Erik Rodner*, Marcel Simon and Joachim Denzler

Deep bilinear features for Her2 scoring in digital pathology

Abstract: We present an automated approach for rating HER2 over-expressions in given whole-slide images of breast cancer histology slides. The slides have a very high resolution and only a small part of it is relevant for the rating.

Our approach is based on Convolutional Neural Networks (CNN), which are directly modelling the whole computer vision pipeline, from feature extraction to classification, with a single parameterized model. CNN models have led to a significant breakthrough in a lot of vision applications and showed promising results for medical tasks. However, the required size of training data is still an issue. Our CNN models are pre-trained on a large set of datasets of non-medical images, which prevents over-fitting to the small annotated dataset available in our case. We assume the selection of the probe in the data with just a single mouse click defining a point of interest. This is reasonable especially for slices acquired together with another sample. We sample image patches around the point of interest and obtain bilinear features by passing them through a CNN and encoding the output of the last convolutional layer with its second-order statistics.

Our approach ranked second in the Her2 contest held by the University of Warwick achieving 345 points compared to 348 points of the winning team. In addition to pure classification, our approach would also allow for localization of parts of the slice relevant for visual detection of Her2 over-expression.

Keywords: Digital pathology, automatic microscopy analysis, visual recognition, machine learning

https://doi.org/10.1515/cdbme-2017-0171

1 Algorithm overview

Our algorithm for automatic Her2 scoring is based on classifying regions of the slice into 4 different categories (0,1,2,3) corresponding to the possible scores. For classification, we use convolutional neural networks (CNN [3,4]), which are directly modelling the whole computer vision pipeline with a huge parameterized model. This has led to a significant breakthrough in a lot of vision applications and also showed promising results for the task under consideration.

Since the given slides have a high resolution and only a small part of it is related to the task, we assume the selection of the probe in the data with just a single click. This is necessary especially for slices acquired together with ``calibration probe", which itself is always scored with 3. Although this is reasonable for manual scoring, it is an avoidable obstacle for automatic scoring, which we right now circumvent with our one-click assumption and can be handled for future applications by simply excluding the calibration probe.

Bilinear features Our CNN architecture is based on AlexNet as introduced in [3] and widely used in computer vision. In particular, we use a CNN pre-learned from ImageNet. Although the ImageNet dataset is comprised of natural object categories not related to biomedical applications, it has been shown that pre-training CNNs is important to deal with the huge number of their model parameters. We can think about ImageNet pre-training for Her2 scoring as initializing the first layers such that typical structural elements in images can be detected, such as edges, corners, colored patches, and contour fragments.

***Corresponding author: Erik Rodner:** Corporate Research and Technology, Carl Zeiss AG, e-mail: Erik.Rodner@zeiss.com
Marcel Simon, Joachim Denzler: Computer Vision Group, Friedrich-Schiller-Universität Jena, Germany, e-mail: {marcel.simon, joachim.denzler}@uni-jena.de

Table 1: Results of our approach on the Her2 challenge dataset.

Approach	Leave-one-out Accuracy	Avg. Points	Conf. Assess.	Weighted Points
AlexNet, conv5 features	73%	13.80	0.73	10.96
AlexNet, conv5 with bilinear, 227x227 region sampling	75%	13.80	0.75	11.25
AlexNet, conv5+bilinear, 1024x1024 region without sampling	67%	13.46	0.67	10.10
ResNet50, pool5 features, 227x227 region sampling	71%	13.70	0.71	10.67
ResNet50, pool5, 1024x1024 region without sampling	60%	13.12	0.60	8.94
Approach	Accuracy on a **Fixed Split**	Avg. Points	Conv. Assess.	Weighted Points
AlexNet, conv5 with bilinear, 227x227 region sampling	70%	14.00	0.70	10.50
AlexNet, conv5 with bilinear, 227x227 region sampling, with fine-tuning	100%	15.00	1.00	15.00

Although distinguishing between score 3 and 0 is rather easy, discriminating between the other classes can be hard and fine-details in the cell structure are necessary. Therefore, the Her2 scoring class is related to the area of fine-grained recognition [5,6,8] as well as texture classification [7]. The recently successful concept of bilinear features [1,5] of this area is used in this paper. We randomly[1] crop several 227 x 227 patches at resolution level 1 (highest resolution is at level 0) within a 1024 x 1024 window around the click location. This technique allows us our approach to be rather robust with respect to the location, requiring only a rough selection of the main probe in the slice. Furthermore, it also allows for data augmentation during training, which is important given the small number of given slices in the contest.

A single 227 x 227 image patch is passed to the CNN and we obtain activations at layer conv5. The activations can be represented as a tensor $x \in \mathbb{R}^{(wxhxd)}$comprised of d-dimensional vectors in a w x h spatial grid. Bilinear features can now be computed as the Gramian matrix **G** calculated by summing up dyadic products along the spatial dimensions: $G = \sum x_{i,j} \cdot x_{i,j}^T$. The matrix **G** simply contains second-order statistics of the CNN features and have been shown to be extremely useful for fine-grained recognition tasks. As far as we know, they have not been used for biomedical applications before. In our current approach, we use signed square root and L_2 normalization of **G** to increase numerical stability of further processing steps [5].

After computing the features, we use a multi-class logistic regression classifier to differentiate between the 4 scoring classes.

To obtain a classification decision during testing for a single slide, we average all the classification scores of each random crop. We decided not to focus on directly predicting the percentage of tumor cells and simply use the mean tumor cell percentage seen in the training set for a particular scoring class as an estimate. The confidence is set to 100\%, since the probabilities achieved by our CNN are likely biased due to the small training set.

Fine-tuning the convolutional neural network The above algorithm achieved a leave-one-out recognition rate of 75% in our experiments (see next section for details). Using a pre-trained network from ImageNet is essential, since estimating millions of CNN parameters from scratch with only a couple of training examples is infeasible and leads to dramatic overfitting. However, the representations learned from ImageNet are likely not well-tuned to the Her2 scoring task.

Due to this reason, we fine-tune our CNN on the training task. Fine-tuning refers to the process of learned a neural network with back-propagation and a pre-trained network as initialization for the optimization. Hyperparameter values used during fine-tuning will be part of our source code release and are not described here for brevity.

We also experimented with a couple of other network architectures, such as the recent ResNet proposed in [2].

[1] We used a fixed random seed to ensure reproducibility.

However, we could not see any benefits in recognition performance. These ideas might be more reasonable if more training data is available or the scoring scheme is more complex.

Table 2: Results of the Her2 competition ranked by weighted points.

Rank	Approach	Avg. Points	Conf. Assess.	**Weighted Points**
1	VISILAB CNN	13,66	0.84	12.43
2	Ours (Deep bilinear features)	13.21	0.82	12.32
3	HUANGCH	13.48	0.81	11.99
4	MTB NLP	13.93	0.82	11.99
5	Team Indus	14.38	0.78	11.48
(13 further submissions)				

2 Experimental results

The dataset of the Her2 competition[2] consists of high resolution slices and is split into training and test data. While there are ground-truth labels for the 52 training slices, the labels for the 28 test slices are only known to the competition organizers. We hence evaluate different versions of our algorithms using leave-one-out accuracy on the training set. The best performing algorithm was then used for our submission in the competition.

The results using a pre-trained and non-fine-tuned CNN are shown in the upper half of Table 1. We observed an advantage of the simpler AlexNet [3] compared to the more complex ResNet [2] and an advantage of bilinear features compared to raw CNN activations. Our evaluation of the fine-tuning is given in lower half of Table 1 using a single split of the given labeled data into training and validation. Leave-one-out evaluation was not applicable here due to the increased computation demands of fine-tuning. As can be seen, the accuracy easily goes up to 100%. Although overfitting might be the case, the results also suggest good results on the test set.

The results of the competition are shown in Table 2. As can be seen, our algorithm ranked second among 18 submissions with a weighted point score of 345. The approaches of the other teams will be presented in an upcoming journal article.

3 Conclusions

We briefly described our entry to the Her2 scoring contest, which is based on convolutional neural networks and bilinear features. Our algorithm was implemented in python using the caffe-framework[3]. We plan to release the source code to allow for reproducibility of our results and it's use for automatic scoring.

It is important to note that our approach also allows for localization of relevant Her2 scoring areas. An interesting property which is beyond the scope of the contest, but perfectly suitable for semi-automatic scoring applications that allow feedback to medical experts.

Acknowledgment: The authors thank Nvidia for GPU hardware donations.

Author's Statement

Research funding: Part of this research was supported by grant RO 5093/1-1 of the German Research Foundation (DFG). Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent is not applicable. Ethical approval: The conducted research is not related to either human or animals use

References

[1] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. arXiv preprint arXiv:1511.06062, 2015.
[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. arXiv preprint arXiv:1603.05027, 2016.
[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.
[4] B Boser Le Cun, John S Denker, D Henderson, Richard EHoward, W Hubbard, and Lawrence D Jackel. Handwritten digit recognition with a back-propagation network. In Advances in neural information processing systems. Citeseer, 1990.
[5] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear CNN models for fine-grained visual

[2] http://www2.warwick.ac.uk/fac/sci/dcs/research/tia/her2contest/

[3] http://caffe.berkeleyvision.org

recognition. In Proceedings of the IEEE International Conference on Computer Vision, pages 1449–1457, 2015.

[6] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In International Conference on Computer Vision (ICCV), 2015.

[7] Marcel Simon, Erik Rodner, and Joachim Denzler. Fine-grained classification of identity document types with only one example. In Machine Vision Applications (MVA), pages 126 – 129, 2015.

[8] Marcel Simon, Yang Gao, Trevor Darrell, Joachim Denzler, and Erik Rodner. Generalized orderless pooling performs implicit salient matching. arXiv preprint arXiv:1705.00487, 2017.