TRANSFER LEARNING WITH SAS-IMAGE CONVOLUTIONAL NEURAL NETWORKS FOR IMPROVED UNDERWATER TARGET CLASSIFICATION

David P. Williams

NATO STO Centre for Maritime Research and Experimentation (CMRE)
La Spezia, Italy

ABSTRACT

The value of transferring convolutional neural networks (CNNs) trained with synthetic aperture sonar (SAS) imagery is demonstrated in the context of an underwater unexploded ordnance (UXO) classification task. Specifically, it is shown that CNNs designed for, and trained on, a mine classification task can be transferred across sensors of the same modality – but different frequency bands and sensor resolutions – and also across target concept (from mines to UXO). Importantly, it is shown that this transfer learning outperforms simply training the CNNs "from scratch" using the limited available data that pertains to the ultimate task. A key element underlying this approach is that the CNNs be specially tailored to the particularities of the sensor modality and its data. These findings are valuable because they illustrate how training-data requirements can be eased for data-limited remote-sensing applications.

Index Terms— Transfer learning, synthetic aperture sonar (SAS), convolutional neural networks (CNNs), classification, unexploded ordnance (UXO)

1. INTRODUCTION

In remote-sensing applications, data collections can be time-consuming and prohibitively expensive. Moreover, when a new sensor is introduced, there is a desire to be able to still leverage historical data collected by similar predecessor systems. In general, it is not feasible to wait for the execution of numerous onerous data collections before being able to accurately assess the utility of the new system. For these reasons, the idea of transfer learning is particularly appealing.

Convolutional neural networks (CNNs) [1] are quickly becoming the preferred image-classification method for any domain characterized by vast amounts of labeled data. And even in remote-sensing applications, where the sensor imagery available is typically limited, CNNs are also increasingly being employed. In the context of CNNs, transfer learning entails taking a CNN trained for one task and re-using it for a related but distinct task. The basic idea exploited in this transfer is that CNN filters that capture lower-level features, such as edges and corners, are useful in a wide range of object-recognition tasks. As a result, one can intelligently initialize (early layers of) a CNN with weights learned for a different task, and then refine with a modest amount of data related to the new task at hand. This approach greatly reduces training time, but also lightens training-data requirements.

This form of transfer learning is now standard for tasks involving optical imagery (*i.e.,* photographs), such that work on a new task usually begins with the architecture and pre-trained weights of a CNN from some famous "named" family, such as VGG-Net [2], Inception [3], or ResNet [4]. But the fact that these "named" nets were designed for, and trained on, optical images creates a dilemma for remote-sensing tasks that employ data of a fundamentally different sensor modality.

The aforementioned "named" CNNs were designed for the ImageNet challenge [5], a thousand-class classification problem involving three-channel (RGB) optical images. In contrast, remote-sensing applications typically address a binary classification task (*e.g.,* discriminating targets from clutter) or consider only a very small number of classes. Moreover, the imaging phenomena is fundamentally different. A prominent characteristic of SAS imagery is speckle, which manifests due to the presence of multiple scatterers within a resolution cell; optical photographs do not exhibit this property. Furthermore, the input imagery for optical data, which usually comprises three color channels, differs from remote-sensing imagery. For example, SAS imagery is single-channel (*i.e.,* gray-scale) but complex-valued, while other modalities feature multiple channels of data owing to multi-band sensors, such as hyperspectral or dual-band SAS systems.

Despite these differences, it has been shown that the inter-modality transfer, to sonar imagery, of CNNs trained on optical imagery can perform surprisingly well. But the use of these huge networks for relatively simple binary classification tasks is akin to using a high-order polynomial to fit samples generated by a quadratic function. The enormous capacities of the CNNs effectively mask the data differences in a brute-force manner. In fact, recent studies involving SAS imagery have shown that instead training the networks "from scratch" achieves comparable performance [6–8]. Moreover, it has also been shown that significantly smaller networks, with many orders of magnitude fewer free parameters, can perform just as well [9].

Motivated by these findings, we attempt to show the feasibility of *intra*-modality transfer learning when the imaging modality is SAS, which has not been demonstrated before. To do so, we argue that a key element underlying this approach is that the CNNs should be tailored to the particularities of

the sensor modality and its data. The main contribution of this work is showing that CNNs designed and trained for a mine-classification task using SAS imagery can be transferred across sensors of the same modality – but different frequency bands and sensor resolutions – and also across target concept (from mines to UXO), and importantly that this outperforms simply training the CNNs from scratch. This insight is valuable because it illustrates how training-data requirements can be eased for data-limited remote-sensing applications.

The remainder of this paper is organized as follows. The real, measured SAS data used in this work is described in Sec. 2, while the CNN design and training procedure is outlined in Sec. 3. Experimental results of the proposed transfer learning are shown in Sec. 4, before concluding remarks are made in Sec. 5.

2. SAS DATA

SAS [10] relies on the coherent processing of acoustic returns to produce high-resolution imagery of underwater environments that can be exploited for object classification and other tasks. We intentionally present the SAS data involved in this study before discussing CNNs because we believe that the sensor data should play a key role in informing CNN design.

The ultimate objective in this work is to discriminate UXO from clutter, but the challenge is that very limited training data – and specifically target views – are available. For this reason, we resort to the idea of intra-modality transfer learning with SAS data. Specifically, we make use of two databases of imagery, one collected by CMRE's MUSCLE autonomous underwater vehicle (AUV) and the other collected by the SeaOtter Mk II AUV.

A relatively large database of 2234 scene-level SAS images (each of which typically spans $50\text{ m} \times 110\text{ m}$ of seafloor) was collected by the MUSCLE AUV during eight sea expeditions conducted between 2007 and 2013 in various geographical locations around the Mediterranean Sea and in Latvian waters of the Baltic Sea. The center frequency of the SAS is 300 kHz and the bandwidth is 60 kHz. The imagery has an along-track resolution of 2.5 cm and a range resolution of 1.5 cm. The target class corresponds to man-made objects mimicking mine shapes that were purposely deployed prior to the surveys. The clutter class comprises rocks, seafloor anomalies, other man-made objects, and all other alarms.

A smaller database of 476 scene-level SAS images (each of which spans $50\text{ m} \times 110\text{ m}$ of seafloor) was collected by the SeaOtter Mk II AUV during a sea expedition conducted in September 2016 in German waters of the Baltic Sea. The center frequency of the SAS is 150 kHz and the bandwidth is 30 kHz. The imagery has an along-track resolution of 2.0 cm and a range resolution of 2.62 cm. The target class corresponds to real, historical UXO present from World War II. Thus, the SAS data in the two databases differ in several notable ways, including the target classes, the types of clutter, seafloor characteristics, and image resolution.

The Mondrian detection algorithm [11] was applied to the complex scene-level sonar images from each database, with this resulting in sets of image "chips" (each approximately $5\text{ m} \times 5\text{ m}$) of objects to be classified as targets (class 1) or clutter (class 0) by the CNNs. To form relatively balanced training and test data sets from the SeaOtter data, the alarms generated from the port sonar were treated as training data, while the alarms generated from the starboard sonar were treated as test data. Details of the data sets are shown in Table 1.

Table 1. Details of the data sets

Sensor Platform	Data Set	Target Class	Number of Targets	Number of Clutter
MUSCLE	–	Mines	2912	29280
SeaOtter	Training	UXO	65	13746
SeaOtter	Test	UXO	74	15324

To create consistent input data for the CNNs, all of the image chips were re-sampled (via bilinear interpolation) at a resolution of $1.5\text{ cm} \times 1.5\text{ cm}$. The pixel-values of each chip, $\mathbf{X}$, were also normalized (*i.e.,* rescaled) from $[0, 40]$ to $[-1, 1]$, as $\mathbf{X}' = (\mathbf{X}/20) - 1$.

3. CNN WITH SAS DATA

A standard CNN [1] is a sequence of convolutional layers, nonlinear activation functions, and pooling operations that collectively transform input data (*i.e.,* imagery) into a new representation space in which the classes are easily separable.

This work demonstrates the feasibility of using CNNs for UXO classification when faced with limited amounts of SAS training data. This is achieved by recognizing that a crucial quantity in determining the success of CNNs for classification tasks is not the amount of training data *per se*, but rather the relationship between training data and network capacity. As the capacity of a CNN – loosely speaking, the number of free trainable parameters in the model – grows, so too do the training data requirements. Therefore, when faced with extremely limited training data, it is imperative to constrain the CNN's capacity by employing small networks.

Each input image to the CNNs is assumed to be 267 pixels $\times$ 267 pixels. We design four CNN architectures, each of which consists of a sequence of 4 alternating convolution *block* layers and pooling layers, followed by a single dense layer. The filter sizes and pooling factors are provided in Table 2.

Each CNN contains 4 convolution blocks; each block contains a specific number of convolutional layers (equal to the number of rows in Table 2's filter-size column). Thus, the four CNNs contain 4, 4, 8, and 12 convolutional layers, respectively. Each filter is square, and only 4 filters are used in each convolutional layer. Rectified linear unit (ReLU) activations are used after each convolutional layer, while a sigmoid

Table 2. CNN architecture details

CNN	Filter Sizes (Pixels Per Side)	Pooling Factors	Number of Parameters
A	[4] [3] [3] [4]	4, 4, 2, 4	629
B	[8] [6] [4] [5]	4, 4, 2, 2	1509
C	$\begin{bmatrix}6\\3\end{bmatrix}$ $\begin{bmatrix}6\\3\end{bmatrix}$ $\begin{bmatrix}6\\3\end{bmatrix}$ $\begin{bmatrix}6\\3\end{bmatrix}$	4, 2, 2, 4	2485
D	$\begin{bmatrix}8\\7\\5\end{bmatrix}$ $\begin{bmatrix}8\\7\\5\end{bmatrix}$ $\begin{bmatrix}7\\7\\5\end{bmatrix}$ $\begin{bmatrix}8\\7\\5\end{bmatrix}$	2, 2, 2, 2	7877

activation is used at the output. All pooling layers use average pooling (rather than max-pooling) because the former approach has been observed [6] to better handle the speckle phenomenon that characterizes sonar imagery. The design of the architecture (and specifically the final pooling layer) ensures that the dense layer always contains 4 nodes.

As can be seen in Table 2, the number of free trainable parameters in each CNN we design is tiny compared to those of popular "named" CNNs designed for optical images – which rely on upwards of 10^7 parameters – and also much lower than our previous SAS-based CNNs [12]. The capacities of the CNNs are intentionally kept so low in order to scrutinize the feasibility of intra-modality transfer learning when the data available is extremely limited, as it is for the UXO task.

CNN training was performed using the RMSprop optimizer with a learning rate of $\eta = 0.001$, in conjunction with a binary-cross-entropy loss function, until the loss on a small validation set converged. A batch size of $b = 64$ was used, with equal numbers selected from each class to combat the severe class-imbalance of the training data. Data augmentation that respected the inviolable geometry of the sonar data-collection procedure was employed during training; this meant a random range translation $i_{tx} \in [0, 0.50\text{ m}]$, along-track translation $i_{ty} \in [-0.25\text{ m}, 0.25\text{ m}]$, and along-track reflection $i_{ry} \in \{1, -1\}$ was applied to each sonar image chip selected for the batch. No attempt was made to optimize the learning rate or batch size.

4. EXPERIMENTAL RESULTS

To investigate the feasibility and value of transfer learning with SAS data, three sets of CNNs were trained, distinguished by the data used for training. First, the 4 CNNs designed in Sec. 3 were trained using only data from the SeaOtter training set (*i.e.,* "from scratch"). Next, the 4 CNNs were trained using only data from the MUSCLE data set; this corresponds to the "No Refinement" case. Lastly, the MUSCLE-trained CNNs were refined using data from the SeaOtter training set; this corresponds to the "With Refinement" case, and represents the transfer-learning scenario. It should be emphasized that the targets present in the MUSCLE and SeaOtter data sets are different classes of objects, namely mines and UXO, respectively.

The classification performance, measured in terms of the area under the curve (AUC), on the SeaOtter test data for the CNNs of each of these three cases is shown in Table 3. The performance of the ensemble prediction from the four CNNs, denoted $\mathcal{E}(\text{ABCD})$, is also shown. For reference, the performance of the Mondrian detector [11], a baseline "shallow" classification approach that makes predictions using a set of 5 features with fixed weights, achieves an AUC of 0.803.

Table 3. AUC on SeaOtter test data with and without transfer

CNN	Trained on SeaOtter Only	Trained on MUSCLE	
		No Refinement	With Refinement
A	0.744	0.752	**0.790**
B	0.839	0.742	**0.851**
C	0.801	0.729	**0.827**
D	0.818	0.701	**0.823**
$\mathcal{E}(\text{ABCD})$	0.842	0.753	**0.866**

As can be seen from the table, the transfer-learning scenario in which the related MUSCLE data is leveraged consistently achieves the best performance across the different CNN architectures considered. That is, despite the sensor and target-class differences, exploiting the related data enabled better performance. This can likely be attributed to the fact that the intermediate representations learned by the CNNs trained on MUSCLE mine data were also useful for the UXO task. The necessity of refinement, given the target-class differences, is also important to note.

From Table 3, it can be observed that there exists an inverse relationship between CNN capacity and performance *without* refinement. This trend suggests that larger nets become more tailored to the (MUSCLE) training data, and hence less *directly* applicable to a new task (*i.e.,* without refinement). Consequently, larger nets will require more data belonging to the new task. Given the modest (UXO) training data available here, it is likely that a large net would not be able to generalize well via transfer learning. It should be noted that the transfer of huge optical-image-trained CNNs to mine classification tasks in SAS images that was considered in [6–8] had thousands of target views (prior to data augmentation) – *i.e.,* sufficient data in the new domain to support significant refinement.

To illustrate how the CNN filters differ depending on whether transfer learning is employed, the filters of the first convolutional layer for two of the CNN architectures are shown in Fig. 1. From the figure, it can be observed that drastically different filters are learned in the two scenarios. We posit that using the MUSCLE mine examples, in the transfer learning scenario, effectively acts as a regularizer that prevents the CNNs from overfitting on the small amount of UXO examples in the (SeaOtter) training set.

To further examine how the CNNs with and without transfer learning differ, Fig. 2 shows the intermediate representations at each layer of CNN C when a typical target is the input image. It is readily evident that the two approaches key on different characteristics in the image to generate predictions.

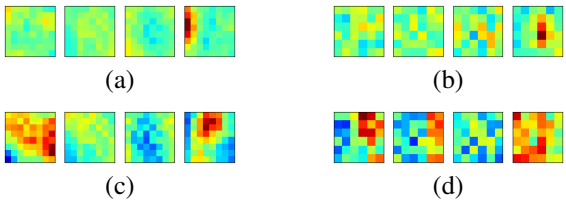

Fig. 1. When training on MUSCLE data and then refining with SeaOtter training data, the first convolutional layer's filters of (a) CNN B and (b) CNN C. When training on only SeaOtter training data, the first convolutional layer's filters of (c) CNN B and (d) CNN C. The filters of a given subfigure use the same colorscale; green corresponds to zero.

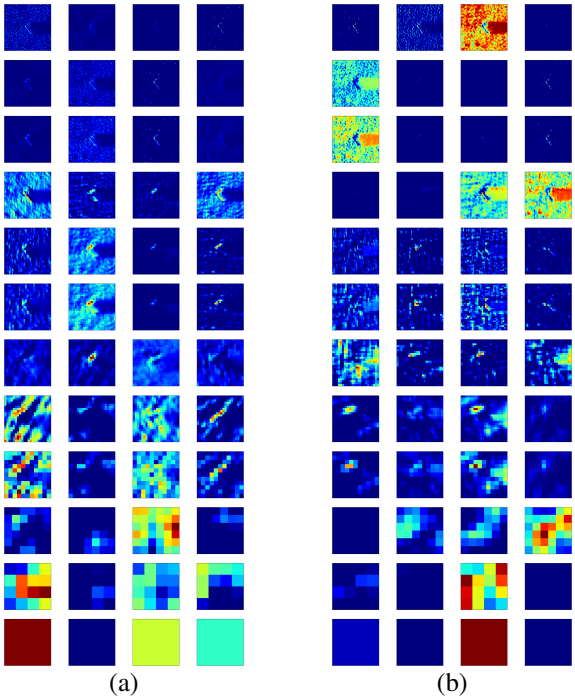

Fig. 2. Intermediate layer responses of CNN C for an example target when (a) the CNN is refined after transfer learning and (b) the CNN is trained from scratch.

5. CONCLUSION

In this work, several new CNNs were specially designed for remote-sensing classification tasks. These much smaller nets, tailored to the SAS sensor modality under study, allowed deep-learning approaches to be used even with very limited training data. It was demonstrated that intra-modality transfer learning enabled superior classification performance compared to training from scratch with only training data related to the final task. These findings can be leveraged in the future to reduce training-data requirements in a wide range of remote-sensing tasks characterized by limited data.

Future work will examine the relative benefit of transfer as a function of training data set size; undertaking this study, however, requires a larger data set than the one currently available. The limits of intra-modality transfer vis-à-vis frequency band differences (*e.g.,* transferring high-frequency SAS-image CNNs to low-frequency SAS imagery) will also be explored.

Acknowledgments

The author thanks Holger Schmaljohann from WTD-71 for providing the SeaOtter data, and Isaac Gerg from PSU-ARL for assisting with the manual labeling of that data. This work was partially supported by the Strategic Environmental Research and Development Program (SERDP) and the NATO Allied Command Transformation (ACT).

6. REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[6] J. Chen, J. Summers, and J. Trader, "Interpretable semi-supervised deep learning with synthetic aperture sonar for automatic target recognition," in *SAS/SAR*, vol. 40, 2018, pp. 132–139.

[7] M. Emigh, B. Marchand, M. Cook, and J. Prater, "Supervised deep learning classification for multi-band synthetic aperture sonar," in *SAS/SAR*, vol. 40, 2018, pp. 140–147.

[8] N. Warakagoda and Ø. Midtgaard, "Transfer-learning with deep neural networks for mine recognition in sonar images," in *SAS/SAR*, vol. 40, 2018, pp. 115–122.

[9] I. Gerg and D. Williams, "Additional representations for improving synthetic aperture sonar classification using convolutional neural networks," in *SAS/SAR*, vol. 40, 2018, pp. 11–22.

[10] M. Hayes and P. Gough, "Broad-band synthetic aperture sonar," *IEEE Journal of Oceanic Engineering*, vol. 17, no. 1, pp. 80–94, 1992.

[11] D. Williams, "The Mondrian detection algorithm for sonar imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1091–1102, 2018.

[12] ——, "Demystifying deep convolutional neural networks for sonar image classification," in *Proceedings of the Underwater Acoustics Conference*, 2017.