Generalized multiple instance learning for cancer detection in digital histopathology*

Jan Hering[1] and Jan Kybic[1][0000−0002−9363−4947]

Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic, {jan.hering, kybic}@fel.cvut.cz

Abstract. We address the task of detecting cancer in histological slide images based on training with weak, slide- and patch-level annotations, which are considerably easier to obtain than pixel-level annotations. we use CNN based patch-level descriptors and formulate the image classification task as a generalized multiple instance learning (MIL) problem. The generalization consists of requiring a certain number of positive instances in positive bags, instead of just one as in standard MIL. The descriptors are learned on a small number of patch-level annotations, while the MIL layer uses only image-level patches for training.
We evaluate multiple generalized MIL methods on the H&E stained images of lymphatic nodes from the CAMELYON dataset and show that generalized MIL methods improve the classification results and outperform no-MIL methods in terms of slide-level AUC. Best classification results were achieved by the MI-SVM(k) classifier in combination with simple spatial Gaussian aggregation, achieving AUC 0.962.
However, MIL did not outperform methods trained on pixel-level segmentations.

Keywords: Multiple-instance learning · histopathology image classification · computer-aided diagnosis

1 Introduction

Training state-of-the-art deep-learning methods in computer-aided diagnosis (CAD) often requires a large image database with pixel-level annotations [9]. When provided with such data, deep-learning computer-aided diagnosis (CAD) methods are the state-of-the-art and can reach or surpass human expert performance. However, obtaining such precise manual annotations is tedious and expensivein terms of time and resources. Therefore, there is a lot of interest in methods capable of learning from weak annotations, such as image or patient level labels. This data, e.g. whether a patient is healthy or not, can be often extracted from the hospital information system automatically, with no or very little additional cost.

* The project was supported by the Czech Science Foundation project 17-15361S and the OP VVV funded project "CZ.02.1.01/0.0/0.0/16_019/0000765 Research Center for Informatics."

One popular class of weakly-supervised learning methods is Multiple-Instance Learning (MIL), which considers image as a collection (bag) of instances (pixels or pixel regions) and requires only image-level labels for training [7]. In standard MIL, a bag is positive iff at least one of its instances is positive.

The task of the CAMELYON challenge [9] is to detect metastases in stained breast lymph node images (see examples in Fig. 1 and 2). Each image is to be assigned a score between 0.0 and 1.0 measuring the likelihood of containing a tumor. The best-ranked submissions use a two-stage approach [2]. First a convolutional neural network (CNN) is learned in a fully-supervised manner to classify rectangular patches, yielding a *tumor probability map*. The second, *aggregation stage*, classifies the whole image based on geometrical properties of detected regions in the prediction map [2].

In this work, we use a CNN only to extract patch descriptors, which are then considered as instances for the MIL approach to classify images. The fact that an image (bag) is positive (contains cancer) iff any of its patches (instances) is positive corresponds exactly to the MIL formulation. However, due to the high number of patches and imperfections of the patch (instance) classifier, applying the standard MIL methods is very sensitive to false positive detections. We alleviate this problem by applying the generalized MIL [6] method, increasing the number of required positive instances for positive bags. This correspond to common histopathological guidelines, where the size of the lesion is one of the important factors of the classification.

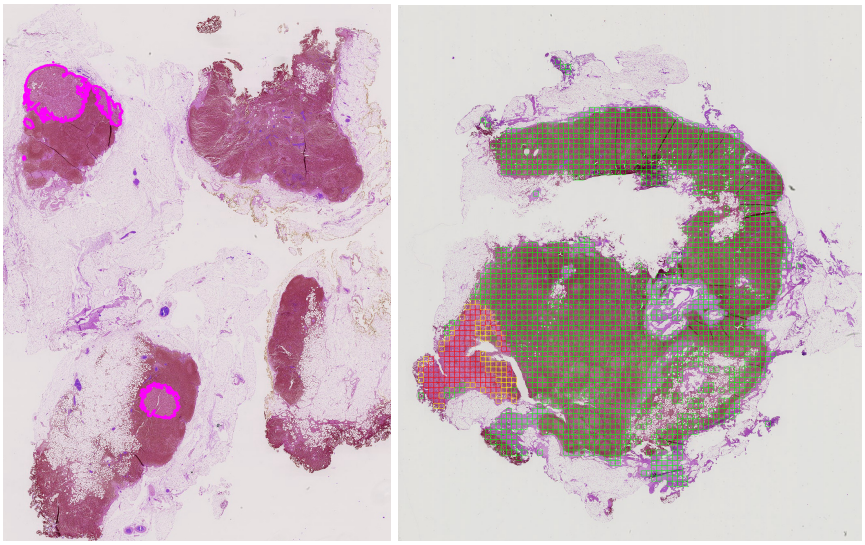
Existing methods combining MIL and CNN are mainly based on region proposals, like R-CNN [11, 5]. However, because of high memory consumption, they have only been applied to much smaller images than ours [12, 4].

2 Method

Let us describe the basic sequential blocks of the proposed method — patch descriptor calculation (Section 2.2), generalized MIL learning from image-level annotations (Section 2.3), and spatial aggregation (Section 2.4). We also describe alternative techniques, used as a baseline.

2.1 Patch extraction

We operate on 256×256 pixel square image patches extracted from the whole slide image (WSI) at the 20× magnification level. We use a random forest classifier on color channels of the down-sampled images to distinguish tissue and background. A patch is used only if it contains at least 80% of tissue. The patch label y_i is set to *tumor* ($y_i = 1$) if at least 60% of its area is tumor tissue and to *normal* ($y_i = -1$) if at most 10% of the tissue are from tumor class. The remaining patches are omitted during training.



(a) (b)

Fig. 1. *(a)* Example whole-slide image (WSI) from the CAMELYON'16 dataset. Tumor annotation boundaries are shown in magenta. *(b)* Another WSI with superimposed tissue patch boundaries — green for healthy, red for tumor — based on human expert annotations. Indeterminate (mostly boundary) yellow patches will not be used. Non-tissue patches (not-shown) were determined automatically.

2.2 Patch descriptors

The goal of this step is to provide a low dimensional descriptor for each patch. The descriptor is learnt from a limited amount of patch-level labels, which we obtain by aggregating the pixel-level segmentations provided by the CAMELYON'16 dataset. It would also be possible to ask the expert to annotate the patches directly, which would be much easier than to create full pixel-level annotations. The hope is that even when trained on limited data, the descriptor gives us a useful embedding for the MIL block. Note that the pixel level segmentations are not used directly at all and this is the only place where patch-level segmentations are needed.

We use a VGG'16 deep network, variant D, with 16 weight layers and the binary cross-entropy loss function [13]. We apply implicit color-normalization by adding a color-normalization layer [10]. We used the following augmentation techniques — random crop to the input size of 224×224, 90°, 180° and 270° rotations as well as random up-down and left-right flips. The accuracy of the predicted patch labels is shown as 'CNN' in Table 1.

We then insert another fully-connected (FC) layer before the output layer, which reduces the dimensionality at the input of the last layer from 4096 to some much smaller D ($D = 32$ was used in the experiments). This augmented CNN

is retrained and the output of the added intermediary layer is used as a patch descriptor f_{CNN}.

As an alternative to the CNN last layer, we have also trained a gradient boosting XGBoost classifier [3] (shown as 'CNN + XGBoost' in Table 1) to predict patch labels from patch descriptors f_{CNN}.

2.3 Generalized MIL

The next block takes the f_{CNN} patch descriptors and produces both patch and image level labels. We have taken two most promising generalized MIL methods based on earlier experiments [6]. Both methods evaluate a scalar patch scoring function $\phi(f_{\mathrm{CNN}})$, which is thresholded to obtain the patch labels $\hat{y}_i$. We introduce a parameter k, the minimum number of patches assumed to be positive in a positive image, with $k = 1$ corresponding to the standard MIL formulation [1].

The first method, MI-SVM(k) [6], is an extension of the MI-SVM classifier [1]. It acts iteratively and repeatedly trains an SVM classifier that calculates ϕ using all instances from negative bags and selected instances (the *witnesses*) from positive bags. After each iteration, the set of witnesses is recalculated by taking the top k positive instances from each positive bag. The iteration ends when the set of witnesses does not change.

The second method, MIL-ARF(k) [6], is an extension of MIL-ARF [8]. It implements ϕ using a random forest classifier and applies deterministic annealing. In each iteration, the instances are first classified using the current instance-level classifier. Then the instance labels are modified to enforce at least k positive instances in each positive image and no positive instances in any negative image. The patch labels are randomly perturbed, with probability decreasing as a function of the iteration number. The random forest is incrementally relearned from the updated labels and the process is repeated until convergence.

Finally, image labels are obtained by thresholding the number of positive patches with k. The results of this method are shown in the 'MIL' column in Table 1.

2.4 Spatial aggregation

The patch-level output from all previously described methods is fed into an aggregator to obtain an image-level prediction. In the simplest but surprisingly efficient case (denoted 'Gaussian' in Table 1), we project patch prediction to pixels to obtain a pixel-level tumor probability map T, apply Gaussian smoothing to obtain $T_\sigma(\mathbf{x}) = G_\sigma * T$ with $\sigma = 2\sqrt{2}$, take a maximum

$$m_\sigma = \max_{\mathbf{x}} T_\sigma(\mathbf{x}) \tag{1}$$

and threshold, $m_\sigma > \tau$, where τ is the threshold parameter, which can be user-specified or learned from data by cross-validation.

A more sophisticated procedure [2] consists of combining the maximal Gaussian response m_σ for $\sigma \in \sqrt{2}[1, 2, 4]$ with properties of the largest 2-connected

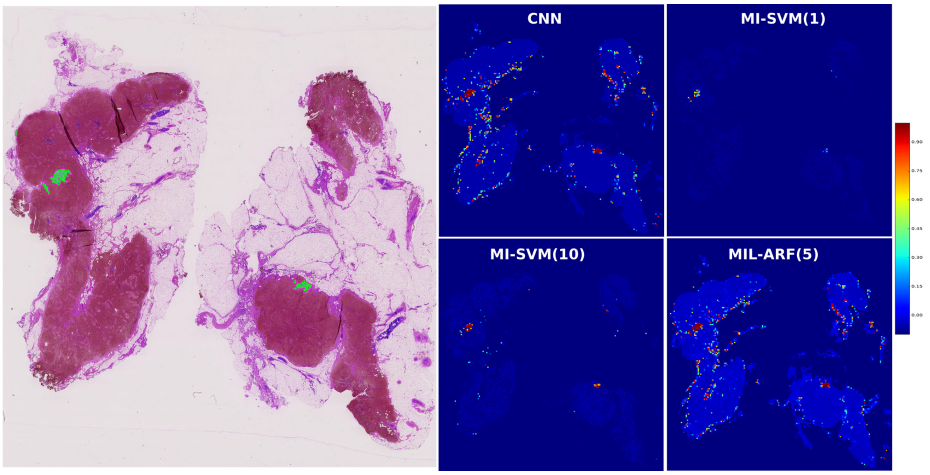


Fig. 2. Example image from the CAMELYON dataset and tumor prediction maps computed by the CNN, the MI-SVM($k = 1$), MI-SVM($k = 10$) and MIL-ARF($k = 5$). The output is scaled between 0.0 and 1.0 (tumor tissue) with the indicated color map. Ground truth annotations are shown as green overlay over the original image.

component for each binary image $T_\sigma > t$ for thresholds $t \in \{0.5, 0.8\}$. The properties are area, extent, solidity and eccentricity, as well as the mean of T_σ within the area. A random-forest classifier is trained on the resulting 30 dimensional descriptors. This is denoted as 'RF' in Table 1.

Figure 2 shows the tumor prediction maps (patch-scores) computed by the VGG net (CNN) and the various MIL methods. We see that MI-SVM with $k = 10$ is closest to the ground truth annotation.

3 Experiments and Results

Experiments were performed on all available training (159 healthy, 110 tumor), respective testing (49 healthy, 78 tumor) whole-slide images as provided within the CAMELYON'16 challenge. In all cases, parameters were found using cross-validation, taking out 20 % of the training dataset for validation.

We trained the VGG16 network with descriptor dimensionality $D = 32$. We use $k = 1, 5, 10, 15$ for MI-SVM(k) and $k = 1, 5, 10$ (required number of positive instances) for the MIL-ARF(k).

The first experiment evaluates the effect of initialization on the two MIL methods. We have taken a fraction ($l \in \{0, 0.25, 0.5, 1.0\}$) of the patch labels in the training set and used them to initialize the generalized MIL classifiers in their first iterations, initializing the remaining patch labels by the bag labels. We can see in Fig. 3 that unlike MI-SVM, MIL-ARF is very sensitive to this type of initialization and that generalized MI-SVM(k) with $k > 1$ provide robust results even when all instance labels are initialized with bag labels.

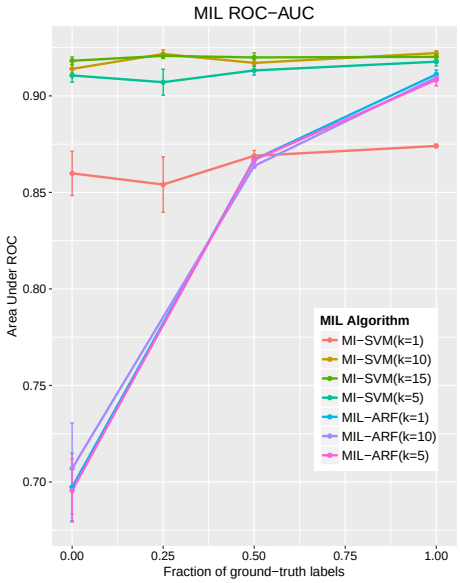


Fig. 3. Bag classification score as a function of the fraction of revealed training labels during initialization. Each line represents the mean ROC-AUC score of the MIL classifier with whiskers denoting ±SD.

The experiment also evaluates the effect of the parameter k (Figure 3). For MI-SVM(k), the parameter k affects the overall performance significantly. The performance improves with higher values with an optimum around $k = 10$ on our data.

The main results are summarized in Table 1, while the ROC curves are shown in Figure 4. The pure CNN approach yielded the best patch-level AUC score of 0.973, which resulted in an image-level AUC of 0.941 after the aggregation phase. Plugging-in the supervised XGBoost classifier yielded almost the same results. Also the MIL-ARF(k) for both $k = 1$ and 5 reached similar patch-level AUC.

In terms of image level accuracy, MI-SVM(k) with $k = 10$ performed the best. When considering achieved ROC-AUC for all initialized fractions l together, MI-SVM(k) with $k = 15$ performed in the most consistent way.

Interestingly, combining patch-level predictions from MI-SVM(k) and Gaussian aggregation to obtain image-level results outperformed all other variants.

To evaluate whether using MIL can help reduce the number of required images with detailed (pixel or patch level) annotations, we trained the CNN on patches from 25% of available training images. It turns out that while using MIL helps, it cannot yet compensate for the lack of detailed annotation. After Gaussian spatial aggregation, we get an image-level AUC of 0.831 for the CNN only and 0.838 for MI-SVM($k = 5$), the AUC for the direct output of the MIL ('MIL' score) is 0.815.

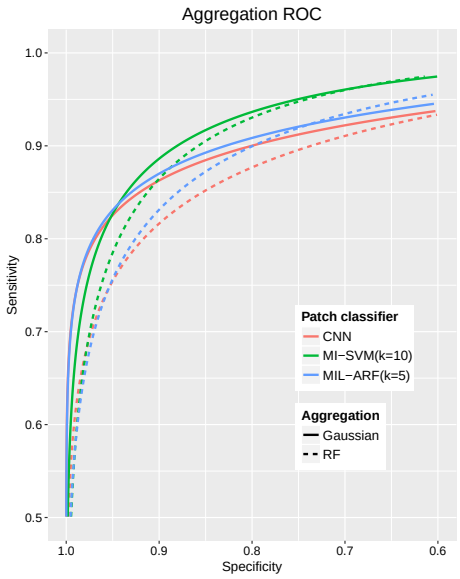


Fig. 4. ROC curves of the spatial aggregation phase. Both *Gaussian* (solid) and *RF* (dashed) [2] aggregation outcomes are shown for the MI-SVM(k=10), MIL-ARF(k=5) and the CNN classifiers.

Interestingly, the more sophisticated aggregation procedure [2] never outperformed the simpler but more robust aggregation based on Gaussian smoothing.

4 Conclusion

We have demonstrated that generalized MIL approaches can boost the performance of fully-supervised methods in the task of classifying histopathology images. We have also shown that it is possible to reach a good level accuracy by training only on a limited amount of patch data. The MI-SVM(k) method was shown to be robust to label initialization.

The direct output of the generalized MIL methods in terms of image-level AUC was lower than for the CNN methods with no MIL, but the high specificity, and thus minimal amount of false positives (see Fig. 2), enabled an important improvement through spatial aggregation, especially in the high specificity regime. The best image-level result (AUC 0.962) is comparable with the pathologist interpreting the slides in the absence of time constraints [2].

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support Vector Machines for Multiple-Instance Learning. In: Advances in Neural Information Processing Systems. pp. 561–568 (2002)

Algorithm	Image-level AUC			Patch-level metrics		
	MIL	Gaussian	RF	AUC	Specificity	Sensitivity
MI-SVM, k=15	0.920	0.945	0.941	0.871	0.991	0.707
MI-SVM, k=10	**0.923**	**0.962**	**0.953**	0.862	**0.992**	0.685
MI-SVM, k=5	0.915	0.955	0.944	0.847	0.990	0.647
MI-SVM, k=1	0.874	0.934	0.904	0.835	0.979	0.529
MIL-ARF, k=1	0.911	0.944	0.931	0.972	0.952	0.916
MIL-ARF, k=5	0.908	0.943	0.931	0.972	0.951	0.917
CNN	*n/a*	0.941	0.935	**0.973**	0.986	0.815
CNN + XGBoost	*n/a*	0.943	0.928	0.972	0.951	0.916

Table 1. Patch- and image-level classification scores. The image-level AUC is either a direct output (*MIL*) or the result of spatial aggregation with global features (*Gaussian*) or with a random forest classifier (*RF*) [2]. Best results in each column are shown in bold.

2. Bejnordi, B.E., Veta, M., van Diest, P.J., van Ginneken, B., et al.: Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA **318**(22), 2199–2210 (Dec 2017). https://doi.org/10.1001/jama.2017.14585
3. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 785–794. ACM Press, San Francisco, California, USA (2016). https://doi.org/10.1145/2939672.2939785
4. Das, K., Conjeti, S., Roy, A.G., Chatterjee, J., Sheet, D.: Multiple instance learning of deep convolutional neural networks for breast histopathology whole slide classification. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 578–581. IEEE, Washington, DC (Apr 2018). https://doi.org/10.1109/ISBI.2018.8363642
5. Durand, T., Thome, N., Cord, M.: WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks. In: 2016 IEEE CVPR. pp. 4743–4752 (Jun 2016). https://doi.org/10.1109/CVPR.2016.513
6. Hering, J., Kybic, J., Lambert, L.: Detecting multiple myeloma via generalized multiple-instance learning. In: Proceedings of SPIE. p. 22. SPIE (Mar 2018). https://doi.org/10.1117/12.2293112
7. Kandemir, M., Hamprecht, F.A.: Computer-aided diagnosis from weak supervision: A benchmarking study. Computerized Medical Imaging and Graphics **42**, 44–50 (Jun 2015). https://doi.org/10.1016/j.compmedimag.2014.11.010
8. Leistner, C., Saffari, A., Bischof, H.: MIForests: Multiple-Instance Learning with Randomized Trees. In: Computer Vision – ECCV 2010, vol. 6316, pp. 29–42. Springer-Verlag Berlin, Berlin (2010)
9. Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., Manson, Q.F., Stathonikos, N., Baidoshvili, A., van Diest, P., Wauters, C., van Dijk,

M., van der Laak, J.: 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. GigaScience **7**(6) (Jun 2018). https://doi.org/10.1093/gigascience/giy065
10. Mishkin, D., Sergievskiy, N., Matas, J.: Systematic evaluation of convolution neural network advances on the Imagenet. Computer Vision and Image Understanding **161**, 11–19 (Aug 2017). https://doi.org/10.1016/j.cviu.2017.05.007
11. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Advances in Neural Information Processing Systems 28, pp. 91–99. Curran Associates, Inc. (2015)
12. Ribli, D., Horváth, A., Unger, Z., Pollner, P., Csabai, I.: Detecting and classifying lesions in mammograms with Deep Learning. Scientific Reports **8**(1), 4165 (Mar 2018). https://doi.org/10.1038/s41598-018-22437-z
13. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: International Conference on Learning Representations (2015)