

Transfer Learning for Cell Nuclei Classification in Histopathology Images

Neslihan Bayramoglu and Janne Heikkilä

Center for Machine Vision and Signal Analysis, University of Oulu, Finland
{nyalcinb,jth}@ee.oulu.fi

Abstract. In histopathological image assessment, there is a high demand to obtain fast and precise quantification automatically. Such automation could be beneficial to find clinical assessment clues to produce correct diagnoses, to reduce observer variability, and to increase objectivity. Due to its success in other areas, deep learning could be the key method to obtain clinical acceptance. However, the major bottleneck is how to train a deep CNN model with a limited amount of training data. There is one important question of critical importance: Could it be possible to use transfer learning and fine-tuning in biomedical image analysis to reduce the effort of manual data labeling and still obtain a full deep representation for the target task? In this study, we address this question quantitatively by comparing the performances of transfer learning and learning from scratch for cell nuclei classification. We evaluate four different CNN architectures trained on natural images and facial images.

1 Introduction

There are two key concepts that makes neural networks powerful in various applications. First, unlike conventional machine learning techniques, deep convolutional neural networks (CNNs) extract features automatically only by using the training data. Second, deep learning methods discover image features at multiple levels (layers) which is called “feature hierarchies”. Features at each layer are computed from the previous layer representations and it was shown that features are learned gradually from low-level to high-level. Multi-level abstraction enables deep learning networks to handle very complex functions and high dimensional data.

While deep learning algorithms achieves state-of-the-art results in different machine learning applications, there are several challenges in their application in biomedical domain. First, training deep CNN requires large amount of annotated images to learn millions of parameters. Although large-scale annotated databases are available for generic object recognition task (e.g. ImageNet), it is currently lacking in biomedical domain. Annotating biomedical data requires expertise therefore it is expensive, time consuming, and subject to observer variability. Second, limited amount of training data leads “overfitting” and features can not generalize well on data. Overfitting becomes more serious when the data contain high variability in the image appearance which is usually the case in biomedical domain. Third, training deep CNNs from scratch requires high computational power, extensive memory resources, and time. Such approaches have practical limitations in biomedical field.

In generic object recognition tasks, “transfer learning” and “fine-tuning” methods are proposed to overcome these challenges [13]. Transfer learning and fine-tuning aims at reusing image representations learned from a source task and a dataset with large amount of labeled data on a second target dataset and a task [20]. It is shown to be an effective tool to overcome overfitting when the target dataset has limited amount of labeled data [13, 20]. However, transferability of deep networks across unrelated domains have been found to be limited due to data bias [13, 20, 8]. In this study, we exploit the transfer learning properties of CNNs for cell nuclei classification in histopathological images. Although, there are significant differences in image statistics between biomedical image datasets and natural images, this study evaluates whether the features learned from deep CNNs trained on generic recognition tasks could generalize to biomedical tasks with limited training data. We evaluate four different CNN architectures trained on natural images (ImageNet [6]) and facial images [7]. We compare the performances of CNN models learned from scratch and fine-tuned from pre-trained models for cell nuclei classification. We present an empirical validation that initializing the network parameters with transferred features can improve the classification performance for any model and learning from pre-trained network model requires less training time than learning from scratch.

2 Related Work and Method

2.1 Nuclei Classification in Histopathology Images

Although there has been a progress in the development of image analysis algorithms in histopathological image assessment [4, 19, 10], there is a still high demand to obtain fast and precise quantification automatically. Such techniques could be beneficial to find clinical assessment clues to produce correct diagnoses, to reduce observer variability, and to increase objectivity. Due to its success in other fields, deep learning could be the key method to obtain clinical acceptance. However, the major bottleneck is how to train a deep CNN model with a limited amount of training data. There is one important question of critical importance: Could it be possible to use transfer learning and fine-tuning in biomedical image analysis to reduce the effort of manual data labeling and still obtain a full deep representation for the target task? In this study, we address this question quantitatively by comparing the performances of transfer learning and learning from scratch for nuclei classification using unrelated source tasks and datasets from different distributions.

Cancer is still one of the top leading cause of death worldwide. In order to develop better cancer treatments, it is important to analyse tumors at cellular level to understand disease development and progression. In cancer histopathology image analysis, convolutional neural networks are used for region of interest detection, segmentation, and also for mitosis detection [2]. On the other hand, there is a relatively little work on cell nuclei classification for histopathology images. However, the analysis of nuclei types provides deeper understanding about the state of the disease[4] which has a critical importance for treatment strategies. Hand-crafted (morphological and intensity) features are often employed

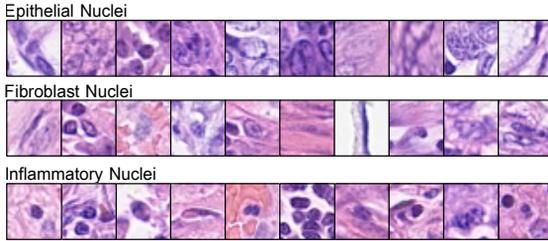


Fig. 1: Examples of different classes of cell nuclei in routine hematoxylin and eosin (H&E) stained histopathology images of colorectal adenocarcinoma. (From top to bottom) Epithelial nuclei, Fibroblasts, and Inflammatory nuclei.

for classification purposes [4, 11]. They include complex preprocessing pipeline including stain normalization, nucleus detection, and region of interest segmentation. This is mainly due to the heterogeneous structure of histopathology images. In this study, we evaluate performances of convolutional neural network models to classify cell nuclei in hematoxylin and eosin (H&E) stained histopathology images of colorectal adenocarcinoma. We used a dataset of H&E stained histology images of size 500×500 cropped from non-overlapping areas of whole slide images from 9 patients. The database (HistoPhenotypes) is published recently by Sirinukunwattana *et al.* in [16] where they also follow a CNN approach for detection and classification purposes. Example patches of different classes of cell nuclei in the dataset are shown in Figure 1.

2.2 Transfer Learning, Fine Tuning, and Full Training

Transfer learning has been explored in many problems including character recognition [3], generic object recognition [13], computer aided diagnosis of lymph node detection and interstitial lung disease classification [14], polyp detection and image quality assessment in colonoscopy videos, human epithelial type 2 cell classification in indirect immunofluorescence images [1], embolism detection [18], and segmentation [12, 18]. Earlier works in deep convolutional neural networks studied ‘learning from related tasks’ whereas recent studies follow domain adaptation by learning shallow representation models to minimize the negative effects of domain discrepancy [9]. Yosinski *et al.* [20] show that the feature transferability drops significantly when the domain discrepancy increases. They also confirm that modern deep neural networks learn *general* features on the first layers and features learned on the last layers depend greatly on the data and therefore, they are *specific*. While *shallower* networks suppress domain specific features and reduce domain discrepancy they have limited capacity to explore and learn more complex features. On the other hand, there is a high feature variability in the domain of biomedical imaging which requires *deeper* architectures to obtain full representation for the target data. In this study, we used CNN architectures with different depths and structures to explore their effects on transfer learning and fine-tuning. We investigate four different CNN architectures: AlexNet [6], GenderNet [7], GoogLeNet [17], and VGG-16 [15].

AlexNet [6] is the winner of ImageNet 2012 challenge that popularized CNNs. It contains five convolutional and pooling layers and three fully connected layers including local response normalization (ReLU) layers and dropouts. It operates on $227 \times 227 \times 3$ input images which are cropped randomly from 256×256 images.

GoogLeNet [17] architecture achieved the state-of-the-art results in ImageNet challenge in 2014. It has 22 layers with 9 inception units and finally a fully connected layer before the output. The inception module has two layers and 6 convolutional blocks which is an intrinsic component of GoogLeNet. The main contribution of this architecture is reducing the number of parameters of neural networks. Although GoogLeNet is very deep, it has $12 \times$ fewer parameters than AlexNet which makes it computationally efficient to train.

VGG-16 [15] has a similar architecture with AlexNet with more convolutional layers. It has 13 convolutional layers followed by rectification and pooling layers, and 3 fully connected layers. All convolutional layers use small 3×3 filters and the network performs only 2×2 pooling. VGG-16 has a receptive field of size 224×224 . Although VGG-16 performs better than AlexNet and has a simpler architectural design, it has $3 \times$ more parameters which requires more computation.

GenderNet [7], this small network contains three convolutional layers followed by rectified linear operation and pooling layers, and two fully connected layers. GenderNet is used for both age and gender classification from real-world, unconstrained facial images which comes from a much smaller dataset than ImageNet.

There are basically two techniques in transfer learning: fine-tuned features and frozen features [20]. When the layers are frozen and initialized from a pre-trained network model, there is no need to back-propagate through them during training and they behave as fixed features without changing on the new task. Fine-tuning involves back-propagating the errors from the new tasks into the copied layers [20]. In this study, we adopt a different strategy by choosing different learning rates for the layers coming from the source network and the new layers in the target network. We allow layers copied from the source network change slowly whereas we learn features at higher layers (last fully connected layers) with higher learning rates (fine-tuning). We propose to utilize this to train the network data specific features while tuning the well learned features from the source task without overfitting. On the other hand, full training requires learning from scratch with all the network layers initialized randomly.

3 Experiments and Results

HistoPhenotypes dataset involves 29,756 manually marked cell nuclei from 100 H&E stained images. Out of these, 22,444 nuclei are classified into four labels: epithelial, inflammatory, fibroblast, and miscellaneous. Miscellaneous category consists of mixed cell nuclei therefore, we have excluded it from this study. There are 7,772 epithelial, 5,712 fibroblast, and 6,971 inflammatory cell nuclei. In our experiments, a total number of 20,405 cell nuclei are divided randomly into training (17,004 nuclei) and testing (3,401 nuclei) set. We cropped small patches of sizes 32×32 around the cell nuclei centers which is large enough to contain

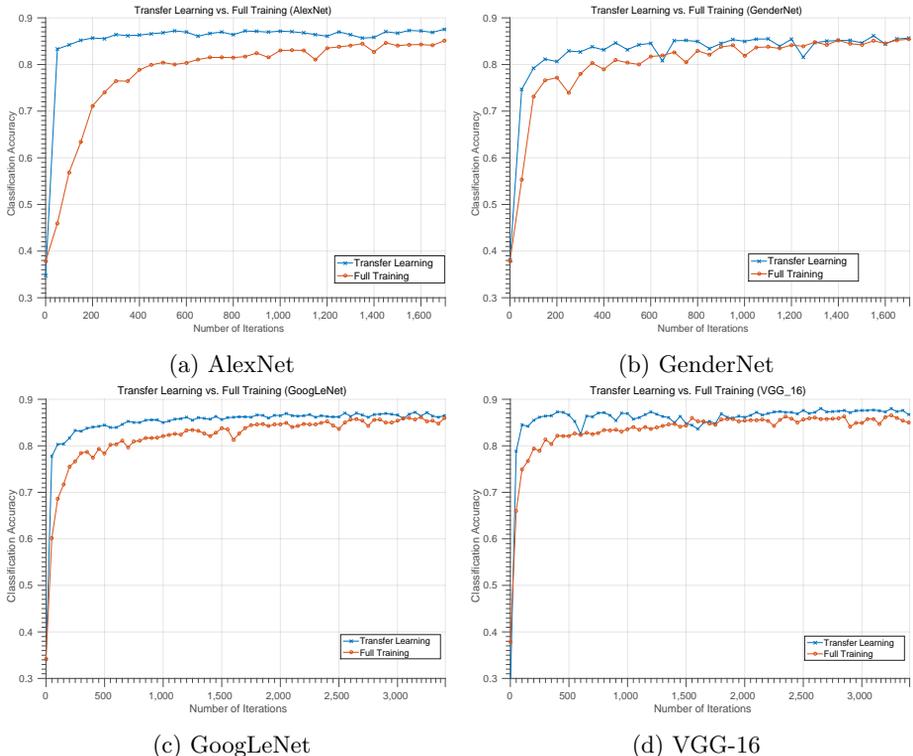


Fig. 2: Comparison of transfer learning with fine tuning and full training for networks a) AlexNet, b) GenderNer, c) GoogLeNet, and d)VGG-16.

whole nucleus. However, the network models we use in our experiments have larger receptive fields ($\sim 256 \times 256$). Therefore, we upsampled cell nuclei patches to 256×256 images. Raw images are then used without any other preprocessing or data augmentation. During training, mean intensity subtraction is employed to normalize illumination changes.

In our experiments, all networks are trained using the minibatch stochastic gradient descent with a momentum factor of 0.9. We initialize the base learning rate (lr) as 0.001 and decrease the learning rate as follows: $lr_{new} = lr_{base} \times (1 + \gamma \times iteration_number)^{power}$ with power=0.75 and $\gamma=0.001$. All network models are trained for 10 epochs either learned from scratch or fine-tuned from pre-trained models. We utilize a batch size of 100 images for AlexNet and GenderNet, whereas GoogLeNet and VGG-16 operates on a minibatch of 50 images due to memory constraints. Our implementation is based on the *Caffe* library [5].

When we learn models from scratch, the network parameters are initialized randomly either from Gaussian distributions (AlexNet and GenderNet) or with Xavier algorithm (GoogLeNet and VGG-16) which is provided in Caffe. For transfer learning, we used models pre-trained on ImageNet database for generic object recognition except GenderNet architecture which is trained on a much smaller facial dataset [7] for gender classification. For all the network architec-

tures we experimented, firstly we copied all the layers from source models to our target networks except the last fully connected layer. Then we modified the last fully connected layers for adapting models to our nuclei classification task in which the output is fed into a 3-way softmax, initialized randomly and trained from scratch.

Figure 2 shows classification accuracies of test set for AlexNet (Figure 2a), GenderNet (Figure 2b), GoogLeNet (Figure 2c), and VGG-16 (Figure 2d). For simplicity and to avoid clutter we present only test set accuracies. In each figure, we plot the performance of transfer learning and fine-tuning, and full training against the number of iterations. Because the batch sizes are smaller in GoogLeNet and VGG-16, required number of iterations are higher to train them for 10 epochs. First, we can observe from the comparisons that the transfer learning outperforms full training. In the GenderNet model, the classification accuracy of the test set for transfer learning and full training are comparable. This difference could be due to the size of the feature space of the source task and the depth of the network. Second, we also observed that fine-tuned models converges much earlier than their fully trained counterparts which concludes that transfer learning requires less training time to achieve the maximum performance. After the first epoch, the classification accuracies for fine-tuned AlexNet, GenderNet, GoogLeNet and VGG-16 are 85.68%, 80.62%, 84.03%, and 87.27% respectively. Accuracies at the same time for full training are 71.13%, 77.18%, 77.42%, and 82.18% in the same order. Deeper architectures AlexNet and VGG-16 convergence faster in the fine-tuned settings which is an indication that they can handle more complex features. A maximum of 88.03% accuracy is achieved with fine-tuned VGG-16 model.

Although there is a great difference between natural/facial images and biomedical images, transfer learning and fine tuning provides much better results than learning from scratch. We confirm that the feature transferability is affected by the depth of the network, source task, and the diversity of the source data. Experimental results are promising that the features learned from deep CNN networks trained on generic recognition tasks could generalize to biomedical tasks and they could be used to fine-tune new tasks having small datasets.

4 Conclusion

Deep learning opened a new era in the field of image analysis including the biomedical domain. Although learning parameters in deep architectures requires a lot of labeled training data which is difficult to obtain in the biomedical domain, transfer learning provides promising results in reducing the effort of manual data labeling by reusing the learned features from a different source task and data. In this study, we compared four different CNN models with depths ranging from 3 to 13 convolutional layers. Firstly, our empirical results show that initializing the network parameters with transferred features can improve the classification performance for any model. However, deeper architectures trained on bigger datasets converges quickly. Secondly, learning from pre-trained network model requires less training time than learning from scratch.

References

1. Bayramoglu, N., Kannala, J., Heikkilä, J.: Human epithelial type 2 cell classification with convolutional neural networks. In: International Conference on Bioinformatics and Bioengineering. pp. 1–6. IEEE (2015)
2. Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: MICCAI. pp. 411–418. Springer (2013)
3. Cireşan, D.C., Meier, U., Schmidhuber, J.: Transfer learning for latin and chinese characters with deep neural networks. In: The International Joint Conference on Neural Networks. pp. 1–6. IEEE (2012)
4. Irshad, H., Veillard, A., Roux, L., Racoceanu, D.: Methods for nuclei detection, segmentation, and classification in digital histopathology: a reviewcurrent status and future potential. *IEEE reviews in biomedical engineering* 7, 97–114 (2014)
5. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guaradarama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
7. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 34–42 (2015)
8. Long, M., Wang, J.: Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791 (2015)
9. Long, M., Wang, J., Jordan, M.I.: Deep transfer learning with joint adaptation networks. arXiv preprint arXiv:1605.06636 (2016)
10. McCann, M.T., Ozolek, J.A., Castro, C.A., Parvin, B., Kovacevic, J.: Automated histology analysis: Opportunities for signal processing. *IEEE Signal Processing Magazine* 32(1), 78 (2015)
11. Nguyen, K., Bredno, J., Knowles, D.A.: Using contextual information to classify nuclei in histology images. In: IEEE 12th International Symposium on Biomedical Imaging (ISBI). pp. 995–998. IEEE (2015)
12. van Opbroek, A., Ikram, M.A., Vernooij, M.W., De Bruijne, M.: Transfer learning improves supervised image segmentation across imaging protocols. *IEEE, TMI* 34(5), 1018–1030 (2015)
13. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: CVPR, IEEE. pp. 1717–1724 (2014)
14. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging* 35(5), 1285–1298 (2016)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Sirinukunwattana, K., Raza, S.E.A., Tsang, Y.W., Snead, D.R., Cree, I.A., Rajpoot, N.M.: Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging* 35(5), 1196–1206 (2016)

17. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE, CVPR. pp. 1–9 (2015)
18. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35(5), 1299–1312 (2016)
19. Veta, M., Pluim, J.P., van Diest, P.J., Viergever, M.A.: Breast cancer histopathology image analysis: A review. IEEE Transactions on Biomedical Engineering 61(5), 1400–1411 (2014)
20. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in Neural Information Processing Systems. pp. 3320–3328 (2014)