Weakly-Supervised Convolutional Learning for Detection of Inflammatory Gastrointestinal Lesions

Spiros V. Georgakopoulos, Dimitris K. Iakovidis,
Michael Vasilakakis, Vassilis P. Plagianakos
Dept. of Computer Science and Biomedical Informatics
University of Thessaly
Lamia, Greece

Anastasios Koulaouzidis
Endoscopy Unit
The Royal Infirmary of Edinburgh
Edinburgh, UK

***Abstract*—Graphic image annotations provide the necessary ground truth information for supervised machine learning in image-based computer-aided medical diagnosis. Performing such annotations is usually a time-consuming and cost-inefficient process requiring knowledge from domain experts. To cope with this problem we propose a novel weakly-supervised learning method based on a Convolutional Neural Network (CNN) architecture. The advantage of the proposed method over conventional supervised approaches is that only image-level semantic annotations are used in the training process, instead of pixel-level graphic annotations. This can drastically reduce the required annotation effort. Its advantage over the few state-of-the-art weakly-supervised CNN architectures is its simplicity. The performance of the proposed method is evaluated in the context of computer-aided detection of inflammatory gastrointestinal lesions in wireless capsule endoscopy videos. This is a broad category of lesions, for which early detection and treatment can be of vital importance. The results show that the proposed weakly-supervised learning method can be more effective than the conventional supervised learning, with an accuracy of 90%.**

Keywords—medical image analysis; weakly supervised learning; convolutional neural networks; inflammatory lesions; lesion detection

I. INTRODUCTION

Supervised machine learning is the prime methodological approach for image-based computer-aided medical diagnosis [1]. It requires a set of annotated images for the training of a supervised system capable of classifying image contents into semantically relevant categories, e.g. normal tissue or lesions. Conventionally, image annotations are provided graphically, with image maps indicating the belongingness of each pixel to the considered categories. However, in order to obtain these annotations a considerable human effort is usually needed for manual delineation of the objects of interest, ideally by more than one expert for increased validity. The amount of effort required can be discouraging for the expert annotators, affecting their productivity, as well as the quality and quantity of the annotated images. In the context of Wireless Capsule Endoscopy (WCE), which is a demanding medical imaging modality in terms of annotation effort as it produces thousands of output video frames, it has been noted that the limited availability of annotated images and videos is an obstacle for essential progress [2].

To cope with this problem some studies indicate that the annotation task could also be assigned to non-experts, without a significant impact in the overall quality of the annotated dataset, given a sufficiently large dataset [3]. In this spirit, a study exploiting crowdsourcing for annotation of endoscopic images has also provided convincing results [4]. However, even in that case, a lot of human effort is still necessary for annotation. A more promising approach requiring less human effort is weakly-supervised learning [5]. This approach does not require a detailed, pixel-wise annotation of the training images. It requires only image-level semantic annotations indicating only the classes in which their contents belong to. For example, in the case of an abnormality detection problem, the images can be annotated with only a label indicating whether they contain an abnormality or they contain only normal tissue.

The utility of weakly-supervised learning in biomedicine is apparent; however, only a few works have been reported [6], [7]. Recent works have exploited the bag-of-words (BoW) approach to obtain image representations from hand-crafted features, which are subsequently classified by supervised algorithms. In [7] color and Local Binary Pattern (LBP) histograms were considered, using a *k*-Nearest Neighbor (kNN) classification with online metric learning to optimize the bag-level classification. In [6], extended multiresolution local patterns were considered as features, with a conventional Support Vector Machine (SVM) classification scheme.

A remarkable classification performance has been reported with supervised convolutional neural networks (CNNs) [8]. They have been proposed as a generic image classification scheme with large learning capacity, incorporating an intrinsic mechanism to automatically extract image features. To date they have been utilized in various medical image analysis tasks, including classification of interstitial lung disease in high-resolution computed tomography [9], breast lesions [10], and large gastrointestinal (GI) lesions in gastroscopy [11] and colonoscopy [12].

In all these previous studies, CNNs were used in a 'strongly'-supervised manner, using pixelwise annotations as learning targets. Recently a weakly-supervised CNN-based

This research was supported by the special account of research grants of the Technological Educational Institute of Central Greece, Lamia, Greece.

architecture was proposed for the classification and localization of objects in natural scenes [13]; however, it includes several modifications that increase the overall complexity and requires a pre-training stage using pixel-level annotations. Inspired by all these latest studies on weakly-supervised learning, we propose a novel CNN-based weakly-supervised learning strategy. The advantage of the proposed over the state-of-the-art approaches is its simplicity. It is based on a CNN architecture, which receives a single WCE video frame as input and after processing it performs binary decision making regarding the presence of an inflammatory lesion in the input frame (i.e. two-neuron output layer).

The rest of this paper consists of six sections. Section II provides a brief medical background on WCE and inflammatory lesions. The proposed strategy is described in section III, and a BoW-based classification approach is outlined as a representative state-of-the-art weakly-supervised method in section IV. The experiments performed along with the parameters used for each of the compared methods are described in section V and the results obtained are presented in section V. The last section summarizes the conclusions of this study.

II. Medical Background

WCE has become the prime diagnostic modality for the investigation of small-bowel diseases, and its application extends for the examination of other parts of the GI tract, including the esophagus and the colon [14]. However, one of the major limitations of WCE in clinical practice is the amount of time that is required for the review of the WCE videos [2]. For instance, each WCE study creates data equivalent to ~100,000 image frames. Although efforts have been made to delegate this task to nurse endoscopists, nurses [15] or other healthcare professionals [16], the final goal remains automated lesion recognition and diagnosis [17].

One of prime indications for performing WCE is the diagnosis or topographic mapping of GI lesions in known or suspected inflammatory bowel disease (IBD) [18]. The most common inflammatory lesions are ulcers, aphthae, mucosal breaks with surrounding erythema, cobblestone mucosa, stenoses and/or fibrotic strictures, and significant mucosal/villous oedema. Representative, graphically annotated datasets are available in our database KID [2], [19] and some of them have already been used in our previous works [20].

III. Weakly-Supervised CNNs

In this section, we provide a brief description of the Convolutional Neural Networks (CNNs) and the weekly supervised classification methodology, which is utilized to cope with the problem of medical image characterization as normal or abnormal, i.e. containing an inflammatory lesion.

A. Convolutional Neural Networks

CNNs have been successfully applied on diverse problems in computer vision. They are multistage trainable architectures, in which each stage consists of three types of layers [21]:

- *Convolutional Layers*, which are the major components of the CNNs. A convolutional layer consists of a number of kernel matrices, which perform convolution on their input and produce an output matrix (feature image) where a bias value is added. The learning procedures aim to train the kernel weights and biases as shared neuron connection weights.
- *Pooling Layers*, which are also integral components of the CNNs. The purpose of a pooling layer is to perform dimensionality reduction of the input feature images. Pooling layers make a subsampling to the output of the convolutional layer matrices combing neighboring elements. The most common pooling function is the max-pooling function, taking the maximum value of the local neighborhoods.
- *Fully-Connected Layer* is a classic Feed-forward Neural Network (FNN) hidden layer. It can be interpreted as a special case of the convolutional layer with kernel size 1×1. This type of layer belongs to the class of trainable layer weights and it is used in the final stages of CNNs.

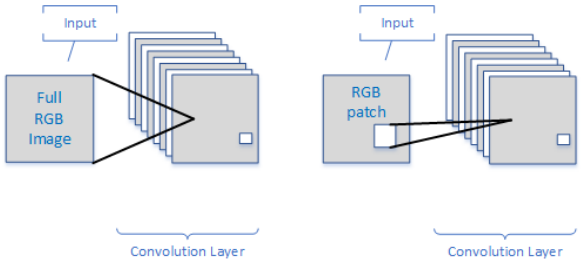
Recent studies have shown that while the input increases, the gradient drops to a value close to zero [8]. To tackle this problem, it is suggested in the convolutional layers utilize the Rectified Linear Unit (ReLU) function f(x)=max(0,x), instead of the more established sigmoid or hyperbolic tangent activation functions.

To train this type of networks, the mini-batch version of the Stochastic Gradient Descent (SGD) is usually utilized [22]. However, a drawback of the CNNs is the high training time required, because of their complex architecture and the large number of weights. However, their inherit parallelism has led to the extensive utilization of graphics processing units (GPUs) instead of the central processing units (CPUs) of the computer platforms. Nowadays, GPU training of CNNs has been established as the only feasible hardware architecture for the training phase [8].

B. Proposed Weakly-Supervised CNN Learning Strategy

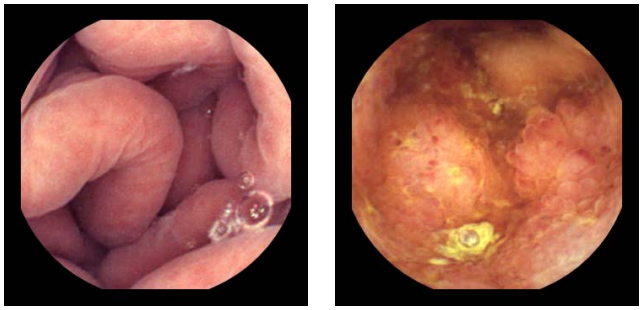
Many approaches in literature deal with medical image classification, as normal or abnormal. The majority of them divide the image into sub-images (patches) and utilize an expert system or model analyze each of these sub-images [9–12], [23], [24]. These sub-images, which must only contain normal or abnormal regions, which are designated by the experts with detailed graphic annotations of the images, are used for system's training. Subsequently, any new image is divided in sub-images characterized by the system one-by-one. This describes the conventional 'strongly'-supervised learning approach.

In this paper we follow a different weakly supervised approach, where the whole medical images are provided as input to the expert system, which is instantiated by a CNN (Fig. 1). The images used for system's training are accompanied by image-level (semantic), instead of pixel-level (graphic), annotations. Another challenge, is that unlike previous approaches [8], [13], the abnormality regions are of variable size and position, in both the training and the test sets.



(a) (b)

Fig. 1. Different strategies for CNNs training. (a) Full RGB image input (proposed). (b). Patch RGB input.



(a) (b)

Fig. 2. Example images from the dataset used in this study. (a) normal image. (b) Abnormal image with an inflammatory lesion.

The proposed CNN learning strategy is simpler than the one used in state-of-the-art CNN-based classification approaches, with the only pre-processing step it may require to be the rescaling of the images from their original to a common size before to feeding the CNN.

IV. BoW-Based Weakly-Supervised Classifiers

BoW is a widespread method in the context of image retrieval and categorization [25]. The main idea behind this method comes from analysis of text document data, which can be characterized by the frequency that same words appear in the text. The first step of this method is the creation of words equivalent to the visual content of every image. For this purpose the set of extracted features is quantized by a clustering algorithm, such as the *k*-means algorithm [26]. The centroids of each cluster represent the words, which compose a vocabulary of visual words. The second step involves a histogram construction, which describes the appearance frequency of each word in every image, and as a result it characterizes the visual content of the image. This simple technique efficiently achieves to reduce the problem of classifying a large number of high dimensional vectors from local point descriptors to a fixed size one dimensional vector without significant loss of visual information.

A conventional classifier, such as an SVM [27] can be used for the classification of the histogram vectors produced by the bag of visual words. At the end of SVM training process a model capable of assigning any test histogram vector – therefore every image – into normal or abnormal.

V. Experiments

In order to investigate the effectiveness of the proposed weakly-supervised CNN strategy, we performed experiments using a dataset available in KID [2], [19]. The dataset is part of "Dataset 2" and it consists of a total of 227 graphically annotated images of inflammatory lesions and 599 normal images of the GI tract. All graphic annotations were performed by GI medicine experts. The access to KID is upon free registration and this dataset can also be used by other researchers for algorithm comparisons.

In this study we performed three experiments using: a) the proposed weakly-supervised CNN learning approach; b) the conventional supervised CNN learning approach utilizing image patches; and c) the BoW-based approach described in section IV using an SVM classifier.

A random partitioning of the dataset was performed in order to obtain a balanced class distribution in the training and test sets. All three methodologies were evaluated on the same training and test sets. The training set consists of 200 abnormal and 200 normal images, whereas the test set consists of 27 abnormal and 27 normal images. All abnormal images contain one or more inflammatory lesions. In the following we provide details about the experimental setup and the results obtained by each evaluated method.

A. Weakly-Supervised CNN Learning

In order to improve the generalization accuracy, the training set was expanded. More specifically, the training images were rotated by 90, 180, and 270 degrees, and flipped. Considering the limitations posed by the GPU memory, we downscaled the images to 320×320 pixels in order to make it possible to store all the weights in the GPU RAM. The input of the CNN is a 3 channel, RGB, image with size 320×320. More specifically, we created a CNN network with five convolutional layers; each one of the first four layers is followed by a max-pooling layer, while the fifth is followed by a fully-connected FNN with two hidden layers. The first four convolutional layers composed of 4×4 kernel filters with stride 1 and pad 2, while the fifth convolutional layer has 4×4 kernel filters with stride 1 and pad 1. The pooling filters are 2×2 with stride 2 and without padding. Each of the first two convolutional layers consists of 16 convolutional filters followed by 16 max-pooling filters and next two convolutional and pooling layers have 32 convolutional and max-pooling filters, respectively. The FNN consists of one layer having 32 neurons followed by one hidden layer with 20 neurons and the output layer of softmax functions with two neurons. The available graphic annotations of the abnormal images were not used in the training process. The only information provided to the system were labels semantically characterizing the images as normal or abnormal.

B. Patch-based Supervised CNN Learning

For the patch-based learning approach, all the available graphic annotations were necessary for the training of the CNN. The images were divided into sub-images, which were labeled upon the respective graphic annotation with the region of interest. The size of the sub-image patches were 64×64 and

32×32 pixels in RGB color space. This resulted in a training and a test set of 7,000 (balanced normal – abnormal ratio) and 952 patches (432 normal and 520 abnormal), respectively. The same number of sub-images was used in both experiments.

We created two CNNs, one for the 64×64-pixel patches and one for the 32×32-pixel patches. The networks consist of three convolutional layers where the first two are followed by a max-pooling layer while the third is followed by a fully-connected connected FNN with two hidden layers. The first two convolutional layers are composed of 16 5×5 kernel filters, while the third convolutional layers using 32 3×3 kernel filters. The differences between the two networks are the stride and padding value of first and second convolutional layers. Each of the two first convolutional layers consists of 16 convolutional filters followed by 16 filters of max-pooling while the next convolutional layers are conduced by 32 convolutional filters. The FNN consists of one layer with 32 neurons followed by one with 20 neurons and the output layers of a softmax functions with two neurons.

C. BoW-based Weakly-Supervised Learning

The speeded-up robust feature extraction algorithm (SURF) [28] was used for automatic key-point detection and description for medical images. The original SURF algorithm considers that images are represented in grey-scale. The SURF feature set was clustered using the *k*-means algorithm. We performed experiments using different numbers of *k*-means centers, which resulted in visual vocabularies of different sizes. More specifically, the sizes of the vocabularies tested include 200, 600, 1000 and 1200 words per experiment.

VI. Results

The results of the experiments described in the previous section are summarized in Tables I and II. In order to investigate the performance of the patch-based supervised CNN approach for image-level classification, i.e., classification of the images into normal or abnormal, we adopted the following rule: an image is characterized as normal when all its patches are classified as normal, whereas it is classified as abnormal if an abnormal patch is found in the image. From Table I it becomes evident that the CNN trained with the weakly-supervised strategy exhibited the best performance in terms of accuracy, sensitivity and specificity. The BoW-SVM weakly-supervised classification approach presented a better classification performance than the CNNs trained using the conventional patch-based supervised learning strategy. For the sake of completeness, in Table II we provide the patch-level classification results of the patch-based supervised learning strategy. These results indicate the ability of the CNN to classify individual sub-images into normal or abnormal.

The best results demonstrated by the CNN weakly-supervised learning strategy were achieved using the SGD algorithm, with a learning rate of 0.001 and a momentum constant of 0.9. The training of the CNN was performed on an NVIDIA GeForce GTX 970 GPU, with 4GB of RAM and the Convolutional Architecture for Fast Feature Embedding (CAFFE) library [28].

TABLE I. Image-Level Classification Results

Results (%)	Method			
	CNN Weakly-Supervised	***CNN Supervised 64x64 patches***	***CNN Supervised 32x32 patches***	***BoW-SVM Weakly-Supervised***
Accuracy	**90.2**	68.5	66.7	72.2
Sensitivity	**92.6**	44.4	37.0	88.9
Specificity	**88.9**	92.6	96.3	55.6

TABLE II. Patch-Level Classification Results

Results (%)	Method	
	CNN Supervised 64x64 patches	***CNN Supervised 32x32 patches***
Accuracy	82.0	75.8
Sensitivity	80.4	66.9
Specificity	84.0	86.6

In the case of the BoW-based approach, the results presented in Table I, were obtained for a vocabulary composed of 1000 words, using the radial basis function (RBF) as a kernel in the SVM classifier [27].

VII. Conclusions

We presented a novel strategy to weakly-supervised learning for CNNs, applied for inflammatory lesion detection in WCE. We performed several experiments for inflammatory lesion recognition on a dataset from a public database, comparing the results of the proposed with other, state-of-the-art supervised and weakly-supervised learning approaches.

The results showed that although the proposed approach does not use pixel-level, graphic annotations for CNN training, it can perform better than the conventional 'strongly'-supervised approach. This finding is very important because one of the main drawbacks for essential progress in the development of computer-aided diagnosis systems is the limited availability of such annotations [2]. Especially in the case of WCE the application of the proposed weakly-supervised learning strategy is expected to have a significant impact, since large-scale image analysis is involved.

Future work includes investigation of the performance of the proposed strategy in larger and more diverse datasets in WCE and other medical imaging domains.

References

[1] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[2] D. K. Iakovidis and A. Koulaouzidis, "Software for enhanced video capsule endoscopy: challenges for essential progress," *Nature Reviews Gastroenterology & Hepatology*, vol. 12, no. 3, pp. 172–186, 2015.

[3] R. Kwitt, S. Hegenbart, N. Rasiwasia, A. Vécsei, and A. Uhl, "Do we need annotation experts? A case study in celiac disease classification," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, Springer, 2014, pp. 454–461.

[4] L. Maier-Hein, S. Mersmann, D. Kondermann, C. Stock, H. G. Kenngott, A. Sanchez, M. Wagner, A. Preukschas, A.-L. Wekerle, S. Helfert, and others, "Crowdsourcing for reference correspondence generation in endoscopic images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*, Springer, 2014, pp. 349–356.

[5] M. Hoai, L. Torresani, F. D. la Torre, and C. Rother, "Learning discriminative localization from weakly labeled data," *Pattern Recognition*, vol. 47, no. 3, pp. 1523–1534, 2014.

[6] S. Manivannan and E. Trucco, "Learning discriminative local features from image-level labelled data for colonoscopy image classification," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, 2015, pp. 420–423.

[7] S. Wang, Y. Cong, H. Fan, L. Liu, X. Li, S. Yang, Y. Tang, H. Zhao, and others, "Computer-Aided Endoscopic Diagnosis Without Human Specific Labeling," 2016.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[9] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical image classification with convolutional neural network," in *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on*, 2014, pp. 844–848.

[10] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen, "Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans," *Scientific Reports*, vol. 6, 2016.

[11] R. Zhu, R. Zhang, and D. Xue, "Lesion Detection of Endoscopy Images Based on Convolutional Neural Network Features," *8th International Congress on Image and Signal Processing (CISP)*, pp. 372–376, 2015.

[12] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automatic polyp detection in colonoscopy videos using an ensemble of convolutional neural networks," in *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*, 2015, pp. 79–83.

[13] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.

[14] A. Koulaouzidis, D. K. Iakovidis, A. Karargyris, and E. Rondonotti, "Wireless endoscopy in 2020: Will it still be a capsule?," *World journal of gastroenterology: WJG*, vol. 21, no. 17, p. 5119, 2015.

[15] A. Guarini, F. De Marinis, C. Hassan, C. Spada, V. Bruzzese, and A. Zullo, "Accuracy of trained nurses in finding small bowel lesions at video capsule endoscopy," *Gastroenterology Nursing*, vol. 38, no. 2, pp. 107–110, 2015.

[16] A. Riphaus, S. Richter, M. Vonderach, and T. Wehrmann, "Capsule endoscopy interpretation by an endoscopy nurse-a comparative trial", *Zeitschrift fur Gastroenterologie*, vol. 47, no. 3, pp. 273–276, 2009.

[17] D. K. Iakovidis, R. Sarmiento, J. S. Silva, A. Histace, O. Romain, A. Koulaouzidis, C. Dehollain, A. Pinna, B. Granado, and X. Dray, "Towards intelligent capsules for robust wireless endoscopic imaging of the gut," in *IEEE International Conference on Imaging Systems and Techniques*, 2014, pp. 95–100.

[18] A. Koulaouzidis, E. Rondonotti, and A. Karargyris, "Small-bowel capsule endoscopy: A ten-point contemporary," *World J Gastroenterol*, vol. 19, no. 24, pp. 3726–3746, 2013.

[19] A. Koulaouzidis and D. K. Iakovidis, "KID: Koulaouzidis-Iakovidis Database for Capsule Endoscopy." 2015.

[20] D. K. Iakovidis and A. Koulaouzidis, "Automatic lesion detection in capsule endoscopy based on color saliency: closer to an essential adjunct for reviewing software," *Gastrointestinal endoscopy*, vol. 80, no. 5, pp. 877–883, 2014.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," in *Proceedings of the IEEE*, 1998, vol. 86, no. 11, pp. 2278–2324.

[22] L. Bottou, "On-line Learning in Neural Networks," D. Saad, Ed. New York, NY, USA: Cambridge University Press, 1998, pp. 9–42.

[23] I. Maglogiannis, S. V. Georgakopoulos, S. K. Tasoulis, and V. P. Plagianakos, "A Software Tool for the Automatic Detection and Quantification of Fibrotic Tissues in Microscopy Images," *Information Science*, vol. 308, no. C, pp. 125–139, Jul. 2015.

[24] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks," in *MICCAI*, 2013, vol. 2, pp. 411–418.

[25] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 591–606, 2009.

[26] J. Drake and G. Hamerly, "Accelerated k-means with adaptive distance bounds," in *5th NIPS workshop on optimization for machine learning*, 2012.

[27] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[28] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional Architecture for Fast Feature Embedding," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, 2014, pp. 675–678.