U-Net adaptation for multiple instance learning

M V Kots and V S Chukanov
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

Abstract. Multiple instance learning (MIL) is a weakly supervised learning method where a single label is assigned to a group of instances. Recent advancement in neural networks makes it possible to achieve great results but the training requires many annotated examples which can be difficult to obtain. In case of medical imaging, such a method can theoretically provide voxel-level annotations basing on the image-level annotations. More precisely, taking a training dataset where each image is given with a label indicating a presence of a region of interest (ROI), the model can be trained to produce voxel labels of the object. We propose a modification of the U-Net architecture for image segmentation that can be trained on weakly labeled datasets and solves the MIL problem. The U-Net architecture is famous for being able to train on small number of samples which can further improve benefits of MIL approach. In this paper we also present results of experiments on synthetic data with investigation of stability to noise and medical images. Experimental results prove that suggested modifications improve the noise stability of the weakly supervised method.

1. Introduction
Over the last few years, deep learning has become a standard solution for various pattern recognition and computer vision problems. These methods usually imply supervised learning, which requires a large set of annotated samples. This is a major constrain for applying deep learning to problems where gathering many training examples may be very expensive or time consuming. This is especially important for medical imaging applications since annotations for the segmentation task are prepared manually by experts. This is not only wasteful, but may create ambiguity since two experts may disagree on labeling of some difficult examples.

Alternatively, weakly supervised methods may overcome this issue by introducing weak annotations that usually provide less information but can be obtained more efficiently. One of the popular approaches is to group training instances into sets (bags) and assign annotation to each bag. During training, the bag labels are available, but the labels of the instances are unknown. Often assumptions are made about the instance labels and their relationship with the bag labels. This group of methods called the multiple instance, or multi-instance learning (MIL) [1].

This paper presents a multiple instance approach to image segmentation based on a convolutional neural network. We used U-Net as a base architecture since it is able to train on small number of samples which can further improve benefits of MIL approach. We present a modification of U-Net architecture to improve segmentation quality on noisy data as well as experimental results on synthetic data and medical images of various organs.

The rest of this paper is organized as follows. The next section formalizes MIL problem. Section 3 presents literature review over MIL methods. Sections 4 describe suggested approach. Experiments are presented in Section 5, followed by a discussion in Section 6.

2. Problem statement
In multiple instance learning training, set \( X = \{(x_i, y_i) | x_i \in \mathbb{R}^n, y_i \in \{0, 1\}\} \) is divided into bags \( B_i = \{x_k | k \in 1..n_i\} \) with labels \( Y_i \in \{0, 1\} \). The relation between bag labels and instance labels follows from the standard assumption: bag is assigned with label \( Y_i = 1 \) if it contains at least one instance with \( y_i = 1 \), otherwise \( Y_i = 0 \) [1].
Various applications can use different semantics for bags and instances. In case of image segmentation instances, \( x_i \) are image pixels with labels \( y_i \in \{0, 1\} \) denoting a pixel relation to object of interest or background, bags \( B_i \) are images with labels \( Y_i \in \{0, 1\} \), which indicate whether the object of interest is present in the image [1].

3. Related work
A wide variety of the multi-instance learning classification methods can be roughly divided into two groups: bag-level and instance-level. The bag classification assigns a class label to a set of instances while the instance classification derives labels of samples from bag label. As it is shown in [2], the best bag-level classifier is rarely the best instance-level classifier, and vice versa. Recent advancements in bag classification allows for reliable quality in medical imaging [3], however, instance classification appears to be much more challenging [4]. In this review, we primarily focus on the instance classification since it is more relevant for image segmentation tasks.

Diverse Density is one of the earliest approaches for solving MIL problem [5]. General concept is to find a single representative instance from a training set. This instance has to be much closer to instances from positive bags rather than negative (density) and to be close to at least one instance from each positive bag. For a given point \( t \) in this space, a measure \( DD(t) \) is defined as the ratio between the number of positive bags which have instances near \( t \), and the distance of the negative instances to \( t \). The point of maximum Diverse Density should therefore correspond to the target concept. The target instance label is based on its distance to the representative instance.

The Diverse Density approach has many modifications. An example for the instance classification is EM-DD, an expectation-maximization algorithm based on Diverse Density which models instance labels hidden variables [6]. After an initial guess for the concept \( t \), the expectation step selects the most positive instance from each bag according to \( t \). The maximization step then finds a new concept \( t' \) by maximizing \( DD \) on the selected, most positive instances. The steps are repeated until the algorithm converges.

There are also a few regular supervised classifiers extended to work with MIL task. One example is m-SVM, an extension of the support vector machine, which attempts to find hidden labels of the instances under constraints posed by the bag labels [4, 7]. The positive region is delimited by a hyperplane with respect to standard assumptions. Another example is MCIL, which defines several clusters for positive instances [8, 9]. Each cluster is defined using AnyBoost. At each boosting iteration, one weak classifier is added to each cluster-specific strong classifier.

Recent success of convolutional neural networks for various image processing tasks resulted in a few algorithms have been proposed for training CNNs to solve multiple instance learning problem: MIL-FCN [10] and MIL-ILP [11]. Both rely on a CNN that produces one heat map per class label which can be used for spatial predictions. During training, heat maps are fine-tuned, and each pixel location in the heat maps is regarded as an instance. In that purpose, one MIL layer is added at the end of the network during the training phase which outputs the maximum patch-level prediction for each class. Note that these algorithms are designed for the bag-level classification. However, their goal is to fine-tune instance-level predictions, so a primary instance-level solution may lead to even better features.

4. Suggested approach
A key component of the suggested approach is MIL Pooling Layer (MPL), which aggregates instance probability distribution vectors into a bag feature vector [12]. It bridges MIL training data with conventional neural networks. Since it must be differentiable, there are a few choices, such as global max pooling or global mean pooling [12]. Both aggregate outputs for all instances in feature map across all feature maps, first operation picks the highest output while second picks an average. As shown in [13] global average pooling is beneficial for instance-level classification since its produces bag features related to all instance features rather than one discriminative feature.

CNN architecture is based on U-Net [14]. The network is based on the fully convolutional network and its architecture was modified and extended to work with fewer training images and to yield more precise segmentations. This allows for numerous successful applications to medical imaging. In order to apply this architecture to MIL segmentation task last convolution is followed by global average pooling and 1x1 convolution for training on bag-level annotations. During prediction step global average pooling is removed which allows to produce segmentation map as a weighed filter sum [12].
U-Net architecture was originally designed as supervised learning model for image segmentation. In order to adapt to MIL task, we are introducing several changes. MPL derives bag feature from the last convolution layer, therefore we increase the number of features on upsampling branch of network. To balance out increased number of parameters one had to remove a single convolution from each block on upsampling path. Batch normalization [15] is also proposed after each convolution as a regularization method. Network architecture is presented on figure 1. In the next section, we will demonstrate that suggested changes improve stability to noise.

**Figure 1.** Architecture of modified U-Net. Proposed U-Net modification: increased number of filters on upsampling path

## 5. Experiments
We have tested our approach on various image segmentation tasks: both synthetic and medical images. To estimate segmentation quality, 5-2 cross-validation is performed and DICE measure for each validation sample is calculated.

Generated synthetic dataset consists of grayscale images containing various geometric figures: circles, triangles and squares with random size, position and color. The MIL task for that dataset is to segment a circle. Images are generated in a way that approximately half of them contains single circle and another half doesn’t to make the training set balanced. Every image contains at most one circle which cannot be covered by other figures. Dataset contains 5000 samples, and two different versions with progressive complexity have been generated:

a. A single circle
b. One circle, triangle and square

Experiment results with unmodified model are shown in figure 2. Most of images from validation subsets are segmented with DICE metric value higher than 0.95 aside from several spikes which indicates that proposed methods allow to achieve high quality segmentation on these datasets.

**Figure 2.** Results of experiments on artificial images. Y-axis shows DICE value.
In order to test suggested approach against noisy data we use the same synthetic images as for the first experiment but also with Gaussian noise added. The noise variance has been chosen to be only 1% of the original signal. The results (figure 3) indicate that proposed approach is very sensitive to noise. Suggested modifications improve quality on noisy dataset as shown in figure 3(b).

Medical images for these experiments are provided by medical segmentation decathlon challenge 2018 (http://medicaldecathlon.com/). Segmentation quality measured over three datasets with different organs:

a. MRI Heart images. Chest images with labeled left atrium. Contains total of 2059 slices
b. CT liver images. Chest images with labeled liver. For testing purposes 10000 slices were chosen for cross-validation with equal balance between slices with or without ROI
c. CT spleen images. Chest images with labeled spleen. Testing images were chosen the same way as liver.

Experiment results with modified model are shown in figure 4. The experiment shows that proposed methods perform well on simple synthetic examples, but segmentation quality starts to decrease when image has a bigger variety in background, thus making it hard to apply to medical images segmentation.

6. Conclusion
In this paper we presented a multiple instance learning approach based on modified U-Net CNN architecture as well as experiment results with synthetic data and medical images. Results show that suggested method
demonstrates a good quality for grayscale synthetic examples and improved U-Net architecture allows better noise stability but results on medical images are more likely unacceptable since this approach proved to be incapable of learning complex features. In the future, we are planning to improve the quality of the proposed method by introducing segmentation within sliding window to focus the network on ROI more than on background features.

Acknowledgments
This work was supported by Philips Research.

References
[1] Dietterich T G, Lathrop R H and Lozano-Pérez T 1997 Solving the multiple instance problem with axis-parallel rectangles Artificial intelligence 89.1 31-71
[2] Vanwinckelen G, Fiers D and Blockeel H 2016 Instance-level accuracy versus bag-level accuracy in multi-instance learning Data mining and knowledge discovery 30.2 313-41
[3] Sudharshan P J, Petijean C, Spanhol F, Oliveira L E, Heutte L and Honeine P 2019 Multiple instance learning for histopathological breast cancer image classification Expert Systems with Applications 117 103-11
[4] Carbonneau M A, Cheplygina V, Granger E and Gagnon G 2018 Multiple instance learning: a survey of problem characteristics and applications Pattern Recognition 77 329-53
[5] Maron O and Lozano-Pérez T 1998 A framework for multiple-instance learning Advances in neural information processing systems 570-576
[6] Zhang Q and Sally A G 2002 Em-dd: an improved multiple-instance learning technique Advances in neural information processing systems 1073-1080
[7] Stuart A, Tsochantaridis I and Hofmann T 2003 Support vector machines for multiple-instance learning Advances in neural information processing systems 577-84
[8] Xu Y, Zhu J-Y, Chang E and Tu Z 2012 Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering Proc. IEEE Comput Soc Conf Comput Vis Pattern Recognit, (USA: Providence, RI) 964–71
[9] Cheplygina V, de Bruijne M and Pluim J P 2018 Not-so-supervised: a survey of semi supervised, multi-instance, and transfer learning in medical image analysis arXiv preprint arXiv:1804.06353.
[10] Pathak D, Shelhamer E, Long J and Darrell T 2014 Fully convolutional multi-class multiple instance learning arXiv preprint arXiv:1412.7144
[11] Pinheiro P O and Collobert R 2015 From image-level to pixel-level labeling with convolutional networks Proc. IEEE Comput Soc Conf Comput Vis Pattern Recognit, (Boston, MA, USA) 1713–1721
[12] Wann X, Yan Y, Tang P, Bai X and Liu W 2018 Revisiting multiple instance neural networks Pattern Recognition 74 15-24
[13] Zhou B, Khosla A, Lapedriza A, Oliva A and Torralba A 2016 Learning deep features for discriminative localization.” Proceedings of the IEEE conference on computer vision and pattern recognition 2921-2929
[14] Ronneberger, Olaf, Fischer P and Brox T 2015 U-net: convolutional networks for biomedical image segmentation International Conference on Medical image computing and computer-assisted intervention 234-241
[15] Ioffe S and Christian S 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift arXiv preprint arXiv:1502.03167