Self-Supervised Convolutional Feature Training for Medical Volume Scans

Max Blendowski¹, Hannes Nickisch², and Mattias P. Heinrich¹

¹Institute of Medical Informatics, University of Lübeck, DE
²Philips Research, Hamburg, DE
blendowski@imi.uni-luebeck.de

Abstract

With their breakthrough achievements in computer vision and afterwards also in the medical image domain, convolutional neural networks (CNNs) have become a standard choice for several machine learning tasks like image segmentation and classification. Their success is due to the incorporation of the problem specific feature learning process in form of adaptive filter kernels. However, more than in other areas, annotated data is scarce in medical imaging despite the fact that the number of acquired images grows rapidly. We propose to take advantage of these unlabeled images. Using readily available spatial cues in medical volume scans without expert input, we train local image descriptors in an unsupervised manner and show that they can easily be employed in subsequent tasks with very few labeled training images. Our experiments demonstrate that predicting simple positional relationships between pairs of non-overlapping subvolumes in medical images constitutes a sufficient auxiliary pretraining task to provide expressive feature descriptors within their receptive field.

1 Introduction

Recognising and classifying multiple objects in a single scene is a common problem in computer vision and characterises the fundamental tasks of segmentation and detection in medical image processing [1]. Their clinical environment poses hard constraints on the latter methods in terms of robustness and runtime to be valuable. However, to delineate organs or structures of interest manually, is - if at all - only feasible in daily routine for selected scans. Yet, scenarios like follow-up assessment of tumor growth rates or image guided interventions either rely on or can benefit from the supply of these additional label information. The handling of a growing amount of medical image data creates an urgent demand of automation that advanced machine learning techniques could meet. But so far, most methods that are successfully applied in the medical image domain are supervised learning algorithms that heavily depend on large training data sets, i.e. image data with voxelwise defined class labels.

Recently, convolutional neural networks (CNNs) spearhead the stage of developments in computer vision and medical image processing. Unfortunately, CNNs need even more labeled training data compared to e.g. Random Decision Forests and learning them from scratch is very time consuming. To tackle the data problem, various training strategies have emerged. Because our problem is facing the lack of corresponding ground truth annotations and not the availability of image data, we follow a different strategy based on the core idea of transfer learning. Deep networks extensively trained on the very large ImageNet dataset have been shown to work as generic and expressive feature extractors when applied to new tasks by replacing and finetuning only the weights of the last network layers [10]. The feasibility of this approach is often explained by referring to an analogy: Inspired by the basic motivation of neural networks in their natural counterparts, the first layers in a network are often
said to learn primitive features and more complex concepts will associate to neurons in deeper layers. Following this argument, the network first has "learned to see", i.e. its kernel weights in the lower layers were successfully trained to describe low level features on a large dataset. Using this ability to see, deeper layers now "recognize" more problem-specific and complex patterns. Considering the fact that the label information on ImageNet is rather weak, because for every example only the global information of the correct class is known, but not its location inside the image (image of a "car"), the generated descriptors perform astonishingly well on much more difficult tasks like semantic segmentation after finetuning. An obvious step with regard to image segmentation therefore could be the application of finetuned CNNs to specific medical imaging modalities. To make use of pretrained 2D models, [12] explore in a 2.5D approach the transfer learning abilities of different 2D CNNs in order to classify lung nodules in 3D CT volumes. But one major drawback prevents the pursuit of a true 3D transfer learning strategy for medical volume data: simply no commonly accepted 3D CNN architecture has emerged yet, although there is successful research employing 3D-CNNS e.g. by [8] for a prostate depicting task.

Related work: In general, CNNs trained in a supervised manner achieve excellent results across vision tasks. With their main limitation of requiring extensive annotations, different strategies were developed using weak or incomplete labels [11] or relaxing the task by incorporating interactive labels [14]. Lately there is growing attention for unsupervised or self-supervised learning [2], a direction our approach in this work is also following. Several research groups developed approaches for two-dimensional images that we comprehend as manifestations of the aforementioned process of "learning to see". The appealing point of these methods is to manage the learning in a self-supervised fashion, i.e. there is no need for even only weak global ground truth labels. This absence of ground truth information nevertheless needs to be compensated. Therefore all these methods share a common paradigm in defining different annotation-free auxiliary tasks.

For example [13] train autoencoder networks for image denoising. During this encoder-decoder scheme, noise-corrupted versions of the input are mapped to a lower dimensional representation in order to generate robust features. On their basis, a noise-free image should be reconstructed, thus the pretext task during training is to recover the known input from a distorted version of itself. In contrast to this end-to-end regression approach [4] propose a strategy based on classification. Randomly drawn image patches (from regions with high gradients) are altered by a variety of transformations (e.g. change in scaling or colorization). The network is now trained to assign these altered versions to their initial image as auxiliary task. However, the number of classes therefore grows directly with the number of base patches. More recently, [5] use image colorization as pretext task. They argue that in order to plausibly colorize a greyscale image, robust understanding of depicted objects and their parts is needed. Unfortunately, a direct transfer of this approach to the medical image domain is not feasible since 3D scans only contain greyvalue information. [9] propose a strategy coined context encoders. In this work, the pretext task for learning expressive feature extractors is - again similar to the autoencoder approach - to reconstruct known images by inpainting entire missing image regions. In [5], the authors pretrain networks to recognise if a given image patch underwent a rotation. In the medical context, e.g. [7] propose an unsupervised learning strategy for spinal MRI scans to predict intervertebral disc degeneration gradings. However, their approach needs the information which scans belong to the same patient. This is a rather strict limitation to unsupervised learning, since in general, one can not expect to have several scans of the same patient.

Lastly, an unsupervised learning strategy proposed in [3] serves as starting point for our work. They introduce a simple auxiliary learning task based on classification and purely relying on the image content itself. Given a pair of non-overlapping image patches, the aim is to predict their spatial constellation from a small predefined set of possible elementary relations. We will show in our experiments that despite the approach seems trivial, the translation of this procedure to medical volume data is straightforward and provides a robust, unsupervised pretraining scheme.

2 Method

The overall goal of our work is to provide an unsupervised pretraining scheme for expressive feature extraction using simple 3D CNNs. To this end, we straightforwardly expand the proposed procedure from Doersch et al. to the third image dimension. With this method at hand, we are able to train a CNN in a self-supervised manner. In a second stage, we pursue the ideas of transfer learning, i.e. supervised finetuning will subsequently guide the network's focus to its actual intended task.
Figure 1: Illustration of the proposed pretraining scheme by [3]: the auxiliary pretraining task is to predict the correct spatial relationship of patch $B_i$ with respect to patch $A_i$. In 2D they propose 4 simple relations: left, right, top and bottom. By adding front and back, this approach is easily expanded to 3D. To prevent the network from learning trivial geometric hints like continuing lines, a margin between the patches and random jitter to their positions is added. In this example the spatial relation from $B_1$ to $A_1$ is bottom. Readers may convince themselves that this auxiliary task is fairly simple for humans by guessing the relation between $B_2$ and $A_2$ (answer: left).

Unsupervised pretraining: The basic idea of our method’s first stage effectively implements the makeshift learning target proposed by [3]. Since we want to operate on 3D medical volume data, in extension to the already defined relations top, left, bottom and right, we add front and back as remaining positional cues for the added depth dimension. The reason to perform unsupervised pretraining in the first place is to train expressive local features. CNNs form today’s standard choice for this task, and we have to start by determining the unsupervised network’s structure. The resulting receptive field corresponds to a subvolume of the image and it condenses the considered number of neighboring image intensity information per voxel - growing cubically with respect to the side length - to a vector representation with a size equal to the number of channels in the last layer of the network. On that account the resulting feature dimension usually is much smaller than using all neighboring intensities as descriptor. Subsequent to the feature extracting CNN part, the network’s remaining fully connected layers perform the auxiliary classification. Thus, using the spatial relation prediction as auxiliary task - which the latter section is assumed to focus on - is enabling the desired “learning to see” adaption of filter kernels inside the preceding descriptor layers.

Training the joint CNN parts can be performed by feeding subvolume pairs from a training scan into a siamese structure. Using relations between pairs instead of e.g. predicting the position of a single subvolume inside a scan, removes the necessity of a detailed alignment of all patients to a reference system. In order to prevent the descriptor network from improper adaption to trivial geometric content, e.g. lines that should continue appropriately, subvolumes used during the unsupervised training keep a small margin in between and are also slightly displaced by random jitter added to their center coordinate. Figure illustrates the spatial relation prediction task for the sake of clarity in 2D: given the $i$-th pair of image patches $P_i = (A_i, B_i, s_i)$, where $A$ and $B$ denote the extracted image parts and $s_i$ the known spatial relation from $B_i$ to $A_i$, it is obvious that the correct output for $P_i$ is bottom. This unsupervised training scheme allows to tackle the lack of annotated medical images and to fully exploit the rapidly growing amount of volume scans by generating a virtually unlimited number of weak but meaningful labeled image pairs. While any desired network architecture can serve as feature extraction CNN part, only its receptive field needs to be known for the generation of paired training subvolumes, that are sufficiently separated by a margin.
Supervised finetuning: After the pretraining’s completion, supervised finetuning forms the second stage of our proposed approach. To this end, parts of the unsupervised pretrained network that are assumed to mainly have adapted to a substitute task will be replaced. Next, under re-usage of the descriptor layers, the new network parts should adapt to the actual machine learning task. During this final, now supervised step, the network’s parameters are finetuned by providing images and their corresponding label maps from the very small set of manually annotated training data.

Figure 2 illustrates the whole procedure and points out another detail that we investigate: during finetuning, the learning rates for the task specific layers are known to be chosen orders of magnitude larger than their counterparts for the pretrained layers, where weights sometimes even are frozen by zero learning rates. Additionally, we analyze if inverting this proportion while performing the unsupervised pretraining might be helpful in comparison to using equal learning rates for all network parameters.

Figure 2: Two staged segmentation strategy: while performing pretraining of the descriptor network part (blue box) we have access to a larger set of patient volumes compared to the supervised segmentation learning (red box). After pretraining completion we transfer the descriptor weights to the actual segmentation CNN. While here we pursue common finetuning with low learning rates for pretrained parts, we also investigate during the first stage if inverting this ratio is beneficial - implied by differently thick red arrows.

3 Experiments and Results

To verify our hypothesis that task-specific CNNs can benefit from the transfer of unsupervised pretrained feature extractors into their pipeline, we perform experiments on the task of multi-organ segmentation for medical volume images. These were acquired by Computed Tomography (CT) scanners. The used VISCERAL Anatomy3 public dataset contains 20 contrast enhanced abdominal CT scans that we downsample to a resolution of $156 \times 160 \times 115$ voxels. Ground truth segmentations are provided for a large variety of organs. To design a clear experimental setup, we pick 7 structures that occur in all datasets (liver, spleen, bladder, kidneys and psoas major muscles). We also choose to implement rather simple network designs in order to prevent that possible effects on the learning task are largely based on the network’s architecture.

As indicated in the Methods section, we employ two-tier respectively three-tier networks. In the following, we describe our network architectures as shown in Figure 3. The descriptor part consists of a four-layered convolutional network. Subsequently, we employ two linear layers during the unsupervised pretraining as task specific layers. Moreover, during the segmentation task learning,
three additional convolutional layers follow the descriptor part before again three fully connected layers perform the organ classification. A basic layer in the first part consists of the commonly applied sequence \textit{\{convolution - batch normalization - ReLU activation\}}. In order to inject fine-grained spatial information to deeper network parts, we concatenate appropriately cropped input volumes as additional channel to layers marked with an indicator variable $i$. Let a tuple $(k, d, c, i)_l$ define the kernel size, the dilation, the output channel number and the indicator variable respectively for a 3-dimensional convolution in layer $l$ of the descriptor network. Then the applied setup for the feature extractor during all experiments is as follows: \{(3, 1, 16, 0)_1, (3, 2, 32, 0)_2, (3, 4, 64, 0)_3, (3, 8, 128, 1)_4\}. The input channel number is defined by $c_{l-1}$ if $i_l = 0$ or $c_{l-1} + 1$ if $i_l = 1$. Since the image intensities are the input to the first layer, $c_0 = 1$. The resulting receptive field for this feature descriptor part corresponds to subvolumes of size $3^1$.

![Network Architecture](image)

Figure 3: Network architecture: All convolutional kernels are of size $3^3$, with the exception of layer 7 with size $5^3$. The parameters C, D and RF describe the output channel number, the dilation and the current layers receptive field respectively. During the pretraining stage, only the blue descriptor parts and the fully connected layers are used to predict the 6 possible spatial relations. When training the network for the segmentation task, the additional yellow layers are added and 8 classes (background + 7 organs) have to be separated. In order to inject fine-grained spatial information to deeper layers, appropriately cropped subvolumes are concatenated to the predecessor’s output on layers indicated by the arrows.

While performing finetuning during the segmentation task we expand our network as aforementioned by subsequently adding the following layer configuration after the feature descriptor part: \{(3, 4, 128, 1)_5, (3, 8, 256, 0)_6, (5, 1, 512, 1)_7\}. In this case the receptive field of a voxel position to be classified is $5^3$.

Additionally to the descriptor part whose structure is fixed, both during pretraining and segmentation task learning, the final task-specific part consists of two fully connected linear layers. The number of neurons in the first linear layer corresponds to half the number of output channels of its predecessor (pretraining: $c_4 = 128$, segmentation: $c_7 = 512$). Finally, the number of output features from the last layers corresponds to the number of trainable classes for each intended task. In practice, this means during the unsupervised pretraining, there are 6 output classes for the different spatial arrangements and 8 classes respectively for the actual intended multi-organ segmentation (7 organs + background).

We perform our experiments by cross validating the different patients in a leave-two-out manner. During each experiment, we iteratively hold out 2 test patients. We randomly split the remaining 18 patients into a validation set of size 2 and use the remaining 16 volumes in our training method. While the unsupervised stage is allowed to fully exploit this set of 16 patients, the following supervised finetuning procedure only has access to a smaller subset of 3 labeled training volumes - now alongside with their labels and again randomly sampled. In doing so, we account for the commonly seen imbalance between images and their costly to generate labelings.

In total, we compare 3 different experimental settings based on the used learning rates and pretraining scheme. As a baseline we train a segmentation network from scratch with weights initialized by the
Figure 4: Left: Training and validation accuracy rate evolution per epoch during the pretraining task. Right: Foreground organ labeling accuracy evolution while finetuning for all three experiments. There is no significant difference between the pretrained methods (d-(different/equal)-lr), but both outperform the method trained from scratch (xelr). Both plots are averaged over all patients.

xavier method. Here employ equal learning rates of $1 \times 10^{-2}$ throughout all network parameters (xelr - xavier equal learning rates). For the remaining two experiments we make use of the pretraining strategy inspired by [3]. During the first pretrain experiment again all network parts employ equal learning rates of $1 \times 10^{-2}$ (delr - inspired by Doersch with equal learning rates), while in the second one we invert the learning rates compared to a finetuning strategy: the deeper fully connected layers, which are assumed to be more task specific now have a different and much smaller learning rate of $1 \times 10^{-4}$ (ddlr - inspired by Doersch with different learning rates). Comparing both pretraining approaches could provide insights if lowering the learning rates for the fully connected layers enforces the descriptor CNN parts to extract even more expressive features. In a second stage, we transfer the pretrained descriptor part weights to the segmentation networks. We finetune the pretrained layers ($lr = 1 \times 10^{-4}$) together with the additional xavier-initialized convolutions ($lr = 1 \times 10^{-2}$) and fully connected layers ($lr = 1 \times 10^{-2}$) to the actual task.

While performing experiments including the auxiliary pretext task, we pretrain the first stage for 50 epochs with a batch size of $2^{5}$ and $2^{12}$ randomly generated pairs per epoch. In all three experiments, finetuning is performed for 50 epochs, with a batch size of $2^{4}$ and $2^{10}$ randomly drawn positions per epoch. In both scenarios we make use of the cross entropy as loss function. When performing the finetuning, additional weights corresponding to the organ label occurrences are passed to this loss to tackle the imbalance between background and organs. To train our models we use the Adam optimizer and implement our experiments in PyTorch. Overall, on one NVidia GeForce GTX 1080 Ti training and testing takes about one day.

In Figure 5, the key results of our experiments are illustrated. The left image shows the training and validation accuracy rates during the learning of the auxiliary task for the ddlr-experiment using all 16 unlabeled training patients. The right image contains the evolution of correctly classified foreground examples during the finetuning task averaged over all patients. Both pretrained methods constantly outperform the xelr-approach, where only the small labeled training data set of 3 patients was available while learning the network parameters. However, although during the ddlr-experiment the learning rates for the task specific fully connected layers are 2 orders of magnitude smaller, there is no difference in performance compared to the delr-experiment with equal learning rates for all network parameters during pretraining.

Figure 6 shows tSNE-plots of the feature space generated by the descriptor part of the network. The figure on the left depicts the initial representation gained only by using the xavier method without any training. After unsupervised pretraining on the auxiliary spatial cues task, the figure on the right demonstrates that already meaningful clusters emerge corresponding to different organ labels.

In order to also provide more graphical insights, in Figure 6 exemplary representations for abdominal multi-organ segmentations are shown. Again, pretrained methods are more similar to the ground
4 Discussion and Conclusion

In our experiments, we demonstrate that already a simple auxiliary pretext task can be sufficient to learn expressive low-level basis features. The proposed strategy by [3] turns out to be an appropriate starting point to pretrain the first layers of a CNN in the medical image context. Because its paradigm can be straightforwardly expanded by an additional dimension and no operations apart from extracting subvolume pairs out of the images have to be executed, the approach does not require expensive preprocessing tasks e.g. for further data augmentation as in [4].
Figure 7: Dice values per organ for all three conducted experiments based on all patients. Both pretrained methods achieve constantly better results than training the segmentation CNN from scratch.

Since our focus in this work is to examine the impact of the proposed pretraining strategy, we do not use a very elaborate CNN-architecture for our segmentation task. Nevertheless a clear trend is visible. Using unsupervised pretraining schemes and therefore making a much broader set of unlabeled training data available, clearly is beneficial with regard to the actual multi-organ segmentation task. As the tSNE-plots indicate, the emergence of spatially plausible clusters facilitates the challenging differentiation of fore- and background. The resulting segmentations do not claim to be comparable to start-of-the-art approaches on the same dataset [6], but in synopsis with the method trained from scratch the result is much better. We assume, that due to the limited number of training images, the xavier initialized segmentation network is much more prone to overfitting. Even the first layers of this method have no choice but to adapt to the available small image set. In contrast, only learning the task specific layers and finetuning expressive low-level descriptor extractors is helping to generalize better on the actual task. However, we could not observe any difference when employing lower learning rates to the fully connected layers during pretraining which we assume to be more task specific. Our assumption that lower layers learn even more expressive features this way could at least in this setup not be verified.

In conclusion, we could demonstrate that using unsupervised pretraining strategies rises the possibility to harness unlabeled image data and therefore to tackle a major bottleneck for CNN-based machine learning approaches in medical image processing, since not the availability of image data poses the problem here, but rather the lack of labeled image data.

Acknowledgments

This work was supported by the German Research Foundation (DFG) under the grant number 320997906 (HE 7364/2-1).

References


