USING 3D RESIDUAL NETWORK FOR SPATIO-TEMPORAL ANALYSIS OF REMOTE SENSING DATA

Muhammad Ahmed Bhimra, Usman Nazir and Murtaza Taj

Department of Computer Science, Syed Babar Ali School of Science and Engineering
Lahore University of Management Sciences (LUMS), Lahore, Pakistan
{17030015, 17030059, murtaza.taj}@lums.edu.pk

ABSTRACT

In this paper, we propose an approach to recognize spatio-temporal changes from remote sensing data. Instead of performing independent analysis on each instance of satellite imagery, we proposed a 3D Convolutional Neural Network (CNN) based on the ResNet architecture. Our approach takes as input a 3D spatio-temporal block comprising of spatial as well as temporal data from multiple years. We predict four key transition classes namely construction, destruction, cultivation and decultivation. In our proposed architecture, we introduced Leaky ReLU instead of ReLU which improves the overall performance as it solves the dying ReLU problem. We also provided dataset and annotations[1] for these four classes and have evaluated the efficacy of our approach on data from three different cities.

Index Terms— Spatial-Temporal, Convolutional Neural Network (CNN), ResNet, Satellite Imagery, Remote Sensing Data.

1. INTRODUCTION

The increase in human population has resulted in multiple changes in our ecosystem. Major changes could be seen in construction and farming which are two of the oldest professions, since the dawn of civilization. The analysis of construction and cultivation using spatial data of past and present, is regarded as one of the basic requirement for future planning and geographical studies [1]. Satellite imagery is one such spatio-temporal data that provide an opportunity to estimate the changes in land use over time. Along with construction and cultivation, it also allows estimation of destruction caused by war, genocide, deforestation as well as natural calamities.

Availability of high resolution satellite imagery along with recent advancements in machine learning particularly deep convolutional neural networks (DCNN) have paved a way for large scale analysis of wide variety of parameters across the globe. Jean et al. [2] proposed a transfer learning method for poverty estimation using a combination of day-time and night-time imagery. They only used one instance each for day and night time for a single year and were thus unable to exploit any temporal relationship between imagery. Crop yields before harvest was predicted by You et al. [3]. Instead of using only two temporal instances (day and night), they used 30 instances per year (at 8-days interval) each from 2003 to 2015. To cater for this dense temporal data, they proposed a Long Short Term Memory (LSTM) based solution. Similarly, Rubwurm et. al. [4] have also used LSTMs for temporal vegetation classification from satellite images using data collected at monthly interval. Although LSTMs are well suited for dense temporal data, they are inherently complex and require much higher amount of annotations, which is scarce in case of remote sensing.

Unlike crop or vegetation, certain transitions such as construction and overall change in land use from barren to cultivated requires using only few temporal instances over multiple years. Such data provides an overall change in socio-economic conditions over a longer period. Such changes include impact of mass transit on urbanization, industrial growth due to access via road and increase in agriculture due to construction of canals. However, such sparse temporal data also poses several additional challenges such as unwanted variations due to atmosphere, weather, and change in imaging sensor over time. Such temporal analysis requires a hybrid approach that is based neither on single temporal instance nor on dense sampling across the year.

We showed that in order to filter these unwanted variations and extract relevant change in land use, 3D convolutions are needed. Thus unlike existing literature, instead of processing imagery for each year separately or using complex LSTM's, we proposed a 34 layered 3D ResNet [5] based solution for spatio-temporal analysis of four key transitions in satellite imagery namely construction, destruction, cultivation and harvesting (decultivation). Furthermore, we also introduced a new dataset and annotations for these transition classes. We evaluated our approach in three different cities including Aleppo, Kathmandu and Lahore and showed that our proposed model outperformed all the existing 2D convolution based techniques.

[1] https://cvlab.lums.edu.pk/?p=1742

978-1-5386-4658-8/18/$31.00 ©2019 IEEE

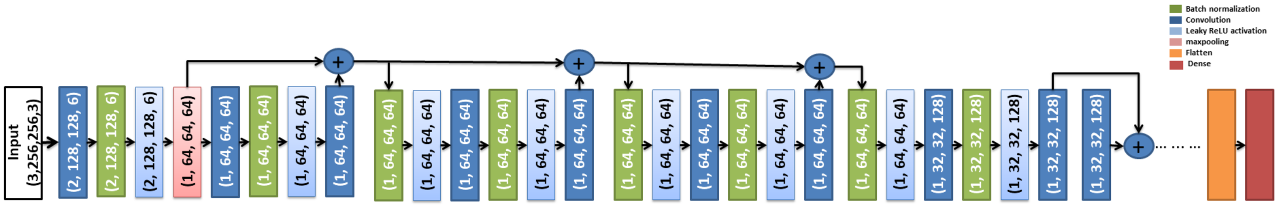


Fig. 1. Proposed 34 layered 3D ResNet architecture showing introduced Leaky ReLU activations.

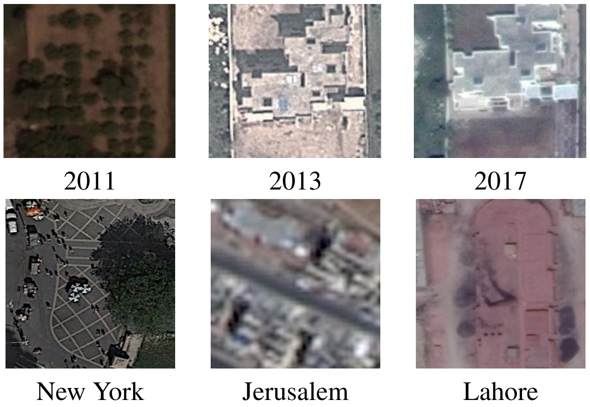


Fig. 2. (Row 1) Example satellite imagery from same spatial location showing variation in quality and color profile along with those due to urbanization. (Row 2) Example satellite imagery showing variation in quality over different spatial locations.

2. PROPOSED MODEL

2.1. Challenges

When different temporal instance of the same location is observed on satellite imagery, several type of variations can be observed. These include variations due to atmosphere, sensory variations and man-made spatio-temporal variations. Atmospheric variations include cloud cover, pollution, effect due to time of the day for e.g. shadows and varying lighting condition.

Sensor variations are usually caused by the fact that different imaging sensors may capture the same imagery. E.g. from 1997 till today Digital Globe which provide data for Google Earth has launched 8 different satellites. In 1997, their EarlyBird-1 satellite included a panchromatic (black-and-white) camera with 3 meters per pixel resolution whereas in 1999 in IKONOS they upgraded their camera to 0.8 meter resolution and also included a multispectral (color) camera. Similarly, their current WorldView-4 satellite provide panchromatic images at a highest resolution of 0.31 meters per pixel, and multispectral images at 1.24 meters per pixel. This suggest that structures (such as motorbikes and narrow roads) which are smaller than 3 meters were not visible in 1997 but can be easily seen in the current imagery. When spatio-temporal analysis is spread across multiple years, these changes in sensors can be observed as huge variations in the quality, resolution and color profile of the imagery (see Fig. 2, row 1).

The third type of variations in the imagery is due to human interventions which can be broadly classified into four fundamental transitions namely i) construction, ii) destruction, iii) cultivation and iv) decultivation. As the world population is increasing we can see rapid increase in urbanization, similarly more and more land is being utilized for agriculture. Furthermore, war, genocide, deforestation can also be seen through satellite imagery in the form of destruction. Some change due to weather conditions like flood, cyclone, earthquake can also be seen as destruction, however in such studies their adverse effect on man-made structures is of more interest.

Thus, a learning model is required to first project the data in a normalized space where noise due to these unwanted variations such as due to sensor and atmosphere are filtered out and only those due to human interventions remain. In the context of deep learning, this requires a much deeper network to learn the four fundamental transitions from the noisy set. Furthermore, the learning should be performed simultaneously on the spatial and temporal dimensions i.e. convolution should be applied along space and time simultaneously (3D convolution).

2.2. Residual Connections

In remote sensing images, many times a feature is more visible in an imagery of a particular year than in the imagery of subsequent years and vice versa (see Fig. 2, row 1). Thus, sometimes features from earlier layers of the network may contribute more information towards the final decision as compared to the later layers. Such flexibility can be incorporated in CNNs by introducing residual layers (see Fig. 3) which can be expressed as:

$$\mathbf{r} = F(\mathbf{p}, W_i) + \mathbf{p}, \tag{1}$$

where $\mathbf{p}$ is input to the residual connection, $F(\mathbf{p}, W_i)$ is residual mapping to be learned. The dimensions of $\mathbf{p}$ and F must be equal. If this is not the case, we used residual connection: $\mathbf{r} = F(\mathbf{p}, W_i) + W_j\mathbf{p}$ where W_j is a matrix formed using convolution to match the size.

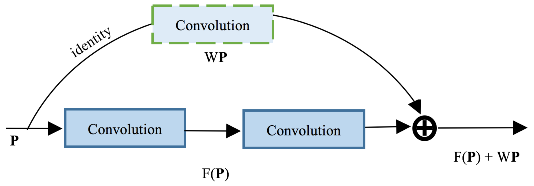


Fig. 3. Residual connection used in all variants as short skip connection. Convolution layer (dotted) is optional to change the dimension of features.

2.3. 3D-ResNet-34-LeakyReLU

Putting together above considerations our proposed solution is thus an extension of 3D-ResNet architecture [6]. In our proposed model we used 34 layered network with regular residual connections (see Fig. 1). We also used Leaky ReLU instead of ReLU that attempts to fix the dying ReLU problem. Instead of the function being zero when $x < 0$, a leaky ReLU will instead have a small negative slope (of 0.01, or so). That is, the function computes:

$$f(x) = 1(x < 0)(\alpha x) + 1(x >= 0)(x), \tag{2}$$

where α is a small constant. We observed that Leaky ReLU is more feasible for datasets with such a high variations in pixel values at the same spatial location. The proposed 3DResNet architecture is shown in Fig. 1.

Our proposed 3D ResNet exploits temporal variations to learn actual transition as shown in Fig. 5. Some filters learn the common structures in transition e.g. roads, while other filters learn the temporal structures e.g. buildings. If we closely look at Fig. 5, corner filters learn the longitudinal changes while the filters in center learn the yearly changes.

3. RESULTS AND EVALUATION

3.1. Dataset Annotation

We prepared a one of its kind spatio-temporal dataset for four fundamental transitions in satellite imagery. We visited over $5,50,000$ random locations each for year 2011, 2013 and 2017 (approximately $5310km^2$) on imagery from Digital Globe. These annotations were done in a semi-automated manner using a 152 layered 2D ResNet (2D-ResNet-152) [7]. We first trained 2D-ResNet-152 on 14 spatial classes including houses, roads, farms, ground etc. Training was performed by collecting 1000 samples for each class using annotations from OpenStreetMap (OSM). Using this land use model we selected most likely locations for a class, these spatial annotations for multiple years are then manually analyzed and to generate the desired 3D spatio-temporal annotations for the four key transition classes.

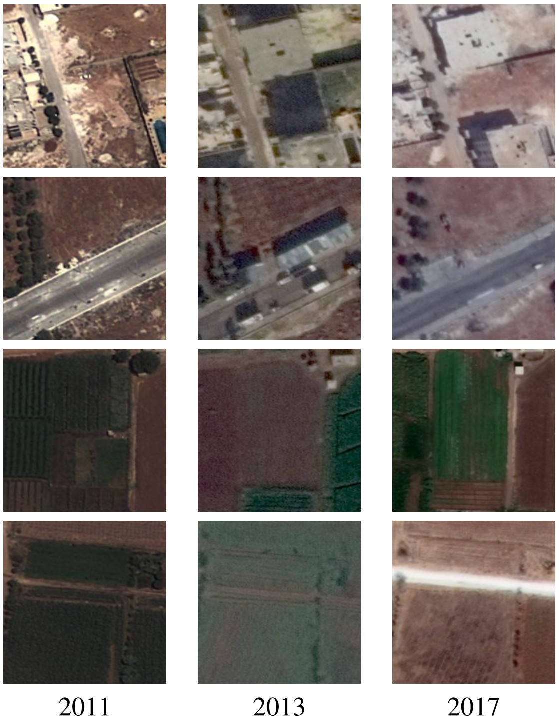


Fig. 4. Sample Annotations for four key transition classes. (Row 1) Construction. (Row 2) Destruction. (Row 3) Cultivation and (Row 4) Decultivation.

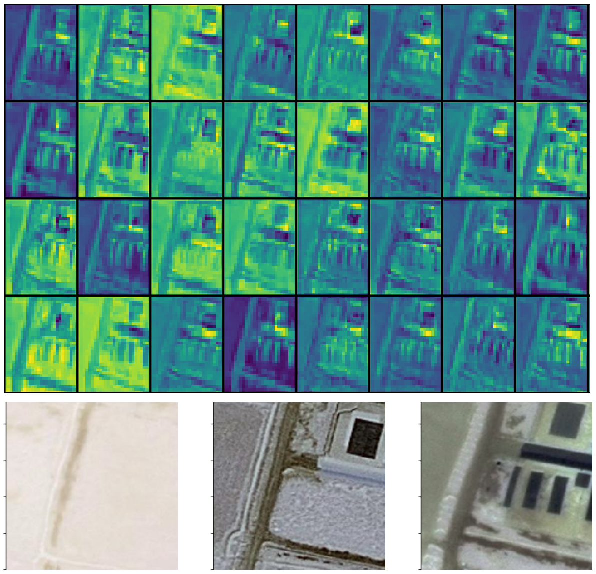
Using this method we recorded 1813 locations from the city of Lahore consisting of 488, 488, 123, 154 and 560 instances of construction, destruction, cultivation, decultivation and no transition of interest respectively. Along with lat-long, at each location we cropped an image patch of resolution 256×256 at zoom level 20 (i.e. 0.149 pixel per meter on equator). We used this entire dataset for training. Figure 4 shows sample annotations for all transition classes.

3.2. Comparison

We evaluated our approach on 836 additional annotations from 3 cities namely Aleppo, Kathmandu and Lahore. We compared our proposed (3D-ResNet-34-LeakyReLU) with two state of the art methods namely Inception-ResNet-v2 [8] and 2D-ResNet-50 [7]. All three network architectures were trained using the same data. In case of 2D models we trained the network to predict 4 classes namely houses, farms, barren lands and others for each year separately which are then converted into transitions using simple voting strategy. Inception-ResNet-v2 is trained on 23 epochs, 2D-ResNet is trained on 30 epochs and 3D-ResNet-34 is trained on 36 epochs and proposed is trained on 35 epochs. Table 1 shows the comparison between proposed with the state-of-the-art network architectures. It can be seen that the proposed approach outperformed both Inception-ResNet-v2 and 2D-ResNet-50 for transition

Table 1. Table showing quantitative evaluation of the proposed 3D-ResNet-34 architecture and its comparison with state-of-the-art approaches namely Inception-ResNet-v2 [8] and 2D-ResNet-50 [7].

		(Spatio-Temporal Classes) **Accuracy**					
Cities	**Network Architectures**	Construction	Destruction	Cultivation	Decultivation	No trans. of interest	**Overall**
Aleppo	Inception-ResNet-v2	16.6%	20%	100%	33%	80%	65.95%
	2D-ResNet-50	25%	25%	100%	100%	69%	58.87%
	3D-ResNet-34	**71%**	**75%**	**100%**	**100%**	**84%**	**80.14%**
Kathmandu	Inception-ResNet-v2	24.5%	25%	100%	100%	**77.7 %**	57.70%
	2D-ResNet-50	31.9%	33%	100%	100%	70.7%	56.45%
	3D-ResNet-34	**48.9%**	**50%**	**100%**	**100%**	65%	**57.72%**
Lahore	Inception-ResNet-v2	44.8%	45%	**85.7%**	**78.3%**	57.8%	58%
	2D-ResNet-50	53%	55%	71.4%	74%	51.23%	55%
	3D-ResNet-34	**57%**	**58%**	57%	48%	**88.4%**	**75%**

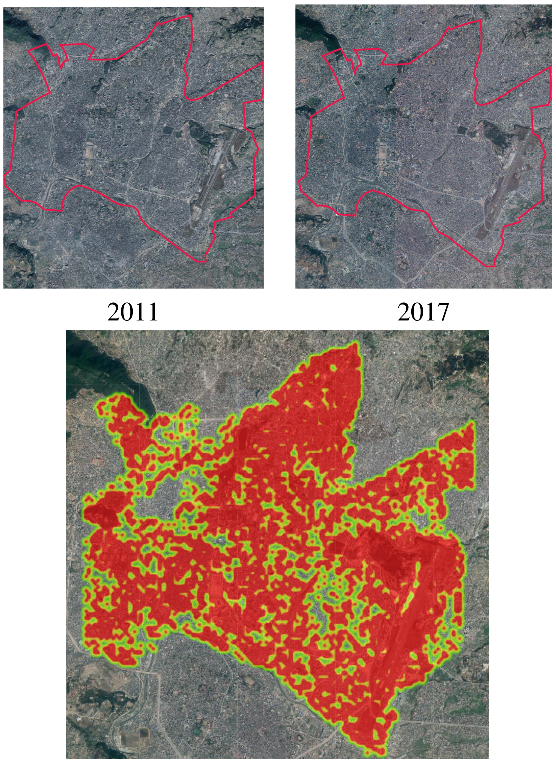


2011 2013 2017

Fig. 5. 3D ResNet representative features shown with its actual image. Out of 128 features, we have visualized only 24 features each of size 8×8. It can be seen that these features maintain the structural information while encoding the image.

classification. We also observed that other networks failed to classify changes due to increase in house density as construction but our proposed network was more successful in identifying them. We were able to identify classes that were not annotated on OSM like construction, destruction, cultivation, decultivation and no transition of interest with high accuracy. We obtained 71% construction testing accuracy for Aleppo, 57% for Lahore and 49% for Kathmandu on our proposed 3D ResNet as shown in Table 1.

We also performed a qualitative evaluation on $1,884,000$ 3D spatial-temporal blocks from above mentioned 3 cities. Evaluation result for Kathmandu city is shown in Fig. 6.



2011 2017

Fig. 6. Heat map showing *construction* transition class for Kathmandu city from year 2011 to 2017.

4. CONCLUSION

In this paper we demonstrated that 3D convolutions can learn desired variations in spatio-temporal data. Our proposed 3D ResNet architecture used much less number of layers as compared to its 2D counterparts, still it outperformed the existing state-of-the-art methods. Furthermore, we provided data for four key transitions in remote sensing namely construction, destruction, cultivation and decultivation which will serve as a valuable resource for further such analysis.

5. REFERENCES

[1] M. Dadras, H. Shafri, N. Ahmad, B. Pradhan, and S. Safarpour, “Spatio-temporal analysis of urban growth from remote sensing data in bandar abbas city, Iran,” *The Egyptian Journal of Remote Sensing and Space Science*, vol. 18, no. 1, pp. 35–52, 2015.

[2] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, “Combining satellite imagery and machine learning to predict poverty,” *Science*, vol. 353, no. 6301, pp. 790–794, 2016.

[3] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, “Deep gaussian process for crop yield prediction based on remote sensing data,” in *Association for the Advancement of Artificial Intelligence*, 2017, pp. 4559–4566.

[4] Marc RuBwurm and Marco Körner, “Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multispectral satellite images.,” in *CVPR Workshops*, 2017, pp. 1496–1504.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatio-temporal features with 3D convolutional networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[6] K. Hara, H. Kataoka, and Y. Satoh, “Learning spatio-temporal features with 3d residual networks for action recognition,” in *Proceedings of the ICCV Workshop on Action, Gesture, and Emotion Recognition*, 2017, vol. 2, p. 4.

[7] K. He, X. Zhang, Sh. 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.

[8] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the impact of residual connections on learning,” in *Association for the Advancement of Artificial Intelligence*, 2017, vol. 4, p. 12.