Published in Proceedings of IPMI'15, 2015

Predicting Semantic Descriptions from Medical Images with Convolutional Neural Networks

Thomas Schlegl[1] *, Sebastian Waldstein[2], Wolf-Dieter Vogl[1,2], Ursula Schmidt-Erfurth[2], and Georg Langs[1]

[1]Computational Imaging Research Lab, Department of Biomedical Imaging and Image-guided Therapy, Medical University Vienna, Austria
thomas.schlegl@meduniwien.ac.at, georg.langs@meduniwien.ac.at
[2]Christian Doppler Laboratory for Ophthalmic Image Analysis, Vienna Reading Center, Department of Ophthalmology and Optometry, Medical University Vienna, Austria

Abstract. Learning representative computational models from medical imaging data requires large training data sets. Often, voxel-level annotation is unfeasible for sufficient amounts of data. An alternative to manual annotation, is to use the enormous amount of knowledge encoded in imaging data and corresponding reports generated during clinical routine. Weakly supervised learning approaches can link volume-level labels to image content but suffer from the typical label distributions in medical imaging data where only a small part consists of clinically relevant abnormal structures. In this paper we propose to use a semantic representation of clinical reports as a learning target that is predicted from imaging data by a convolutional neural network. We demonstrate how we can learn accurate voxel-level classifiers based on weak volume-level semantic descriptions on a set of 157 optical coherence tomography (OCT) volumes. We specifically show how semantic information increases classification accuracy for intraretinal cystoid fluid (IRC), subretinal fluid (SRF) and normal retinal tissue, and how the learning algorithm links semantic concepts to image content and geometry.

1 Introduction

Medical image analysis extracts diagnostically relevant information such as position and segmentations of abnormalities, or the quantitative characteristics of appearance markers from imaging data. To this end, algorithms are typically trained on annotated data. While the detection of subtle disease characteristics requires large training data sets, annotation does not scale well, since it is costly, time consuming and error-prone. An alternative for learning classifiers or predictors from very large data sets is to rely on existing data generated during clinical routine, such as imaging data, and corresponding reports. We propose

* This work has received funding from the European Union FP7 (KHRESMOI FP7-257528, VISCERAL FP7-318068) and the Austrian Federal Ministry of Science, Research and Economy.

to learn the link between semantic information in textual clinical reports and imaging data, by training convolutional neural networks that predict semantic descriptions from images without additional annotation. Experiments show that the inclusion of semantic representations has advantages over standard multiple-instance learning with independent labels. It increases the classification accuracy of pathologies, and learns semantic concepts such as spatial position.

Learning from Medical Imaging Data Typical clinical imaging departments generate hundreds of thousands of image volumes per year that are assessed by clinical experts. The image and textual information comprise a rich source of knowledge about epidemiology and imaging markers that are a crucial reference during clinical routine and treatment guidance. They promise to serve as basis for the detection of imaging biomarkers, co-morbidities, and subtle signatures that are relevant for treatment decisions. Training reliable classifiers or segmentation algorithms is crucial in their processing, but requires annotated training data. While supervised learning based on annotated imaging data yields accurate classification results, annotation becomes unfeasible for large data. At the same time, expert reports created during clinical routine offer detailed descriptions of observations in the imaging data, and could fill this gap. They are currently largely unexploited.

Contribution In this paper we propose a method to use the semantic content of textual reports linked to the image data instead of voxel-wise annotations for the training of an image classifier. We evaluate if the semantic information in medical reports can improve weakly supervised learning of abnormality detectors over standard multiple-instance learning with independent labels. Since medical reports not only list observations (pathologies) but also semantic concepts of their locations, we learn the relationship between these semantic terms and specific local entities in the imaging data, together with their location. The algorithm has to learn a mapping from image location to semantic location information encoded in a semantic target vector. The benefits of this algorithm are two-fold. First, we can estimate semantic descriptions from imaging data, second, we learn an accurate voxel-wise classifier without the need for voxel-wise classification in the training data.

Related Work Weakly supervised learning approaches use binary class labels that indicate the presence or absence of an object of the corresponding class in an image or volume. Multiple-instance learning [1] is a form of weakly supervised learning to solve this problem, and views images as labeled bags of instances. A positive class label is assigned to the bag of examples if at least one example belongs to the class. The negative class label is assigned to all examples of the bag if no example belongs to the class. The corresponding situation in medical imaging consists of information that *somewhere* in the image there is a certain abnormality. Weakly supervised approaches learn from these weak or noisy labels. Examples are the *Diverse Density* Framework by Maron and Lozano-Pérez

[2], mappings among images and captions [3] using algorithms, such as Random Forests [4] or Support Vector Machines [5]. A recently published work [6] presents a multi-fold multiple-instance learning approach for weakly supervised object category localization. The work of Verbeek et al.[7] is another weakly supervised learning example, wherein semantic segmentation models are learned using image-wide class labels. While these methods yield lower classification accuracy compared to voxel-wise training set labels, they have proven useful in computer vision on natural images. Unfortunately, this does not translate directly to medical imaging data. For example, a large part of data such as retinal spectral-domain optical coherence tomography (SD-OCT) images show normal tissue. While abnormalities typically cover only a tiny fraction of the volume, they are the focus of diagnostic attention and observations encoded in the textual report. Furthermore, abnormality appearance can be modulated by location. This puts standard multiple-instance learning at a disadvantage. Our work differs from these weakly supervised learning approaches in two aspects: *(i)* We do not extract local or global image descriptors but the visual input representation is learned by our network and adapts to imaging and tissue characteristics. *(ii)* We do not only use class labels depicting the global presence or absence of objects in the entire image but use semantic information from clinical reports.

We use convolutional neural networks (CNN) to learn representations and classifiers. They were introduced in 1980 by Fukushima [8], and have been used to solve various classification problems (cf. [9,10,11]). They can automatically learn translation invariant visual input representations, enabling the adaptation of visual feature extractors to data, instead of manual feature engineering. The application of CNN in the domain of medical image analysis ranges from manifold learning in the frequency domain of 3D brain *magnetic resonance* (*MR*) imaging data [12] to domain adaptation via unsupervised pre-training of CNN to improve lung tissue classification accuracy on *computed tomography* (*CT*) imaging data [13]. A weakly supervised approach using CNN performed on natural images was presented in [14]. Our work differs from the aforementioned approaches as: *(i)* we do not perform supervised (pre-) training, *(ii)* we use medical images and *(iii)* our classifier does not only predict global image labels indicating the presence or absence of object classes in an image but also corresponding location information.

2 Weakly Supervised Learning of Semantic Descriptions

A CNN is a hierarchically structured feed-forward neural network comprising one or more pairs of *convolution layers* and succeeding *max-pooling layers* (cf. [10,11,13]). The stack of convolution and max-pooling layer pairs is typically followed by one or more fully-connected layers and a terminal classification layer. We can train more than one stack of pairs of convolution and max-pooling layers feeding into the first fully-connected layer. This enables training CNNs based on multiple scales. We use a CNN to perform voxel-wise classification on visual inputs and corresponding quantitative spatial location information. Figure 1

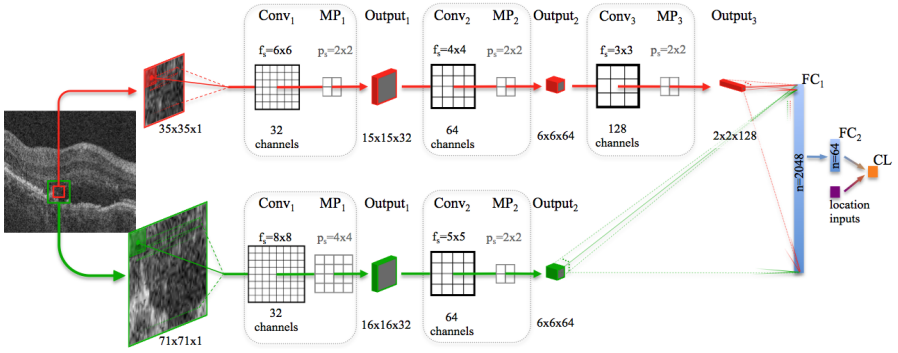

Fig. 1. Multi-scale CNN architecture used in our experiments. One stack of three pairs of convolution (*Conv*) and max-pooling layers (*MP*) uses input image patches of size 35×35. The second stack comprises two pairs of convolution and max-pooling layers and uses input image patches of size 71×71 (centered at the same position). The resulting outputs $Output_k$ of the k pairs of convolution $Conv_k$ and max-pooling MP_k layers are the inputs for the succeeding layers, with corresponding convolution filter sizes f_s and sizes of the pooling regions p_s. Our CNN also comprises two fully connected layers (*FC*) and a terminal classification layer (*CL*). The outputs of both stacks are connected densely with all neurons of the first fully-connected layer. The location parameters are fed jointly with the activations of the second fully-connected layer into the classification layer.

shows the architecture of the CNN. In the following we explain the specific representation of visual inputs and semantic targets.

2.1 Representing Inputs and Targets

The overall data comprises M tuples of medical imaging data, corresponding clinical reports and voxel-wise ground-truth class labels $\langle \mathbf{I}^m, \mathbf{T}^m, \mathbf{L}^m \rangle$, with $m = 1, 2, \ldots, M$, where $\mathbf{I}^m \in \mathbb{R}^{n \times n}$ is an intensity image (e.g., a slice of an SD-OCT volume scan of the retina) of size $n \times n$, $\mathbf{L}^m \in \{1, ..., K+1\}^{n \times n}$ is an array of the same size containing the corresponding ground-truth class labels and $\mathbf{T}^m$ is the corresponding textual report. During training we are only given $\langle \mathbf{I}^m, \mathbf{T}^m \rangle$ and train a classifier to predict $\mathbf{L}^m$ from $\mathbf{I}^m$ on new testing data. In this paper we propose a weakly supervised learning approach using semantic descriptions, where the voxel-level ground-truth class labels $\mathbf{L}^m$ are not used for training but only for evaluation of the voxel-wise prediction accuracy.

Visual and Coordinate Input Information To capture visual information at different levels of detail we extract small square-shaped image patches $\dot{\mathbf{x}}_i^m \in \mathbb{R}^{\alpha \times \alpha}$ of size α and larger square-shaped image patches $\ddot{\mathbf{x}}_i^m \in \mathbb{R}^{\beta \times \beta}$ of size β with $\alpha < \beta < n$ centered at the same spatial position $\mathbf{c}_i^m$ in volume $\mathbf{I}^m$, where i is the index of the centroid of the image patch. For each image patch, we provide

two additional quantitative location parameters to the network: *(i)* the 3D spatial coordinates $\mathbf{c}_i^m \in \Omega \subset \mathbb{R}^3$ of the centroid i of the image patches and *(ii)* the Euclidean distance $d_i^m \in \Omega \subset \mathbb{R}$ of the patch center i to a given reference structure (in our case: fovea) within the volume. We do not need to integrate these location parameters in the deep feature representation computation but inject them below the classification layer by concatenating the location parameters and activations of the fully-connected layer representing visual information (see Figure 1). The same input information is provided for all experiments.

Semantic Target Labels We assume that objects (e.g. pathology) are reported together with a textual description of their approximate spatial location. Thus a report $\mathbf{T}^m$ consists of K pairs of text snippets $\langle t_P^{m,k}, t_{Loc}^{m,k} \rangle$, with $k = 1, 2, \ldots, K$, where $t_P^{m,k} \in \mathcal{P}$ describes the occurrence of a specific object class term and $t_{Loc}^{m,k} \in \mathcal{L}$ represents the semantic description of its spatial locations. These spatial locations can be both abstract subregions (e.g., centrally located) of the volume or concrete anatomical structures. Note that $t_{Loc}^{m,k}$ does not contain quantitative values, and we do not know the link between these descriptions and image coordinate information. This semantic information can come in Γ orthogonal semantic groups (e.g., in (1) the lowest layer and (2) close to the fovea). That is, different groups represent different location concepts found in clinical reports. The extraction of these pairs from the textual document is based on semantic parsing [15] and is not subject of this paper. We decompose the textual report $\mathbf{T}^m$ into the corresponding **semantic target label** $\mathbf{s}^m \in \{0,1\}^{K \cdot \sum_\gamma n_\gamma}$, with $\gamma = 1, 2, \ldots, \Gamma$, where K is the number of different object classes which should be classified (e.g. cyst), and n_γ is the number of nominal region classes in one semantic group γ of descriptions (e.g., $n_\gamma = 3$ for upper vs. central vs. lower layer, $n_\gamma = 2$ for close vs. far from reference structure). I.e., lets assume we have two groups, then $\mathbf{s}^m$ is a K-fold concatenation of pairs of a binary *layer group* $g_1^k \in \{0,1\}^{n_1}$ with n_1 bits representing different layer classes and a binary *reference location group* $g_2^k \in \{0,1\}^{n_2}$ with n_2 bits representing relative locations to a reference structure. For all object classes, all bits of the layer group, and all bits of the reference location group are set to 1, if they are mentioned mutually with the respective object class in the textual report. All bits of the corresponding layer group and all bits of the corresponding reference location group are set to 0, where the respective object class is not mentioned in the report. The vector $\mathbf{s}^m$ of semantic target labels is assigned to all input tuples $\langle \dot{\mathbf{x}}_i^m, \ddot{\mathbf{x}}_i^m, \mathbf{c}_i^m, d_i^m \rangle$ extracted from the corresponding volume $\mathbf{I}^m$. Figure 2a shows an example of a semantic target label representation comprising two object classes. According to this binary representation the first object is mentioned mutually with layer classes 1, 2 and 3 and with reference location class 1 in the textual report. Figure 2c shows the corresponding volume information.

Voxel-wise Ground Truth Labels To evaluate the algorithm, we use the ground-truth class label $l_i \in \{1, ..., K+1\}$ from $\mathbf{L}^m$ at the center position $\mathbf{c}_i^m$ of the patches for every multi-scale image patch pair $\langle \dot{\mathbf{x}}_i^m, \ddot{\mathbf{x}}_i^m \rangle$. Labels include the

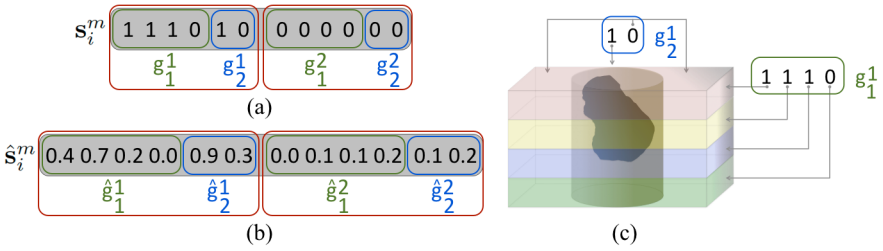

Fig. 2. (a) Example of a semantic target label comprising two object classes (K=2). Each of which comprises a *layer group* g_1^k with 4 bits (layer 1,2,3, or 4) and a *reference location group* g_2^k with 2 bits (close or distant). (b) Prediction of a semantic description $\hat{\mathbf{s}}_i^m$ that would lead to a corresponding object class label prediction $\hat{l}_i = 1$. (c) Visualization of the volume information which could lead to the given semantic target label shown in (a). (Best viewed in color)

reported observations $t_P^{m,k}$ and a healthy background label. l_i is assigned to the whole multi-scale image patch pair $\langle \dot{\mathbf{x}}_i^m, \ddot{\mathbf{x}}_i^m \rangle$ centered at voxel position i.

2.2 Training to Predict Semantic Descriptors

We train a CNN to predict the semantic description from the imaging data and the corresponding location information provided for the patch center voxels. We use tuples $\langle \dot{\mathbf{x}}_i^m, \ddot{\mathbf{x}}_i^m, \mathbf{c}_i^m, d_i^m, \mathbf{s}_i^m \rangle$ for weakly supervised training of our model. The training objective is to learn the mapping

$$f : \langle \dot{\mathbf{x}}_i^m, \ddot{\mathbf{x}}_i^m, \mathbf{c}_i^m, d_i^m \rangle \longmapsto \mathbf{s}_i^m \tag{1}$$

from multi-scale image patch pairs $\langle \dot{\mathbf{x}}_i^m, \ddot{\mathbf{x}}_i^m \rangle$, 3D spatial coordinates $\mathbf{c}_i^m$ and a distance value d_i^m to corresponding location specific noisy semantic targets $\mathbf{s}_i^m$ in a weakly supervised fashion. During testing we apply the mapping to new image patches in the test set. During classification, an unseen tuple $\langle \dot{\mathbf{x}}_i^m, \ddot{\mathbf{x}}_i^m, \mathbf{c}_i^m, d_i^m \rangle$ of multi-scale image patch pairs $\langle \dot{\mathbf{x}}_i^m, \ddot{\mathbf{x}}_i^m \rangle$ centered at voxel position i and corresponding location parameters $\mathbf{c}_i^m$ and d_i^m causes activations of the classification layer, which are the predictions of the semantic descriptions $\hat{\mathbf{s}}_i^m$. During training, all model parameters θ (weights and bias terms) of the whole model are optimized by minimizing the *mean squared error* between the actual volume-level semantic target labels $\mathbf{s}_i^m$ and the voxel-level probabilities of semantic descriptions $\hat{\mathbf{s}}_i^m$ predicted by the model.

2.3 Evaluation of Local Image Content Classification

We want to know if the proposed approach learns a link between semantic concepts and image content and location. To this end, we perform weakly supervised learning as described above. During testing, we apply the trained CNN to new data. We transform the voxel-wise predictions of semantic descriptions into

voxel-level class labels and compare these labels with ground-truth labels on the testing data. Specifically we are interested in the increase of accuracy caused by the inclusion of semantic information into the training procedure. An object class may occur simultaneously in a number of layer classes within the layer group and in a number of reference location classes within the reference location group. But if an object class is present in an image, then this occurrence has to be reflected by the predictions of both groups. So, based on the predictions of the semantic descriptions $\hat{\mathbf{s}}_i^m$ we compute the mean activation $\bar{a}_i^k$ of class k over the maximum activation within each semantic location group $\hat{g}_\gamma^k$:

$$\bar{a}_i^k = \frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} max(\hat{g}_\gamma^k) \tag{2}$$

Now we can compute location-adjusted predictions $\hat{l}_i$ for the class k having the highest mean activation:

$$\hat{l}_i = \begin{cases} 0, & \text{if } \bar{a}_i^k < 0.5, \forall \bar{a}_i^k, k = 1, 2, \ldots, K \\ \underset{k}{\operatorname{argmax}} \, (\bar{a}_i^k), & \text{otherwise} \end{cases} \tag{3}$$

If the mean activations of all classes are less than 0.5, the label 0 (*background class*) is assigned to the corresponding patch center. The resulting class label predictions $\hat{l}_i$ are assigned to the voxels in the center of the image patches. Based on these class label predictions $\hat{l}_i$ we now can measure the performance of our model in terms of misclassification errors on object class labels. Figure 2b shows an example of a prediction $\hat{\mathbf{s}}_i^m$ of a semantic description comprising two object classes and two semantic groups. The corresponding object class-wise mean activations would evaluate to $\bar{a}_i^1 = 0.8$ and $\bar{a}_i^2 = 0.2$. According to equation (3) that would lead to the object class label prediction $\hat{l}_i = 1$.

3 Experiments

Data, Data Selection and Preprocessing We evaluate the method on 157 clinical high resolution SD-OCT volumes of the retina with resolutions of $512 \times 128 \times 1024$ voxels (voxel dimensions $12 \times 47 \times 2\mu m$). The OCT data we use is not generated instantly but single slices (in the z/x-plane) are acquired sequentially to form the volume. Due to relatively strong anisotropy of the imaging data, we work with 2D in-plane (z/x) patches. From these volumes we extract pairs of 2D image patches (see Figure 3a and Figure 3b) with scales 35×35 and 71×71 for 300,000 positions. The positions of the patch centers within an image slice as well as the slice number within a volume are sampled randomly. A fast patch based image denoising related to non-local-means is applied as preprocessing step. Additionally, the intensity values of the image patches are normalized by transforming the data to zero-mean and unit variance. The human retina can be subdivided into different layers. We use an implementation of an automatic layer

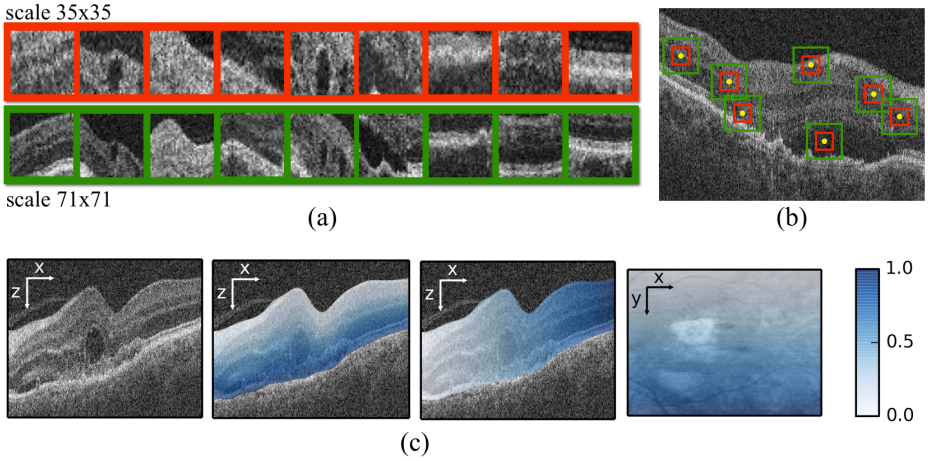

Fig. 3. (a) Visual inputs with two different scales. Patches in the same column share the centroid position. (b) Patch extraction at random positions from an SD-OCT intensity image: patches of size 35×35 (red), patches of size 71×71 (green) and corresponding patch centers (yellow). (c) Normalized coordinates of voxels within the retina. (Best viewed in color)

segmentation algorithm following [16] and based on the top and bottom layer we compute a retina mask. The voxel positions within this mask are normalized into the range $[0, 1]$, where the voxels at the top and bottom layer (z-axis), the voxels in the first and last column of the image (x-axis) and the voxels in the first and last slice of the volume (y-axis) are assigned to the marginals 0 and 1 respectively (see Figure 3c). These normalized 3D coordinates are used as location specific inputs. In every SD-OCT volume the position of the fovea is also annotated. We use the annotated position of the fovea as reference structure and provide the Euclidean distance of every image patch center as additional location specific input.

Evaluation. For the purposes of evaluation of voxel-wise classification performance we extract ground-truth class labels at the patch center positions from the corresponding volume with voxel-wise annotations. These labels are assigned to the whole corresponding image patches. In our data, 73.43% of the patches are labeled as healthy tissue, 8.63% are labeled as IRC and 17.94% are labeled as SRF. Pairs of patches sampled at different positions within the same volume may partially overlap. We split the image patches on a patient basis into training and test set to perform 4-fold cross-validation, so that there is no patient both in the training, and the test set.

We train a classifier to perform 3-class classification between IRC, SRF and normal retinal tissue. Normal retinal tissue is handled as background class. We compare three approaches:

(1) Naïve weakly supervised learning: We perform weakly supervised learning that links volume-level class labels to image content. We only use the information which object class (pathology) is present in the volume. This results in a 3-class multiple-instance classification problem. The volume-level target label is assigned to all image patches of the corresponding volume.
(2) Learning semantic descriptions: We evaluate the performance of our proposed learning strategy. We use the volume-level semantic representation of the reported pathologies. We use semantic target labels encoding two pathologies (IRC and SRF) each of which comprises four bits for the layer group and two bits for the reference structure group resulting in a 12 bit semantic target vector (see Figure 2). The semantic equivalent in a textual report for the four classes in the layer group would be *"ganglion cell complex"* (top layer of the retina), *"inner nuclear and plexiform layers"*, *"outer nuclear and plexiform layers"* and *"photoreceptor layers"* (bottom layers of the retina). The semantic equivalent for the two classes in the reference location group are *"foveal"* (in the vicinity of the fovea) and *"extrafoveal"* (at a distance from the fovea). This volume-level semantic representation is assigned to all image patches of the corresponding volume.
(3) Supervised learning: We perform fully supervised learning using the voxel-level annotations of class labels. We evaluate what classification accuracy can be obtained when the maximum information at every single voxel-position - namely voxel-wise class labels - is available.

All experiments are performed using Python 2.6 with the Theano [17] library and run on a graphics processing unit (GPU) using CUDA 5.5.

3.1 Model Parameters

For every approach training of the CNN is performed for 200 epochs. We choose a multi-scale CNN architecture with two parallel stacks of pairs of convolution and max-pooling layers. These stacks take as input image patches of size 35×35 and 71×71 and comprise 3 and 2 pairs of convolution and max-pooling layers respectively (see Figure 1). The outputs of the max-pooling layers on top of both stacks are concatenated and fed into a fully-connected layer with 2048 neurons. This layer is followed by a second fully-connected layer with 64 neurons. The activations of this layer are concatenated with the spatial location parameters of the patch centers and fed into the terminal classification layer. All layer parameters are learned during classifier training. The architecture of our multi-scale CNN and the detailed model parameters are shown in Figure 1. The model parameters were found empirically due to preceding experiments using OCT data that differs in visual appearance from the data used in our presented experiments. The model parameters were tuned in these preceding experiments solely on the supervised training task to be a good trade-off between attainable classification accuracy and runtime efficiency. Thereafter they were fixed and used in all of our presented experiments to ensure comparability between the results of the different experiments.

Table 1. Confusion matrix of classification results and corresponding class-wise accuracies on (a) the naïve weakly supervised learning approach, (b) the weakly supervised learning approach using semantic descriptions and (c) the supervised learning approach.

		prediction			
		healthy	IRC	SRF	accuracy
(a)	healthy	**144329**	4587	70994	0.6563
	IRC	10391	**5653**	9718	0.2194
	SRF	4978	231	**48511**	0.9030
(b)	healthy	**173121**	10603	36186	0.7872
	IRC	2230	**22102**	1430	0.8579
	SRF	2963	1285	**49472**	**0.9209**
(c)	healthy	**214848**	2303	2759	**0.9770**
	IRC	2222	**23086**	454	**0.8961**
	SRF	3670	638	**49412**	0.9198

3.2 Classification Results

Experiment (1) The naïve weakly supervised learning approach represents the most restricted learning approach and serves as reference scenario. Classification results are shown in Table 1a. This approach yields a classification accuracy over all three classes of 66.30%. Only 21.94% of samples showing IRC are classified correctly, while the SRF class is classified relatively accurately (90.30% of all patches showing SRF are correctly classified).
Experiment (2) The classification results of our proposed weakly supervised learning approach using semantic descriptions are shown in Table 1b. This approach yields a classification accuracy over all three classes of 81.73% with lower accuracy for the healthy class (78.72%) compared to SRF (92.09% accuracy) which is also the best performance on SRF over all three approaches.
Experiment (3) As expected, the supervised learning approach performs best. This approach yields a overall classification accuracy over all three classes of 95.98%. Classification results are shown in Table 1c. While it has most difficulties with IRC (89.61% accuracy) it still obtains the highest accuracy for IRC over all three approaches. This approach also performs best for the healthy class (97.70% accuracy). Figure 4 shows a comparison of voxel-wise classification results obtained by the different approaches. For each of the three training approaches the computation of the voxel-wise map took below 10 seconds.

4 Discussion

We propose a weakly supervised learning method using semantic descriptions to improve classification performance when no voxel-wise annotations but only textual descriptions linked to image data are available. A CNN learns optimal multi-scale visual representations and integrates them with location specific inputs to perform multi-class classification. We evaluated the accuracy of the

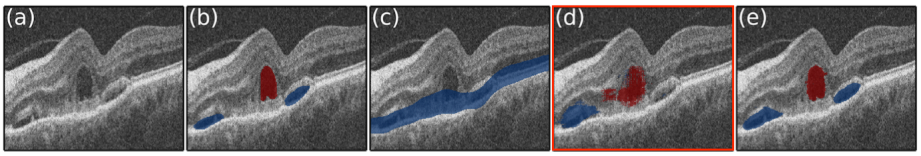

Fig. 4. (a) Intensity image of a single slice (zx-view) of a clinical SD-OCT scan of the retina. (b) Voxel-wise ground-truth annotations of IRC (red) and SRF (blue). Automatic segmentation results corresponding to voxel-wise class label predictions obtained with (c) the naïve weakly supervised learning approach, (d) our weakly supervised learning approach using semantic descriptions and (e) the supervised learning approach. On the class label predictions (c-e) no post-processing was performed. (Best viewed in color)

proposed approach on clinical SD-OCT data of the retina and compared its performance on class label prediction accuracy with naïve weakly supervised learning and with fully supervised learning. Experiments demonstrate that based on volume-level semantic target labels the model learns voxel-level predictions of object classes. Including semantic information substantially improves classification performance. In addition to capturing the structure in intensity image patches and building a pathology specific model, the algorithm learns a mapping from 3D spatial coordinates and Euclidean distances to the fovea to semantic description classes found in reports. That is, the CNN learns diverse abstract concepts of "location". The learning approach can be applied to automatic classification and segmentation on medical imaging data for which corresponding report holds semantic descriptions in the form of pathology - anatomical location pairs.

Exploiting semantics over naïve weakly supervised learning has several benefits. The latter performs poorly for classes occurring only in few volumes or in a vanishingly low amount of voxels. While this is a characteristic of many diagnostically relevant structures in medical imaging, weakly supervised learning performed particularly poorly on them (e.g., IRC), while exhibiting the strongest bias towards the SRF class in our experiments. This can be explained by the fact that many volumes show SRF resulting in a large amount of patches having the (false) noisy SRF class label. Our approach achieves higher classification accuracy on class labels over all three classes by approximately 15%. Results indicate that semantic descriptions which provide class occurrences and corresponding abstract location information provide a rich source to improve classification tasks in medical image analysis where no voxel-wise annotations are available. This is important, because in many cases medical images have associated textual descriptions generated and used during clinical routine. The proposed approach enables the use of these data on a scale for which annotation would not be possible.

References

1. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence **89**(1) (1997) 31–71
2. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. Advances in Neural Information Processing Systems (1998) 570–576
3. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep boltzmann machines. In: Advances in Neural Information Processing Systems 25. (2012) 2231–2239
4. Leistner, C., Saffari, A., Santner, J., Bischof, H.: Semi-supervised random forests. In: 12th International Conference on Computer Vision, IEEE (2009) 506–513
5. Zhou, Z.H., Zhang, M.L.: Multi-instance multi-label learning with application to scene classification. Advances in NIPS **19** (2007) 1609–1616
6. Cinbis, R.G., Verbeek, J., Schmid, C.: Multi-fold MIL training for weakly supervised object localization. In: Conference on Computer Vision and Pattern Recognition, IEEE (2014)
7. Verbeek, J., Triggs, B.: Region classification with markov field aspect models. In: Conference on Computer Vision and Pattern Recognition, IEEE (2007) 1–8
8. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics **36**(4) (1980) 193–202
9. Lee, H., Grosse, R., Ranganath, R., Ng, A.Y.: Unsupervised learning of hierarchical representations with convolutional deep belief networks. Communications of the ACM **54**(10) (2011) 95–103
10. Ciresan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. In: Conference on Computer Vision and Pattern Recognition, IEEE (2012) 3642–3649
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. Volume 1. (2012) 4
12. Brosch, T., Tam, R.: Manifold learning of brain MRIs by deep learning. Medical Image Computing and Computer-Assisted Intervention (2013) 633–640
13. Schlegl, T., Ofner, J., Langs, G.: Unsupervised pre-training across domains improves lung tissue classification. In: Medical Computer Vision. Large Data in Medical Imaging. Springer (2014)
14. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Weakly supervised object recognition with convolutional neural networks. Technical Report HAL-01015140, INRIA (2014)
15. Pradhan, S., Ward, W., Hacioglu, K., Martin, J., Jurafsky, D.: Shallow semantic parsing using support vector machines. In: Proceedings of HLT/NAACL. (2004) 233–240
16. Garvin, M.K., Abràmoff, M.D., Wu, X., Russell, S.R., Burns, T.L., Sonka, M.: Automated 3-D intraretinal layer segmentation of macular spectral-domain optical coherence tomography images. Transactions on Medical Imaging, IEEE **28**(9) (2009) 1436–1447
17. Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y.: Theano: A CPU and GPU math expression compiler. In: Proceedings of the Python for scientific computing conference (SciPy). Volume 4. (2010)