

**A MACHINE VISION CASE STUDY OF U-NET NETWORKS FOR SUPERFICIAL
CORROSION AND DIRT IMAGE SEGMENTATION ON INDUSTRIAL COATED STEEL
STRUCTURES**

Wesley Kaizer,
Vidya Technology
Av. Comendador Franco, 1341
Curitiba, Paraná
Brazil

Otavio Correa
Vidya Technology
Av. Comendador Franco, 1341
Curitiba, Paraná
Brazil

ABSTRACT

Corrosion detection in industrial assets and components is an important broad problem in the industries since it allows the temporal tracking of possible issues and the execution of preventive maintenance actions, such as protective coating. However, solving this problem using modern machine learning methods usually demands a careful design of artificial intelligence tools, such as neural networks, high computational resources for training and inference, and a large and adequate dataset. In this work we investigate the application of deep convolutional neural networks to the problem of image semantic segmentation of superficial corrosion and dirt present in mining industrial assets, using a set of images collected in place by corrosion inspectors and manually labeled by a data team. We compare two networks based on the popular U-Net model, in which one of them uses the transferred features from a pre-trained VGG-16 image classification model. Our main contribution is to provide insights about the application of deep neural networks in this particular domain, mostly regarding the size and quality of the constructed dataset, the existing computational resources constraints, and the observed benefits of using a pre-trained model, as well as discuss some preliminary segmentation results obtained. Our main results show the practical usefulness of transfer learning approaches, which presents significantly better results in our dataset, and also highlight the importance of constructing a dataset with appropriate size and sample quality.

Keywords: Industrial Assets Management, Corrosion Detection, Corrosion Localization, Machine Vision, Artificial Intelligence, Deep Learning, Semantic Image Segmentation, Transfer Learning

INTRODUCTION

Convolutional deep neural networks are one of the main machine learning techniques applied to computer vision and object recognition tasks. Currently, they are very popular due to their proven effectiveness in solving image classification tasks and their significant theoretical and practical

importance to the advancement of the deep learning field. Examples of successful image classification networks developed are AlexNet, VGG, and GoogLeNet.^{1,2,3} Following the popularization of artificial intelligence accelerators that significantly improve the performance of these methods, such as Graphics Processing Units (GPUs) and more recently Tensor Processing Units (TPUs), the convolutional networks approach gained even more attention from the academic community, with numerous works published every year, as well as experienced impressive growth in their potential application in industrial real-world problems.

Semantic segmentation is a prominent and interesting fundamental problem among the wide range of existing ones covered by the deep learning and computer vision fields. It is a high-level machine learning task concerning the identification and localization of the objects present in a scene. The proposed computational methods for solving these problems usually predict one class for every pixel present in the input image, attempting to generate dense regions enclosing the objects detected. This allows the predictive model to not only provide information about which objects are present but also where they are located in the scene. This problem is a natural evolution of the image classification problem, whose objective is to predict only one class for the whole image, identifying the most distinguishable object present in it. However, it is important to note that, in the semantic segmentation problem, the objective is not to identify each distinct instance separately, that occur when multiple objects of the same class appear in the scene, or even identify parts and components of assembled or composite objects, as in the instance segmentation or part-based segmentation problems, respectively. Due to its generality and therefore high applicability, the semantic segmentation problem comprises an increasing number of real-world tasks that generally aim to automatically extract knowledge from images, such as disease detection, autonomous driving, computational photography, intelligent systems assisted by augmented reality, and many others.⁴

Nowadays, Fully Convolutional Networks (FNCs) are widely researched and applied to solve semantic segmentation tasks.⁵ These networks emerged from the idea of adapting existing convolutional neural networks used for image classification to also perform semantic segmentation. Essentially, they use the hierarchy of learned features from the convolutional layers as encoders and replace the fully connected layers at the end of the net with fully convolutional ones. In traditional convolutional layers, features are learned by using a sliding window technique to transverse the input images and extract useful visual information. Also, this process occurs repeatedly in the net, forming a hierarchy of low-level features, that are closer to the input images, and high-level features, that are constructed from the low-level ones. This hierarchy of features transferred from the original convolutional layers can be fine-tuned to the particular semantic segmentation task later. These networks gained attention in the deep learning community due to their faster inference time, lower number of learnable parameters, the ability to generate segmentation masks from input images with any resolution, and introducing the concept of skip connections to combine low and high-level information.⁶

There are several successful deep networks based on FNCs specialized for solving semantic segmentation. One of the most prominent and relevant to this work is the U-Net model, which was designed to work with very small training datasets yet providing precise segmentations.⁷ This network won the Cell Tracking Challenge at ISBI 2015 using a training dataset containing only around 30 image samples and applying aggressive techniques of data augmentation relevant to the medical domain. It has gained this name due to its architectural shape, and is divided into the so-called contracting and expansive paths. The contracting path successively downsamples the input image while doubling the number of features learned. The expansive path performs in a reverse way, upsampling the feature maps but halving the number of features learned. The network also contains skip connections from the contracting path to the expansive path, allowing the combination of low and high-level feature maps. U-Net is also a very popular architecture applied to real-world problems due to its fast inference: the authors showed that the net is able to predict segmentation masks for 512x512 images in less than one second with the support of a GPU.

Generally, in practical applications, the weights from deep neural networks are not trained from scratch, since a dataset of sufficient size is not always available and the learning process can take too much time to converge. However, it is difficult to obtain good and large datasets for semantic segmentation tasks, mainly due to the costs related to collecting and pixel-wise labeling images. Thus, usually, it is more appropriate to use a set of pre-trained weights from another network and continue the training process

by fine-tuning the weights to the particular dataset available. The weights transference between networks can be beneficial even if their image domains are very distinct from each other since using low and high-level features extracted from one domain to initialize the weights for another domain is generally better than using random initialized ones. The application of this transfer learning technique depends on several architectural decisions that are taken during the network project. However, it is not always possible to directly use a network as a building block to another due to the disposition of layers and the general network configuration. A commonly applied approach is to extend the base transfer learning model to the new task, choosing a subset of the original layers to fine-tune later in the new domain by training with very small learning rates.

In this work, we present our insights and preliminary experimental results on the application of two FNC models in the context of superficial corrosion and dirt semantic segmentation in industrial assets. The deep networks are based on the U-Net architecture and one of them uses the convolutional blocks from a pre-trained VGG-16 image classification model as a features encoder. This encoder network was previously trained using the well-known ImageNet dataset, which currently contains over a million of high-quality labeled images from several distinct object classes.⁸ Our experimental results show that using the transferred features extracted from the ImageNet dataset considerably improves the model performance to detect corrosion and dirt patterns, even if the two data domains in which the models were trained are very distinct from each other. We also provide insights into the difficulty to apply these methods in real industrial problems, especially regarding the dataset construction and the required computational resources to train and predict segmentation masks using deep learning models. The remainder of the paper is organized as follows: in Section 2 we present relevant related work from the literature. In Section 3 it is described in detail the deep networks and other relevant implementation and training aspects. A comparison between the models, the obtained evaluation metrics, and the segmentation results are presented in Section 4. Section 5 concludes the paper.

EXPERIMENTAL PROCEDURE

In the last few years, several works investigated the use of deep convolutional neural networks to the automatic detection and localization of defects present in industrial structures and components. In ⁹, the authors developed a multiscale deep neural network, intended to perform damage localization and classification, being able to differentiate between six types of common defects encountered in civil structures. This work addresses the lack of contextual information that was common in other deep learning approaches at the time, providing the algorithm with a larger context around the areas of damage occurrence. The proposed pipeline is quite distinct since there are two networks called damage classifier and damage segmenter that are applied in a parallel way to obtain the final segmentation results. Despite the unusual machinery, the reported experimental results using images of a variety of structures show high pixel-wise accuracies on the test dataset.

In ¹⁰, the authors compared three semantic image segmentation models: FNC, U-Net, and Mask R-CNN ¹¹, to perform general corrosion identification in metal structures, considering that these models require a smaller training dataset and perform predictions efficiently. However, despite these characteristics, the authors noted that the final predictions given by the models generally were not sufficiently good to perform the necessary structural analyses. To mitigate this problem, they proposed a post-prediction refinement method based on morphological operators that improves the final model results and presents more accurate predictions in general but exhibiting a tendency of over-segmenting regions of corrosion. Besides that, their experimental results using the European H2020 Panoptics project images show that the combined use of the refinement procedure yielded better performance than traditional deep networks alone.

Lastly, another important open question in the semantic segmentation field lies in the construction of image datasets of sufficient size and quality, aiming to achieve some target model predictive performance when integrated into production systems or reduce collection and labeling efforts and costs. In ¹², the authors addressed this problem by evaluating the performance of the same FCN model trained using a large but poorly labeled dataset against a dataset with high-quality labels but a limited number of samples. The reported experimental results show that the model trained using the larger dataset outperforms the

model trained using the high-quality label images. In ¹³, the authors implemented a simplified U-Net model to detect corrosion in dams, penstocks, and locks using a set of images automatically collected by a Micro Aerial Vehicle (MAV). The set of images used in their study were captured from very difficult access areas, being very noisy and presenting several lighting issues such as high exposure on reflective areas and low exposure under non-reflective areas. However, despite the fact that the model was trained with less than 40 noisy images, the study reports good overall segmentation results and highlights the model's ability to perform predictions in a production environment at a rate of 12 fps with the support of a single GPU.

Model Architecture

The use of layers from pre-trained models as building blocks to define new deep neural networks is a very commonly applied approach in practice since their layers and weights provide a quick start to the new model architecture design and usually improve the training convergence time and the final results. Visual Geometry Group (VGG) is a successful convolutional neural network for the image classification task that was submitted to the ImageNet Large Scale Visual Recognition Challenge (ILSVRC-2013). Due to its high 92.7% top-5 test performance in the competition and the hierarchical organization of the convolutional layers, the VGG-16 network became one of the most used models to perform transfer learning and define new architectures based on FNCs to perform semantic image segmentation.

Our base model architecture is heavily based on the original U-Net model, depicted in Figure 1. It is kept the original number of feature maps, the kernel sizes used, and the general organization of components, according to the described in the paper. To be able to perform comparisons with the original U-Net, we defined a second deep learning model that uses a subset of the convolutional layers from a VGG-16 network trained in the ImageNet dataset as a features encoder. In this network, we replaced the five groups of two sequential convolutional layers existent in the contracting path, before the max-pooling layers, with the convolutional layers at the end of each block in the encoder part of VGG-16. After each reused convolutional layer, we appended a ReLu activation and a batch normalization layer ¹⁵. Effectively in the implementation, we use the last convolutional layer of each one of the five blocks, being these layers originally named block1_conv2, block2_conv2, block3_conv3, block4_conv3, and block5_conv3 in the available TensorFlow implementation of VGG-16. These layers have similar dimensions to the original U-Net layers, facilitating their direct reuse to define new models using transfer learning. The expansive path organization is the same as the original U-Net model and the skip connections between the contracting and expansive paths were also kept as is. Note that, in the transfer learning model, the skip connections copy the features from the selected convolutional layers of the VGG-16 model to the expansive path. The pre-trained weights reused from the convolutional layers of VGG-16 are frozen during the training process of the second model, aiming to preserve the hierarchy of features extracted from the ImageNet dataset.