Multiple Instance Learning through Explanation by Using a Histopathology Example

Andrei Konstantinov, Lev Utkin
Peter the Great Saint-Petersburg Polytechnic University,
Saint-Petersburg, Russia
andrue.konst@gmail.com, lev.utkin@gmail.com

***Abstract*—A new method for solving the multiple instance learning (MIL) problem, which is based on ideas of the black-box model prediction explanation, is proposed. The explanation aims to show instances (pixels, patches) which have the highest contribution into the image (bag) classes and to automatically annotate instances in bags. Three ideas behind the method are used. First, the surrogate black-box model is implemented as the Siamese neural network which is trained on pairs of whole images. Second, patches in each image are changed by using their dynamic fill or noise. Third, noisy images are compared with the original image by using the Siamese neural network such that Euclidean distances between outputs of the network depending on the noise level form a shape function for every patch. The shape function is interpreted from its contribution into the image class. Numerical experiments with the real Breast Cancer Cell Segmentation dataset illustrate the method.**

I. Introduction

Many applied real-life machine learning problems can be successfully performed in the framework of the Multiple Instance Learning (MIL) [1]. MIL was introduced for drug activity prediction [2], and it aims to train a model using a set of weakly labeled data. In the original MIL, a training set consists of bags, labeled as positive or negative; and each bag includes many instances, whose labels are unknown. From the above statement of MIL, it can be regarded as a type of weakly supervised learning which covers many machine learning problems. Several surveys consider various MIL problem statements related to different applications, including medical imaging and diagnosis, tracking, computer vision, system safety, etc. [1], [3]–[8].

At the present time, there exist many MIL methods solving MIL tasks in their different statements, for example, the citation-kNN [9], mi-SVM and MI-SVM [10], the Multiple Instance Learning Convolutional Neural Network [11]–[13], Deep Attention Multiple Instance Survival Learning [7], MILD [14], ProtoMIL [15]. This is a small part of all MIL methods and models developed in the last years.

One of the important applied areas, where MIL can be viewed as a main and inherent tool, is the computational histopathology. The histopathology is a significant part of the disease approval because it aims to detect whether cancer exists or no. A common approach to the computational histopathology is the generation of digital images from glass microscope slides and then obtaining meaningful information from the images. The histology images are very large and often represented as a set of small parts (patches, cells). If to use the machine learning terminology, every histology image with a label indicating a disease, for example, cancer or non-cancer, can be viewed as a "bag" consisting of patches extracted from the image which are referred to as "instances". Depending on prior information available about labels of patches or whole images, four learning scheme are proposed [16]: the supervised learning when patches are annotated by the pathologist, for example, as cancerous or normal; the weakly supervised learning when image-level annotations are available, but patches of each image have to be annotated by a learning algorithm, but not by the pathologist; the unsupervised learning as a worse case when no labels are available for patches as well as for whole images; the transfer learning aiming to transfer knowledge from a source domain, for example, with annotated data, to another target domain, for example, with unannotated data.

We pay attention on the weakly supervised learning, i.e., we assume that bags (images) have class labels, but instances (individual segments, patches, subsets) do not. This type of learning is very common in medicine practice. It should be noted that the lack of labels for instances is a key peculiarity of MIL. From this point of view, MIL can be regarded as a type of the weakly supervised learning problem. A huge number of methods and models have been proposed in order to solve the problem of the instance annotation in the computational histopathology. These methods are comprehensively studied and discussed in several survey papers [6], [16]–[18].

We propose a new method which is based on ideas of the black-box model prediction explanation. Many new explanation methods have been developed in the last years due to requirements to explain the machine learning model predictions [19]–[23]. The problem is that most powerful machine learning models, for example, deep neural networks, are very complex. They are actually black boxes. Therefore, an explanation component has to be supplemented these models in order to give a user to understand the corresponding model predictions [24].

However, by applying the explanation, we are pursuing a double object. First, we aim to show features (pixels, patches) which have the largest contribution into the image class (cancer/non-cancer or malignant/benign). But the most important object is to use the explanation to automatically annotate instances in images. In order to solve the above tasks, we propose the following scheme.

1) A surrogate black-box model implemented as the Siamese neural network [25], [26] is trained on whole images (bags).

ISSN 2305-7254

2) Each bag is divided into a grid of patches which are changed by using their dynamic fill or noise. The procedure of the image fill is similar to the algorithm "Hide-and-Seek" [27], but not the same. As a result, we get new bags which are compared by means of the trained Siamese neural network with the initial "clean" bags. Results of comparison are distances between embeddings (outputs of the trained Siamese neural network) obtained for each "clean" bag and the corresponding changed noising bags.
3) By using the distances, we construct shape functions which show how the distance between embeddings depends on the noise values. The noise is defined by the image dynamic filling.
4) A rapid change of a shape function constructed for a patch indicates that the corresponding patch is important and is annotated in accordance with a class of the whole bag containing this patch.

It is important to note that the proposed scheme can be applied to various weakly supervised MIL problems. But we illustrate it on the histology images. We do not claim that the proposed method outperforms many existing approaches (see, for example, [28]). Its performance depends on the considered dataset. However, our numerical experiments show that it is comparable with other methods and can be applied to many problems. Moreover, its tuning may lead to outperforming results.

In summary, the following contributions are made in this paper:

1) We propose a new explanation and annotation MIL method in terms of the computational histopathology.
2) The method is illustrated by means of numerical experiments with real histological data.

The paper is organized as follows. A brief introduction to MIL, the Hide-and-Seek approach and to Siamese neural networks can be found in Section 2. The proposed method and the algorithm implementing it are described in Section 3. Numerical experiments are provided in Section 4. Concluding remarks can be found in Section 5.

II. Preliminary

A. MIL

It is supposed in MIL that bags, for example, images in a data set have class labels; however, instances (individual segments, patches, subsets) do not. The lack of labels for instances is a key peculiarity of MIL. From this point of view, MIL can be regarded as a type of the weakly supervised learning problem.

Let X be a bag defined as a set of feature vectors $X = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$. Each instance (i.e. feature vector) $\mathbf{x}_i$ in feature space $\mathcal{X}$ can be mapped to a class by some process $f : \mathcal{X} \to \{0, 1\}$, where the negative and positive classes denoted as $y_1, ..., y_m$ correspond to 0 and 1, respectively. Values $y_1, ..., y_m$ remain unknown during training. Generally, the number of instances N can vary for different bags. However, for simplicity purposes, we will assume that the number of instances m is the same for all bags. We will denote bags by capitals and instances by bold letters.

One of the important assumptions accepted in the MIL stems from the fact that all negative bags contain only negative instances, and that positive bags contain at least one positive instance [1]. Hence, the bag classifier $g(X)$ is defined by

$$g(X) = \begin{cases} 1, & \exists \mathbf{x} \in X : f(\mathbf{x}) = 1, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

It should be noted that the above rule is not a unique one. Sometimes, a threshold θ can be introduced to define the bag classifier $g(X)$, i.e., there holds

$$g(X) = \begin{cases} 1, & \theta \leq \sum_{\mathbf{x} \in X} f(\mathbf{x}), \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Taking into account that the problem is the weakly supervised learning, different goals exist to train the classifier $g(X)$. In particular, three types of the classification problems can be pointed out [5]: (1) global detection which identifies a class of every instance in a bag (a histology image); (2) local detection aims to identify some subsets of interest, for example, cancerous patterns from a patch $\mathbf{x}_i$ belonging to the bag X; (3) global and local detection aims to detect whether an image has a pattern of a class of interest, for example, cancer, and also to identify the location where it occurs within an image.

We mainly solve a task corresponding to the first type of the classification problem. However, the proposed method is simply extended on the second and the third types.

B. Hide-and-Seek approach

An interesting approach to get the accurate classification performance called Hide-and-Seek was proposed by Singh and Lee [27]. According to this approach, an image is divided into a grid of patches. The image patches are hidden during training such that the model seeks the relevant object parts from remaining elements of the image. If some patches are randomly removed from an image, for example, with a dog, then there is a possibility that the dog's face, which is the most discriminative, will not be visible to the model. In this case, the model must seek other relevant parts like the tail and legs in order to do well on the classification task.

The idea to hide patches can be applied to quite different problems, namely, to explanation of the image patches and to constructing shape functions explaining each patch. Suppose there is a trained black-box model predicting one of two classes corresponding to an input feature vector. It is assumed that the probability of the class will be smaller when a patch is hidden. By the patch dynamic fill, we can get different probabilities of each class.

C. Siamese neural networks

The Siamese neural network realizes a non-linear embedding of data [29]. It consists of two identical subnetworks with shared parameters. Every subnetwork models a function

$f(X)$ of input feature vector $X = (x_1, ..., x_n) \in \mathbb{R}^n$, which maps X to embedding vector $\mathbf{h} = (h_1, ..., h_D) \in \mathbb{R}^D$ in a low-dimensional space.

If there are two feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ being similar (dissimilar), then the Euclidean distance $d(\mathbf{h}_i, \mathbf{h}_j)$ between the output vectors should be as small (large) as possible. In order to train the Siamese neural network with the above property, a specific loss function should be used. One of the functions is the contrastive loss function defined as:

$$l(X_i, X_j, z_{ij}) = \begin{cases} \|\mathbf{h}_i - \mathbf{h}_j\|_2^2, & z_{ij} = 0, \\ \max(0, \tau - \|\mathbf{h}_i - \mathbf{h}_j\|_2^2), & z_{ij} = 1, \end{cases} \tag{3}$$

where τ is a predefined threshold which is regarded as a tuning parameter; z_{ij} is the indicator function taking the value 0, if vectors X_i and X_j are similar, and value 1 otherwise.

Hence, the total loss function is of the form:

$$L_{\text{Siam}}(W) = \sum_{(i,j)\in K} l(X_i, X_j, z_{ij}) + \mu R(W). \tag{4}$$

where $R(W)$ is a regularization term added to improve generalization of the neural network; W is the matrix of the neural parameters; μ is a hyper-parameter which controls the strength of the regularization; loss functions l are summed over all pairs of input vectors.

We apply the Siamese neural network in order to implement the black-box model. It should be noted that the Siamese neural network has been used in the problem of the histopathological image classification [7]. However, Yao et al. [7] used the Siamese neural network to learn features from different phenotype clusters. Moreover, the Siamese network has been used in self-supervised multi-instance learning for autonomous driving [30] to improve the learning generalization. In our case, the Siamese neural network has two goals. First, it is used to increase the training set due to considering pairs of images and to perform the back-box model. However, the main goal of the Siamese neural network is to study how the distance between embeddings, corresponding to images and obtained by two identical neural networks, is changed with dynamic fill of patches of one of the images. The distances are used for constructing the shape functions characterizing the impact of the corresponding patches on the class of the whole image and, in fact, providing their interpretation. If the distance is rapidly changed, then we can conclude that the corresponding patch is important and can be annotated depending on the considered class of the whole image.

III. The proposed method

Suppose that there are N bags $X_1, ..., X_N$ with labels $Y_1, ..., Y_N$. Every bag consists of m patches. Let us consider the scheme given in the introductory section in detail in order to describe the proposed method.

1) In order to use an explanation method, we need to have a black-box surrogate model whose prediction is explained. The black-box model solves the supervised learning task, namely, classification task. For simplicity purposes, we consider the binary classification when two classes are defined by the bag classes (cancer and non-cancer). The main difficulty of training the black-box model is that bags may have a very large dimensionality, for example, the number of pixels in the corresponding images, whereas the number of bags may be very small. Therefore, it is difficult to construct an accurate classifier under the above conditions. At the same time, we do not need to implement the accurate classifier. It is important for us to study how predictions are changed when bags are under noise in the form of dynamic fill of some patches of the bag, for example, how probabilities of cancer or non-cancer are changed under the noise. Nevertheless, the classifier can be improved if we take the Siamese neural network as the black-box model and consider distances between predictions corresponding to "clean" bag and noisy bag. In this case, the distance is determined not for bags, but for the corresponding embeddings. In fact, distances play the role of probabilities in this scheme. Moreover, to train the network, we use pairs of bags instead of single images. As a results, the number of training examples significantly increases. In sum, the Siamese neural network is trained on pairs of bags. Then it is tested by using pairs consisting of a bag and its noisy variants. Suppose that X is a tested bag, and $X(t)$ are its noisy variants depending on parameter of the noise $t \in \mathcal{T}$. The set $\mathcal{T}$ will be defined below. Then by using the trained Siamese neural network, we determine the distances $d(t, \mathbf{h}, \mathbf{h}(t))$, where $\mathbf{h}$ and $\mathbf{h}(t)$ are embeddings corresponding to X and $X(t)$, respectively. The distances as a function of t produce a set of shape functions for every bag.
2) In order to implement the procedure of dynamic fill of each bag, we partially use ideas behind the algorithm "Hide-and-Seek" [27]. However, in contrast to this algorithm, we propose the following procedure whose final goal is to construct shape functions of distances $d(t, \mathbf{h}, \mathbf{h}(t))$. Each bag is divided into a grid of m patches. Each patch makes to be dynamically noisy in accordance with the expression $c_t = tc_o + (1-t)c_b$, where c is the noisy color value; c_o is the original "clean" patch color value; c_b is the black color value; $t \in [0, 1]$ is a noisy parameter which controls the color change of the patch. It should be noted that the above expression for the color values is applied to every pixel of the analyzed patch because c_o is different inside the original patch. It can be also seen from the above expression that we get images $X(t)$ with the noisy patch having color c_t. The color value c_t is the linear combination of the black color value and the original color values of every pixel in the patch. By using the trained Siamese neural network, the bag X is compared with the bag $X(t)$ with the noisy patch through the distance between the corresponding embeddings $\mathbf{h}$ and $\mathbf{h}(t)$. The comparison result is the distance $d(t, \mathbf{h}, \mathbf{h}(t))$ as a function of t. Distances indicate how the noisy

patch changes the original image. If this change is significant, then we can say that the corresponding patch is important. The distances by different $t \in [0, 1]$ can be viewed as a shape function for the selected patch. The shape function with the largest change corresponds to the important patch. This implies that the next step is to select a shape function with the largest change.

3) After using the Siamese neural network and the procedure of dynamic fill of each patch in the analyzed bag, we have m shape functions (one function for each patch). To make decision whether a patch with the corresponding shape function is important for explaining the bag label "cancer", we have to consider how rapidly the shape function is changed. The rapid change of the shape function says that small changes of noise significantly change the class of the bag. This implies that the patch impacts on the class of the bag, and it is important. On the other hand, if the shape function of a patch is slowly changed, then the patch is not important because its change does not impact on the class of the bag. Therefore, one of the ways for annotating patches is to measure how rapidly the shape function is changed. There are several ways to solve this task. We have the original patch by $t = 1$ and $d(1, \mathbf{h}, \mathbf{h}) = 0$, and we have the black patch by $t = 0$ and the largest distance $d(0, \mathbf{h}, \mathbf{h}(0))$. A simple way for analyzing changes of the shape function is to approximate it by a linear function $d^*(t) = at + b$. Then coefficient a can be regarded as the quantitative measure of the patch impact. It should be noted that t takes a finite number s of values, i.e., we obtain s colors of the analyzed patch. Therefore, we have s distances for every patch at $t_1, ..., t_s$. Hence, coefficient a and bias b in the linear approximation can be obtained by solving the simplest optimization problem $\min_{a,b} \sum_{i=1}^{s} (at_i + b - d(t, \mathbf{h}, \mathbf{h}(t_i)))^2$. As a matter of fact, the linear approximation is one of the procedures which are very simple and can be used for analyzing shape functions.
4) In sum, after computing coefficients $a_1, ..., a_m$ for all patches of the bag and their normalizing, we construct a heatmap illustrating importance of features. If to introduce a threshold ω for normalized coefficients $a_1, ..., a_m$ as a tuning parameter, then the obtained heatmap also shows the classes of each patch, namely, the i-th patch belongs to class "cancer" if $|a_i| \geq \omega$, otherwise it belongs to class "non-cancer".

Algorithm 1 can be viewed as a formal scheme for computing the weights and shape functions.

A testing scheme of the Siamese neural network is depicted in Fig. 1. Pairs of bags consisting of the "clean" image and images with a single noisy patch are fed to the network input. The patch is dynamically filled with colors defined by $tc_o + (1 - t)c_b$. Four points $t_1 = 0$, $t_2 = 0.33$, $t_3 = 0.66$, $t_4 = 1$, four corresponding noisy images, and four pairs of input data are used to construct the shape function for the patch. Distances between embeddings corresponding to the pairs produce the shape function shown in Fig. 1. Linear approximation allows us to get coefficient a which indicates how the analyzed patch is important. A large absolute value of a says that the patch is important.

Algorithm 1 The algorithm of the image explanation and the patch annotation

Require: Training set of bags with labels $\{(X_1, Y_1), ..., (X_N, Y_N)\}$, $X_i = \{\mathbf{x}_1^{(i)}, ..., \mathbf{x}_m^{(i)}\}$, parameters τ, ω, s.
Ensure: Classes of patches $y_1, ..., y_m$

```
Train the Siamese neural network on annotated bags (large images) (X_1, Y_1), ..., (X_N, Y_N) with parameter τ
for j = 1, j ≤ N do
    for i = 1, i ≤ m do
        Select the i-th patch from the j-th bag
        for k = 1, k ≤ s do
            Compute c = t_k c_o + (1 − t_k) c_b for every pixel of the i-th patch and get the image X_j(t_k) with noisy patch
            Test the Siamese neural network with the pair of images (X_j, X_j(t_k)) and compute the distance d(t_k, h_j, h_j(t_k)) between the corresponding network outputs
        end for
        Solve the optimization problem min_{a_ji, b_ji} Σ_{k=1}^{s} (a_ji t_k + b_ji − d(t_k, h, h(t_k)))^2
        if |a_ji| ≥ ω then
            Patch x_i^(j) is annotated as "cancer"
        else
            Patch x_i^(j) is annotated as "non-cancer"
        end if
    end for
end for
```

IV. Numerical experiments

In order to study the proposed method, we use the Breast Cancer Cell Segmentation dataset [31] which consists of 58 histopathology images with expert annotations. Images are used in breast cancer cell detection with associated ground truth data available. The dataset aims to validate methods for cell segmentation and their classification. The dataset can be downloaded from https://www.kaggle.com/andrewmvd/breast-cancer-cell-segmentation.

Each image from the Breast Cancer Cell Segmentation dataset has the size 896×768 pixels and is divided into 672 patches of size 32×32. It can be seen from the dataset that the number of images is very small in order to train a black-box classifier. However, we use the Siamese neural network which allows us to form many pairs of images. Moreover, we aim to study how the distance between "clean" and noisy images is changed for different noise values. It turns out that this small dataset allows us to get correct explanations and annotations

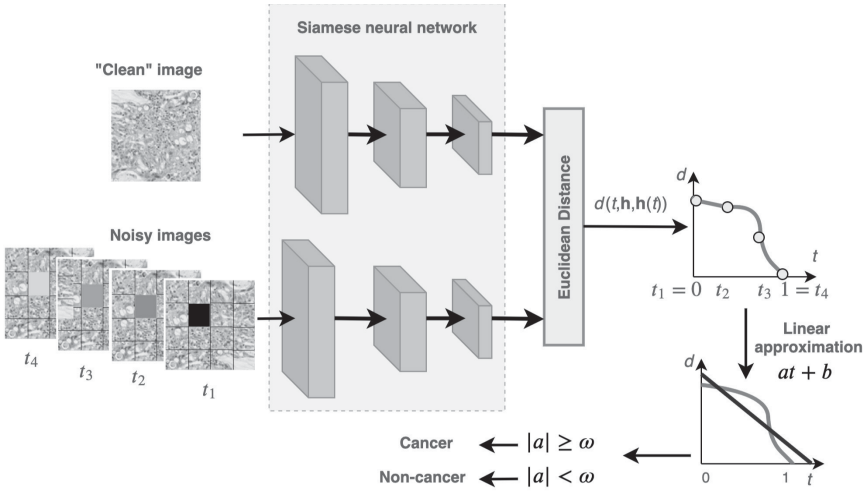

Fig. 1. The Siamese neural network testing and constructing the shape function of distances by using pairs of bags consisting of the "clean" image and noisy images

TABLE I. AN EXAMPLE OF THE ARCHITECTURE OF THE BLACK-BOX SIAMESE NEURAL NETWORK

1	Convolution layer 5×5; input: 3 channels; output: 32 channels
2	Parametric ReLU activation
3	MaxPooling layer 2x
4	Convolution layer 5×5; input: 32 channels; output: 64 channels
5	Parametric ReLU activation
6	Global Average Pooling to $4 \times 4 \times 64$ tensor
7	Linear layer with PReLU activation; input: $4 \times 4 \times 64$; output: 256
8	Linear layer with PReLU activation; input: 256; output: 256
9	Linear layer with PReLU activation; input: 256; output: embedding

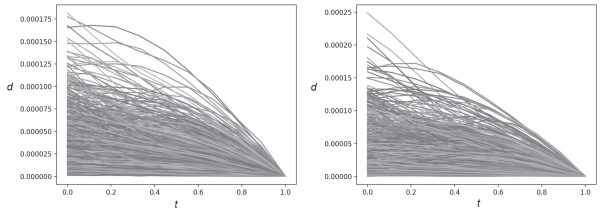

Fig. 2. Examples of instance shape functions for two random bags

of instances. An architecture of the black-box Siamese neural network is shown in Table I.

Examples of shape functions computed for two bags are shown in Fig. 2. 672 shape functions corresponding to 672 patches are illustrated in each picture. One can see that values of all functions by $t = 1$ are equal to 0. This is due to the fact that patches by $t = 1$ are not noisy, and the pair of patches, which is fed to the Siamese neural network, consists of two identical images. At the same time, one can see from Fig. 2 that the black patch in an image ($t = 0$) produces different distances between embeddings. Most functions have small changes of distances. They are located at the bottom of the coordinate quadrant. However, there are a few functions that stand out from most functions. They actually correspond to important patches which significantly impact on the predicted class of the whole image. The linear approximations of these functions corresponding to important patches have the largest values of coefficients a.

Four randomly selected pairs of pictures illustrating the true mask (the left picture in each pair) and the corresponding heatmap obtained by using the proposed method (the right picture in each pair) are shown in Fig. 3. One can see from Fig. 3 that the heatmaps clearly indicate the cancer cells (patches), and their locations coincide with the corresponding masks.

We compare the proposed method with the well-known method proposed by Yamamoto et al. [28]. This method uses a typical scheme of several methods. First, an autoencoder is trained on all patches of all available bags in order to get a low-dimensional representation (embedding) of the patches. Second, the embeddings are divided into some number of clusters by using, for example, the k-means clusterization method. Third, for every cluster, probabilities of the cancer and non-cancer patches are computed. This procedure consists in determining the proportion of patches that belong to cancerous images among all patches in the cluster and the proportion of patches belonging to non-cancerous images. Fourth, the obtained probabilities are compared with a threshold, and decisions are made for patches about their class in accordance with results of the comparison. This is a very simple and efficient method. However, its accuracy strongly depends on the number of bags and the number of clusters. A small number of clusters leads to the low sensitivity of the method because quite different patches may be in the same cluster. On the other hand, a large number of clusters leads to inaccurate probabilities. Moreover, it is difficult to make decisions when probabilities of cancerous or non-cancerous patches are close to 0.5. It should be noted that embeddings of patches do

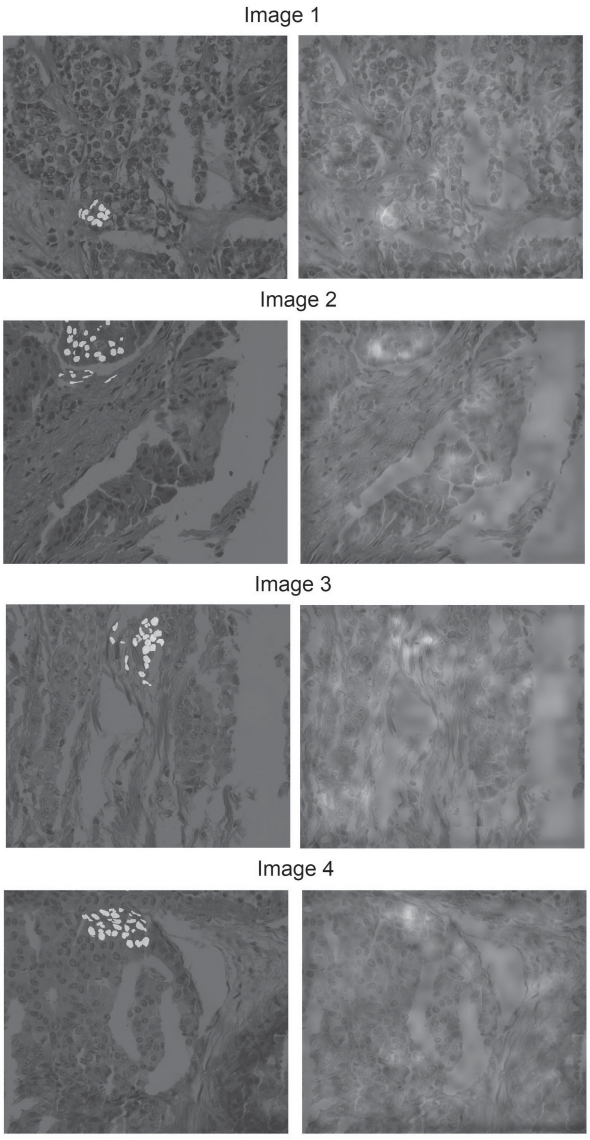

Fig. 3. Four pairs of pictures illustrating the true mask (the left picture in each pair) and the corresponding heatmap (the right picture in each pair)

not take into account neighboring patches which may impact on the classification. Fig. 4 illustrates F-score measures of two methods as functions of threshold ω. The first method is proposed by Yamamoto et al. [28]. It is depicted by the line with square markers. Since the method does not depend on ω, its F-score is constant. The proposed method is depicted by the line with triangle markers. One can see from Fig. 4 that the proposed method provides outperforming results in comparison with the method [28] for some values of the threshold. The largest value of F-score for the proposed method is 0.71 whereas the method [28] provides the F-score equal to 0.64.

It should be noted that the obtained results are valid for

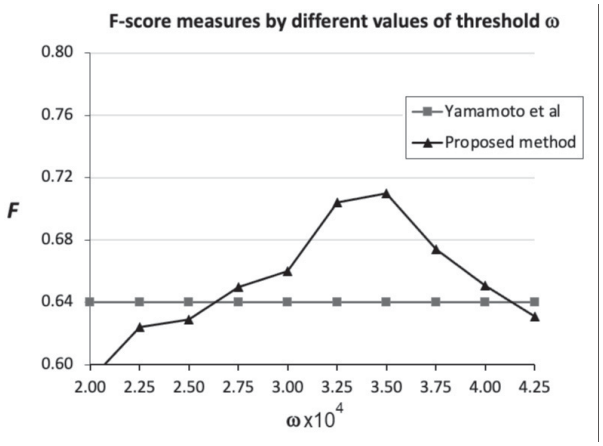

Fig. 4. F-score measures of two methods by different values of threshold ω

the Breast Cancer Cell Segmentation dataset which is very small. It is impossible to train a machine learning model by using only 58 images of size 896×768. The model [28] tries to overcome this difficulty by using low-dimensional representation of patches. However, computing the probabilities of the cancer and non-cancer patches in clusters can be also regarded as a classification problem for which 58 images can provide only inaccurate results. The proposed method overcomes the problem of small dataset by using the Siamese neural network and the weakly supervised learning. Moreover, the proposed method does not try to classify pairs of images. It computes distances between embeddings of images. It could seem that embedding may hide some important information, for example, information about the neighboring patches. However, in contrast to embedding obtained by means of the autoencoder [28], the Siamese neural network considers each whole image, and outputs of the network take into account all image information.

V. Conclusion

A new approach for solving the MIL problem by using the explanation methods has been proposed. The following advantages can be pointed out. First of all, the method deals with small datasets due to usage of the Siamese neural network. We train the network to solve the weakly supervised classification problem, but we use it to study how the distance between embeddings of the corresponding "clean" and noisy bags depends on the noise level. This implies that we do not need to train a powerful network to get desirable shape functions. Second, the method also allows us to explain why each analyzed image belongs to one of the classes. If we look at heatmaps given, for example, in Fig. 3, then we see areas of important patches. Moreover, it can be seen from the heatmaps that some areas are close to explained patches, and a doctor can pay attention to these areas in order to state a more correct diagnosis. Third, the proposed method has many ways to be extended and modified. For example, the linear approximation has been used to analyze shape functions and to select the

most important patches. However, different approaches can be applied to implement this procedure. We have used a specific scheme for producing noisy patches which are the linear combination of the black color value and the original patch. This scheme was used due to its simplicity because this procedure has to be repeated many times. However, many other schemes can be applied to implement the algorithm "Hide-and-Seek". These schemes as well as the analysis of shape functions are directions for further research.

ACKNOWLEDGEMENT

The research results have been obtained in December of 2021. This work is supported by the Russian Science Foundation under grant 21-11-00116.

REFERENCES

[1] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognition*, vol. 77, pp. 329–353, 2018.

[2] T. Dietterich, R. Lathrop, and T. Lozano-Perez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, pp. 31–71, 1997.

[3] J. Amores, "Multiple instance classification: review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, 2013.

[4] B. Babenko, "Multiple instance learning: Algorithms and applications," University of California, San Diego, Tech. Rep., 2008.

[5] V. Cheplygina, M. de Bruijne, and J. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.

[6] G. Quellec, G. Cazuguel, B. Cochener, and M. Lamard, "Multiple-instance learning for medical image and video analysis," *IEEE Reviews in Biomedical Engineering*, vol. 10, pp. 213–234, 2017.

[7] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang., "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning network," *Medical Image Analysis*, vol. 65, no. 101789, pp. 1–14, 2020.

[8] Z.-H. Zhou, "Multi-instance learning: A survey," National Laboratory for Novel Software Technology, Nanjing University, Tech. Rep., 2004.

[9] J. Wang and J.-D. Zucker, "Solving the multiple-instance problem: A lazy learning approach," in *Proceedings of the seventeenth international conference on machine learning, ICML*, 2000, pp. 1119–1126.

[10] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proceedings of the 15th international conference on neural information processing systems, NIPS'02*. MIT Press, Cambridge, MA, USA, 2002, pp. 577–584.

[11] O. Kraus, J. Ba, and B. Frey, "Classifying and segmenting microscopy images with deep multiple instance learning," *Bioinformatics*, vol. 32, no. 12, pp. i52–i59, 2016.

[12] M. Sun, T. Han, M.-C. Liu, and A. Khodayari-Rostamabad, "Multiple instance learning convolutional neural networks for object recognition," in *International conference on pattern recognition (ICPR)*, 2016, pp. 3270–3275.

[17] M. Hagele, P. Seegerer, S. Lapuschkin, M. Bockmayr, W. Samek, F. Klauschen, K.-R. Muller, and A. Binder, "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods," *Scientific Report*, vol. 10, no. 6423, pp. 1–12, 2020.

[13] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[14] W. Li and D. Yeung, "MILD: Multiple-instance learning via disambiguation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 1, pp. 76–89, 2010.

[15] D. Rymarczyk, A. Kaczynska, J. Kraus, A. Pardyl, and B. Zielinski, "ProtoMIL: Multiple instance learning with prototypical parts for fine-grained interpretability," Aug 2021, arXiv:2108.10612.

[16] C. Srinidhi, O. Ciga, and A.L.Martel, "Deep neural network models for computational histopathology: A survey," *Medical Image Analysis Volume 67, January 2021, 101813*, vol. 67, p. 101813, 2021.

[18] J. van der Laak, G. Litjens, and F. Ciompi, "Deep learning in histopathology: the path to the clinic," *Nature Medicine*, vol. 27, pp. 775–784, 2021.

[19] A. Arrieta, N. Diaz-Rodriguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[20] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys*, vol. 51, no. 5, p. 93, 2019.

[21] R. Guidotti, A. Monreale, D. Pedreschi, and F. Giannotti, "Principles of explainable artificial intelligence," in *Explainable AI Within the Digital Transformation and Cyber Physical Systems*. Springer, Cham, 2021, pp. 9–31.

[22] X. Li, H. Xiong, X. Li, X. Wu, X. Zhang, J. Liu, J. Bian, and D. Dou, "Interpretable deep learning: Interpretations, interpretability, trustworthiness, and beyond," Mar 2021, arXiv:2103.10689.

[23] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," March 2021, arXiv:2103.11251.

[24] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Muller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[25] J. Bromley, J. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Sackinger, and R. Shah, "Signature verification using a siamese time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 4, pp. 737–744, 1993.

[26] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.

[27] K. Singh and Y. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 13 524–3533.

[28] Y. Yamamoto, T. Tsuzuki, and J. Akatsuka, "Automated acquisition of explainable knowledge from unannotated histopathology images," *Nature Communications*, vol. 10, no. 5642, pp. 1–9, 2019.

[29] S. Roy, M. Harandi, R. Nock, and R. Hartley, "Siamese networks: The tale of two manifolds," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, vol. 2, 2019, pp. 3046–3055.

[30] K. Chen, L. Hong, H. Xu, Z. Li, and D.-Y. Yeung, "Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving," Aug 2021, arXiv:2108.12178.

[31] E. Gelasca, J. Byun, B. Obara, and B. Manjunath, "Evaluation and benchmark for biological image segmentation," in *IEEE International Conference on Image Processing*. IEEE, Oct 2008, pp. 1816–1819.