Rotation Equivariant CNNs for Digital Pathology

Bastiaan S. Veeling*, Jasper Linmans*, Jim Winkens*, Taco Cohen, and Max Welling

University of Amsterdam, The Netherlands

Abstract. We propose a new model for digital pathology segmentation, based on the observation that histopathology images are inherently symmetric under rotation and reflection. Utilizing recent findings on rotation equivariant CNNs, the proposed model leverages these symmetries in a principled manner. We present a visual analysis showing improved stability on predictions, and demonstrate that exploiting rotation equivariance significantly improves tumor detection performance on a challenging lymph node metastases dataset. We further present a novel derived dataset to enable principled comparison of machine learning models, in combination with an initial benchmark. Through this dataset, the task of histopathology diagnosis becomes accessible as a challenging benchmark for fundamental machine learning research.

1 Introduction

The field of digital pathology is developing rapidly, following recent advancements in microscopic imaging hardware that allow digitizing glass slides into whole-slide images (WSIs). This digitization has facilitated image analysis algorithms to assist and automate diagnostic tasks. A proven approach is to use convolutional neural networks (CNNs), a type of deep learning model, trained on patches extracted from whole-slide images. The aggregate of these patch-based predictions serves as a slide-level representation used by models to identify metastases, stage cancer or diagnose complications. This approach has been shown to outperform pathologists in a variety of tasks[1,2,3].

This performance is achieved using off-the-shelf CNN architectures originally designed for natural images [2]. The effectiveness of these models can be largely attributed to the efficient sharing of parameters in convolutional layers. As a result, local patterns are encoded independently of their spatial location, and shifting the input leads to a predictable shift in the output. This property, known as translational equivariance, effectively exploits the translational symmetry inherent in natural images leading to strong generalization.

In contrast to natural images, WSIs exhibit not only translational symmetry but rotation and reflection symmetry as well. CNNs do not exploit these symmetries, and as a result are found empirically to spend a large part of their parameter budget on multiple rotated and reflected copies of filters [4]. Additionally,

*Equal contribution.

we find that CNNs trained on histopathology data exhibit erratic fluctuations in predictions under input rotation and reflection. Enforcing equivariance in the model under these transformations is expected to reduce such instabilities, and lower the risk of overfitting by improving parameter sharing.

To encode these symmetries, we leverage recent findings in rotation equivariant CNNs [5,6,7], a current topic of interest in the machine learning community. These methods show strong generalization under limited dataset size and are more robust under adversarial perturbations in rotation, translation and local geometric distortions [8]. We propose a fully-convolutional patch-classification model that is equivariant to 90° rotations and reflection, using the method proposed by [5]. We evaluate the model on the Camelyon16 benchmark [9], showing significant improvement over a comparable CNN on slide level classification and tumor localization tasks.

As slide-level metrics potentially obscure the relative performance of patch-level models, we further validate on a patch-level task. In this regime, there is currently no benchmark that harbors the high volume, quality and variety of Camelyon16. Thus, we present *PatchCamelyon*(PCam), a large-scale patch-level dataset derived from Camelyon16 data. Through this dataset, we demonstrate that the proposed model is more accurate and more robust under input rotation and reflection, compared to an equivalent standard CNN.

The contributions of this work are as follows: **(1)** we propose a novel deep learning model that utilizes symmetries inherent to histopathology[1], **(2)** demonstrate that rotation equivariance improves model reliability and **(3)** present a new large-scale histopathology dataset that enables precise model evaluation.

Related Work A common approach to improve orientation robustness is to train CNNs using extensive *data augmentation*, perturbing data with random transformations [1,2]. Although this may improve generalization, it fails to capture local symmetries and does not guarantee equivariance at every layer. As CNNs have to learn rotation equivariance from data, they require a larger model capacity to hold copies of identical filters. Even if rotation equivariance is achieved on training data, there is no guarantee that this generalizes to a test set. Orthogonally, [1,10] propose a test-time augmentation strategy that averages the predictions of 90°-rotated and mirrored versions to improve robustness to orientation-induced instability. As a downside, this comes at 8 times the computational cost and does not provide guarantees on equivariance [11].

Methods that enable equivariance under rotations and other transformations include Harmonic Networks [6], which constrain the set of filters to circular harmonics, allowing for full 360°-equivariance. [7] employs steerable filters and evenly samples a small number of rotations. In this work, we focus on the straight-forward G-CNN method from [5] applied on discrete rotation/reflection groups. Although these groups do not cover the full continuous rotational symmetry inherent in WSIs, the empirical evidence gathered so far shows that 90° rotation equivariance improves performance significantly[7].

[1]PCam details and data at `https://github.com/basveeling/pcam`. Implementations of equivariant layers available at `https://github.com/basveeling/keras_gcnn`.

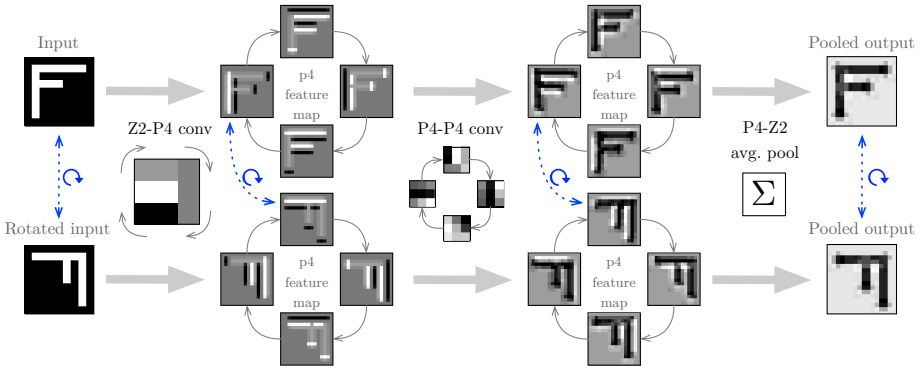

Fig. 1: Given a canonical input and a rotated duplicate, we demonstrate how a 2-layer G-CNN is equivariant in $p4$. Feature maps of one kernel per layer are shown, and the dashed blue arrows indicate how (intermediate) representations of the two inputs correspond. The $\mathbb{Z}^2 \to p4$ convolution correlates the input with 4 rotated versions of the same kernel. The $p4 \to p4$ convolution correlates the resulting feature map with the $p4$-kernel, cyclically-shifting and rotating the kernel for each orientation. The final layer demonstrates how average-pooling over the orientations produces a representation that is locally invariant and globally equivariant to rotation. *Global* average pooling over $p4$ would result in a representation globally invariant to both translation and rotation.

2 Methods

2.1 Background

In the mathematical model of CNNs and G-CNNs introduced in [5], input images and output segmentation masks are considered to be functions $f : \mathbb{Z}^2 \to \mathbb{R}^K$, where K denotes the number of channels, and f is implicitly assumed to be zero outside of some rectangular domain.

A standard convolution[2] (denoted $*$) of an input f with filter ψ is defined as:

$$[f * \psi](x) = \sum_{y \in \mathbb{Z}^2} \sum_{k=1}^{K} f_k(y) \psi_k(x - y). \tag{1}$$

G-CNNs are a generalization of CNNs that are equivariant under more general symmetry groups, such as the group $G = p4$ of 90° rotations, or $G = p4m$ which additionally includes reflection. In a G-CNN, the feature maps are thought of as functions on this group. For $p4$ and $p4m$, this simply means that feature channels come in groups of 4 or 8, corresponding to the 4 pure rotations in $p4$ or the 8 roto-reflections in $p4m$. In the first layer, these are produced using the $(\mathbb{Z}^2 \to G)$-convolution:

$$[f * \psi](g) = \sum_{y \in \mathbb{Z}^2} \sum_{k=1}^{K} f_k(y) \psi_k(g^{-1} y), \tag{2}$$

[2]Technically, this is a cross-correlation

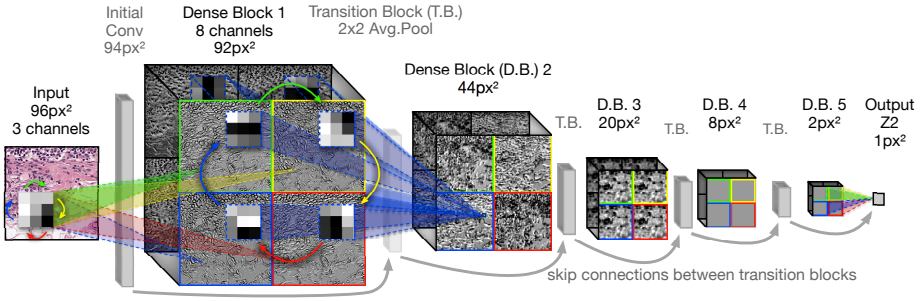

Fig. 2: The proposed equivariant DenseNet architecture for the $p4$ group, consisting of 5 Dense Blocks (D.B.) alternated with Transition Blocks (T.B.). The final layer of the model is a $p4 \to \mathbb{Z}^2$ group pooling layer followed by a sigmoid activation. The four orientations in $p4$ are illustrated through primary colors. A $\mathbb{Z}^2 \to p4$ kernel (*left*), $p4 \to p4$ kernel (*middle*) and $p4 \to \mathbb{Z}^2$ kernel (*right*) illustrate how equivariance arises in the model.

where $g = (r, t)$ is a roto-translation (in case $G = p4$) or roto-reflection-translation (in case $G = p4m$).

In further layers, both feature maps and filters are functions on G, and these are combined using the $(G \to G)$-convolution:

$$[f * \psi](g) = \sum_{h \in G} \sum_{k=1}^{K} f_k(h) \psi_k(g^{-1}h). \tag{3}$$

In the final layer, a group-pooling layer is used to ensure that the output is either invariant (for classification tasks) or equivariant as a function on the plane (for segmentation tasks, where the output is supposed to transform together with the input). In Fig. 1 we demonstrate how equivariance is achieved through this process. Non-linear activations and pooling operations are equivariant in $p4m$[5], allowing layers to be freely stacked to enable deep architectures.

2.2 G-CNN DenseNet architecture

The proposed patch-classification model is shown in Fig. 2 for $p4$ (the $p4m$-variant is a trivial extension). The architecture is based on the densely connected convolutional network (DenseNet) [12], which consist of dense blocks with layers that use the stack of all previous layers as input, alternated with transition blocks consisting of a 1×1 convolutional layer and 2×2 strided average pooling. We use one layer per dense block due to the limited receptive field of the model, with 5 dense-block/transition-block pairs. The model spatially-pools the input by a factor of 2^5, the output of which resembles the segmentation resolution used in [1].

Full-model group equivariance is achieved by replacing all convolution layers with group-equivariant versions [5]. Batch normalization layers[13] are made

group-equivariant by aggregating moments per *group* feature map rather than spatial feature map (as proposed by [5]). Zero-padding is removed to prevent boundary-effects. The final layer consists of a group-pooling layer followed by a sigmoid activation, resulting in tumor-probability output on the plane $\mathbb{Z}^2$. As the model is fully convolutional, efficient inference can be achieved at test time by reusing computation of neighbouring patches, reducing segmentation time of a full WSI from hours to ~ 2 minutes on a NVIDIA Titan XP.

3 Experimental results

3.1 Datasets and Evaluation

To evaluate the proposed model, we use Camelyon16 [9] and PCam. Additional testing is performed on BreakHis [14]. **(1)** The Camelyon16 dataset contains 400 H&E stained WSIs of sentinel lymph node sections split into 270 slides with pixel-level annotations for training and 130 unlabeled slides for testing. The slides were acquired and digitized at 2 different centers using a $40\times$ objective (resultant pixel resolution of 0.243 microns). In the Camelyon16 challenge, model performance is evaluated using the FROC curve for tumor localization. **(2)** The PCam dataset contains 327,680 patches extracted from Camelyon16 at a size of 96×96 pixels @ $10\times$ magnification, with a 75/12.5/12.5% train/validate/test split, selected using a hard-negative mining regime[1]. **(3)** The BreakHis dataset contains 7909 H&E stained microscopy images at a size of 700×460 pixels. The task is to classify the images into benign or malignant cases for multiple magnification factors. We limit our evaluation to the images at $4\times$ magnification, for which previous approaches [14,15] have reported the highest accuracy.

For the evaluation on the WSI-level Camelyon16 benchmarks, we largely follow the pipeline proposed in [1], uniformly sampling WSIs and drawing tumor/non-tumor patches with equal probability. To prevent overrepresentation of background and non-tissue patches, slides are converted to HSV, blurred, and rejected if the max. pixel saturation lies below 0.07 (range [0,1]) and value above 0.1. This was empirically verified to not drop tissue patches. For computing the FROC score, tumor location candidates are selected with an efficient square non-maximum suppression window rather than radial. The window-size is tuned per model on the validation set. FROC score confidence bounds are computed using 2000 bootstrap samples [1]. Train and validation splits are created by dividing the available WSIs randomly, maintaining equal tumor/normal ratio. We focus on the WSI data at $10\times$ magnification (4 times smaller than the original dataset, at 0.972 microns per pixel) to fit the compute budget available for this work. Following [1], we focus on the more-granular tumor-detection FROC metric in favor of slide-level AUC.

Training Details: Models are optimized using Adam[16] with batch size 64 and initial learning rate $1e{-}3$ (halved after 20 epochs of no improvement in validation loss). Epochs consists of 312 batches with a batch size of 64. Validation loss is computed using 40.000 sampled patches. Weights with lowest validation loss are selected for test evaluation.

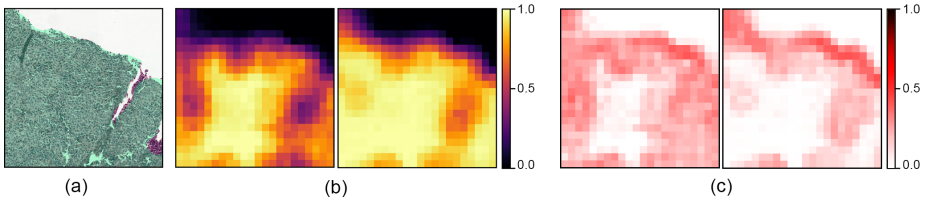

Fig. 3: **(a)** shows a large input region spanning multiple patches, with the tumor ground truth overlayed in green. The region is predicted under 32 evenly spaced sub-90° rotations, and prediction maps rotated back to original orientation. **(b)** shows the mean prediction and **(c)** shows the standard deviation of the predictions across all rotations, using DenseNet (*left*) and P4M-DenseNet (*right*). Both networks are trained on the 12.5% data regime.

3.2 Model reliability

We evaluate stability of predictions under rotation of the input. We present a visual analysis in Fig. 3. For a comparable baseline we use an equivalent model with standard convolutions. For a fair model comparison, we keep the number of parameters consistent by multiplying the growth rate of the baseline model by the square root of the group size [5]. Bar the expected fluctuation around the tumor boundary (that arises due to the sub-sampled segmentation), the $p4m$-model is more robust to transformations even outside the group (sub-90° rotations). In addition, we observe a higher confidence for predictions inside the tumor regions for P4M-DenseNet as compared to the baseline.

3.3 P4M-DenseNet Performance

Table 1: Performance on PCam, measured by negative log-likelihood, accuracy and AUC. Experiments with additional data augmentation with 90° rotations and reflections are marked by +. M indicates matching number of $\mathbb{Z}^2$ maps, $\#W$ number of weights, K number of $\mathbb{Z}^2$ maps per layer.

Network	K	$\#W$	NLL	Acc	AUC
P4M-DenseNet	64	119K	**0.260**	**89.8**	**96.3**
P4M-DenseNet M	24	19K	0.273	89.3	95.8
P4-DenseNet	48	125K	0.329	89.0	94.5
DenseNet+	24	128K	0.306	88.1	95.1
DenseNet+ M	64	902K	0.365	87.2	94.6
DenseNet	24	128K	0.315	87.6	95.5

PatchCamelyon (PCam) We assess the performance of our main contribution, the P4M-DenseNet architecture, on the PCam dataset. Table 1 reports the performance. P4M-DenseNet outperforms other models, closely followed by the P4-DenseNet, indicating that both rotation and reflection are useful inductive biases, that can not be learned by data augmentation alone. Keeping the number of $\mathbb{Z}^2$ maps fixed in the baseline degrades performance further, demonstrating the sample-efficiency of the P4M model.

Model	Data	FROC
P4M-DenseNet 123k params	100%	84.0 (75.5, 91.5)
	50%	81.5 (72.2, 89.3)
	25%	72.6 (58.7, 84.6)
	12.5%	60.7 (46.0, 74.1)
DenseNet 126k params	100%	81.7 (72.1, 90.3)
	50%	80.0 (69.3, 89.1)
	25%	71.0 (57.7, 82.0)
	12.5%	55.4 (42.6, 68.5)
Liu *et al.* [1] 10×	100%	79.3 (74.2, 84.1)
Wang *et al.* [17]	100%	80.7*
Pathologist [9]	–	73.3

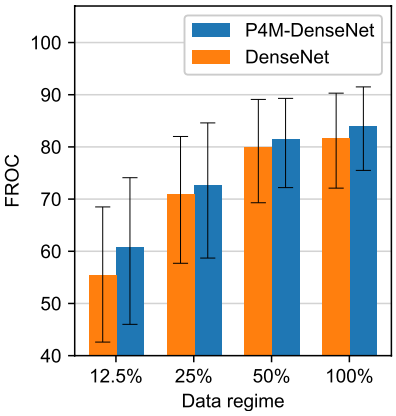

Fig. 4 & Table 2: Performance on the Camelyon16 test set. The confidence bounds are obtained using a 2000-fold bootstrap regime. *Challenge winner [17] uses 40× resolution and is not directly comparable.

Camelyon16 We evaluate our patch-based model on the slide-level tumor localization task of the Camelyon16 challenge. Fig. 4 reports the performance on the FROC score, next to those of a pathologist [9] and the state-of-the-art approaches reported on this dataset, including [1,17]. For the baseline DenseNet, the training data is augmented with 90° rotations and reflection. We experiment with multiple data regimes, where the number of WSIs in the training set is incrementally reduced by a factor of two.

The results indicate that the proposed method performs consistently better than all compared methods in terms of the FROC metric. Comparing to the baseline DenseNet results, we see that the superiority of our proposed architecture is predominantly due to the increased parameter sharing by the $p4m$-equivariance, which frees up model capacity and reduces the redundancy of detecting the same histological patterns in different orientations.

We also observe that the performance gap between our model and the baseline increases when we limit the dataset size by removing WSIs. This seems to indicate that the performance in the small-data regime benefits significantly from the sample efficiency of P4M-DenseNet, with diminishing returns when the amount of data is sufficient for the baseline network to achieve (approximate) rotation equivariance. This performance gap remains for the full data set.

BreakHis As an additional evaluation method, we assess the performance of the proposed model on the binary classification task of BreakHis as described in Section 3.1. As training the model from scratch is impractical given the small dataset, we pre-train on Camelyon16 at a similar pixel resolution. Similar to [14], we predict the malignancy of a test image by using the maximum activation of 1000 random crops. We obtain an accuracy of 96.1 ± 3.2 and 93.5 ± 4.7 for

P4M-Densenet and the baseline respectively, outperforming previous approaches [14][15].

4 Conclusion

We present a novel histopathology patch-classification model that outperforms a competitive traditional CNN by enforcing rotation and reflection equivariance. A derived patch-level dataset is presented, allowing straightforward and precise evaluation on a challenging histopathology task. We demonstrate that rotation equivariance improves reliability of the model, motivating the application and further research of rotation equivariant models in the medical image analysis domain.

Acknowledgements We thank Geert Litjens, Jakub Tomczak, Dimitrios Mavroeidis and the anonymous reviewers especially for their insightful comments. This research was supported by Philips Research, the SURFSara Lisa cluster and the NVIDIA GPU Grant.

References

1. Liu, Y., et al.: Detecting cancer metastases on gigapixel pathology images. (2017)
2. Litjens, G., et al.: A survey on deep learning in medical image analysis. (2017)
3. Bejnordi, B.E., et al.: Stain specific standardization of Whole-Slide histopathological images. IEEE Trans. Med. Imaging **35**(2) (2016) 404–415
4. Zeiler, M., et al.: Visualizing and understanding convolutional networks. (2014) 818–833
5. Cohen, T.S., et al.: Group equivariant convolutional networks. (2016)
6. Worrall, D.E., et al.: Harmonic networks: Deep translation and rotation equivariance. In: Proc. IEEE CVPR. Volume 2., openaccess.thecvf.com (2017)
7. Weiler, M., et al.: Learning steerable filters for rotation equivariant CNNs. (2017)
8. Dumont, B., et al.: Robustness of Rotation-Equivariant networks to adversarial perturbations. (2018)
9. B, E.B., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA **318**(22) (2017) 2199–2210
10. Cireşan, D.C., et al.: Mitosis detection in breast cancer histology images with deep neural networks. MICCAI **16**(Pt 2) (2013) 411–418
11. Lenc, K., et al.: Understanding image representations by measuring their equivariance and equivalence. In: 2015 IEEE CVPR. (2015) 991–999
12. Huang, G., et al.: Densely connected convolutional networks. (2016)
13. Ioffe, S., et al.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. (2015) 448–456
14. Spanhol, F.A., et al.: A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng. **63**(7) (2016) 1455–1462
15. Song, Y., et al.: Supervised intra-embedding of fisher vectors for histopathology image classification. In: MICCAI 2017, Cham (2017) 99–106
16. Kingma, D.P., et al.: Adam: A method for stochastic optimization. (2014)
17. Wang, D., et al.: Deep learning for identifying metastatic breast cancer. (2016)