Towards Grading Gleason Score using Generically Trained Deep convolutional Neural Networks

Källén, Hanna; Molin, Jesper; Heyden, Anders; Lundström, Claes; Åström, Karl

Published in:
2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)

DOI:
10.1109/ISBI.2016.7493473

2016

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):
Källén, H., Molin, J., Heyden, A., Lundström, C., & Åström, K. (2016). Towards Grading Gleason Score using Generically Trained Deep convolutional Neural Networks. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (pp. 1163-1167). IEEE - Institute of Electrical and Electronics Engineers Inc.. https://doi.org/10.1109/ISBI.2016.7493473

Total number of authors:
5

TOWARDS GRADING GLEASON SCORE USING GENERICALLY TRAINED DEEP CONVOLUTIONAL NEURAL NETWORKS

Hanna Källén[⋆], *Jesper Molin*[†‡§], *Anders Heyden*[⋆], *Claes Lundström*[†§], *Kalle Åström*[⋆]

[⋆]Centre for Mathematical Sciences,
Lund University
[†]Center for Medical Image Science and Visualization,
Linköping University
[‡]t2iLab, Chalmers University of Technology
[§]Sectra AB

ABSTRACT

We developed an automatic algorithm with the purpose to assist pathologists to report Gleason score on malignant prostatic adenocarcinoma specimen. In order to detect and classify the cancerous tissue, a deep convolutional neural network that had been pre-trained on a large set of photographic images was used. A specific aim was to support intuitive interaction with the result, to let pathologists adjust and correct the output. Therefore, we have designed an algorithm that makes a spatial classification of the whole slide into the same growth patterns as pathologists do. The 22-layer network was cut at an earlier layer and the output from that layer was used to train both a random forest classifier and a support vector machines classifier. At a specific layer a small patch of the image was used to calculate a feature vector and an image is represented by a number of those vectors. We have classified both the individual patches and the entire images. The classification results were compared for different scales of the images and feature vectors from two different layers from the network. Testing was made on a dataset consisting of 213 images, all containing a single class, benign tissue or Gleason score 3-5. Using 10-fold cross validation the accuracy per patch was 81 %. For whole images, the accuracy was increased to 89 %.

Index Terms— Prostate cancer, Gleason Score, Deep Learning, Convolutional Neural Networks

1. INTRODUCTION

The Gleason grading system is a widely used classification system of malignant prostate adenocarcinomas based on the growth patterns of the cancer cell population [**?**]. In the current revision from the 2005 consensus meeting [**?**], visually detectable malignant growth patterns are organized into three main groups (3,4,5), which are summarized into overall scores based on patient outcome. The meeting also recommended that benign patterns in group 1 and 2 should not be reported separately. In the newly proposed revision [**?**], the same growth patterns should be detected but are now organized into new overall scores. This makes it possible to develop the same image analysis system for both revisions, by just changing the way that the scores are summarized and grouped.

Automatic image analysis methods for Gleason grading have already been proposed. Doyle et al. [**?**] described an approach using hand-crafted features based mainly on detecting individual nuclei, whereas both Gorelick et al. [**?**] and Jacobs et al. [**?**] used features derived from super-pixels. Another method was proposed in [**?**] where histograms of SIFT-features were used to classify the images.

With the advent of deep learning techniques [**?**], it might be possible to reach high classification accuracy without using hand-crafted features, since the features can be derived during the training. One important technique is convolutional neural networks (CNN), [**?**, **?**]. These networks consist of multiple layers with different functions. The first layer performs convolution on the different color channels of the image, following layers consist of interleaved subsampling and convolutional layers. A subsampling layer reduces the spatial resolution and a convolutional layer combines information using different kernels. With each new layer, a network of feature vectors that represent the images is produced. The length and number of these vectors can either increase or decrease with each layer depending on the design of the network.

In this paper we used a pre-trained CNN, trained on a dataset consisting of photographic images, and applied it on pathology images, similar to the idea presented in [**?**]. For this purpose the classification step in the network was removed and replaced with other machine learning techniques to classify the feature vectors extracted from the network.

The algorithm development that this paper describes is a part of a larger project envisioning a semi-automatic human-computer system that pathologists could use to increase the efficiency and accuracy when reporting the Gleason grade, which implies that both the resolution and the speed of the algorithms could be equally important as the accuracy. In this paper, we evaluate the overall accuracy on a set of images as well as the accuracy when the images are divided into small patches, while keeping in mind that we want to keep the processing time as low as possible.

2. MATERIAL

During the development phase, self-annotated images generated by the TCGA Research Network[1], were used. The algorithm was then evaluated using cross-validation on an independent set of images that was used by Lippolis [**?**]. The images came from Beaumont Hospital in Dublin, Ireland and PathXL Ltd in Belfast, UK and consisted of homogeneous single class images classified by one or more pathologists. In total we had 52 images of benign glands, 52 images of Gleason grade 3, 52 images of Gleason grade 4 and 57 images of Gleason grade 5. The images were scanned in 40x magnification, which were downsampled to both 10x and 5x magnification to reduce the processing time and investigate how the scale of the images affects the classification result.

3. METHODS

The analysis method can be divided into three parts: Feature extraction, patch classification, and whole image classification.

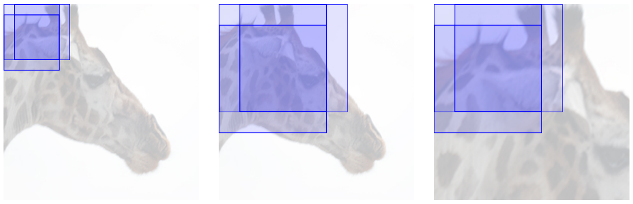
3.1. Feature Extraction

To extract the initial features, an pre-trained convolutional neural network, OverFeat, was used [**?**]. Overfeat is a 22-layer network in several stages. The first stages include convolutional layers followed by a piecewise linear function, defined as $f(x) = \max(0, x)$, and, in some of the stages, max-pooling layers. Later stages include fully connected layers and classification layer but these are not interesting for our purpose. A summary of the first 5 stages in the fast version of the network is shown in Table **??**, the table shows the size of the windows used for convolution and max-pooling. In the first two stages the convolution is performed only on valid pixels, later stages use zero-padding. The last column shows the length of the feature vectors that the different stages outputs.

We have extracted features from both layer 9 and layer 16 corresponding to the output from stage 3 and stage 5 respectively. In layer 9, a window of 87 × 87 pixels was used to compute the feature vector. A new feature vector was computed for the square 16 pixels to the right of the first one, a schematic image of this is shown in Figure **??**. This way, we got several feature vectors representing the image and each feature vector was represented by 512 features. The number of feature vectors depends on the size of the input image. In layer 16, these windows were larger, 167 × 167 pixels, with 32 pixels between the squares, see Figure **??**, and the feature vectors consist of 1024 features each. The window size of layer 9 is approximately half of the size of the window in layer 16, this enables the comparison of the effect depth while maintaining the same spatial resolution if the we feed the layer 9 version with a 5x image and and the layer 16 version with a 10x image, see Figures **??** and **??**.

	convoultion	maxpool	depth
stage 1	11×11 window, 4×4 stride	2×2 window, 2×2 stride	96
stage 2	5×5 window, 1×1 stride	2×2 window, 2×2 stride	256
stage 3	3×3 window (full), 1×1 stride	none	512
stage 4	3×3 window (full), 1×1 stride	none	1024
stage 5	3×3 window (full), 1×1 stride	2×2 window, 2×2 stride	1024

Table 1: Summary of the first 5 stages in the OverFeat network, with the window sizes used for convolution and max-pooling. The last column shows the layer depth after each stage.



(a) Layer 9 in 5x (b) Layer 16 in 5x (c) Layer 16 in 10x

Fig. 1: Window sizes, image magnifications, and step sizes used in our experiments. Note that feature vectors in (a) and (c) covers the same spatial location, and that (b) and (c) use the same window size.

3.2. Patch classification

Two different classifiers, Random Forest [**?**] and Support Vector Machines [**?**] were used to classify the feature vectors obtained by OverFeat.

In Random Forest an ensemble of decision trees is trained from the training data. Each tree uses a number of randomly chosen features to build the decision tree. The classification result depends on the number of trees in the forest and the number of features used.

In Support Vector Machines, SVM, a separating hyperplane that separates the classes in the training set is found.

[1] http://cancergenome.nih.gov/

A drawback of the SVM is that it is a binary classifier, so it has to be modified to support multiclass classification. Here we have used the one-versus-all strategy [?]. This is done by training four different classifiers and combining them. First, we build a classifier for benign tissue versus Gleason grade 3, 4 and 5, then another one for Gleason grade 3 versus benign tissue and Gleason grade 4 and 5 and so on. To classify a new vector we look at the distances between the separating hyperplane and the feature vector for all models. In the classifier benign versus all other classes the vectors that were classified as benign will have positive distances and the vectors that were classified as one of the other classes will have negative distances. To classify a vector we compute the distance to the separating plane for all models and choose the class that has the largest positive distance to the plane. Different kernel functions can be applied to the data to make the SVM classifier non-linear.

3.3. Classification of whole images

The patch classification above classifies all individual vectors in the images, where the vectors corresponds to patches of size 87 × 87 or 167 × 167 pixels. To classify the whole image we let all patches in an image vote for the different classes and the class with the highest number of votes is chosen.

4. EXPERIMENTS

The proposed method was evaluated for different resolutions of the images, 5x and 10x magnification. For 10x magnification, only features obtained from layer 16 were evaluated. For 5x magnification, features from both layer 9 and layer 16 were investigated. The area used to compute a feature vector in layer nine at 5x magnification roughly corresponds to the area used to compute a feature vector in layer 16 at 10x magnification.

We used 10-fold cross validation to evaluate the random forest and the support vector machines. The number of trees and the number of variables used was optimized during the cross validation of random forest. Different kernel functions were applied to the SVM classifier and the one with the lowest test error was chosen. The kernel function that performed the best differs between different experiments, but it is always either a first or second order polynomial. Figure **??** shows the mean training and mean test errors from the cross validation for different magnifications and layers. The plot shows the errors for the random forest classifiers and the error for the best kernel functions of the support vector machines classifier. The training error is shown in blue and the test error in red.

Confusion matrices for the the best classifier for the different magnifications and layers are shown in Table **??** - **??**. Table **??** shows the result for 10x magnification and layer 16, Table **??** the result for 5x magnification and layer 9 and Table **??** the result for 5x magnification and layer 16.

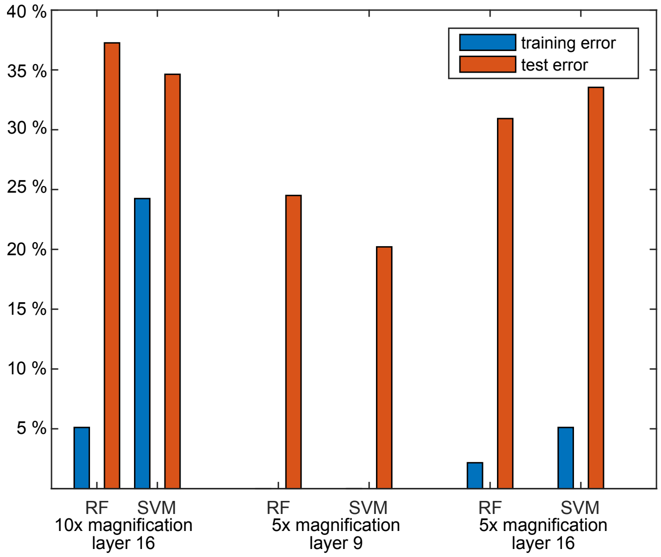


Fig. 2: Training and test error per small patch for both random forest and support vector machines for different magnifications and layers. Some of the training errors are very close to zero which could be an indication of overfitting.

		estimated class			
		benign	3	4	5
true class	benign	4812	843	665	511
	3	1440	2167	1635	1126
	4	474	738	4579	1361
	5	269	438	778	9097

Table 2: Confusion matrix for 10x magnification, layer 16, classified with support vector machines. The matrix shows the classified patches. The overall error is 33.2 %.

		estimated class			
		benign	3	4	5
true class	benign	3987	632	194	78
	3	716	2620	696	353
	4	213	636	4099	306
	5	58	239	176	7772

Table 3: Confusion matrix for 5x magnification, layer 9, classified with support vector machines. The matrix shows the classified patches. The overall error is 18.9 %.

For the binary classification of benign tissue versus cancerous tissue we discovered a true positive rate of 94.5 % and a false negative rate of 5.5 % using Table **??**.

Figure **??** shows the training and test errors for different magnifications and layers when whole images were classified. The plot shows the errors for the random forest classifiers and the error for the best kernel functions of the support vector machines classifier. The training error is shown in blue and the test error in red.

		estimated class			
		benign	3	4	5
	benign	774	50	45	38
true class	3	178	244	178	154
	4	56	82	539	278
	5	23	29	81	1493

Table 4: Confusion matrix for 5x magnification, layer 16, classified with random forest. The matrix shows the classified patches. The overall error is 28.1 %.

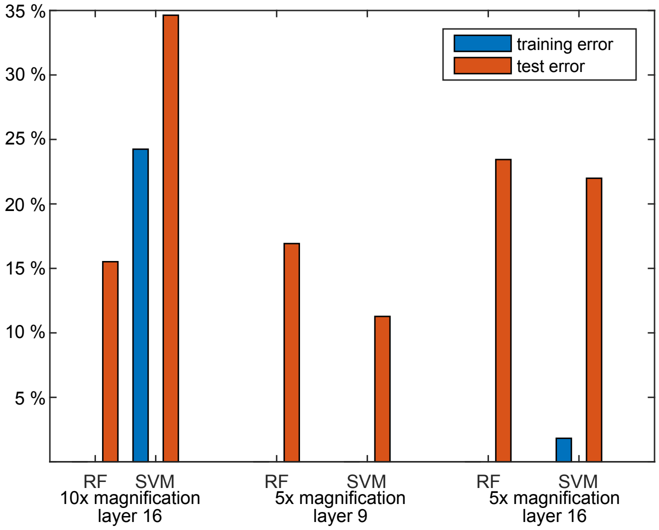


Fig. 3: Training and test error per whole image for both random forest and support vector machines for different magnifications and layers. Also here some of the training errors are very close to zero.

Confusion matrices for the best classifier for the different magnifications and layers when whole images were classified are shown in Table **??** - **??**. Table **??** shows the result for 10x magnification and layer 16, Table **??** the result for 5x magnification and layer 9 and Table **??** the result for 5x magnification and layer 16.

		estimated class			
		benign	3	4	5
	benign	52	0	0	0
true class	3	9	26	5	12
	4	0	5	37	10
	5	0	0	3	54

Table 5: Confusion matrix for 10x magnification, layer 16, classified with random forest. The matrix shows the classified images. The overall error is 20.7 %.

For the binary classification benign tissue versus cancerous tissue in whole images we discovered a true positive rate of 96.3 % and a false negative rate of 3.7 % using Table **??**.

		estimated class			
		benign	3	4	5
	benign	52	0	0	0
true class	3	4	40	3	5
	4	2	6	42	2
	5	0	0	1	56

Table 6: Confusion matrix for 5x magnification, layer 9, classified with support vector machines. The matrix shows the classified images. The overall error is 10.8 %.

		estimated class			
		benign	3	4	5
	benign	47	4	1	0
true outcome	3	7	24	14	7
	4	4	8	34	6
	5	0	3	3	51

Table 7: Confusion matrix for 5x magnification, layer 16, classified with support vector machines. The matrix shows the classified images. The overall error is 26.8 %.

5. CONCLUSIONS AND FUTURE WORK

As a first step towards a semi-automatic tool for analyzing prostate biopsies we have presented a method to automatically classify images into benign tissue and Gleason score 3-5. The framework perform with an accuracy of 81.1 % when analyzing small patches of the image, retaining the spatial resolution of the classification. When classifying entire images the accuracy was 89.2 %. This level of accuracy is on the same level as previous work, but without using hand-crafted features or other pre-processing of the images. In the future, it would be very interesting to see if a network trained on pathologiy images could perform even better.

6. ACKNOWLEDGEMENTS

Thanks to Giuseppe Lippolis for sharing this dataset with us. We would also like to thank Beaumont Hospital in Dublin, Ireland and PathXL Ltd in Belfast, UK for annotating the images. This work was supported by VINNOVA (2013-03906) and the Swedish Research Council (2011-4138).