MitoDet: Simple and robust mitosis detection

Jakob Dexl, Michaela Benz, Volker Bruns, Petr Kuritcyn, Thomas Wittenberg

Fraunhofer Institute for Integrated Circuits IIS

Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

`jakob.dexl@fau.de`

September 2, 2021

ABSTRACT

Mitotic figure detection is a challenging task in digital pathology that has a direct impact on therapeutic decisions. While automated methods often achieve acceptable results under laboratory conditions, they frequently fail in the clinical deployment phase. This problem can be mainly attributed to a phenomenon called domain shift. An important source of a domain shift is introduced by different microscopes and their camera systems, which noticeably change the color representation of digitized images. In this method description we present our submitted algorithm for the Mitosis Domain Generalization Challenge [1], which employs a RetinaNet [2] trained with strong data augmentation and achieves an F1 score of 0.7138 on the preliminary test set.

Keywords Mitosis detection · Domain generalization · Digital pathology

1 Methods

Motivated by recent data centric approaches we use a RetinaNet [2] trained with strong data augmentation to enforce prediction consistency.

1.1 Dataset

We use the publicly available Mitosis Domain Generalization Challenge (MIDOG) dataset [1]. The data consists of 200 Whole Slide Images (WSIs) from hematoxylin and eosin (HE) stained breast cancer cases. Furthermore the dataset can be divided into subsets of 50 images, which were acquired and digitized with four different scanners (Aperio ScanScope CS2, Hamamatsu S360, Hamamatsu XR NanoZoomer 2.0, Leica GT450). For three scanners annotations for mitotic figures and hard negatives (imposters) are provided. The disclosed preliminary and final test sets contain samples of two known scanners and two unknown ones.

1.2 Model

Our object detection algorithm consists of a RetinaNet [2] with an EfficientNet B0 [3] backbone. The backbone is initialized with state of the art ImageNet weights, which were trained using RandAugment [4] and Noisy Student [5]. We did not change the feature pyramid and used all five pyramid levels. The network's heads consist of 4 layers with a channel size of 128. Anchor ratios are set to one while the differential evolution search algorithm introduced by [6] is employed to determine three anchor scales (0.781, 1.435, 1.578).

1.3 Domain generalization through augmentation

Our main method to approach domain generalization is data augmentation. Data driven approaches such as RandAugment [4] have been proven to increase model robustness and have been used in recent state of the art models.
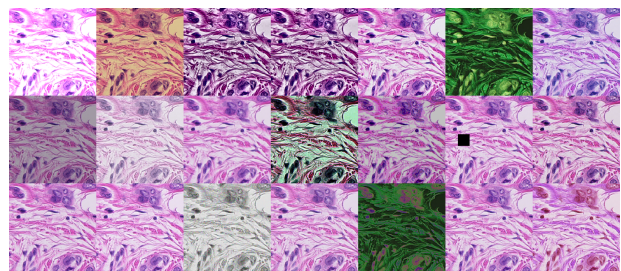

Figure 1: Used augmentations with different strengths.

Inspired by Trivial Augment [7] a very simple random augmentation strategy is used, where a single augmentation is applied to each image. The augmentations are drawn uniformly from a set of color, noise and special transformations while the augmentation strength is random to some defined degree. The pool of augmentations consists of color jitter, HE [8], fancy PCA, hue, saturation, equalize, random contrast, auto-contrast, contrast limited adaptive histogram equalization (CLAHE), solarize, solarize-add, sharpness, Gaussian blur, posterize, cutout, ISO noise,

JPEG compression artifacts, pixel wise channel shuffle and Gaussian noise. In addition every image is randomly flipped and RGB channels are randomly shuffled.

1.4 Training and evaluation

For experimentation, we divide the dataset into five folds with three training, one validation and one test split for each scanner (test splits are added to the train set for submissions). During the training phase we uniformly sample the images of the train set and randomly select a mitotic figure or an imposter annotation. A patch with a size of 448 pixels is randomly cropped around the selected annotation similar to [9]. The RetinaNet is trained for 100 pseudo epochs with a batchsize of 16 using the super convergence scheme [10]. Adam optimizer with a maximum learning rate of 1e-4 is used. The best models are selected based on the lowest validation loss. After the training phase, we combine the training and validation set and optimize the models confidence threshold with respect to the best F1 score. During inference, incoming WSIs are tiled into overlapping patches of 448 pixels. All models are trained and tested using a Nvidia GeForce RTX 3060 with 12GB GPURAM.

2 Results

Our proposed method achieves a F1 score of 0.7138 on the preliminary test set of the MIDOG challenge.

3 Discussion

Overall, we are able to generalize better across multiple scanner domains with strong data augmentation. The magnitude at which such simple transformations improve generalization at no cost of inference speed is higher than expected. Even models trained with only one scanner reach similar results on our test split, showing only a small performance drop. In the following, we will lay out unsuccessful attempts to improve the quality further. One major issue was the model selection based on the validation loss. The models were not capable of overfitting the data, assumingly due to the sampling and the strong data augmentation, models ended up in an equilibrium mode where performance improvements were wiggling between the different scanners back and forth. Because of that, the representation shift metric proposed by Staake et al. [11] was tested. It was applied to the three convolutional layers, which flow into the feature pyramid, but was found to not help the model selection process. Another strategy was a dual stage attempt with a verification net proposed by Li et al. [12]. The network was trained on the predicted patches of the first stage using the same augmentation and in addition a Gradient Reversal Layer [13] to remove even the last bits of scanner dependent information. Unfortunately this resulted in a performance drop of 12.1% on the preliminary dataset. Finally, the choice of using an EfficientNet originated from the attempt to incorporate the unlabeled data using a self-supervised Student Teacher learning procedure based on the STAC framework [14].

While increasing the performance on our test split, this resulted in a small performance drop of 1% on the preliminary dataset. One problem was that producing pseudo labels with a high confidence threshold resulted in very few labeled samples while self-training reportedly needs a huge amount of pseudo labeled data to make use of it. A second problem arises with false positive pseudo labels. We used a labeled scanner to check the number of wrong labels incorporated in the pseudo labels and found that for mitotic figures pseudo labels were mainly correct while hard negatives actually included a lot mitotic figures. This probably led to more confusion than having a positive effect.

References

[1] Marc Aubreville, Christof Bertram, Mitko Veta, Robert Klopfleisch, Nikolas Stathonikos, Katharina Breininger, Natalie ter Hoeve, Francesco Ciompi, and Andreas Maier. MItosis DOmain Generalization Challenge. March 2021. Publisher: Zenodo.

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *arXiv:1708.02002 [cs]*, February 2018. arXiv: 1708.02002.

[3] Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection. *arXiv:1911.09070 [cs, eess]*, July 2020. arXiv: 1911.09070.

[4] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. RandAugment: Practical automated data augmentation with a reduced search space. *arXiv:1909.13719 [cs]*, November 2019. arXiv: 1909.13719.

[5] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with Noisy Student improves ImageNet classification. *arXiv:1911.04252 [cs, stat]*, June 2020. arXiv: 1911.04252.

[6] Martin Zlocha, Qi Dou, and Ben Glocker. Improving RetinaNet for CT Lesion Detection with Dense Masks from Weak RECIST Labels. *arXiv:1906.02283 [cs, eess]*, June 2019. arXiv: 1906.02283.

[7] Samuel G. Müller and Frank Hutter. TrivialAugment: Tuning-free Yet State-of-the-Art Data Augmentation. *arXiv:2103.10158 [cs]*, August 2021. arXiv: 2103.10158.

[8] David Tellez, Maschenka Balkenhol, Irene Otte-Holler, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Whole-Slide Mitosis Detection in H&E Breast Histology Using PHH3 as a Reference to Train Distilled Stain-Invariant Convolutional Networks. *IEEE Transactions on Medical Imaging*, 37(9):2126–2136, September 2018. arXiv: 1808.05896.

[9] Christian Marzahl, Marc Aubreville, Christof A. Bertram, Jason Stayt, Anne-Katherine Jasensky, Florian Bartenschlager, Marco Fragoso-Garcia, Ann K. Barton, Svenja Elsemann, Samir Jabari, Jens Krauth, Prathmesh Madhu, Jörn Voigt, Jenny Hill, Robert Klopfleisch, and Andreas Maier. Deep Learning-Based Quantification of Pulmonary Hemosiderophages in Cytology Slides. *Scientific Reports*, 10(1):9795, December 2020. arXiv: 1908.04767.

[10] Leslie N. Smith and Nicholay Topin. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. *arXiv:1708.07120 [cs, stat]*, May 2018. arXiv: 1708.07120.

[11] Karin Stacke, Gabriel Eilertsen, Jonas Unger, and Claes Lundstrom. Measuring Domain Shift for Deep Learning in Histopathology. *IEEE journal of biomedical and health informatics*, 25(2):325–336, February 2021.

[12] Chao Li, Xinggang Wang, Wenyu Liu, and Longin Jan Latecki. DeepMitosis: Mitosis detection via deep detection, verification and segmentation networks. *Medical Image Analysis*, 45:121–133, April 2018.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *arXiv:1505.07818 [cs, stat]*, May 2016. arXiv: 1505.07818.

[14] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A Simple Semi-Supervised Learning Framework for Object Detection. *arXiv:2005.04757 [cs]*, December 2020. arXiv: 2005.04757.

Acknowledgments