Visualizing and Interpreting Feature Reuse of Pretrained CNNs for Histopathology

Mara Graziani[1,2], Vincent Andrearczyk[1] and Henning Müller[1,2]

[1]*University of Applied Sciences of Western Switzerland, HES-SO Valais*
[2]*University of Geneva (UNIGE), Geneva, Switzerland*

Abstract

Reusing the parameters of networks pretrained on large scale datasets of natural images, such as ImageNet, is a common technique in the medical imaging domain. The large variability of objects and classes is, however, drastically reduced in most medical applications where images are dominated by repetitive patterns with, at times, subtle differences between the classes. This paper takes the example of finetuning a pretrained convolutional network on a histopathology task. Because of the reduced visual variability in this application domain, the network mostly learns to detect textures and simple patterns. As a result, the complex structures that maximize the channel activations of deep layers in the pretrained network are not present after finetuning. The learned features seem to be used by the network to spot atypical nuclei in the images, as shown by class activation maps. Finally, texture measures appear discriminative after finetuning, as shown by accurate Regression Concept Vectors.

Keywords: Medical Imaging, Deep Learning Interpretability, Activation Maximization, grad-CAM

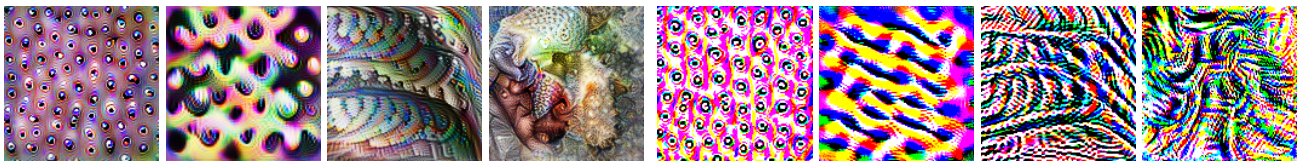
1 Introduction

Visualization techniques can improve our understanding of how concept representations are organized over the layers of deep Convolutional Neural Networks (CNNs). The concept representations change when the CNN is finetuned on the binary classification task of *tumor* and *non-tumor* breast histopathology images. We generate images that maximally activate the network channels as in [Erhan et al., 2009, Olah et al., 2017]and compare the results in pretrained and finetuned networks. We notice that finetuning reduces the complexity and abstraction of the representations learned by the pretrained networks, focusing on texture and simple repeated patterns. Gradient-weighted Class Activation Maps (grad-CAM) [Selvaraju et al., 2017, Chattopadhay et al., 2018] are used to visualize the CNN attention and further demonstrate this idea. Results suggest that the CNN focus is mostly on the atypical nuclei with morphological anomalies (nuclei pleomorphism). The recently developed Regression Concept Vectors (RCVs) quantified the relevance of nuclei pleomorphism in the classification of histopathology images [Graziani et al., 2018, Graziani et al., 2020]. Second-order Haralick descriptors of texture correlation and contrast were shown to influence the classification. In addition to the qualitative analysis of the visualizations, we expand the experiments on RCVs, suggesting that concepts of textures are inherited from the architecture itself and refined during network training and finetuning. Therefore, experimental results in this paper show that feature reuse from ImageNet pretrained CNNs is most meaningful at early layers.

2 Experiments

The Camelyon17[1] dataset was developed to evaluate the classification of breast cancer metastases in lymph node sections. From the gigapixel images, we randomly sample $24,775$ patches of 224×224 pixels. Patch labels are

[1]https://camelyon17.grand-challenge.org



(a) pretrained (b) finetuned

Figure 1: Feature visualization of layers at increasing depths, namely mixed5b, mixed5c, mixed6a, mixed7c. In (a) the pretrained and in (b) the finetuned model. Only a subset of channels is presented due to space restrictions. Visualizations of all channels are available in the online repository[3].

extracted from the regions in the image annotated as *non-tumor* and *tumor*. InceptionV3 [Szegedy et al., 2016] pretrained on ImageNet is finetuned on the binary classification task with stochastic gradient descent (learning rate 10^{-4}, decay 10^{-6}, and Nesterov momentum 0.9) for 30 epochs. Vertical and horizontal flipping and color augmentation, i.e. hue and brightness perturbations, are applied as data augmentation. The model performance is validated on $2,274$ validation patches, with validation accuracy 0.87 and Area under the ROC Curve (AUC) 0.97. The patch-based AUC is comparable to the competition-winning models and sufficient for a meaningful model interpretation analysis.

Filter Visualization We apply the Lucid feature visualization toolbox [Olah et al., 2017] to the CNN before and after finetuning[2]. The toolbox generates an image that maximally activates the filter outputs for a single channel, solving the optimization problem as initially introduced by [Erhan et al., 2009]. Obtaining understandable patterns in the images generated to maximize a given channel is a non-trivial task. The generation of images for the post-finetuning filters required up to $3,584$ steps to converge as well as tedious parameters tweaking [3]. Without preconditioning and parametrization, the generated images contain high-frequency patterns that resemble adversarial images. In the pretrained network, the representations become more abstract and sophisticated in deeper layers (Fig. 1a) as previously shown in [Olah et al., 2017]. The detection of simple textures and patterns of early layers is maintained after finetuning (as shown by the two images on the left in Fig. 1a and b). Complex collages of object-resembling shapes appear, however, only in the deep layers of the pretrained network. The finetuned filters become more and more dissimilar from the pretrained filters in deeper layers. This phenomenon was already observed on the classification of cellular morphological changes by [Kensert et al., 2019], who attributed the lack of high-level abstractions to model overparametrization. The differences between medical images and ImageNet are, in fact, considerable. In histopathology images the variety of color, textures, backgrounds and objects is substantially shrunk to repetitive patterns (nuclei) as opposed to the wide diversity of natural images. Sources of variability are, for the most part, texture, shape irregularities and spatial arrangement of the cells.

Activation Maps Activation maps are particularly useful to directly visualize the attention of the CNN on the input images. Nonetheless, their application in histopathology is likewise challenging. Methods like grad-CAM [Selvaraju et al., 2017] are optimized to give explanations for the predicted class with fine-grained details about the object parts that influenced the decision. When the inputs come from the domain of histopathology, multiple occurrences of small instances (such as nuclei and mitosis) dominate in the image. In this context, grad-CAM fails to localize the multiple occurrences individually and its output is little informative [Chattopadhay et al., 2018]. This limitation is partially solved in grad-CAM++ [Chattopadhay et al., 2018] by replacing the average of the partial derivatives used in standard grad-CAM with the weighted average of the pixel-wise gradients. As a result, localization is more robust to multiple instances of the same class in the image.

[2]Post-finetuning filters were obtained using the lucid4keras wrapper.

[3]Implementation, parameter configurations and all the visualizations are available in the repository `github.com/maragraziani/IMVIP2019`.

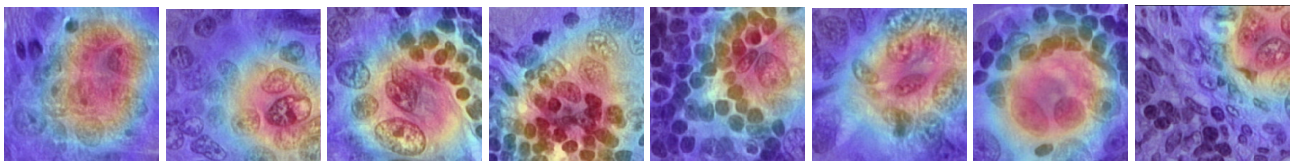


Figure 2: The activation maps of grad-CAM++ show that nuclei pleomorphism captures the attention of the classifier. A subset of the images, for which probabilities of tumor are above 0.99 is presented. All the visualizations are available online[3].

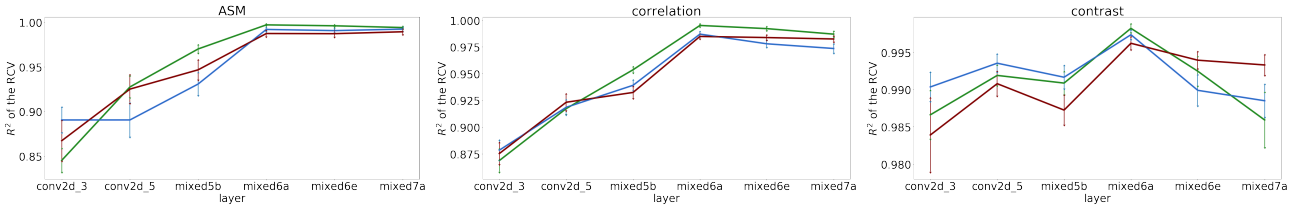


Figure 3: R^2 of the RCV for concept measures of ASM, correlation and contrast over the layers for InceptionV3 with randomized parameters (red), InceptionV3 pretrained (blue) and InceptionV3 finetuned (green). Standard errors were computed over 10 different data splits.

The attention visualizations in Fig. 2 show that the last layer activations focus on nuclei with high pleomorphism, i.e. marked variations in size, shape and texture appearance.

Interpretation with Concept Attribution This experiment on concept attribution aims at quantitatively analyzing the impact of finetuning on the representation of concepts of texture within the CNN layers. RCVs generate quantifiable explanations that do not depend on input features but rather on a set of arbitrarily chosen concepts [Graziani et al., 2018, Graziani et al., 2020]. Linear regression is solved in the activation space of a layer to find the direction of sharpest increase of a continuous measure representing one concept in the image, which is called *concept measure*. RCVs were successfully applied to histopathology applications [Graziani et al., 2018] and retinal fundus images [Graziani et al., 2019a]. In histopathology applications, *concept measures* of nuclei shape and texture were used to represent nuclei morphology and appearance, which are relevant to stage grading. The Haralick texture descriptors were used as *concept measures* and their relevance was evaluated in a CNN classifying *tumor* from *non-tumor* images. Angular Second Moment (ASM), contrast and correlation were found particularly relevant and bidirectional scores showed that ASM and correlation explain the decisions for the *non-tumor* class, while contrast explains the *tumor* class. We further extend this analysis by evaluating the RCVs (evaluation is given by the determination coefficient of the regression, see [Graziani et al., 2020] for details) in three InceptionV3 networks with different parameters: 1) Xavier's random initialization of the parameters 2) pretrained on ImageNet and 3) finetuned on the histopathology task. The best performing *concept measures* of texture in [Graziani et al., 2018, Graziani et al., 2020], i.e. ASM, correlation and contrast, are computed on a subset of 1,000 images. Since nuclei segmentation of the images are not available for this dataset, the concept measures are computed on the entire image. Regression is solved on the global spatial average of features maps at intermediate layers (as recommended in [Graziani et al., 2020]). Fig. 3 shows the determination coefficient, R^2, at six different depths in the three networks. Concepts of texture seem to depend only moderately on the network parameters, as the R^2 for the randomized network is just below the other two networks. Pretraining and finetuning improve the R^2 and reduce the standard error. These results suggest that the architecture itself acts as a prior on the features extracted from the images. This aspect is further analyzed in [Graziani et al., 2019b], where more concepts and datasets are used in the comparison between randomly initialized networks and pretrained networks. Our results, finally, seem in line with the idea that transfer from ImageNet is mostly beneficial for the better scaling of the weights, rather than reuse of deep features [Raghu et al., 2019].

3 Conclusion

This paper summarizes a recipe to interpret feature reuse in ImageNet pretrained models finetuned as classifiers of breast cancer histopathology images. Despite the clear differences between natural and medical images, finetuning is still a common practice. Results on histopathology data show that feature reuse is meaningful in the early layers of the network, which focus on identifying repetitive patterns and textures. These patterns are used by the network to detect nuclei pleomorphism, as shown by class activation maps and further confirmed by results with RCVs.

Future work will address the partial recycling of the pretrained weights of only early layers. Concept attribution appears as a promising tool that generates explanations in terms of arbitrary, human-friendly concepts. Clinicians could, therefore, verify and enhance (or discard) the learning of some concepts during network training.

Acknowledgments

This work was possible thanks to the PROCESS project (grant agreement No 777533).

References

[Chattopadhay et al., 2018] Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *WACV*.

[Erhan et al., 2009] Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1.

[Graziani et al., 2020] Graziani, M., Andrearczyk, V., Marchand-Maillet, S., and Müller, H. (2020). Concept attribution with regression concept vectors. *(submitted)IEEE transactions on Multimedia*.

[Graziani et al., 2018] Graziani, M., Andrearczyk, V., and Müller, H. (2018). Regression concept vectors for bidirectional explanations in histopathology. *iMIMIC at MICCAI*.

[Graziani et al., 2019a] Graziani, M., Brown, J., Andrearczyk, V., Yildiz, V., Campbell, J. P., Erdogmus, D., Ioannidis, S., Chiang, M. F., Kalpathy-Kramer, J., and Müller, H. (2019a). Improved interpretability for computer-aided severity assessment of retinopathy of prematurity. *Medical Imaging 2019: Computer-Aided Diagnosis*.

[Graziani et al., 2019b] Graziani, M., Müller, H., and Andrearczyk, V. (2019b). Interpreting intentionally flawed models with linear probes. *(in press) Statistical Deep Learning for Computer Vision, ICCV 2019*.

[Kensert et al., 2019] Kensert, A., Harrison, P. J., and Spjuth, O. (2019). Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. *SLAS DISCOVERY: Advancing Life Sciences R&D*, 24(4):466–475.

[Olah et al., 2017] Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. *Distill*, 2(11):e7.

[Raghu et al., 2019] Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: Understanding transfer learning with applications to medical imaging. *arXiv preprint arXiv:1902.07208*.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

[Szegedy et al., 2016] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *CVPR*.