Going Deep into Remote Sensing Spatial Feature Learning

Keiller Nogueira[1], William Robson Schwartz[1], Jefersson Alex dos Santos[1]
[1]Department of Computer Science, Universidade Federal de Minas Gerais
31270-901, Belo Horizonte, MG - Brazil
Email: {keiller.nogueira, william, jefersson}@dcc.ufmg.br

***Abstract*—A lot of information may be extracted from the Earth's surface through aerial images. This information may assist in myriad applications, such as urban planning, crop and forest management, disaster relief, etc. However, the process of distilling this information is strongly based on efficiently encoding the spatial features, a challenging task. Facing this, Deep Learning is able to learn specific data-driven features. This PhD thesis[1] introduces deep learning into the remote sensing domain. Specifically, we tackled two main tasks, scene and pixel classification, using Deep Learning to encode spatial features over high-resolution remote sensing images. First, we proposed an architecture and analyze different strategies to exploit Convolutional Networks for image classification. Second, we introduced a network and proposed a new strategy to better exploit multi-context information in order to improve pixelwise classification. Finally, we proposed a new network based on morphological operations towards better learning of some relevant visual features.**

***Keywords*-deep learning; machine learning; remote sensing; image classification; image segmentation;**

I. Introduction

Earth's Planet is constantly being modified due to natural and human interference, including hurricanes, earthquakes, new residential and agricultural areas, landfills, etc. It is costly and almost impractical to understand all these changes and developments via on-the-ground observations. Thus, a lot of effort has been employed for obtaining images from the Earth's surface, i.e., aerial ones. Although a laborious task, it can be justified first by the amount of information that may be extracted from these images and second by the potential usage of this data in several tasks (such as classification and segmentation) assisting in the understanding of a myriad of events. Based on this argument, new technologies have been proposed toward acquiring aerial images with improved quality, resulting in more advanced satellites launched to observe the Earth, as well as, more recently, in drones and unmanned aerial vehicles. These top-notch Remote Sensing Images (RSIs) may provide useful information that could be employed in several Earth Observation applications, including urban planning [1], crop and forest management [2], [3], disaster relief [4], [5], phenological studies [6]–[8], etc.

In general, all the information distilled by these applications are highly dependent on the creation of high quality thematic maps (to establish precise inventories about land cover use [9]) as well as on detection and monitoring of events. However, the development of both tasks, by manual efforts (e.g., using edition tools), is slow and costly, being unfeasible, given the large amount of data. Therefore, automatic methods appear as an appealing alternative for the community. Traditionally, such automatic methods would perform these tasks by using machine-learning based approaches over features encoded by some visual description technique. Therefore, efficiently encoding of these features is one of the most important steps in almost any image-related problem since it is the main key to generate discriminative models. Given this, through years, substantial efforts have been dedicated to develop automatic and discriminative feature techniques [10], commonly called (hand-crafted) descriptors. Some of them [11], [12] were originally proposed and successfully employed in the computer vision scenario and then, experimented into the remote sensing domain, while others [13] were specifically designed for Earth Observation applications.

In the first case, successful descriptors proposed to handle everyday pictures may not have the same favorable outcome for RSIs given the distinct characteristics between these images, which include: (i) perspective, as traditional images typically have a clear definition of fore and background, while, in RSIs, all the pixels should be managed with the same attention, (ii) context, as everyday pictures have a specific notion of context related to the scene while aerial images do not have this concept but have the geographic context, (iii) elementary properties, given that traditional images usually have more complex and rich scenes than aerial ones, mainly when considering low-resolution images, and (iv) channels, as traditional images usually encode only visible information while RSIs may have hundreds or even thousands of bands. Thus, as introduced, based on these specificities and differences, many of these techniques, originally proposed and successfully applied for computer vision applications [14], have not the same success in the remote sensing domain [15]. In the second case, though successfully proposed and employed into the remote sensing domain, each descriptor technique is highly dependent of the intrinsic properties of the image, such as gradient, edges, colors, etc. For instance, a novel descriptor proposed specifically for land-use scenes may not be a good choice for agricultural images. Thus, the development of algorithms for spatial extraction information is still a hot research topic in the remote sensing community [16]. Besides all this, in

[1]This work relates to a Ph.D. thesis.

a typical scenario, different descriptors may produce distinct results depending on the data. Therefore, it is imperative to design and evaluate many descriptor algorithms in order to find the most suitable ones for each application [17]. This process is also expensive and, likewise, does not guarantee an efficient and discriminative representation.

Overcoming aforementioned limitations, deep learning [18], a branch of machine learning that refers to multi-layered interconnected neural networks, can learn features and classifiers at once, i.e., a unique network may be able to learn features and classifiers (in different layers) and adjust the parameters, at running time, based on accuracy, giving more importance to one layer than another depending on the problem. End-to-end feature learning (e.g., from image pixels to semantic labels) is the great advantage of deep learning when compared to previously state-of-the-art methods [19], such as mid-level (Bag of Visual Words (BoVW) [20]) and global low-level color and texture descriptors. Among all deep learning-based networks, a specific type, called Convolutional (Neural) Networks [18], ConvNets or CNNs, is the most popular for learning features in computer vision applications. This sort of network relies on the natural stationary property of an image, i.e., the statistics of one part of the image are the same as any other part and information extracted at one part of the image can also be employed to other parts. Furthermore, ConvNets usually obtain different levels of abstraction for the data, ranging from local low-level information in the initial layers (e.g., corners and edges), to more semantic descriptors, mid-level information in intermediate layers and high level information (e.g., whole objects) in the final layers.

The work developed in this PhD thesis [21] was one of the first to introduce deep learning into the remote sensing domain. Given its pioneering spirit, several works published related to this PhD thesis are widely cited and can be considered one of the main references regarding deep learning and the remote sensing domain. Precisely, the main contributions are:

1) **A novel ConvNet architecture for remote sensing image classification [22].** This network has fewer layers and parameters, being able to converge using a small quantity of data, demonstrating the effectiveness of deep learning methods to encode features even for the remote sensing domain. This is one of the first works to exploit deep learning for remote sensing image classification. Due to its pioneering spirit, it is the seventh most cited article among all those published at the Sibgrapi conference, since 2015, according to Google Scholar[2].
2) **An extensive assessment to define the best training strategy for exploiting ConvNets for RSI classification [23], [31]**. Three distinct training strategies were tested for RSI classification: (i) fully-train, (ii) fine-tuning, and (iii) pre-trained network as feature extractors. Moreover, this evaluation was carried out using six popular ConvNets and three remote sensing datasets. These pioneer works have been widely cited and can be considered some of the main references regarding deep learning for remote sensing applications. Specifically, the work published in [31] has more than 460 citations while the work published in [23] has more than 440 citations (according to Google Scholar[3]). The former work is the sixth most cited article among all those published at the Computer Vision and Pattern Recognition Workshops, since 2015, according to Google Scholar[4] while the latter is the second most cited article among all those published at the renowned Pattern Recognition journal according to the Elsevier website[5].
3) **A new network architecture for pixel classification of RSI [24].** The proposed ConvNet, one of the first for pixel classification of RSI, was evaluated for two distinct datasets, achieving state-of-the-art in both cases.
4) **A new strategy to better exploit multi-context information to perform pixel classification of RSI using ConvNets [25].** The proposed technique is capable of aggregating multi-context information without increasing the number of parameters (and the complexity of the network) while defining, in training time, the best patch size to be used for the inference phase.
5) **A new paradigm for deep networks that exploits non-linear morphological filters to capture the patterns [26]**. This paradigm learns the structure elements of the morphological operation in order to extract the features. With such concept, new morphological layers and networks were created and optimized for RSI classification.

Besides those contributions, many others related to this PhD work were published in [3], [5]–[8], [27]–[30], [32]. The following sections detail each of the contributions obtained from the developed research.

II. ConvNet-Based Scene Classification

As introduced, hand-crafted descriptors are created for a specific domain and may fail when applied in another one. For instance, descriptors created to handle everyday pictures in the computer vision domain may fail to encode features of aerial images as well as descriptor techniques conceived to deal with urban aerial scenes may encounter problems in handling agricultural images. Therefore, a data-driven feature learning step, as in ConvNets, is essential to extract all feasible information from the data and create discriminative models. However, ConvNets are hard to train, because of its high number of parameters, requiring a really large amount of annotated data. Going in the other direction, remote sensing domain has huge amount of data but with only few annotations. Hence, it is essential to evaluate if it is feasible to exploit deep learning for aerial images as well as to define the best strategy to do so.

[2]https://scholar.google.com.br/citations?hl=pt-BR&view_op=list_hcore&venue=sPLlun2OWTwJ.2020

[3]https://scholar.google.com.br/citations?user=cuLQ6NgAAAAJ&hl=pt-BR

[4]https://scholar.google.com.br/citations?hl=pt-BR&view_op=list_hcore&venue=ihIDe6biV1gJ.2020

[5]https://www.journals.elsevier.com/pattern-recognition/most-cited-articles

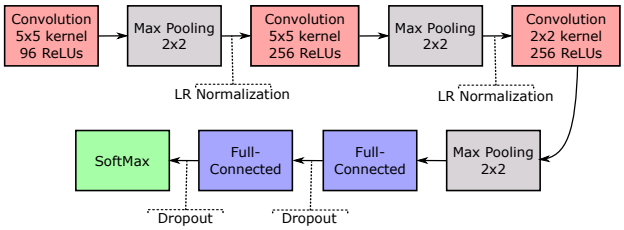


Fig. 1: Architecture of the proposed PatreoNet.

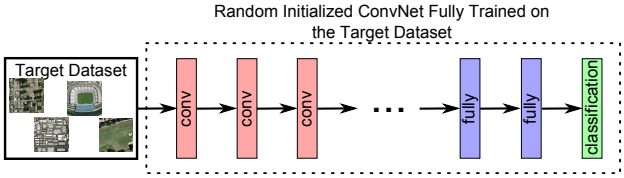


Fig. 2: Illustrative example of a ConvNet being fully trained. Weights from the whole network are randomly initialized and then trained for the target dataset.

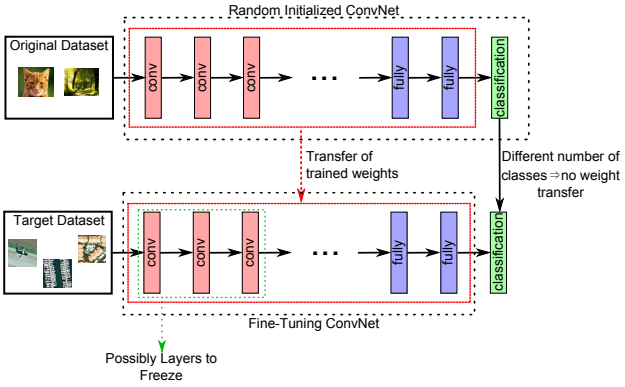


Fig. 3: Illustrative example of two options for the fine-tuning process. In one of them (highlighted in red), all layers are fine-tuned according to the target dataset, but final layers have increased learning rates. In the other option (highlighted in green), weights of initial layers can be frozen and only final layers are tuned.

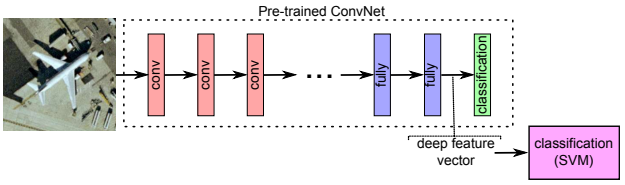


Fig. 4: Illustrative example of the use of a ConvNet as feature extractor. The final classification layer is ignored and one should only choose from which layer to consider the features. The figure shows the use of the features from the last layer before the classification layer.

The contribution published in [22] introduces a novel network architecture fully trained specifically for the remote sensing domain. This novel architecture, presented in Figure 1, has fewer layers and parameters and is able to converge even using a smaller amount of labeled data. Such network outperformed all traditional baselines, an outcome that demonstrated the ability of networks to learn patterns for RSIs which, opened new opportunities towards a better feature representation.

After verifying the effectiveness of deep learning for RSIs, it was fundamental to analyze and define the best strategies to exploit such technique in this domain. The contributions published in [23], [31] carried out a systematic set of experiments to evaluate three different strategies for the remote sensing domain: (i) fully-train, illustrated in Figure 2, which is the strategy to train a network from scratch (with random initialization of the filter weights), (ii) fine-tuning, presented in Figure 3, that consists in performing fine adjustment in the parameters of a pre-trained network by resuming the training of the model from a current set of parameters but considering a new dataset, and (iii) feature extractor, presented in Figure 4, which consists in using a pre-trained network as a feature extractor and then train a standard machine learning with such features. Those strategies were tested considering six popular ConvNets, including the proposed PatreoNet, and three remote sensing datasets. The results point that fine tuning tends to be the best strategy in different situations. Specially, using the features of the fine-tuned network with an external classifier, linear SVM in our case, provides the best results. It is worth mentioning that there these works have been widely cited and can be considered some of the main references regarding deep learning for remote sensing applications.

III. ConvNet-Based Pixel Classification

Although a lot of attention has been given to scene classification, one of the most important application in the remote sensing community is the creation of thematic maps, which may help in understanding events over a specific region, such as new urban areas and deforestation. Essentially, this application is modeled in a supervised manner which should have, as outcome, a class for each and every pixel of the input image. Based on its outcome, this task is commonly called pixel classification (also known, in the computer vision field, as semantic segmentation [1]). Although important, pixel classification is a hard task given that its basic element (the pixel) has not enough information to allow its classification. Therefore, it is essential to research deep learning-based methods that could efficiently exploit the pixel context in order to perform the final classification.

The contribution published in [24] presents a novel network that aggregates the context of the pixel by using overlapping patches, centered on each pixel, that carry the context of the pixel and help understand the spatial patterns around them. This strategy, presented in Figure 5, allows the network to efficiently understand the context around the pixel and correctly classify it.

In the previous work, the input patches delimited the visual context that the network could exploit to learn the spatial features. However, using only one context size, as in the previous and other works [33], could lead to several problems, as distinct classes may require different context sizes. To

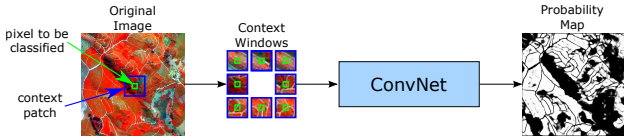


Fig. 5: Each pixel is actually represented by a large context patch, centered on the pixel, in order to include the context of its neighborhood. Those patches are then classified by the network. Note that the predicted class for the context patch is actually the label of the centered pixel.

alleviate this problem, several approaches [33] exploited multi-context information by combining networks or layers. This process increases the number of parameters resulting in a more difficult model to train. The contribution published in [25] presents a novel technique to perform semantic segmentation of remote sensing images that aggregates information from contexts of multiple sizes (without increasing the number of parameters) while defining the best context size for the testing phase. This multi-context strategy allows the network to capture distinct information of the context of the objects, allowing a better understanding of the scene.

The proposed technique receives as input the data and a distribution over the desired patch sizes. During the training procedure, a size is randomly select from this distribution and then is used to create a new batch, composed uniquely of patches of that size. This batch is then employed to train the network, that outputs a score for the current batch, which can be any metric (such as a loss or accuracy) that estimates the performance of the network based on the current batch. This generated score is used to update the patch scores, which accumulate, throughout the training procedure, the scores of the patch sizes and are employed in the selection of the best patch size during the inference stage. All the aforementioned steps are repeated during the training process until the number of iterations is reached. As it can be noticed, the multi-context information is aggregated to the model by allowing the network to be trained using batches composed of patches of multiple sizes. An overview of this procedure is presented in Figure 6.

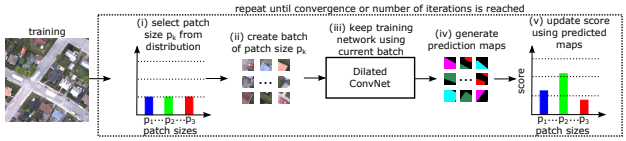


Fig. 6: Overview of the training procedure of the proposed dynamic multi-context strategy.

During the prediction, the accumulated scores over the patch sizes are averaged and analyzed. The best patch size is then selected and used to create patches. The network processes these patches outputting the prediction maps, but no updates in the patch scores are performed. It is important to highlight that the proposed technique can only choose the best patch size within all possible sizes determined by the patch distribution.

This proposed technique was evaluated on four high-resolution remote sensing datasets, achieving state-of-the-art in two and yielding competitive results in the remaining two. Among all evaluated methods independently of the dataset, the proposed one has the least number of parameters and is, therefore, less pruned to overfitting and, consequently, easier to train. At the same time, it produces one of the highest accuracies, which shows the effectiveness of the proposed technique in extracting all feasible information from the data using limited (in terms of parameters) architectures. Aside from this, an interesting aspect of the proposed technique is that the networks trained using such approach can be fine-tuned for any semantic segmentation application, since they do not depend on the patch size to process the data. This allows other applications to benefit from the patterns extracted by our models, a very important process mainly when working with small amounts of labeled data [23].

IV. An Introduction to Deep Morphological Networks

ConvNets are able to efficiently learn distinct patterns, achieving state-of-the-art in several applications. Although this deep learning technique may be composed of several distinct components (such as convolutional and pooling layers, non-linear activation functions, etc), its core operation is the convolution, a linear filtering process whose weights, in this case, are to be learned based on the input data. Easy and fast to implement, convolutions actually play a major role, not only in ConvNets [18], but in digital image processing and analysis [34] as a whole, being effective for many tasks and employed by several techniques [34]. However, aside from convolutions, researchers also proposed and developed non-linear filters, such as operators provided by mathematical morphology. Even though these are not so computationally efficient as the linear filters, in general, they are able to capture different patterns and tackle distinct problems when compared to the convolutions. In fact, supported by this capacity of extracting distinct features, some non-linear filters, such as the morphological operations [35], are still very popular and state-of-the-art in some scenarios [36]. Therefore, it would be interesting to combine morphological filters and deep learning, creating a new framework capable of performing and optimizing these non-linear operations.

This contribution [26] presents a novel paradigm for deep networks where linear convolutions are replaced by the aforementioned non-linear morphological operations. Furthermore, differently from the current literature, wherein distinct morphological filters must be evaluated in order to find the most suitable ones, the proposed technique, called Deep Morphological Network (DeepMorphNet), learns the filters (and consequently the features) based on the input data.

Technically, the proposed basic framework, capable of performing morphological erosion and dilation, uses operations already employed in other existing deep learning-based methods, so it can preserve the end-to-end learning strategy. The processing of this framework can be separated into two steps.

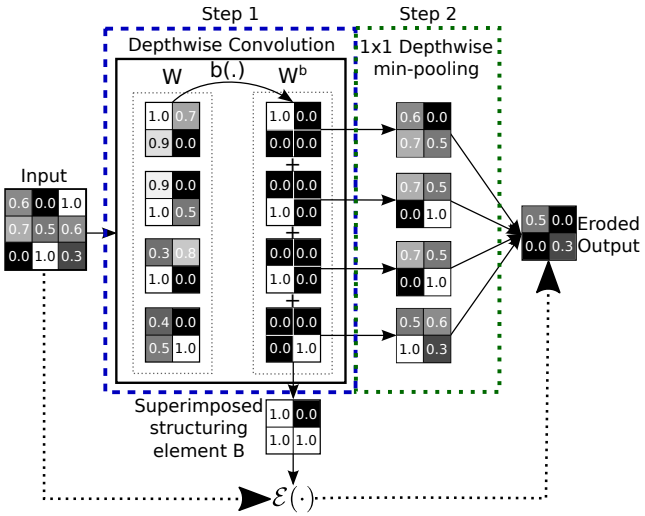


Fig. 7: Example of a morphological erosion based on the proposed framework. The 4 filters W (with size 4×4) actually represent a unique 4×4 structuring element. Each filter W is first converted to binary W^b, and then used to process each input channel (step 1, blue dashed rectangle). The output is then processed via a pixel and depthwise min-pooling to produce the final eroded output (step 2, green dotted rectangle). Note that the binary filters W^b, when superimposed, retrieve the final structuring element B. The dotted line shows that the processing of the input with the superimposed structuring element B using the standard morphological erosion results in the same eroded output image produced by the proposed morphological erosion.

The first one employs depthwise convolution [37] to perform a delimitation of features, based on the neighborhood (or filter). However, just using this type of convolution does not allow the reproduction of morphological transformations, given that a spatial linear combination is still performed by this convolutional operation. To overcome this, we decompose each filter into several ones (one for each weight of the filter), that when superimposed retrieve the final structuring element. Such filters are converted into binary and then used in the depthwise convolution operation. The output of this step is then processed by a depthwise pooling operation that is responsible to retrieve the final outcome, i.e., the eroded or dilated image. This is the **second** step of this proposed framework, which is responsible to extract the relevant information based on the depthwise neighborhood. A visual example of this proposed framework being used for morphological erosion is presented in Figure 7.

This framework is the foundation of five types of morphological processing units (or neurons): (i) composed processing units, presented in Figure 8a, which have a morphological erosion followed by a dilation (or vice-versa), without any constraint on the weights, (ii) opening and closing processing units, presented in Figure 8b, which use morphological erosion followed by a dilation (or vice-versa) with tie weights in order to make them use the same filters, i.e., the same structuring element. (iii) the top-hat processing units, presented in Figure 8c, which use an opening or closing morphological processing unit, a skip connection (that allows the forwarding of the input data), and a subtraction function that operates over the processed and forwarded data, generating the final outcome. (iv) reconstruction processing units, presented in Figure 8d, which approximate the geodesic reconstruction by processing the input data using two basic morphological operations (erosion and dilation or vice-versa) followed by an elementwise max- or min-operation (depending on the operation) over the processed data and the original input data.

These processing units are employed to create the morphological layers, which provide the essential tools for the creation of the DeepMorphNets. Different from the standard convolutional layer, a single morphological layer can be composed of several neurons that may be performing different operations. This process allows the layer to produce distinct outputs, increasing the heterogeneity of the network and, consequently, the generalization capacity.

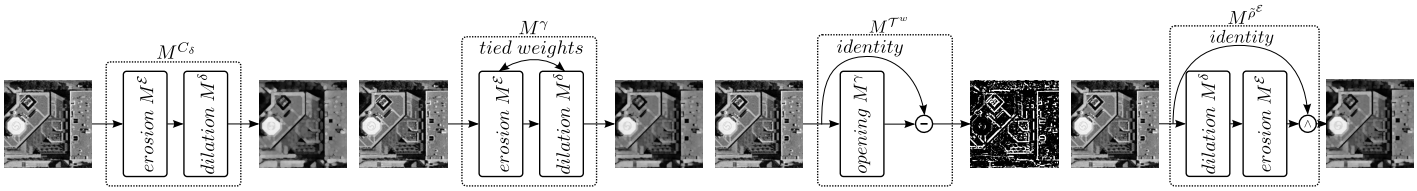
The morphological layers are then used to create the DeepMorphNets. The first proposed network is a morphological version of the AlexNet [38], presented in Figure 9. Such network was evaluated using two image classification datasets. In both cases, the DeepMorphNet produced competitive results when compared to ConvNets with equivalent architectures. A qualitative evaluation showed that the morphological network is capable of learning relevant filters and extracting salient features. Some of these features are presented in Figure 10.

V. Conclusion

In this PhD thesis, we proposed solutions that address important challenges related to the exploitation of deep learning into the remote sensing domain, including data availability, context exploitation, and so on. It was completed in approximately four years (from March 2015 to May 2019) and has resulted in four international journal papers [5], [8], [23], [25], and eleven international conference papers [3], [6], [7], [22], [24], [27]–[32].

Future work includes better analyze the effectiveness of the morphological neurons, combine ConvNets and DeepMorphNets, adapt DeepMorphNets to pixel classification, and implement more moderns architectures:

- *Analyze the effect of each type of morphological neuron.* This topic involves understanding the benefits of each type of morphological neuron and analyzing which ones are the best for each scenario. This is an interesting topic given that each type of neuron produces distinct outcomes. Therefore, this analysis would allow the definition of which neurons are most suitable for each application.
- *Combine ConvNets and DeepMorphNets.* This is a captivating research topic given that it focuses on extracting and combining the benefits of convolutional and morphological networks. As introduced, these techniques are able to capture distinct features. Hence, a combination of these approaches should be able to create a better representation, mainly because of the generated diversity.



(a) Composed Neuron (b) Opening Neuron (c) White Top-hat Neuron (d) Reconstruction by Erosion Neuron

Fig. 8: Definition of morphological neurons based on the proposed framework. Some units have tied weights to force the network to use the same filters (i.e. structuring element) in both operations. Other neurons have skip connections in order to allow an new operation using the original input and the processed data.

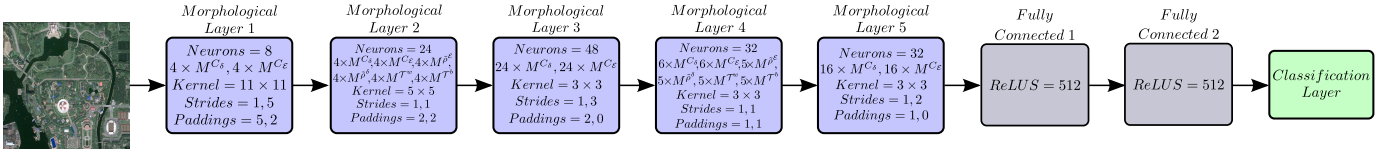


Fig. 9: The proposed morphological network DeepMorphNet conceived inspired by the AlexNet [38]. Since each layer is composed of distinct types of morphological neurons, the number of each type of neuron in each layer is presented as an integer times the symbol that represents that neuron: M^{C_δ} for composed processing units with erosion followed by dilation, $M^{C_\mathcal{E}}$, for composed processing units with dilation followed by erosion, M^{γ}, for opening morphological neuron, M^{φ}, for closing processing units, $M^{\mathcal{T}^w}$, for white top hat neuron, $M^{\mathcal{T}^b}$, for black top hat processing units,, $M^{\tilde{\rho}^{\mathcal{E}}}$, for morphological reconstruction by erosion, and $M^{\tilde{\rho}^{\delta}}$, for morphological reconstruction by dilation. Hence, the total number of processing units in a layer is the sum of all neurons independently of the type.

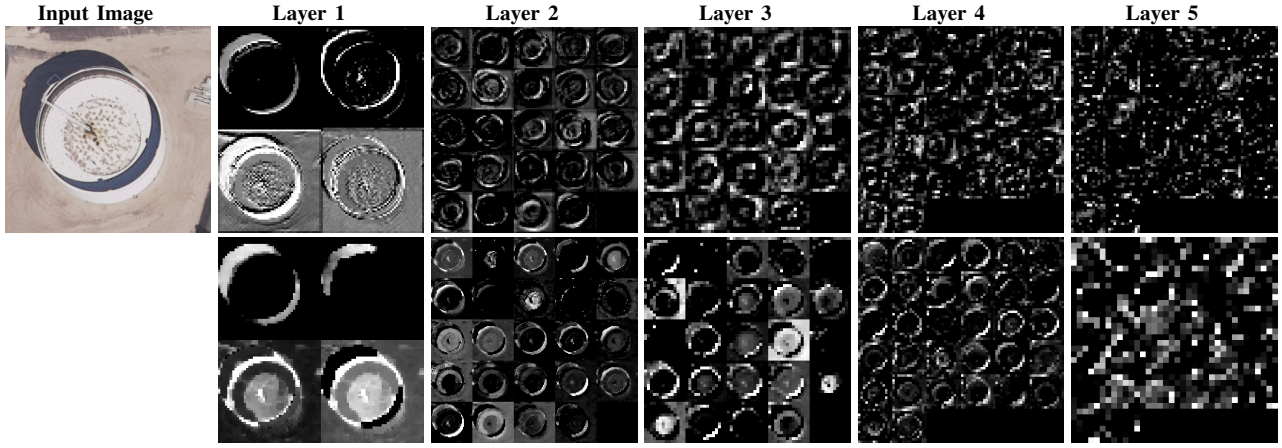


Fig. 10: Input images and some produced (upsampled) feature maps extracted from all layers of the networks for the UCMerced Land-use. The first row presents features from the ConvNet network and the second row presents the features of the proposed morphological network.

- *Adapt the deep morphological networks to perform pixel classification.* This is a direct application of the proposed DeepMorphNets. In this thesis, such network has been evaluated only for remote sensing scene classification. Therefore, it should be natural to apply the proposed networks for remote sensing pixel classification.
- *Implement modern layers and architectures based on the deep morphological networks.* There are several modern layers (such as the ones with dilated filters) and architectures (including ResNets [39] and DenseNets [40]). It is an interesting research topic to analyze if it is feasible and logical to adapt the deep morphological networks to recreate these techniques.

ACKNOWLEDGMENT

This research was partially supported by CNPq (grant 312167/2015-6), CAPES (grant 88881.131682/2016-01), and Fapemig (APQ-00449-17). The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of the GeForce GTX TITAN X GPU used for this research.

REFERENCES

[1] M. Volpi and D. Tuia, "Dense semantic labeling of subdecimeter resolution images with convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 881–893, 2017.

[2] J. A. d. dos Santos, P.-H. Gosselin, S. Philipp-Foliguet, R. d. S. Torres, and A. X. Falao, "Multiscale classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 10, pp. 3764–3775, 2012.

[3] K. Nogueira, W. R. Schwartz, and J. A. dos Santos, "Coffee crop recognition using multi-scale convolutional neural networks," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2015, pp. 67–74.

[4] D. Fustes, D. Cantorna, C. Dafonte, B. Arcay, A. Iglesias, and M. Manteiga, "A cloud-integrated web platform for marine monitoring using gis and remote sensing," *Future Generation Computer Systems*, vol. 34, pp. 155–160, 2014.

[5] K. Nogueira, S. G. Fadel, Í. C. Dourado, R. O. Werneck, J. A. V. Muñoz, O. A. Penatti, R. T. Calumby, L. T. Li, J. A. dos Santos, and R. S. Torres, "Exploiting convnet diversity for flooding identification," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 9, pp. 1446–1450, 2018.

[6] K. Nogueira, J. A. Dos Santos, T. Fornazari, T. S. F. Silva, L. P. Morellato, and R. d. S. Torres, "Towards vegetation species discrimination by using data-driven descriptors," in *Pattern Recogniton in Remote Sensing (PRRS), 2016 9th IAPR Workshop on*. IEEE, 2016, pp. 1–6.

[7] K. Nogueira, J. A. dos Santos, L. Cancian, B. D. Borges, T. S. F. Silva, L. P. Morellato, and R. S. Torres, "Semantic segmentation of vegetation images acquired by unmanned aerial vehicles using an ensemble of convnets," *IEEE International Geoscience & Remote Sensing Symposium*, 2017.

[8] K. Nogueira, J. A. dos Santos, N. Menini, T. S. F. Silva, L. P. C. Morellato, and R. S. Torres, "Spatio-temporal vegetation pixel classification by using convolutional networks," *IEEE Geoscience and Remote Sensing Letters*, 2019.

[9] J. R. Jensen and K. Lulla, "Introductory digital image processing: a remote sensing perspective," 1987.

[10] G. Kumar and P. K. Bhatia, "A detailed review of feature extraction in image processing systems," in *Advanced Computing & Communication Technologies*. IEEE, 2014, pp. 5–12.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] R. de O. Stehling, M. A. Nascimento, and A. X. Falcao, "A compact and efficient image retrieval approach based on border/interior pixel classification," in *International Conference on Information and Knowledge Management*, 2002, pp. 102–109.

[13] F. Hu, G.-S. Xia, J. Hu, Y. Zhong, and K. Xu, "Fast binary coding for the scene classification of high-resolution remote sensing imagery," *Remote Sensing*, vol. 8, no. 7, p. 555, 2016.

[14] C.-h. Chen, L.-F. Pau, and P. S.-p. Wang, *Handbook of pattern recognition and computer vision*. World Scientific, 2010, vol. 27.

[15] J. A. dos Santos, O. A. B. Penatti, and R. da Silva Torres, "Evaluating the potential of texture and color descriptors for remote sensing image retrieval and classification." in *International Conference on Computer Vision Theory and Applications*, 2010, pp. 203–208.

[16] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community," *Journal of Applied Remote Sensing*, vol. 11, no. 4, p. 042609, 2017.

[17] J. dos Santos, O. Penatti, P. Gosselin, A. Falcao, S. Philipp-Foliguet, and R. Torres, "Efficient and effective hierarchical feature propagation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, no. 99, pp. 1–12, 2014.

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, vol. 2, 2003, pp. 1470–1477.

[21] K. Nogueira, "Going deep into remote sensing spatial feature learning," 2019.

[22] K. Nogueira, W. O. Miranda, and J. A. Dos Santos, "Improving spatial feature representation from aerial scenes by using convolutional networks," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 2015, pp. 289–296.

[23] K. Nogueira, O. A. Penatti, and J. A. dos Santos, "Towards better exploiting convolutional neural networks for remote sensing scene classification," *Pattern Recognition*, vol. 61, pp. 539–556, 2017.

[24] K. Nogueira, M. Dalla Mura, J. Chanussot, W. R. Schwartz, and J. A. dos Santos, "Learning to semantically segment high-resolution remote sensing images," in *International Conference on Pattern Recognition*. IEEE, 2016, pp. 3566–3571.

[25] ——, "Dynamic multicontext segmentation of remote sensing images based on convolutional networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7503–7520, 2019.

[26] K. Nogueira, J. Chanussot, M. D. Mura, W. R. Schwartz, and J. A. d. Santos, "An introduction to deep morphological networks," *arXiv preprint arXiv:1906.01751*, 2019.

[27] T. M. Santana, K. Nogueira, A. M. Machado, and J. A. dos Santos, "Deep contextual description of superpixels for aerial urban scenes classification," in *IEEE International Geoscience & Remote Sensing Symposium*, July 2017, pp. 3027–3031.

[28] R. Baeta, K. Nogueira, D. Menotti, and J. A. dos Santos, "Learning deep features on multiple scales for coffee crop recognition," in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2017, pp. 262–268.

[29] K. Nogueira, S. G. Fadel, I. C. Dourado, R. d. O. Werneck, J. A. V. Muñoz, O. A. B. Penatti, R. T. Calumby, L. T. Li, J. A. dos Santos, and R. d. S. Torres, "Data-driven flood detection using neural networks," in *Working Notes Proc. MediaEval Workshop*, 2017, p. 2. [Online]. Available: http://slim-sig.irisa.fr/me17/Mediaeval_2017_paper_39.pdf

[30] J. A. V. Muñoz, L. T. Li, I. C. Dourado, K. Nogueira, S. G. Fadel, O. A. B. Penatti, J. Almeida, L. A. M. Pereira, R. T. Calumby, J. A. dos Santos, and R. d. S. Torres, "A ranking fusion approach for geographic-location prediction of multimedia objects," in *Working Notes Proc. MediaEval Workshop*, 2016, p. 2. [Online]. Available: http://ceur-ws.org/Vol-1739/MediaEval_2016_paper_22.pdf

[31] O. A. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Conference on Computer Vision and Pattern Recognition Workshop*, 2015, pp. 44–51.

[32] L. T. Li, J. A. V. Muñoz, J. Almeida, R. T. Calumby, O. A. B. Penatti, I. C. Dourado, K. Nogueira, P. R. Mendes Júnior, A. M. L. Pereira, D. C. G. Pedronette, M. A. Gonçalves, J. A. dos Santos, and R. d. S. Torres, "Recod @ placing task of mediaeval 2015," in *Working Notes Proc. MediaEval Workshop*, 2015, p. 2. [Online]. Available: http://ceur-ws.org/Vol-1436/Paper49.pdf

[33] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, 2019.

[34] R. Szeliski, *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[35] J. Serra and P. Soille, *Mathematical morphology and its applications to image processing*. Springer Science & Business Media, 2012, vol. 2.

[36] Y. Seo, B. Park, S.-C. Yoon, K. C. Lawrence, and G. R. Gamble, "Morphological image analysis for foodborne bacteria classification," *Transactions of the American Society of Agricultural and Biological Engineers*, vol. 61, pp. 5–13, 2018.

[37] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1800–1807.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Neural Information Processing Systems*, 2012, pp. 1106–1114.

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

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Conference on Computer Vision and Pattern Recognition*, June 2017, pp. 2261–2269.