# Efficient Neural Network Compression via Transfer Learning for Industrial Optical Inspection

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

In this paper, we investigate learning the deep neural networks for automated optical inspection in industrial manufacturing. Our preliminary result has shown the stunning performance improvement by transfer learning from the completely dissimilar source domain: ImageNet. Further study for demystifying this improvement shows that the transfer learning produces a highly compressible network, which was not the case for the network learned from scratch. The experimental result shows that there is a negligible accuracy drop in the network learned by transfer learning until it is compressed to 1/128 reduction of the number of convolution filters. This result is contrary to the compression without transfer learning which loses more than 5% accuracy at the same compression rate.

## 1 Introduction

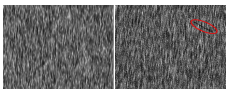
Since every manufactured product must pass the inspection process of detecting defects on the product surface, the fast and accurate inspection algorithm is essential in the manufacturing industry. Contrary to the success in natural image detection using sufficient data [4, 7], the available images of product surfaces, in particular those containing the patterns of defects, are insufficient for the reliable training of the deep structured network. While training the network from scratch only produces 81.07% performance, we previously reported the 99.90% performance when the whole network is fine-tuned from a successfully learned network for imageNet classification [2]. The result is non-trivial because the images in the ImageNet source domain do not overlap with the visual inspection images in the target domain as shown in Figure 1.

In this paper, we investigate the compression of the learned network for fast inference. Once the network is learned by transfer learning, the activation of the neurons is very sparse. Different from the result in [2], the network trained from scratch can also achieve 99.78% accuracy with the extensively augmented data and a long period of training. Surprisingly, however, although the performance of both networks is similar, the network trained from scratch learns much denser features of the input data than the network trained by transfer learning. We experimentally show that using standard teacher-student model compression technique [3], the network trained by transfer learning can be compressed to 1/128 reduction of the number of convolution filters with a negligible accuracy drop. In contrast, the network trained from scratch loses more than 5% accuracy at the same compression rate. Previous works on the network compression have focused on the methods of compression [3, 5, 6, 9], but our work is the first report to the best of our knowledge that the *transfer learning* is related to the network compression.

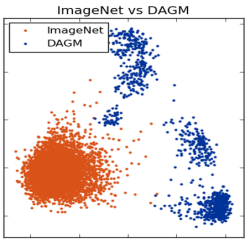
The rest of the paper is organized as follows. In Section 2, we explain the method of experiments and show the compression result in Section 2.3. Finally, we conclude with a brief discussion in Section 3.

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

(a) Source domain (ImageNet)



(b) Target domain (DAGM)




(c) PCA embedding

Figure 1: Sample images of the source and target domains, and their embedding result using PCA. Both domains are clearly dissimilar each other.

Table 1: Teacher network result

| Method | Accuracy | Convergence time (SGD iterations) |
|---|---|---|
| Scratch (no aug) | 81.07% | 35,880 |
| Scratch (aug) | 99.78% | 165,600 |
| TL (no aug) | **99.90%** | **2242** |

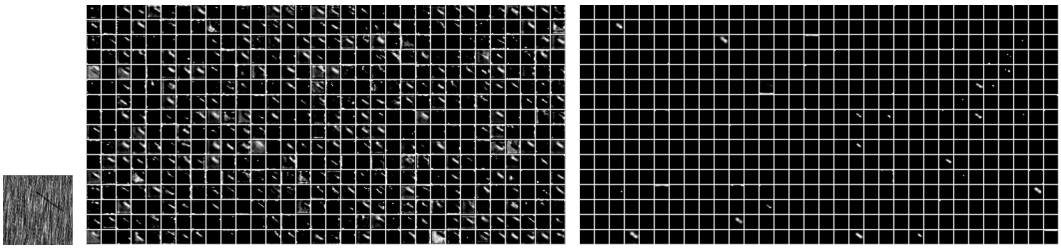


Figure 2: Output activations of the last convolutional layer of the Scratch (middle) and TL (right) teacher networks for a given input image (left).

## 2 Methods and experimental results

Following the neural network compression framework of [3], we first train the high performance deep teacher network and train a shallow student network to mimic it. To summarize the main results, the teacher network trained by transfer learning can be compressed to the shallow student network with 1/156 number of parameters and achieves 40× faster inference speed while showing less than 1% accuracy drop (Table 2). Details of the experiments are explained below.

### 2.1 Target problem

The target problem is a texture inspection problem that the dataset is publicly accessible in the DAGM [1]. The dataset contains six patterns of texture with each pattern containing 1000 non-defective and 150 defective images, resulting in 6900 images of 12 classes (Figure 1b). The entire dataset is randomly divided into 80% for training (5520 images) and 20% for evaluation (1380 images).

### 2.2 Teacher networks: training from scratch vs transfer learning

We use the VGG16 [8] as a teacher network. We train the teacher network in two different ways with the same optimizer setting (SGD with batch size 32 and learning rate $10^{-3}$). In the first case, we randomly initialize all weights in the network and train it directly on the target problem from scratch. We denote this deep teacher network as a *Scratch*. Since the number of training data is insufficient, the accuracy of the network no longer increases at 81.07% without any data augmentation. On the other hand, if we augment the training data by adding rotated and flipped images, the network can be successfully trained to achieve 99.78% accuracy after 165k iterations of training (Table 1).

For the second case of the teacher network, we apply transfer learning to train the network. We denote this teacher network as a *TL*. The source domain of the transfer learning is the ImageNet data [7] (Figure 1a) where the images bear almost no similarity to the target inspection data (Figure 1b). We can clearly see the dissimilarity between each domain by embedding features extracted from both datasets using principal component analysis (PCA) (Figure 1c). We follow the transfer learning method of [2] that weights of the target network initialized with the weights of the source network only up to convolutional layers. Then, the network is fine-tuned using target data without any data augmentation. Even though the source and target domains are completely dissimilar from each other, the network trained by transfer learning converge to 99.90% accuracy 70× faster than training from scratch (Table 1). In addition, when we visualize the activation of the last convolutional layer for

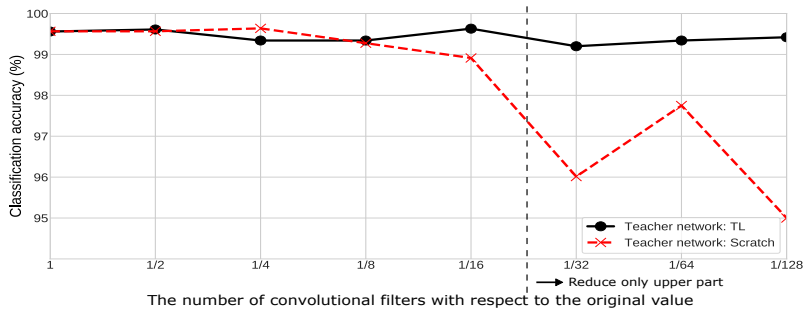



Figure 3: Neural network compression results (Scratch teacher network vs TL teacher network).

Table 2: Summary of neural network compression results

| Model | Number of parameters | Accuracy (%) | Inference speed (ms/image) |
|---|---|---|---|
| Teacher Network (TL) | 134 million | **99.90** | 57.43 |
| Student Network (TL) | **858,380** | 99.42 | **1.46** |
| Student Network (Scratch) | 858,380 | 94.92 | 1.46 |

a given input image, we can see that the TL teacher network learns much sparser features of the input data than the Scratch teacher network (Figure 2). The sparsity of activation in the TL teacher network suddenly increases sharply from the eighth convolutional layer and reaches 99.10% at the last convolutional layer.

### 2.3 Neural network compression: *Scratch* teacher network vs *TL* teacher network

The student network is trained to learn the input-output mapping function of the teacher network by regressing the logits output of the teacher network [3]. For training the student network, we use the data same as the teacher network training set with the augmented version. Focusing on accelerating the inference time, we design the shallow student network as following rules. We remove the first fully-connected layer to reduce the number of parameters in the network. In addition, we eliminate eight of the 13 convolutional layers that account for a large part of the computational cost. The number of filters in remaining convolutional layers is denoted as $N_1, N_2, N_3, N_4$, and $N_5$ respectively. Then, we gradually reduce the number of filters in all convolutional layers by half from the original value $(N_1, N_2, N_3, N_4, N_5) = (64, 128, 256, 512, 512)$. The accuracy of both teacher networks is maintained until we reduce the number of all convolutional filters to 1/16. However, when it is reduced to 1/32, a sudden accuracy drop occurs to 82% accuracy in both networks. Using the fact that the sparsity of the activation grows rapidly from the eighth convolutional layer of the TL teacher network, we further reduce only upper part of the student network $(N_4, N_5)$. The TL teacher network has a negligible accuracy drop until the upper part is compressed to 1/128. In contrast, compressing the Scratch teacher network loses more than 5% accuracy at the same compression point (Figure 3). Combining the results, the final compressed student model is $(N_1, N_2, N_3, N_4, N_5) = (4, 8, 16, 4, 4)$. The student network compressed from TL teacher network has 1/156 number of parameters and achieves 40$\times$ faster inference speed while only 0.48% accuracy drop (Table 2).

## 3 Conclusion

Through experiments using inspection data, we show that the high performance, lightweight shallow network can be obtained for optical inspection via compressing the deep network trained by transfer learning. When the performance of two networks is similar, sparse features learned in the network seems to play a key role in model compression. This result proposes that in addition to the compression method, what the deep network has learned is also important to the model compression framework.

# References


[1] Weakly supervised learning for industrial optical inspection. `https://hci.iwr.uni-heidelberg.de/node/3616`. Accessed: 2018-10-20.

[2] ———. Transfer learning for automated optical inspection. In *International Joint Conference on Neural Networks (IJCNN), 2017*, pages 2517–2524, 2017.

[3] Jimmy Ba and Rich Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014.

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

[5] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] Xiyu Yu, Tongliang Liu, Xinchao Wang, and Dacheng Tao. On compressing deep models by low rank and sparse decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7370–7379, 2017.