Classification of Mitotic Cells
Potentials Beyond the Limits of Small Data Sets

Maximilian Krappmann[1], Marc Aubreville[1],Andereas Maier[1],
Christof Bertram[2], Robert Klopfleisch[2]

[1]Pattern Recognition Lab, Friedrich-Alexander-University Erlangen-Nuernberg, Germany
[2]Institute of Veterinary Pathology, Freie University of Berlin, Germany
maximilian.krappmann@inveox.com

Abstract. Tumor diagnostics are based on histopathological assessments of tissue biopsies of the suspected carcinogen region. One standard task in histopathology is counting of mitotic cells, a task that provides great potential to be improved in speed, accuracy and reproducability. The advent of deep learning methods brought a significant increase in precision of algorithmic detection methods, yet it is dependent on the availability of large amounts of data, completely capturing the natural variability in the material. Fully segmented images are provided by the MITOS dataset with 300 mitotic events. The ICPR2012 dataset provides 326 mitotic cells and in AMIDA2014 dataset, 550 mitotic cells for training and 533 for testing. In contrast to these datasets, a dataset with high number of mitotic events is missing. For this, either one of two pathologist annotated at least 10 thousand cell images for cells of the type mitosis, eosinophilic granulocyte and normal tumor cell from canine mast cell tumor whole-slide images, exceeding all publicly available data sets by approximately one order of magnitude. We tested performance using a standard CNN approach and found accuracies of up to 0.93.

1 Introduction

The number of mitotic figures within a certain area is an important characteristic in tumor grading. Mitotic cell count is a valuable predictor, correlating highly with tumor proliferation [1]. Most of the histopathological grading systems require the number of mitotic figures within a defined area (high power field, HPF). However, this area is not defined uniquely [2]. If the number of mitotic figures exceeds a defined limit within ten consecutive HPFs, the grading of the tumor may change. However, this makes a clear distinction of different grades of illness highly questionable. The consensus among different pathologists in grading was evaluated in [3], revealing an inter-oberver discordance in 50% of the cases. As a reason, the high subjectivity of the pathologists and the flexibility in the choice of the high power fields is reported. In order to counteract these obstacles, computer-aided methods for detecting mitotic figures were taken into account. This process can be clustered into two main tasks. First of all, the detection of cells and at second, the classification of the cells found. The hosts of competitions to detect mitotic cells prepared datasets for this tasks. The ICPR2012 dataset provides 326 mitotic events in fully segmented images. In the AMIDA2014 dataset, 550 mitotic cells for training and 533 events for testing can be found. The MITOS dataset provides about 300 mitotic figures. However, it can be questioned if this amount of training data is sufficient to train accurate classifiers. It is known that mitosis can be distinguished into four different phases, each phase results in high variance in their visual representation. This work focuses on the classification task of the cells not on detection. Due to that a dataset with more than 12000 mitotic cells, 17000 tumor cells and 10000 granulocytes is provided by our dataset. These figures have been extracted by manually reviewing 100000 histological image sections of stained canine mast cell tumors. In order to prove the demand for such a dataset, a simple deep learning approach is trained and evaluated with respect to all its parameters and with respect to the modifications of the input data. To provide the correctness of the data, the cells have been annotated by pathologists.

2 Related work

In [4] a multitask learning approach is presented which is able to detect mitotic events. A deep learning architecture similar to [5] was used. This method achieved an F-score of 0.53 across 550 mitotic events in comparison to an F-score of 0.90 for 3064 lymphatic cells. The detection rate of all cells was 98%. Malon et al. [6] used the same architecture as used in this work achieved an F-score of 0.65 on the MITOS dataset. As input 72 × 72 cell patches were used. Handcrafted features like morphology and run length features were extracted on blue ration images. The method introduced in [7] makes use of a pixel-wise sliding window approach. Based on that, pixels were considered as true positive if a single pixel was in a specific range around a mitotic figure. Furthermore, images around the mitotic figures were cropped and the network was trained. The architecture consists of five 2 × 2 convolutional layers followed by 2 × 2 max pooling layers and two fully connected output layers. This process needs a total of 31 seconds per patch and eight minutes per HPF. The achieved F-score was 0.782. A second approach with a reduced depth of three convolutional layers resulted in an F-score of 0.758. This network was eight times faster than the original method. Albarqouini et al. [8] came up with a special type of network that is able process two types of input data. First of all, ground truth data from established datasets like ICPR2012, MITOS or AMIDA are used. The second type of data was obtained using crowd sourcing layers. These layers incorporate annotations of non-professionals. Doing this needed some sort of reliability measurement which is why an expectation maximization algorithm is used to do some weighting. The authors revealed an improvement of 0.36 in the best and 0.068 in the worst case. The best F score was 0.76. Chen et al. [9] used another deep learning approach that made use of fully convolutional networks in order to detect mitotic candidates. Therefore, a probability map of the input is obtained yielding high values for mitotic candidates. This net was trained on 94 × 94 input images. However, since the spatial information is preserved, the network is able to detect mitotic candidates. These candidates are then fed into a convolutional network with a fully connected output layer for classification. A F score of 0.785 was archieved by this cascaded pipeline. Moreover, the pipeline is able to predict a 2000 × 2000 input image within 0.49 seconds using a NVIDIA GeForce GTX Titan GPU. All current deep learning approaches made us of augmentation due to the fact that they require a comparatively large amount of training data. This is obvious since deep learning requires a comparatively large amount of training data.

3 Material and methods

3.1 Dataset

The dataset was acquired using a semi-automatic annotation GUI. Therefore, the whole slide images are processed using the openslide software [10] and manually cropped by simply clicking at cells. The position of the midpoint for a single cell was saved. In addition, the size of the cell patches can be varied if needed. As advantage, different well known deep learning models can be used without adjusting their parameters too much. The cropped candidates were forwarded to two pathologists for revision. Exemplary results of this process can be seen in Fig. 1.

3.2 Network

Using deep learning in this work was based on the feature extraction ability of this method. Since works with good features for mitosis detection could not be found, it was logical to pick a deep learning approach. The size of the cell patches was 50 × 50. An example of different cell types is found in Fig. 2. To classify the mitotic cells two other classes have been extracted namely tumor cells and granulocytes. Besides mitotic cells, canine mast cell tumors largely consist of those last-mentioned cell types. Moreover, the detection of other cells is required for calculating mitotic index. This index was used in [2] to increase the accordance between pathologists. The idea is to divide the mitotic cell count by the number of other cells because it is assumed that mitotic cells in sparse regions may have higher influence on the diagnosis than in studded regions. Our work will show that it is sufficient to use a simple net in order to achieve good results due to the amount of data. Therefore the LeNet architecture was chosen. To evaluate parameters that may effect the performance of the network, the deep learning pipeline is taken into account. First of all, the manipulation of the input is investigated. Afterwards, the feature extraction ability of the network must be evaluated. The standard CNN in this work consists of two 5×5 convolutional layers followed by 2×2 max pooling layer, nesterov momentum of 0.9 and a step size of 0.01 was applied. The fully connected layer consists of 256 nodes followed by a softmax activation layer. For better generalization a dropout of 0.5 was used.

Table 1. In the first part of the table methods for input manipulations are tested. At next different modifications of the CNN architecture are evaluated.

CNN	Precision	Recall	F1	support
CNN+RGB	0.91	0.91	0.91	9490
CNN+GRAY	0.92	0.92	0.92	9490
CNN+AUG	0.92	0.92	0.92	9490
CNN+RGB+AUG	0.93	0.92	0.93	9490
CNN+Moment	0.92	0.92	0.92	9490
CNN+AdaDelta	0.89	0.89	0.89	9490
CNN+AdaGrad	0.91	0.91	0.91	9490
CNN+ExtraLayer	0.89	0.88	0.88	9490

4 Results

4.1 Manipulation of the input

The augmentation of data is routinely performed to create virtually new samples. This is beneficial when the amount of data is not sufficient or the variance in

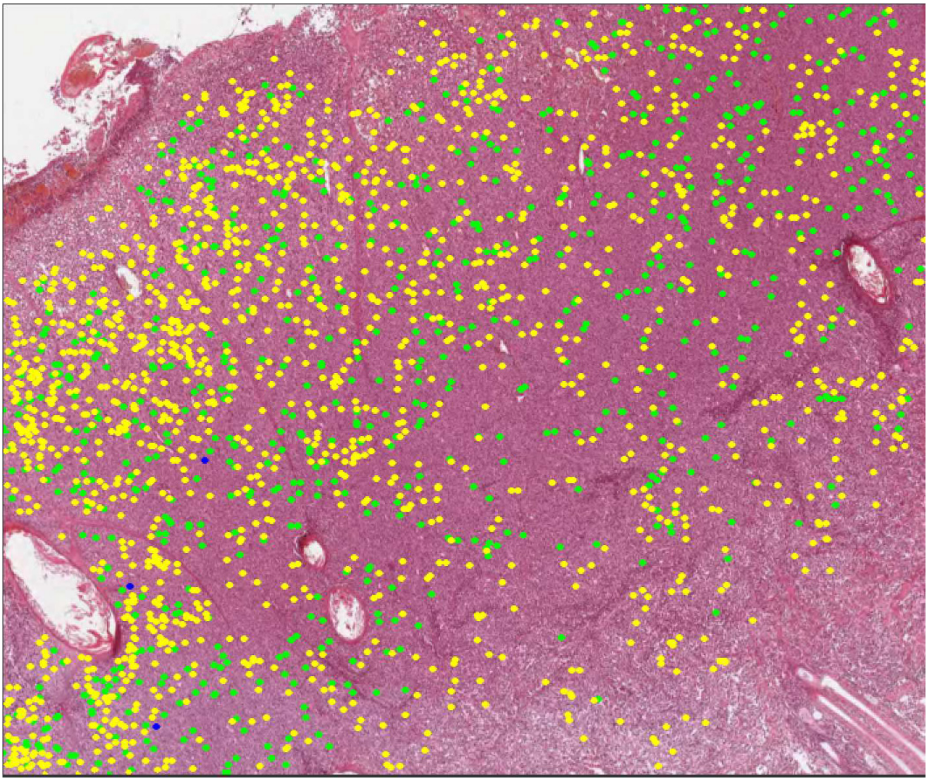

Fig. 1. Cell distribution within a partially segmented H&E slide (green: mitotic cells, blue: granulocytes, yellow: tumor cells).

Table 2. MITOS and our dataset for equal amount of mitotic and non mitotic cells.

CNN	Precision	Recall	F1	support
MITOS2014	0.66	0.65	0.65	172
OURS	0.93	0.93	0.93	4429

the test data is smaller than in real applications. To generate more varying samples, random operations like rotations, scaling, shift in colour channels or noise can be applied to achieve more robust features. If some information or small differences are missing or getting obliterated, this might also have a negative effect. Furthermore, the usage of grey scale images in contrast to RGB images is evaluated. Augmentation in combination with RGB images results in an increased performence. It is likely that the information is increased due to the usage of three channels. In addition, augmentation is used to increase the dataset and capture more variances in the images. The full impact can be seen from the first part in Table 1.

4.2 Parameters affecting the feature extraction

The feature extraction ability of a neural network is based on its adaptation of the filters. To adapt these filters a global loss function is minimized by back-propagating an error. This requires different gradient descent methods. The methods are able to overshot a global minimum or getting stuck in a local minimum. The different methods yield differnces in F-scores between 0.88 and 0.92 due to that behaviour. In addition, the depth of the network is effecting the amount and quality of the filters. If the depth is too large, the network may

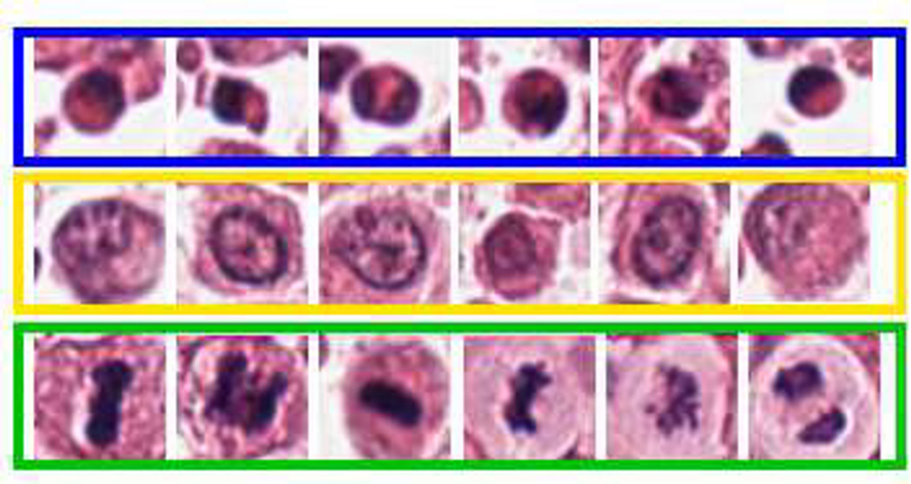

Fig. 2. Different cells of interest (green: mitotic cells, blue: granulocytes, yellow: tumor cells).

run into overfitting the training data which happened in our case. All effects are gathered in the second part in table 1.

4.3 Baseline comparison

As baseline the MITOS2014 dataset of the ICPR14 contest was used. Therefore, our dataset was translated into a two class problem. The network was trained on a $80 - 20$ split. The results in table 2 show a clear improvement on detection accuracies, as is likely induced by the increased dataset size.

5 Conclusion

Our work supports the idea that the key of boosting the classification performance is the amount of input data. The different gradient descent methods reveal their effects on the performance of network as expected. The increased depth of the network has no positive effect on the scores. The limitations of small datasets in the medical field were shown in table 2 and emphasise the need for larger datasets for pattern recognition in the medical context.

References

1. Kiupel M, Webster J, Bailey K, et al. Proposal of a 2-Tier histologic grading system for canine cutaneous mast cell tumors to more accurately predict biological behavior. Veterin Pathology. 2011;48(1):147–155.
2. Meuten D, Moore F, George J. Mitotic count and the field of view area: Time to standardize. SAGE Publications Sage CA: Los Angeles, CA; 2016.
3. Northrup N, Howerth E, Harmon B, et al. Variation among Pathologists in the histologic grading of canine cutaneous mast cell tumors with uniform use of a single grading reference. J Veterin Diagn Invest. 2005;17(6):561–564.
4. Romo-Bucheli D, Janowczyk A, Gilmore H, et al. A deep learning based strategy for identifying and associating mitotic activity with gene expression derived risk categories in estrogen receptor positive breast cancers. Cytometry A. 2017.
5. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems; 2012. p. 1097–1105.
6. Malon CD, Cosatto E, et al. Classification of mitotic figures with convolutional neural networks and seeded blob features. J Pathol Inform. 2013;4(1):9.
7. Cireşan DC, Giusti A, Gambardella LM, et al.; Springer. Mitosis detection in breast cancer histology images with deep neural networks. Proc MICCAI. 2013; p. 411–418.
8. Albarqouni S, Baur C, Achilles F, et al. AggNet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans Med Imaging. 2016;35(5):1313–1321.
9. Chen H, Dou Q, Wang X, et al. Mitosis detection in breast cancer histology images via deep cascaded networks. In: 13th AAAI Conf Artific Intell; 2016.
10. Goode A, Satyanarayanan M. A Vendor-Neutral library and viewer for whole-slide images. Computer Science Department, Carnegie Mellon University. 2008.