SELF SUPERVISED DEEP REPRESENTATION LEARNING FOR FINE-GRAINED BODY PART RECOGNITION

Pengyue Zhang∗† Fusheng Wang† Yefeng Zheng∗

* Medical Imaging Technologies, Siemens Medical Solutions USA Inc., Princeton, NJ 08540, USA
† Stony Brook University, Stony Brook, NY 11794, USA

ABSTRACT

Difficulty on collecting annotated medical images leads to lack of enough supervision and makes discrimination tasks challenging. However, raw data, e.g., spatial context information from 3D CT images, even without annotation, may contain rich useful information. In this paper, we exploit spatial context information as a source of supervision to solve discrimination tasks for fine-grained body part recognition with conventional 3D CT and MR volumes. The proposed pipeline consists of two steps: 1) pre-train a convolutional network for an auxiliary task of 2D slices ordering in a self-supervised manner; 2) transfer and fine-tune the pre-trained network for fine-grained body part recognition. Without any use of human annotation in the first stage, the pre-trained network can still outperform CNN trained from scratch on CT as well as MR data. Moreover, by comparing with pre-trained CNN from ImageNet, we discover that the distance between source and target tasks plays a crucial role in transfer learning. Our experiments demonstrate that our approach can achieve high accuracy with a slice location estimation error of only a few slices on CT and MR data. To the best of our knowledge, our work is the first attempt studying the problem of robust body part recognition at a continuous level.

Index Terms— Self Supervised Learning, Slice Ordering, Body Part Recognition

1. INTRODUCTION

Deep learning techniques have received much attention in last decade. With the increasing computing power thanks to modern GPUs and large labeled datasets such as ImageNet and PASCAL VOC, deep architectures such as convolutional neural networks (CNNs) have improved the state-of-the-art of many computer vision problems, such as image categorization, object detection, image quality assessment and so on [1]. Recently, many efforts are exploited to apply deep learning techniques to medical imaging tasks with promising results [2].

However, although CNNs have achieved impressive progress, the problem becomes much more complicated in medical domain. For CNNs, a large labeled image set is required for satisfying network training. However, collecting large-scale medical data and annotations is time-consuming and requires expertise from domain experts such as radiologists, which makes training a network from scratch infeasible.

One natural intuition is to learn the network in an unsupervised manner. However, until now unsupervised learning methods have limited performance on learning meaningful representations for a discrimination task. In the literature, one successful solution to alleviate the lack of annotated training samples is to pre-train a network on large-scale natural image datasets (e.g., ImageNet) and then fine-tune the network parameters for specific tasks [3]. Such knowledge transfer based approach is proved to be not only feasible but also superior than training a CNN from scratch in terms of accuracy. Nevertheless, though most natural images and medical images share many low-level features, they still considerably differ in object-level structures. Thus transfer learning from natural image data to medical applications may bring substantial bias which can lead to limited accuracy.

For unsupervised learning in natural language processing domain, the relationship between neighboring words or context, is widely used as a cost-effective yet effective source of supervision [4]. By learning vector representation of words and modeling their similarity and co-occurrence probability, context of a query word or sentence can be automatically predicted. In 2D natural images, spatial relationship between neighboring image patches is also proved to contain useful information for object detection and visual data mining [5, 6]. With one additional spatial dimension, a typical 3D CT or MR volume contains much richer context information. Within a 3D volume, the transversal 2D slices can be easily indexed and the order of slices can be acquired for free as natural indicators of spatial positions. Following this idea, we are able to transform the unsupervised learning problem to a self-supervised one, without the need of any manual efforts on annotation.

CNN-based methods have reported promising results and become the state-of-the-art in body part recognition [7, 8]. However, previous CNN-based body part recognition remains
at a coarse level, while real-world applications require more precise recognition. In [7] and [8], the human body is classified into 5 or 12 parts covering different body ranges. However, human body is a coherent and continuous whole object instead of several unrelated object classes. Two neighboring slices across body part border are likely to have even more similar shapes and structures than two distant slices from the same body part, thus a hard division of the body parts is inappropriate. Once trained, the pre-defined body parts in [7, 8] are fixed, which prohibits re-definition of parts or splitting a large part into several fine-grained small parts. In this work we estimate the height (i.e., distance to the toe) of a transversal slice in a normalized body model. With a continuous output, our approach is more flexible and achieves a much finer level recognition compared to the approach based on several disjoint body parts. Moreover, both [7] and [8] consider only the body parts from nose to kidney instead of the full body, which limits the generalization ability to other body parts.

In this paper, we provide an end-to-end convolutional network called P-CNN (for the paired CNN) to predict the spatial order of slices. We then use the network as a knowledge source for other tasks, e.g., fine-grained body part recognition. The idea behind the use of pre-trained slice ordering model is that correctly recognizing the relative position of slices requires good visual understanding of images. Indeed, as shown in [9], the initial several convolutional layers of a CNN serve similarly like filters which automatically learn common low-level features such as edges, corners and textures. Therefore, weights from the pre-trained P-CNN can be used as a better initialization for body part recognition networks than randomized initialization. Compared to the network pre-trained on ImageNet, P-CNN is also a better knowledge source for medical image applications, closing the gap between pre-trained model and target medical tasks. In our experiments, we demonstrate that by pre-training P-CNN and fine-tuning, we can gain a boost for fine-grained body part recognition in terms of both resolution and accuracy, in CT as well as MR images. As far as we are aware, our work is the first attempt studying the problem of fine-grained body part recognition at a continuous level.

2. SLICE ORDERING WITH PAIRED CNN

Conventional 3D medical data, such as CT or MR volumes are usually interpreted as sets of consecutive 2D slices for clinical diagnosis and treatment. Within high resolution 3D volumes, the difference between the slices within a small range may be very limited. The similar appearances of neighboring slices make it hard to differentiate them and estimate their spatial positions. On the other hand, the slices from distinct body parts may have distinct differences in structure. This makes it also difficult to design unified feature extractors in a handcrafted manner. Therefore, we employ a CNN in our method, which is known as an effective representation learning technique without need of any feature designing expertise. We aim to learn a deep visual representation by pre-training a CNN for an auxiliary problem, i.e. predicting the spatial order of transversal slices. Then the CNN serves as a baseline with good generalization power to fine-grained body part recognition.

For a complete 3D volume, a set of transversal slices as well as their order is available. Though every human body is unique in many details, the scanned CT or MR images share similar variation trend from head to toe. We can use their order as a free source of supervision and use a CNN to learn the intrinsic relationship among slices. Since the pre-training does not involve human annotation, we have more flexibility in choosing the size of training set. Using more data only means more computation time, but does not involve any extra human effort except data collection.

We use a paired convolutional network called P-CNN to learn deep representation for slice ordering. The network architecture (Fig. 1) contains two stages: two AlexNet-like sub-networks for the lower six layers and the upper three global layers. We force the two sub-networks to share weights, leading to 50% fewer parameters in the lower six layers. Dropout layers with ratio 0.5 are used to prevent the network from overfitting. A final softmax loss layer is used for the binary classification. The training process is straightforward: 1) randomly sample a slice pair from the same volume; 2) feed the paired slices to the two subnetworks respectively; 3) fuse their outputs and compute the probabilities for two possible slice ordering outcomes (above vs. below); 4) predict the result as the class with higher probability. To solve this binary classification problem, the network must provide good visual understanding of objects and structures. Thus the network can serve as a universal low-level feature learner. Next, we demonstrate that the pre-trained network can be applied to body part recognition with fine-tuning.

Fig. 1. Network architecture of the pair network for slice ordering. Every conv (convolution) and fc (fully connected) layers except fc8 are followed by a ReLU layer. The kernel size, output number and stride are listed after the layer names.

<table>
<thead>
<tr>
<th>Fc8(2)</th>
<th>Dropout3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fc7(4096)</td>
<td>Dropout4</td>
</tr>
<tr>
<td>Pool5(3x3,256,2)</td>
<td>Pool5_p(3x3,256,2)</td>
</tr>
<tr>
<td>Conv5(3x3,256,1)</td>
<td>Conv5_p(3x3,256,1)</td>
</tr>
<tr>
<td>Conv4(3x3,384,1)</td>
<td>Conv4_p(3x3,384,1)</td>
</tr>
<tr>
<td>Conv3(3x3,384,1)</td>
<td>Conv3_p(3x3,384,1)</td>
</tr>
<tr>
<td>LRN2</td>
<td>LRN2_p</td>
</tr>
<tr>
<td>Pool2(3x3,256,2)</td>
<td>Pool2_p(3x3,256,2)</td>
</tr>
<tr>
<td>Conv2(3x3,256,1)</td>
<td>Conv2_p(3x3,256,1)</td>
</tr>
<tr>
<td>LRN1</td>
<td>LRN1_p</td>
</tr>
<tr>
<td>Pool1(3x3,96,2)</td>
<td>Pool1_p(3x3,96,2)</td>
</tr>
<tr>
<td>Conv1(11x11,96,4)</td>
<td>Conv1_p(11x11,96,4)</td>
</tr>
</tbody>
</table>
3. FINE-GRAINED BODY PART RECOGNITION

Transversal slice based body part recognition is an important pre-processing step which can help computer aided detection and diagnosis in several ways. First, precisely recognized body parts can serve as an ideal initialization for further analysis, e.g., detection, segmentation, and classification. Second, by labeling body part slices, organ/structure-based image retrieval can be easily implemented. Moreover, with fast recognition speed, automatic body part recognition can be integrated into current imaging techniques, such as CT and MR, to enable real-time planning and diagnosis. However, previous body part recognition methods remain at a coarse level, i.e., classifying a slice into one of a few region classes.

In this paper, we model body part recognition as a regression problem instead, which is at a much finer recognition resolution. Our task is shown as in Fig. 2: given a slice from any part of the body, we aim to predict a real-number score in $[0, 1]$ (with 0 for the bottom of a foot and 1 for the head top), indicating the normalized height in the body. To compensate the variation in body part ratio, six anatomical landmarks (head top, neck, lung top, spine, knee and foot) are manually annotated on the training set. Each landmark is mapped to its mean position in our normalized body height model and the height of other slices is obtained by linear interpolation using these six landmarks. This serves as ground truth for training and evaluation.

We demonstrate that the learned P-CNN features can be successfully applied to fine-grained body part recognition for whole body images. We pre-train a P-CNN for slice ordering and copy one of the sub-networks of the lower six layers to the regression network (Fig. 3). The following two fully connected layers are randomly initialized from a Gaussian distribution with zero mean and small variation. An Euclidean loss layer replaces the softmax loss layer as the final layer [10]. We adopt the deep fine-tuning strategy [3] during training: we fine-tune all the layers and apply 1/10 original learning rate for the lower six layers, to preserve the power of the pre-trained network and to boost learning speed for the following fully connected layers.

4. EXPERIMENTS

4.1. Slice ordering

For slice ordering, we train a P-CNN to learn features from both slices and predict the spatial relative position. The slice pair is randomly sampled from the same volume to eliminate the effect of body shape variation between different people. With a large set of training samples covering every part of human body, the order of slices can be predicted no matter which part they belong to. Large variations in shape across slices require a good visual understanding of images, which makes the task challenging.

A set of 370 CT volumes containing either full body or partial body is used for pre-training. The scanner table is removed from the image to eliminate the effect of non-body structures. The 2D transversal slices are extracted and resized to $256 \times 256$ pixels. Since the pre-trained AlexNet uses color images as input, we also transform the grayscale CT slices to color ones by duplicating the grayscale image to the three channels. Each pair of slices is randomly sampled from the same volume and labeled automatically with a binary pair label. Mirror images are also included for data augmentation. We subtract the mean from all images and shuffle the training pairs. Overall 83,000 pairs of slices are used for training and 32,400 pairs for testing. With P-CNN, we can achieve a prediction accuracy of 90% on the test set. The method is implemented based on the Caffe framework [10].

4.2. Fine-grained body part recognition

The proposed fine-grained body part recognition method is validated on two datasets (202 unseen CT volumes and 100 MR volumes). We compare the body part recognition results of different methods/settings: training from scratch (AlexNet-S), pre-trained AlexNet on ImageNet with fine-tuning (AlexNet-F) and pre-trained P-CNN on slice ordering with fine-tuning. The same preprocessing is performed to match the pre-training task. Hyperparameters are set as follows: momentum $\mu = 0.9$; weight decay $\omega = 0.0005$; the
base learning rate $\alpha$ is set as high as possible such that the training still converges. Overall 102 CT volumes (18,700 slices) are used to train/fine-tune the network. The other 100 volumes (19,800 slices) are used as the test set. The CT data used for body part recognition are completely different to those used for pre-training. For MR images, we use 50 volumes (7,140 slices) for training and 50 volumes (7,245 slices) for testing. The test results without post processing on CT and MR image are illustrated in Fig. 4. The scatter plot is close to a diagonal line, which is the ideal case. If the input is a 3D volume, we can apply median filtering or line fitting to the raw regression scores of the slices to denoise and remove outliers for better performance. Quantitatively, 90.3% of CT test slices and 91.1% of MR test slices are predicted with an normalized error of less than 0.05.

The recognition results are also compared with other methods. We train/fine-tune on both CT and MR volumes with three different sizes of training set: 100%, 50% and 25% of all training images. For all configurations, we evaluate on the same unseen CT/MR test set. We also provide results of several baseline methods including: 1) SIFT + Bag-of-Word + Support Vector Regression; 2) SURF + Bag-of-Word + Support Vector Regression. As shown in Table 1, by pre-training on slice ordering task and fine-tuning, we can achieve the smallest Euclidean loss for most cases. We also demonstrate the average recognition error in millimeters in Table 2. With an average body height of 1809 mm and 1740 mm in CT and MR test set respectively, we can achieve a low recognition error of 25.3 mm and 20.1 mm, or 1.40% and 1.16% of full body height respectively. Note that a typical gap between neighboring transversal slices is 5 mm for CT and 10 mm for MR, thus our body part recognition error is just within a few slices.

The results also demonstrate that though P-CNN is pre-trained only on CT data, it can generalize well to body part recognition on both CT and MR images. We also observe that the test performance is significantly affected by training set size. This is natural since more training data cover more cases with more information. With fewer training samples, the model tends to overfit after 10-200 epochs: the validation error begins to increase while the training error keeps decreasing. However, it is noticed that training from scratch overfits most easily and the models tend to overfit earlier with less training data. Without any human annotation, the pre-trained P-CNN outperforms pre-trained AlexNet by a large margin, though the training set of slice ordering is significantly smaller than ImageNet ILSVRC 2012 training set (~89K vs 1.3M). Surprisingly, fine-tuning from pre-trained AlexNet performs not as well as training from scratch for this regression task. This indicates that the dissimilarity between natural images and medical images may pose considerable obstacle on the body part recognition task.

5. CONCLUSION

In this paper, we propose a self supervised approach for deep representation learning for slice based body part recognition. Using only context information in 3D volumes, we can effectively solve the problem of slice ordering by P-CNN. The pre-trained P-CNN can also be transferred and fine-tuned for body part recognition. Our experimental results demonstrate that P-CNN can achieve promising results with minimal use of human annotations.

6. ACKNOWLEDGEMENT

This research is supported in part by grants from National Science Foundation ACI 1443054 and IIS 1350885.
7. REFERENCES


