pixel-

wise classification of input images. CNN is rarely used from scratch due to limited amount of training data. Instead, it is common to take a pre-trained model on a dataset that is significantly larger and more complex than those used in prior works, and transfer its relevant knowledge an initialization for a new task. The trained model achieves a superior performance on datasets with promising and better predicted outputs. This is called transfer learning. Fine-tuning pre-trained models used in transfer learning provide not only better image classification performance and segmentation tasks, but also helps achieve training convergence faster.

The current study has been carried out to identify the optimum convolutional neural network that gives efficient classification of a multispectral high resolution satellite image with urban features extracted with maximum possible accuracy on our given dataset. Remote sensing data deals with real-life applications by classifying and detecting objects. The primary objective of the study is to treat the acquired data and provide accurate results such that deep learning plays an important role.

2. **Theoretical Background:** The architecture used in our study is Convolutional Neural Network (CNN). CNN is a type of deep learning model that is used for image processing and segmentation for better result in accuracy. CNN uses the concept of transfer learning, which is the most common approach for most deep learning algorithms for better predicted results. CNNs are comprised of neurons that self-optimize through learning as shown in figure 1. Each neuron receives an input and performs an operation which is used as an input for the next neuron and so on.

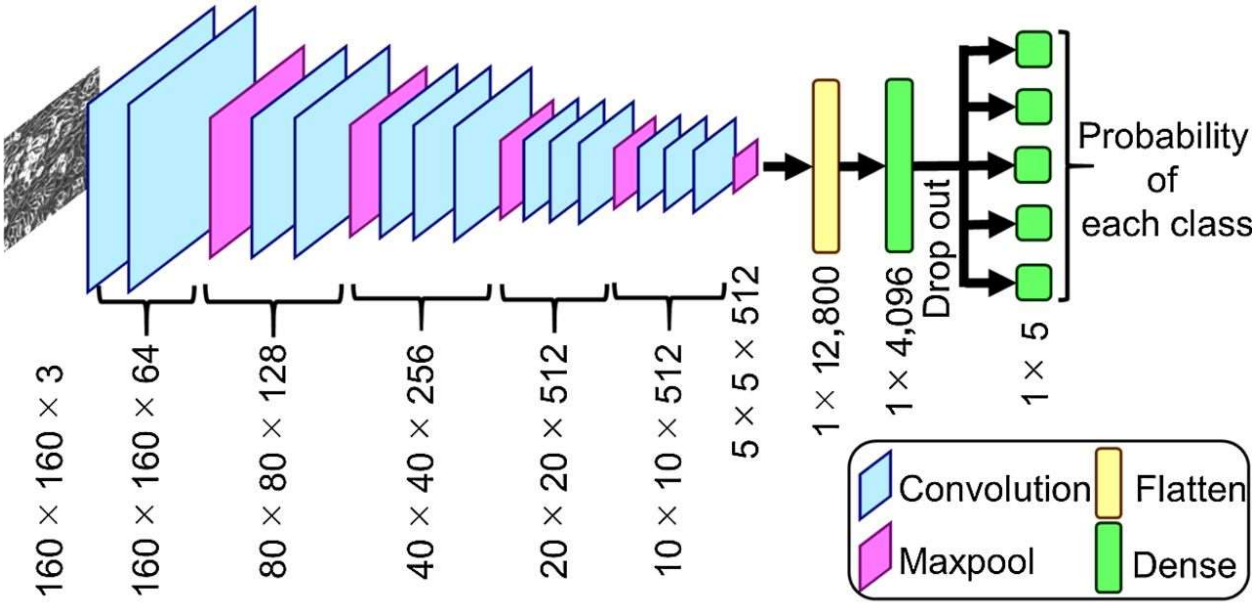


Figure 1: A TYPICAL LAYOUT OF A CONVOLUTIONAL NEURAL NETWORK (CNN)

For better performance, we have applied Resnet50 model in our algorithm. Resnet50 is a convolutional neural network that is trained on more than a million images from the ImageNet database. The network is 50 layers deep and can classify images into 1000 object categories. As a result, the network has learnt rich feature representations for a wide range of images.

There are many hyper parameters used in deep learning algorithms along with Resnet50 which makes the performance more robust, efficient and better in terms of accuracy and speed [7]. Hyper parameters are considered critical in machine learning algorithm, as different hyper parameters on the same training dataset, often result in model with significantly different performance on the testing dataset. Hence, choosing appropriate hyper parameters is important since they directly control the behavior of training algorithm, having significant impact on performance of the model under training. The success of neural network architecture lies in the easier management of a large set of experiments.

- Learning rate (LR): Learning rate is one of the hyper parameter that controls how much a model needs to be changed by determining the step size of iteration while moving toward a minimum loss function, each time the model's weight is updated. If the learning rate is way too smaller than optimal values, it will take a much longer time to reach an ideal state. In this case, overfitting might occur. Overfitting is when a network is unable to learn effectively and the accuracy starts declining because of that. On the other hand, if the learning rate is a lot larger than the optimal value, then it would overshoot the ideal state and the training will diverge.
- Batch size: Batch size is another parameter that controls the number of training samples to work through before any changes are made in any other parameters. Batch size can affect the execution time of training. The batch time is limited by respective hardware's memory. It is often better to use a larger batch size so a larger learning rate can be used.
- Number of epochs: It is a hyper parameter that defines the number of times that the learning algorithm will work through the entire training dataset.
- Intersection over Union (IoU): Intersection over Union is usually used to evaluate the performance of any object category segmentation by measuring the accuracy on a particular dataset. Given an image dataset, the IoU measure gives the similarity between the predicted region and the ground truth region for an object present in the image, and is defined as the size of the intersection divided by the union of the two regions [8].

3. **Practical Approach and Results:** This algorithm has been implemented using the fastai library [10]. Based on top of PyTorch, fastai is a deep learning library that contains some of the most recent and popular algorithms for image classification and segmentation. The fastai library simplifies training fast and accurate neural nets using best practices. We have achieved clear-cut accuracy in predicting the high resolution satellite images by using the Resnet50 model. Here Resnet is used as backbone for UNet architecture [9]. This study has used transfer learning with pre-trained models for better performance result in terms of accuracy. The architecture is used on a multispectral high resolution satellite image of a segment of the town of Gandhinagar, Gujarat. First the training is done by splitting the image to create a minimal dataset of about 75 images. For better accuracy the same image is then split to create a training dataset of about 1200 images. More samples give the learning algorithm more opportunity to understand the underlying pattern of inputs to map to the outputs which results in a better performing model.

- **Training and results from the minimal dataset:** The first dataset consists of approximately 75 images.

CASE 1: The standard library allows us to first train the model for half the images which gives an accuracy of 79.29% and an IoU score of 0.6152 as shown in figure 2. Figure 3 shows a representation of the ground truth vs. the predicted results.

CASE 2: A second run of the training cycle on all the images improves the accuracy further, optimization of other hyper parameters such as unfreezing of layers and change of learning rate helps further to give a final accuracy of 79.71% and an IoU score of 0.6656 as shown in figure 4. Figure 5 represents a ground truth vs. predicted output image for training on complete dataset.

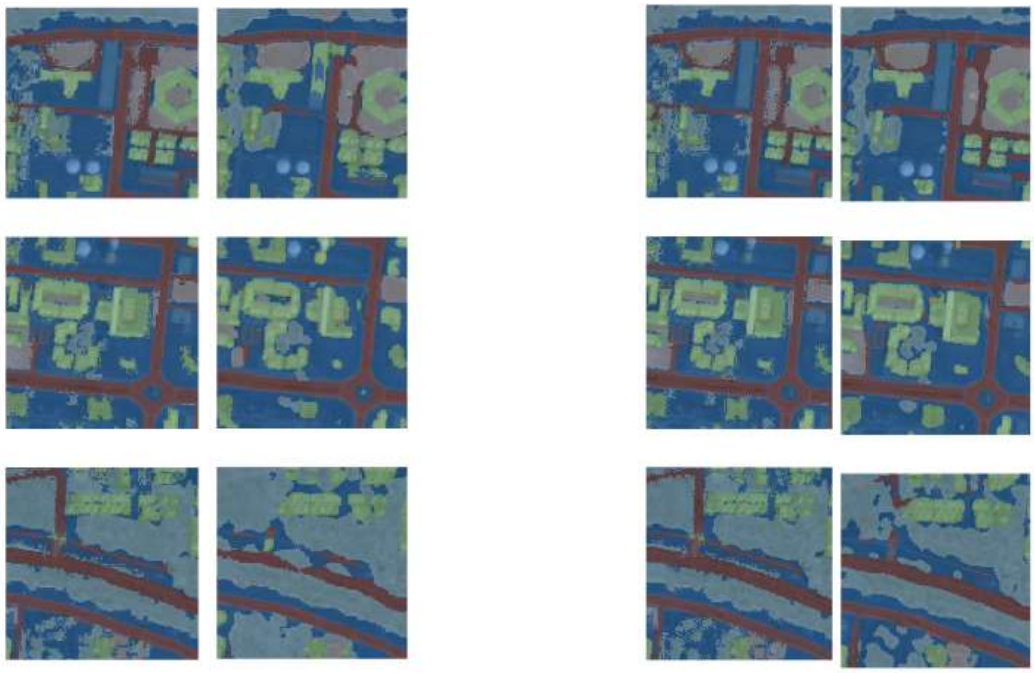
Table 1: First Training Cycle with Minimal Dataset Case 1 (a) And Case 2 (b)

epoch	train_loss	valid_loss	acc_sat	IOU	time
0	0.677835	0.656591	0.754823	0.603066	00:16
1	0.670515	0.642867	0.807028	0.614304	00:16
2	0.655630	0.659024	0.761904	0.595054	00:16
3	0.653816	0.643189	0.772802	0.608458	00:16
4	0.646308	0.642237	0.768515	0.614713	00:16
5	0.638937	0.684711	0.686997	0.585494	00:16
6	0.644110	0.641542	0.821286	0.630504	00:16
7	0.653021	0.691157	0.708926	0.587844	00:16
8	0.654757	0.645202	0.798178	0.621249	00:16
9	0.648310	0.645891	0.792896	0.615268	00:16

a

epoch	train_loss	valid_loss	acc_sat	IOU	time
0	0.401222	0.560886	0.796177	0.659841	01:09
1	0.398694	0.577018	0.781519	0.655343	01:09
2	0.398104	0.569295	0.795956	0.663261	01:09
3	0.399662	0.587315	0.769493	0.653688	01:09
4	0.397944	0.556586	0.801472	0.669183	01:09
5	0.399178	0.585962	0.772071	0.653141	01:09
6	0.398126	0.574289	0.784494	0.657391	01:09
7	0.394850	0.559629	0.800831	0.666192	01:09
8	0.394355	0.559118	0.798384	0.665706	01:09
9	0.393793	0.558819	0.797132	0.665646	01:09

b



a **b**

Figure 2: Ground Truth Vs. Predcitions for Minimal Dataset For Case 1(a) and Case 2 (b)

• **Training results from larger dataset:** To attain a better accuracy and improve the results, we use default functions to split the given satellite image and generate a dataset of approximately 1200 images. As in case with the minimal dataset, here too we train the model in two cases once with half the images and the second time with the complete dataset by optimizing the hyper parameters such as learning rate etc.

CASE 1: Upon training on half the images, we get an accuracy of 83.90% and an IoU score of 0.7022 as shown in figure 6. Figure 7 shows a representation of the ground truth vs. predictions.

CASE 2: On training on the complete dataset and fine tuning the hyper parameter such as changing the learning rate we get a final accuracy of 82.78% and an IoU score of 0.663 as shown in figure 8. Figure 9 shows ground truth vs. predictions for the given case.

Table 2: First Training Cycle with Larger Dataset Case 1 (a) And Case 2 (b)

epoch	train_loss	valid_loss	acc_sat	IOU	time
0	0.485406	0.468214	0.827463	0.686580	01:36
1	0.473072	0.461263	0.829764	0.689467	01:34
2	0.478614	0.455724	0.831078	0.691722	01:34
3	0.475317	0.455709	0.826550	0.691970	01:34
4	0.471195	0.451555	0.833527	0.694163	01:34
5	0.467249	0.449281	0.832162	0.695315	01:34
6	0.464496	0.444601	0.836251	0.697624	01:35
7	0.460936	0.441907	0.834067	0.698990	01:35
8	0.456690	0.437084	0.836517	0.702195	01:34
9	0.451663	0.438770	0.839029	0.702231	01:34

epoch	train_loss	valid_loss	acc_sat	IOU	time
0	0.465492	0.477394	0.816448	0.634170	06:08
1	0.483182	0.470823	0.826105	0.639880	06:06
2	0.456135	0.454189	0.829035	0.649324	06:06
3	0.463918	0.451476	0.831112	0.655860	06:07
4	0.467879	0.440491	0.838952	0.663451	06:08
5	0.456294	0.440985	0.832035	0.660311	06:06
6	0.442290	0.434958	0.837263	0.663306	06:06
7	0.435265	0.437070	0.833454	0.662915	06:06
8	0.440859	0.438084	0.831332	0.661320	06:05
9	0.451794	0.437076	0.827821	0.663192	06:06

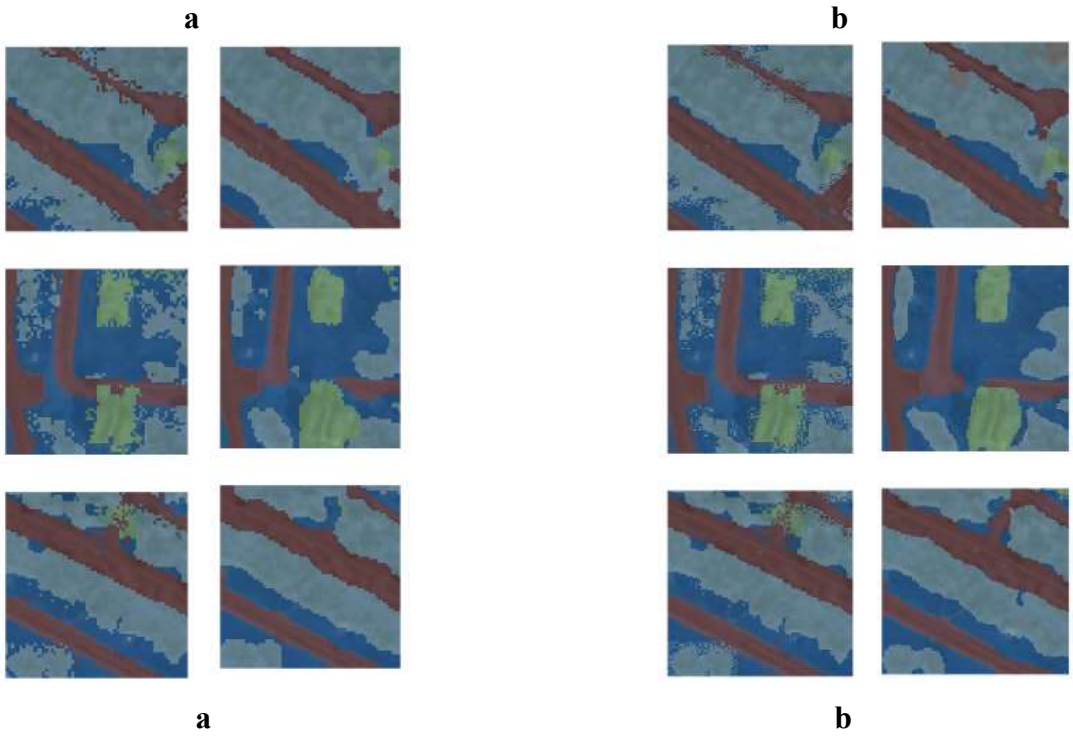


Figure 2: Ground Truth Vs. Predcitions for Large Dataset For Case 1(a) and Case 2 (b)

4. **Conclusion:** As opposed to the minimal dataset, it can be clearly concluded that better results are obtained when the dataset is larger and the satellite image is split up. Not only does it increase the training accuracy greatly, but it also affects the ground truth vs. predictions and the IoU score. Resnet 50 works significantly better than its other lower variants such the 34-layer implementation Resnet 34 and Resnet 18. A further investigation into the study also helps us conclude that, the higher layer implementation, namely Resnet 101 will work better. However, with the current dataset and the image Resnet 101 overfits to the data which in fact decreases the accuracy. This can noticeably be improved using more and better computational resources, increasing the size of the training dataset etc. Thus, for our given system conditions and dataset Resnet 50 works as a possible best variant for the extraction of urban features from the given multispectral satellite image.

References

[1] Building Extraction at Scale using Convolutional Neural Network: Mapping of the United States by Hsiuhan Lexie Yang, Member, IEEE, Jiangye Yuan, Member, IEEE, Dalton Lunga, Senior Member, IEEE, Melanie Laverdiere, Amy Rose, Budhendra Bhaduri

[2] Deep learning in remote sensing applications: A meta-analysis and review by Lei Ma, Yu Liu, Xueliang Zhang, Yuanxin Ye, Gaofei Yin, Brian Alan Johnson

[3] Recent progress in semantic image segmentation by Xiaolong Liu, Zhidong Deng, Yuhan Yang

[4] O'Shea, Keiron & Nash, Ryan. (2015). An Introduction to Convolutional Neural Networks. ArXiv e-prints.

[5] Implementation of Training Convolutional Neural Networks by Tianyi Liu, Shuangsang Fang, Yuehui Zhao, Peng Wang, Jun Zhang

[6] Learning Building Extraction in Aerial Scenes with Convolutional Networks by Jiangye Yuan

[7] Hyperparameters in machine/Deep Learning by Jorge Leonel

[8] Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation by Md Atiqur Rahman and Yang Wang

[9] U-Nets with ResNet Encoders and cross connections by Christopher Thomas BSc Hons. MIAP

[10] Welcome to fastai: https://docs.fast.ai/