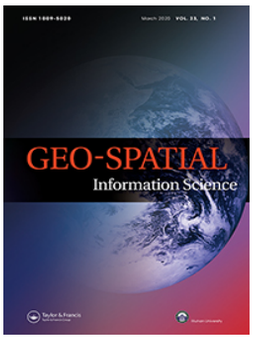

Geo-spatial Information Science

ISSN: 1009-5020 (Print) 1993-5153 (Online) Journal homepage: www.tandfonline.com/journals/tgsi20

Deep learning for geometric and semantic tasks in photogrammetry and remote sensing

Christian Heipke & Franz Rottensteiner

To cite this article: Christian Heipke & Franz Rottensteiner (2020) Deep learning for geometric and semantic tasks in photogrammetry and remote sensing, Geo-spatial Information Science, 23:1, 10-19, DOI: 10.1080/10095020.2020.1718003

To link to this article: https://doi.org/10.1080/10095020.2020.1718003

GEO-SPATIAL INFORMATION SCIENCE
2020, VOL. 23, NO. 1, 10–19
https://doi.org/10.1080/10095020.2020.1718003

Taylor & Francis
Taylor & Francis Group

Deep learning for geometric and semantic tasks in photogrammetry and remote sensing

Christian Heipke and Franz Rottensteiner

Institute of Photogrammetry and GeoInformation (PI), Leibniz University Hannover, Hannover, Germany

ABSTRACT

During the last few years, artificial intelligence based on deep learning, and particularly based on convolutional neural networks, has acted as a game changer in just about all tasks related to photogrammetry and remote sensing. Results have shown partly significant improvements in many projects all across the photogrammetric processing chain from image orientation to surface reconstruction, scene classification as well as change detection, object extraction and object tracking and recognition in image sequences. This paper summarizes the foundations of deep learning for photogrammetry and remote sensing before illustrating, by way of example, different projects being carried out at the Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, in this exciting and fast moving field of research and development.

ARTICLE HISTORY
Received 18 December 2019
Accepted 14 January 2020

KEYWORDS
Deep learning; machine learning; convolutional neural networks(CNN); example project from IPI

1. Introduction

The use of neurons and neural networks for artificial intelligence in general, and for tasks related to image understanding in particular, is not new. Artificial neurons were described by McCulloch and Pitts as early as 1943. Rosenblatt (1958) developed the first computer program, which implemented the so-called concept of perceptrons (see Figure 1) and was able to learn based on trial and error. After Minsky and Papert (1969) proved mathematically that the original concept could not model the important XOR statement (exclusive OR; the result is true only for an odd number of positive inputs), which dealt the research on neural networks a significant blow, the field was revived about two decades later with the introduction of back-propagation (Rummelhart, Hinton, and Williams 1986; LeCun 1987), which allowed the efficient training of multi-layer artificial neural networks (see Figure 2), to which the theoretical restrictions noted by Minsky and Papert (1969) do not apply. Other important steps were the introduction of Convolutional Neural Networks (CNN, LeCun et al. 1989; LeCun and Bengio 1998) and deep belief networks (Hinton, Osindero, and Teh 2006). The breakthrough of deep learning came, when Krizhevsky, Sutskever, and Hinton (2012) won the ImageNet Large-Scale Recognition Challenge, a classification task involving 1000 different classes (Russakovsky et al. 2015) using a CNN-based approach. Their network, called AlexNet, lowered the remaining error by nearly 50% compared to the previous best result.

Since then, deep learning based on neural networks has seen a tremendous success in many different areas including photogrammetry and remote sensing (Zhu et al. 2017). The main reasons are twofold: (a) since a few years, computers are powerful enough to process and store data using large networks with many layers (called "deep" networks), in particular when using GPUs (graphical processing units) during training, and (b) more and more training data became available for the different tasks (it should be noted that AlexNet used some 1,2 million labeled training images to learn a total of some 60 million parameters). The most comprehensive textbook available for deep learning today is the one by Goodfellow, Bengio, and Courville (2016).

This paper is structured as follows: after a brief summary of the principles of deep learning and CNN, by way of example we describe the work carried out along those lines at the Institute of Photogrammetry and GeoInformation (IPI) of Leibniz University Hannover. We subdivide the main chapter into geometric approaches and those used in aerial image analysis and close range. Finally, some conclusions are drawn.

2. Convolutional networks for image analysis

In principle, a CNN can be considered a classifier. In traditional classifiers (random forests, support vector machines, conditional random fields, maximum likelihood estimation, etc.) features representing the different classes are extracted from the data set in a pre-processing step, and classification is then performed based on these features. It is clear then that the results can only be as good as the selected features. CNN overcome this problem by learning the features together with the

CONTACT Christian Heipke heipke@ipi.uni-hannover.de

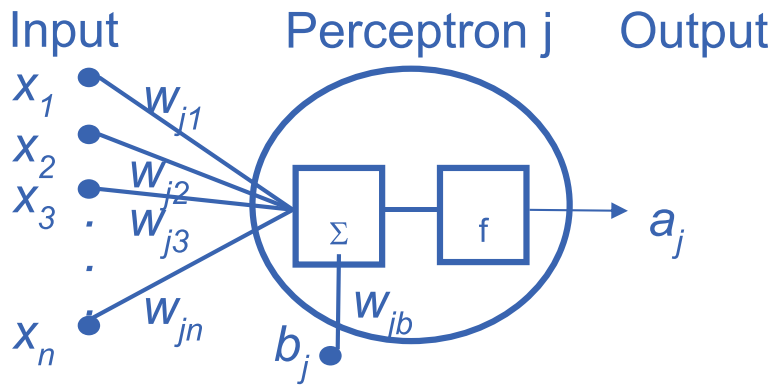

Figure 1. Concept of a perceptron *j*. Depicted are the input x_i, the weight w_{ji}, the bias b_j, the (non-linear) function *f* and the resulting output a_j.

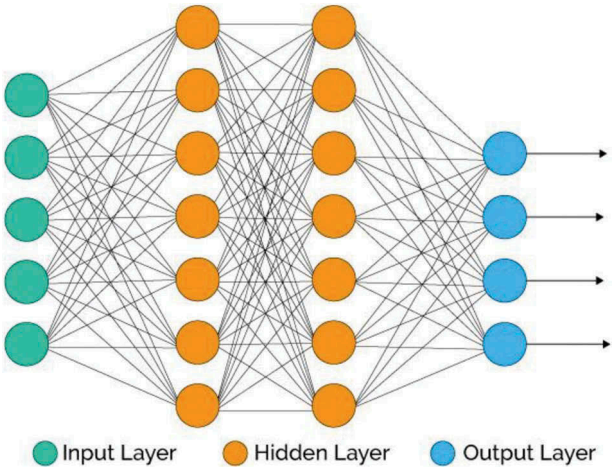

Figure 2. Artificial neural network with input layer, two hidden layers and output layer.

corresponding label for each data sample (see Figure 3). The price to pay is the fact that a very large amount of training data is needed to estimate this largely increased number of unknowns. Since often the required amount of training data is not available, additional data are generated from the available ones (data augmentation) or simulation results are used as a substitute for real training data.

In a CNN architecture, in principle three different steps are carried out in each layer (see Figure 4): (a) the convolution step, where a set of digital filters is applied to an input image of fixed size, (b) a so-called pooling step, where from a larger group of filtered pixels only one (the one with the maximum entry in the case of max-pooling) is retained and (c) an activation step, where the remaining set of pixels is subjected to a non-linear function. In most current works the rectilinear unit (ReLU) has been chosen as an activation function. These steps are followed by processing through a few densely connected layers which eventually results in a feature vector representing the complete input image. This feature vector is then classified using an arbitrary classifier. Typically, the softmax classifier is used as it has several advantages (Kreinovich and Quintana 1991).

Similar to the concept of image pyramids the pooling step is employed to increase the context area considered by each filter. A non-linear activation function must be used, since otherwise, all steps could be substituted by one (linear) layer between input and output, which is known not to be expressive enough for learning any but very simple tasks. The elements of the filters are considered as unknown parameters which are learned from training data via stochastic gradient descent. Initial values can typically be selected arbitrarily and the gradients are computed by back-propagation. Updates for the unknowns are found based on a specially designed loss function, which for the training data minimizes a function of the differences between the class predicted by the network and the known class. Various training strategies are in use regarding the size of the sample set used simultaneously (called batch size) in one parameter update

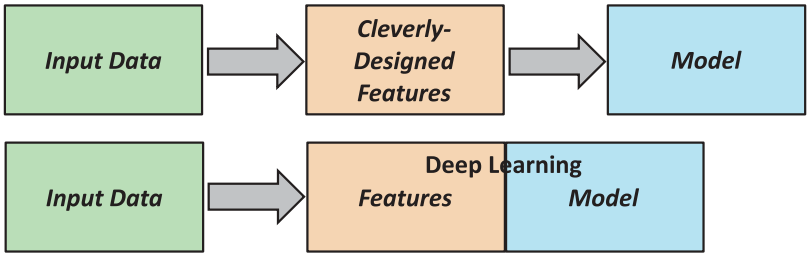

Figure 3. Concept of a standard classifier (top) and a CNN classifier (bottom). The advantage of the latter is that the features and the model parameters are learned simultaneously from the training data.

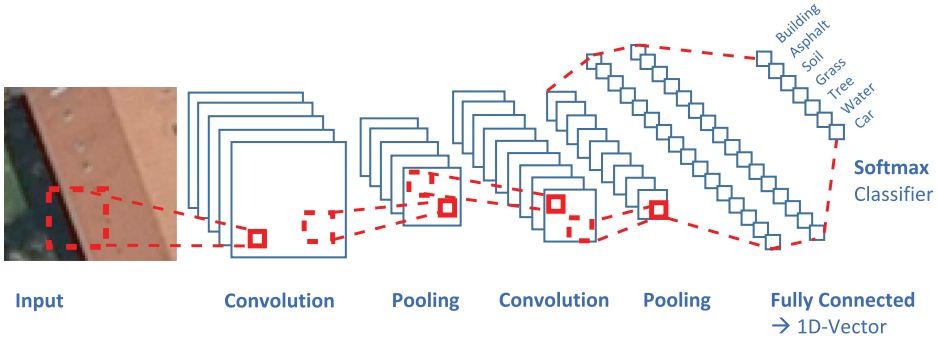

Figure 4. Architecture of a typical Convolutional Neural Network for image analysis. The figure shows the successive steps of convolution and pooling to generate a feature vector which is classified in the final step, typically using the softmax classifier (the non-linear activation function is not depicted).

step and the selection of nodes used for each training sample (in the so-called dropout strategy some of the nodes are not always used to increase the generalization capabilities of the network).

As would have become apparent after this description, when using a CNN, several parameters need to be fixed prior to processing the images. These comprise among others the number of filters and their size, the number of nodes in each layer and the number of layers. The latter one is of particular importance (Baral, Fuentes, and Kreinovich 2018): In principle, a neural network (as any supervised classifier) can be seen as an interpolation function with the training samples serving as support. Each path between input and output through the network represents such a function. In order to increase the accuracy of the overall results, many different functions are needed. However, permutations within a layer lead to the same function being implemented through different paths. Therefore, the number of nodes per layer should be kept reasonably small, and as a consequence, many layers are needed in order to obtain the number of unknowns necessary for complex tasks; this explains the fact that in general deeper networks yield better results (e.g. He et al. 2015).

While the original concept of a CNN would typically learn a feature vector to represent a whole image, other tasks have also been solved using CNN. Among those are pixel-wise classification (called semantic segmentation in Computer Vision (CV) terminology), where Fully Convolutional Networks (FCN, Long, Shelhamer, and Darrell 2015) are employed. Encoder-decoder networks (Hinton and Salakhutdinov 2006; Ronneberger, Fischer, and Brox 2015, see Figure 5) carry out the upsampling required to get pixel-wise class predictions in a series of steps in the decoder part that mirror the structure of the downsampling procedure of the encoder network. The U-net structure of Ronneberger, Fischer, and Brox (2015) includes so-called skip connections to better preserve object boundaries. Also object detection, where objects are described by bounding boxes (Ren et al. 2017) and object delineation (instance segmentation in the CV world, He et al. 2017), where in addition to these

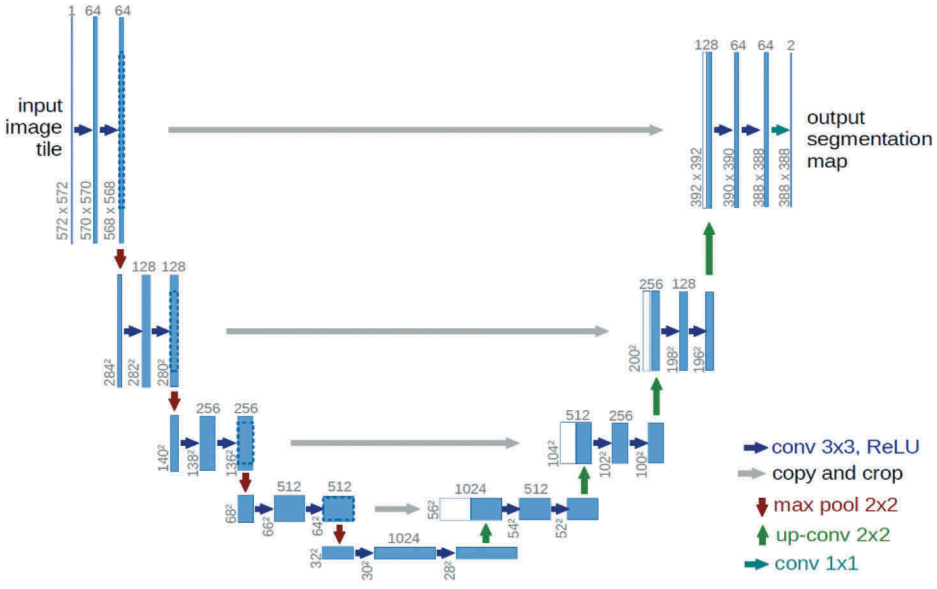

Figure 5. The U-net architecture, an example of an encoder network with skip connections (Ronneberger, Fischer, and Brox 2015).

bounding boxes a mask is computed for each object with pixels belonging to either fore- or background, describe very useful tasks tackled using CNNs. Other network architectures comprise Siamese networks (Bromley 1993), where weights are shared between two different parts of the network, often to determine similarity of two images (e.g. in image matching), Recurrent Neural Networks (RNN, e.g. Grave et al. 2009) for dealing with time-dependent data and Generative Adversarial Networks (GAN, Goodfellow et al. 2014), which can learn new data with the same statistical distribution as a given data set. The latter can be useful, e.g. in transfer learning (Yosinski et al. 2014; Tzeng et al. 2017). Finally, CNN techniques have also been applied to unstructured 3D point data (Landrieu and Simonovsky 2018), e.g. representing depth (Qi et al. 2016).

In particular, for pixel-wise classification and for object delineation it is important in our field to consider the geometric accuracy of the object boundary, as a different label is sought for each pixel. Thus, in some works maximum pooling, which acts as a low path filter and thus blurs the boundary, is not used. In order to still keep the number of filter elements, and thus of unknown parameters to be estimated, at a reasonable number, filter elements are interpolated from a selected number of unknowns in successive layers, or dilated convolution, originally developed for wavelet decomposition (Holschneider et al. 1990; Yu and Koltun 2016), is used, where a number of elements are set to zero. In both cases, care should be taken not to violate the sampling theorem.

3. Deep learning research at IPI

In photogrammetry and remote sensing, and in particular when dealing with aerial or satellite images, some of the conditions which hold true for typical computer vision applications do not apply: (a) the images are much larger and contain a multitude of objects, each often only a few pixels in size; (b) the image orientation and the ground sampling distance are typically known; (c) there is no preferred direction in the image ("up" does not point to the sky); (d) besides 3-channel color images other modalities such as additional bands (e.g. the infrared channel) and depth are often available, sometimes also other data such as maps, social media data or Volunteered Geographical Information (VGI); (e) often, considerable prior knowledge about the scene is available; (f) typically, there is a shortage of training data, while at least in an update scenario outdated map data are given; and finally (g) the accuracy requirements are typically more stringent, both for geometric and for semantic results. Thus, the question did arise a few years ago, in how far deep learning and CNN can be used to advantage also in photogrammetry and remote sensing. This question has also influenced work at the Institute of Photogrammetry and GeoInformation, as will be shown in the following.

3.1. CNN for geometric tasks

Problems relating to image orientation and dense surface reconstruction are considered geometric tasks in this context. We report on projects related to these two tasks.

In image orientation, a specific problem is the detection, description and matching of conjugate point pairs. While in standard cases different operational solutions based on the well-known SIFT (Scale Invariant Feature Transform, Lowe 2004) operator exist, these solutions reach their limits for wide baseline image pairs with largely different viewing directions and different scales. This is for instance the case when oblique aerial images of different viewing directions need to be matched. Chen, Rottensteiner, and Heipke (2016) suggest a Siamese network to learn a feature descriptor to solve this problem. The loss function is designed according to the triplet loss paradigm (Weinberger and Saul 2009): it pulls the descriptors of matching patches closer in feature space while pushing the descriptors for non-matching pairs further away from each other.

Also after decades of research and development, 3D surface reconstruction cannot be considered a problem solved under all circumstances: areas with poor and repetitive texture, as well as sharp depth discontinuities and resulting occlusions continue to pose difficulties. The first solution based on CNN was presented by Zbontar and LeCun (2015). At IPI we deal with this problem on two levels: On the one hand, Kang et al. (2019) developed a new dense stereo method based on dilated convolution, which does not only use depth as training data but includes a depth gradient term into the loss function (see Figure 6). The results show that more detail can be retrieved in particular in the presence of depth discontinuities, if (and only if) the gradients in the training data are reliable. On the other hand, Mehltretter and Heipke (2019) improve the quality of dense stereo matching by analyzing the 3D cost volume of the related disparity space image. In a novel CNN architecture features for confidence estimation are directly learned from the volumetric 3D data.

3.2. Aerial image analysis

The automatic analysis of aerial imagery has been a major focus of research for a number of decades at IPI. We currently work on three different topics with a connection to deep learning: land cover and land use classification, transfer learning and bomb crater detection.

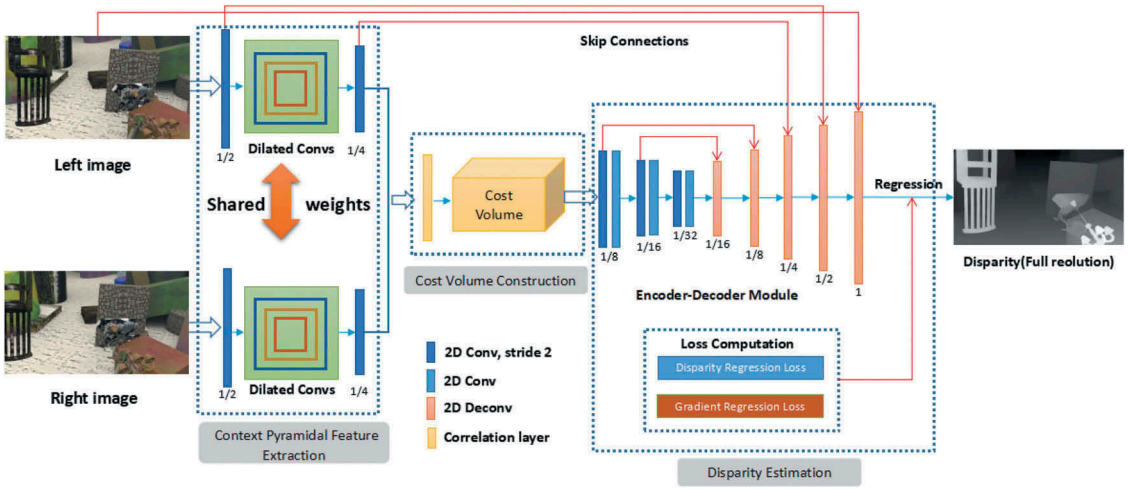

Figure 6. Network architecture for dense matching (Kang et al. 2019).

The first one is concerned with the update of land cover and land use databases. Yang, Rottensteiner, and Heipke (2018, 2019) have suggested two network architectures, one for land cover and another one for land use update. For the land cover, an ensemble classifier combining RGB data with an infrared channel and height in the form of a normalized Digital Terrain Model is being used in an encoder-decoder network structure with skip connections (see Figure 7). In the following land use estimation, the object shapes are taken from the topographic database to stabilize the solution, while for each object the label is estimated using the input information as well as the result of land cover classification. The results confirm that CNN can outperform the best methods employed previously, i.e. Conditional Random Fields (Albert, Rottensteiner, and Heipke 2017).

Another topic we work on is related to transfer learning with the goal of pixel-wise classification of mono-temporal data (Wittich and Rottensteiner 2019). Assuming the availability of labeled training data for existing data (called the *source domain*), we adapt a CNN trained on these data to new data (*target domain*) that have a different joint distribution of class labels and features. In domain adaptation, a specific setting of transfer learning, this adaptation is to be achieved without new hand-labeled training samples from the new domain. For that purpose, we adapt Adversarial Discriminative Domain Adaptation (ADDA; Tzeng et al. 2017) to the prediction of land cover from aerial images and a Digital Surface Model (DSM). Adversarial methods try to train a neural network to produce a feature representation that is independent from the domain from which a sample is drawn; similarity is measured by the capability of another neural network (called discriminator) to predict from which domain a feature vector was drawn. While ADDA gives encouraging results for similar domains, there is clearly room for improvement if the domains are very different, especially with respect to the distribution of class labels.

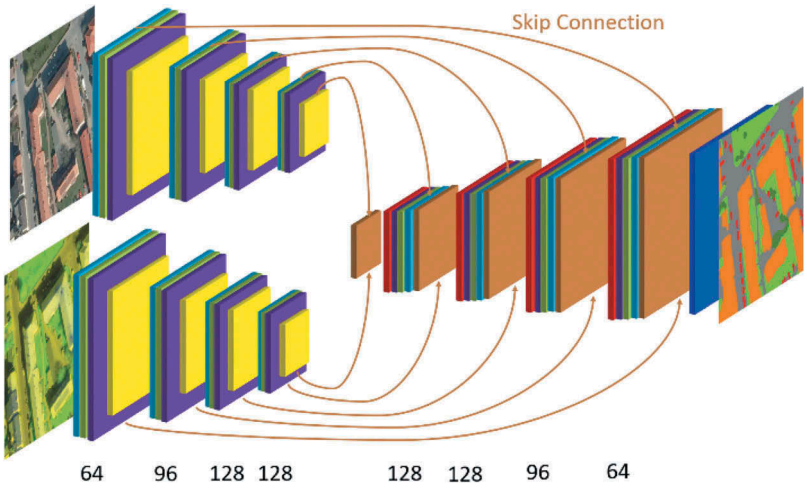

Figure 7. Architecture of ensemble classifier for semantic segmentation of land cover including skip connections. The top encoder part takes color images as input, the bottom part the infrared channel and height information. The encoder part ensures a detailed information for each pixel (Yang et al. 2019).

In a more classical pattern recognition approach Clermont et al. (2019) extract bomb crater from images acquired during the second world war (see Figure 8). The background of this work is the fact that a number of bombs did not explode during the war and are still sitting in the ground, posing a significant danger in particular during ground construction work. The rationale of the project is that finding the bomb craters will give an indication of where unexploded bombs might lie. The work is based on a variant of the ResNet architecture (He et al. 2015), the results show that this seemingly not so difficult problem is indeed challenging, partly because of the lack of a sufficient number of training data.

3.3. Close range applications

In this area, we are concerned with mobility, as well as a project dealing with artwork. In the field of mobility, we have designed and implemented a system that can recognize and determine the relative poses of cars in a stereoscopic image sequence based on adaptive shape models. In a related project, pedestrians are detected and tracked in these sequences. Finally, we are working on the re-identification of persons being viewed from different cameras of a sensor network. All three projects are connected to the German Science Foundation as part of the Research Training Network "Integrity and Collaboration in Dynamic Sensor Networks" funded at our university (i.c.sens 2019).

In the first project (Coenen, Rottensteiner, and Heipke 2019), for every detected object a CAD model is fitted into a stereo image pair and the derived point cloud, allowing to estimate the pose of the car relative of the camera position and, consequently, of the camera relative to the other car. If the detected cars are equipped with a GNSS receiver and can communicate their position to the camera, these cars thus act as dynamic control points for image orientation and, thus, the positioning of the cars. The core of the method is 3D reconstruction by optimizing a probabilistic energy function involving several data and prior terms. A multi-task CNN delivers some of the image-related data terms by predicting the positions of keypoints and model edges in the image while also providing a prior term for the coarse orientation (rotation about the vertical axis) of the car. Figure 9 shows the qualitative results of 3D reconstruction based on Coenen, Rottensteiner, and Heipke (2019).

Pedestrian detection and tracking (Nguyen, Rottensteiner, and Heipke 2019) rely on the Mask R-CNN approach (He et al. 2017) to generate and classify region proposals assumed to contain pedestrians. Since stereo information is available detection and tracking are carried out in 3D space, which allows to employ additional geometric constraints (a position in 3D can only be occupied by one person). Data association is then based on the triplet loss using TriNet (Hermans, Beyer, and Leibe 2017) and takes into account the local context. Experiments indicate the good quality of the results, both when evaluating the geometric accuracy of the resulting trajectories and also when investigating their length: the new approach shows fewer identity switches and thus longer trajectories than comparable solutions.

Person re-identification is tackled by using a fisheye camera in nadir viewing position (Blott, Takami, and Heipke 2018, Blott, Yu, and Heipke 2019). In this way, multiple views of a person (front, side, back) can be extracted from the image sequences, before comparing this 3-view set of images with a database in order to re-identify the person. Classification of the different views uses a ResNet variant (He et al. 2015), while in the matching stage the TriNet is used to extract features. The results are promising and the approach outperforms existing approaches by a significant margin, partly due to the fact that more information is available than in single image solutions.

The last project we want to discuss is related to cultural heritage documentation. There are many museums having collections of silk fabrics. These collections are also documented in digital records, typically consisting of digital images and a corresponding text.

Figure 8. Results of automatic detection of bomb craters in historic wartime images using convolutional neural networks.

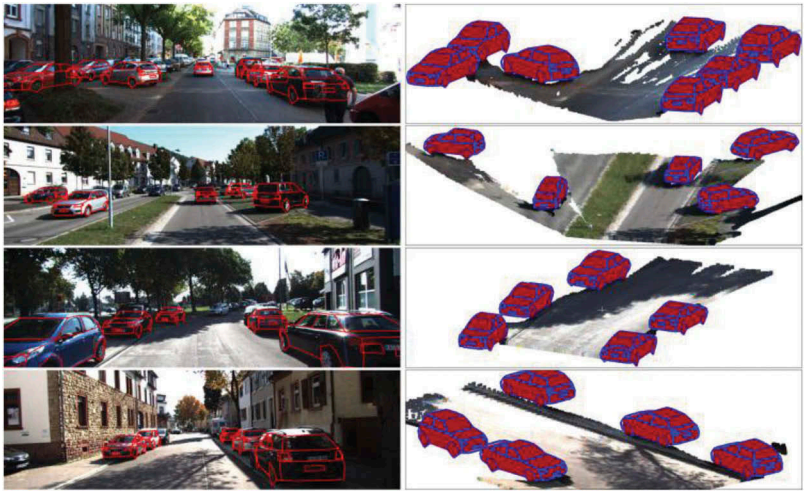

Figure 9. Four qualitative results of 3D vehicle reconstruction based on (Coenen et al. 2019). Left: Input image, superimposed with extracted model wireframes. Right: 3D view on the reconstructed scene.

The information contained in the text, e.g. describing the time or place of production of a fabric, is very important for art historians, but it is not provided in a standardized way, and sometimes important pieces of information are missing. In the context of an EU H2020 project (SILKNOW 2019), a multi-task CNN based on ResNet (He et al. 2015) was developed that simultaneously predicts the production time, the production place and the production technique from a digital image, deriving the training data automatically by analyzing existing collections (Dorozynski, Clermont, and Rottensteiner 2019). The results show that by combining these prediction tasks, the accuracy of prediction is increased if high-quality training samples are used.

4. Conclusions

The short summary of the individual projects had the goal to convince the reader, that indeed, deep learning and CNN-based solutions carry great value in photogrammetry and remote sensing. In both, geometric and semantic tasks, CNN-based solutions outperform those based on more traditional image analysis. The strength of CNN is the combined estimation of the feature representation and the labels during classification, and it seems that deeper networks are practically guaranteed to yield better results than shallow networks, as long as enough training data is available. Open source implementations for CNN exist, and the industry has started to make heavy use of these algorithms.

Having said that, one should not forget that in essence, a CNN (and any deep learning approach) is a classifier. As such it comes with the same general limitations as any other classifier. Therefore, a number of questions need further attention:

- A CNN needs a sufficient number of representative training data, well balanced with respect to the related classes. Otherwise there is a risk of overfitting the classifier to the training data and a bias is likely to be introduced into the results. To increase the amount of training data, data augmentation, transfer learning, approaches which are able to tolerate a certain amount of incorrect labels (label noise), semi-supervised and unsupervised learning (clustering) can be employed and should be studied. In some cases, simulation techniques may also help.
- A CNN "cannot learn the unseen", the generalization capabilities are limited to previously seen training data.
- Incremental learning and forgetting (or "unlearning") data, e.g. those which are not relevant anymore due to a changing environment, is a topic which has received little attention in our field so far, yet this area offers a large potential, in particular for multi-temporal analysis.
- A number of design decisions need to be taken, e.g. with respect to the network architecture and the design of the loss function. It is not clear in general, how different choices influence the results, and how robust the classifiers are. Some works suggest that CNN can be indeed be fooled relatively easily (Nguyen, Yosinski, and Clune 2015).
- A CNN is based on correlations of different data sets. We argue that understanding a task to then reason about possible solutions in a way humans do is far beyond the scope of the currently employed methods (note that this does not mean that reasoning is not done, e.g. in a game of chess or Go. It does mean, however, that CNN does not have an intuition for possibly correct solutions and abstract deductive learning).
- A CNN is largely a black box. While it may deliver very good results, it is largely unknown why and how exactly these results are being reached. Besides

being a little frustrating from a scientific point of view, this means that the limitations of these methods cannot clearly be stated, resulting in some doubts whether the methods can be employed in real-world safety- and security-related areas – autonomous driving is a good example.

Thus, it seems that a number of difficult research questions still exist in our field. Besides taking care of a better geometric and semantic accuracy of the results, improving their reliability is of great importance. This will only be possible by investigating better ways to explain why deep learning approaches give the results they do (see e.g. Roscher et al. 2019). Another important aspect is the integration of deep learning approaches with other learning paradigms and prior knowledge, according to the motto, "Why learn what we already know?". So far, the approaches discussed in this paper are mainly standalone solutions. We believe that in the long run, only a combination of different methods will lead to success.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Christian Heipke is a professor of photogrammetry and remote sensing at Leibniz University Hannover, where he currently leads a group of about 25 researchers. His professional interests comprise all aspects of photogrammetry, remote sensing, image understanding and their connection to computer vision and GIS. His has authored or coauthored more than 300 scientific papers, more than 70 of which appeared in peer-reviewed international journals. He is the recipient of the 1992 ISPRS Otto von Gruber Award, the 2012 ISPRS Fred Doyle Award, and the 2013 ASPRS Photogrammetric (Fairchild) Award. He is an ordinary member of various learnt societies. From 2004 to 2009, he served as vice president of EuroSDR. From 2011-2014 he was chair of the German Geodetic Commission (DGK), from 2012-2016 ISPRS Secretary General. Currently he serves as ISPRS President.

Franz Rottensteiner is an Associate Professor and leader of the research group "Photogrammetric Image Analysis" at Leibniz University Hannover. He received the Dipl.-Ing. degree in surveying and the Ph.D. degree and venia docendi in photogrammetry, all from Vienna University of Technology (TUW), Vienna, Austria. His research interests include all aspects of image orientation, image classification, automated object detection and reconstruction from images and point clouds, and change detection from remote sensing data. Before joining LUH in 2008, he worked at TUW and the Universities of New South Wales and Melbourne, respectively, both in Australia. He has authored or coauthored more than 150 scientific papers, 36 of which have appeared in peer-reviewed international journals. He received the Karl Rinner Award of the Austrian Geodetic Commission in 2004 and the Carl Pulfrich Award for Photogrammetry, sponsored by Leica Geosystems, in 2017. Since 2011, he has been the Associate Editor of the ISI-listed journal "Photogrammetrie Fernerkundung Geoinformation". Being the Chairman of the ISPRS Working Group II/4, he initiated and conducted the ISPRS benchmark on urban object detection and 3D building reconstruction.

ORCID

Christian Heipke http://orcid.org/0000-0002-7007-9549
Franz Rottensteiner http://orcid.org/0000-0003-1942-8210

References

General references

Baral, C., O. Fuentes, and V. Kreinovich. 2018."Why Deep Neural Networks: A Possible Theoretical Explanation". In *Constraint Programming and Decision Making: Theory and Applications*, edited by M. Ceberio and V. Kreinovich, 1–6. Cham, Switzerland: Springer. http://www.cs.utep.edu/vladik/2015/tr15-55.pdf

Bromley, J., J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, and R. Shah. 1993. "Signature Verification Using a "Siamese" Time Delay Neural Network." *International Journal of Pattern Recognition and Artificial Intelligence* 7 (04): 669–688. doi:10.1142/S0218001493000339.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, 2014. "Generative Adversarial Nets." Advances in Neural Information Processing Systems 27 (NIPS'14), Montreal, Quebec, Canada, December 8–13, 2672–2680.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press.

Grave, A., M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber. 2009. "A Novel Connectionist System for Improved Unconstrained Handwriting Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31 (5): 855–868. doi:10.1109/TPAMI.2008.137.

He, K., G. Gkioxari, P. Dollar, and R. Girshick. 2017. "Mask R-CNN." Proc. International Conference on Computer Vision (ICCV), Venice, Italy, 2980–2988.

He, K., X. Zhang, S. Ren, and J. Sun. 2015. "Delving Deep into Rectifiers: Surpassing Human-level Performance on ImageNet Classification." IEEE International Conference on Computer Vision (ICCV), Las Condes, Santiago, Chile, 1026–1034.

Hermans, A., L. Beyer, and B. Leibe. 2017. "In Defence of the Triplet Loss for Person Re-identification." In *CoRR. arXiv:abs/1703.07737*. Ithaca, NY: Cornell University. https://arxiv.org/abs/1703.07737

Hinton, G., and R. Salakhutdinov. 2006. "Reducing the Dimensionality of Data with Neural Networks." *Science* 313 (5786): 504–507. doi:10.1126/science.1127647.

Hinton, G., S. Osindero, and Y. Teh. 2006. "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation* 18: 1527–1554. doi:10.1162/neco.2006.18.7.1527.

Holschneider, M., R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. 1990. "A Real-time Algorithm for Signal Analysis with the Help of the Wavelet Transform." In *Wavelets*, edited by J. M. Combres, A. Grossmann, and P. Tchamitchian, 286–297. Berlin, Heidelberg: Springer.

Kreinovich, V., and C. Quintana. 1991. "Neural Networks: What Non-linearity to Choose?" Proceedings of the 4th

University of New Brunswick Artificial Intelligence Workshop, Fredericton, New Brunswick, 627–637.
Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." Advances in Neural Information Processing Systems 25 (NIPS'12), Lake Tahoe, NV, 1097–1105.
Landrieu, L., and M. Simonovsky. 2018. "Large-scale Point Cloud Semantic Segmentation with Superpoint Graphs." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT.
LeCun, Y. 1987. "Modèles connexionnistes de l'apprentissage." Thèse de Doctorat, Université Paris 6.
LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. "Handwritten Digit Recognition with a Back-propagation Network." 2nd International Conference on Neural Information Processing Systems (NIPS'89), Denver, CO, 396–404.
LeCun, Y., and Y. Bengio. 1998. *Convolutional Networks for Images, Speech, and Time Series, the Handbook of Brain Theory and Neural Networks*. Cambridge, MA: MIT Press.
Long, J., E. Shelhamer, and T. Darrell. 2015. "Fully Convolutional Networks for Semantic Segmentation." IEEE Computer Vision and Pattern Recognition (CVPR '15), Boston, MA.
Lowe, D. G. 2004. "Distinctive Image Features from Scale-invariant Keypoints." *International Journal of Computer Vision* 60 (2): 91–110. doi:10.1023/B:VISI.0000029664.99615.94.
McCulloch, W., and W. Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 5: 115–133. doi:10.1007/BF02478259.
Minsky, M., and S. Papert. 1969. *Perceptrons*. Cambridge, MA: MIT Press.
Nguyen, A., J. Yosinski, and J. Clune. 2015. "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images." IEEE Computer Vision and Pattern Recognition (CVPR '15), Boston, MA.
Qi, C. R., H. Su, K. Mo, and L. Guibas. 2016. "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV.
Ren, S., K. He, R. Girshick, and J. Sun. 2017. "Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (6): 1137–1149. doi:10.1109/TPAMI.2016.2577031.
Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-Net: Convolutional Networks for Biomedical Image Segmentation." 18th Int. Conference on medical image computing and computer assisted intervention, Munich, Germany.
Roscher, R., B. Bohn, M. F. Duarte, and J. Garcke. 2019. "Explainable Machine Learning for Scientific Insights and Discoveries." In *arXiv:1905.08883*. Ithaca, NY: Cornell University.
Rosenblatt, F. 1958. "The Perceptron. A Probabilistic Model for Information Storage and Organization in the Brain." *Psychological Reviews* 65: 386–408. doi:10.1037/h0042519.
Rummelhart, D., G. Hinton, and R. Williams. 1986. "Learning Representations by Back-propagating Errors." *Nature* 323: 533–536. doi:10.1038/323533a0.
Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, et al. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115 (3): 211–252. doi:10.1007/s11263-015-0816-y.
Tzeng, E., J. Hoffman, K. Saenko, and T. Darrell. 2017. "Adversarial Discriminative Domain Adaptation." Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, July 21–26.
Weinberger, K. Q., and L. K. Saul. 2009. "Distance Metric Learning for Large Margin Nearest Neighbor Classification." *Journal of Machine Learning Research* 10: 207–244.
Yosinski, J., J. Clune, Y. Bengio, and H. Lipson. 2014. "How Transferable are Features in Deep Neural Networks?" Advances in Neural Information Processing Systems 27 (NIPS'14), Montreal, Quebec, Canada, December 8–13.
Yu, F., and V. Koltun. 2016. "Multi-Scale Context Aggregation by Dilated Convolutions." 4th International Conference on Learning Representations, Caribe Hilton, San Juan, Puerto Rico, May 2–4.
Zbontar, J., and Y. LeCun. 2015. "Computing the Stereo Matching Cost with a Convolutional Neural Network." *CVPR* 1592–1599. doi:10.1109/CVPR.2015.7298767.
Zhu, X., D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer. 2017. "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources." *IEEE GRSS Magazine* 5 (4): 8–36.

IPI contributions

Albert L., Rottensteiner F., and Heipke C. 2017. "A Higher Order Conditional Random Field Model for Simultaneous Classification of Land Cover and Land Use." *ISPRS Journal for Photogrammetry and Remote Sensing* 130 (2017): 63–80.
Blott G., Yu J., and Heipke C. 2019. "Multi-View Person Re-Identification in a Fisheye Camera Network with Different Viewing Directions." *PFG*. doi:10.1007/s41064-019-00083-y.
Blott G., Takami M., and Heipke C. 2018. "Semantic Segmentation of Fisheye Images." Computer Vision – ECCV 2018 Workshops Part I – 6th Workshop on Computer Vision for Road Scene Understanding and Autonomous Driving, Springer LNCS 11,129, Cham, 181–196.
Chen L., Rottensteiner F., and Heipke C. 2016. "Invariant Descriptor Learning Using a Siamese Convolutional Neural Network." ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences III-3, Prague, Czech Republic, July 12–19.
Clermont D., Kruse C., Rottensteiner F., and Heipke C. 2019. "Supervised Detection of Bomb Craters in Historical Aerial Images Using Convolutional Neural Networks." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W16: 67–74. doi:10.5194/isprs-archives-XLII-2-W16-67-2019.
Coenen M., Rottensteiner F., and Heipke C. 2019. "Precise Vehicle Reconstruction for Autonomous Driving Applications." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2/W5: 21–28. doi:10.5194/isprs-annals-IV-2-W5-21-2019.
Dorozynski M., Clermont D., and Rottensteiner F., 2019. "Multi-task Deep Learning with Incomplete Training Samples for the Image-based Prediction of Variables

Describing Silk Fabrics." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2/W6: 47–54.

i.c.sens. 2019. Accessed 20 November 2019. https://www.icsens.uni-hannover.de/start.html?&L=1

Kang J., Chen L., Deng F., and Heipke C. 2019. "Context Pyramidal Network for Stereo Matching Regularized by Disparity Gradients." *ISPRS Journal of Photogrammetry and Remote Sensing* 157 (2019): 201–215.

Mehltretter M., and Heipke C. 2019. "CNN-based Cost Volume Analysis as Confidence Measure for Dense Matching." ICCV Workshop on 3D Reconstruction in the Wild (3DRW2019). http://openaccess.thecvf.com/content_ICCVW_2019/papers/3DRW/Mehltretter_CNN-Based_Cost_Volume_Analysis_as_Confidence_Measure_for_Dense_Matching_ICCVW_2019_paper.pdf

Nguyen U., Rottensteiner F., and Heipke C. 2019. "Confidence-aware Pedestrian Tracking Using a Stereo Camera." ISPRS Annals IV-2/W5, Enschede, The Netherlands, June 10–14, 53–60.

SILKNOW. 2019. Accessed 20 November 2019. http://silknow.eu/

Wittich D., and Rottensteiner F. 2019. "Adversarial Domain Adaptation for the Classification of Aerial Images and Height Data Using Convolutional Neural Networks." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2/W7: 197–204.

Yang C., Rottensteiner F., and Heipke C. 2018. "Classification of Land Cover and Land Use Based on Convolutional Neural Networks." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-3: 251–258. doi:10.5194/isprs-annals-IV-3-251-2018.

Yang C., Rottensteiner F., and Heipke C. 2019. "Classification of Land Cover and Land Use Based on Convolutional Neural Networks." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* III-3: 251–258.