PCA-aided Fully Convolutional Networks for Semantic Segmentation of Multi-channel fMRI

Lei Tai\textsuperscript{1,3}, Haoyang Ye\textsuperscript{1,3}, Qiong Ye\textsuperscript{2} and Ming Liu\textsuperscript{3}

Abstract—Semantic segmentation of functional magnetic resonance imaging (fMRI) makes great sense for pathology diagnosis and decision system of medical robots. The multi-channel fMRI provides more information of the pathological features. But the increased amount of data causes complexity in feature detections. This paper proposes a principal component analysis (PCA)-aided fully convolutional network to particularly deal with multi-channel fMRI. We transfer the learned weights of contemporary classification networks to the segmentation task by fine-tuning. The results of the convolutional network are compared with various methods e.g. k-NN. A new labeling strategy is proposed to solve the semantic segmentation problem with unclear boundaries. Even with a small-sized training dataset, the test results demonstrate that our model outperforms other pathological feature detection methods. Besides, its forward inference only takes 90 milliseconds for a single set of fMRI data. To our knowledge, this is the first time to realize pixel-wise labeling of multi-channel magnetic resonance image using FCN.

I. INTRODUCTION

Medical-image analysis is indispensable in modern computer-aided diagnosis, therapy planning, and execution. Management of the related pathology relies on several different kinds of imaging processing technologies including positron emission tomography (PET), X-ray computer tomography (CT), ultrasound and magnetic resonance imaging (fMRI) [1]. MRI is amongst the most useful technique for visualization. MRI shows high contrast and high resolution for different body tissues in a non-invasive way. Semantic segmentation of magnetic resonance image is an essential step for diagnosis with the abstraction of relevant features. Through segmentation in the feature space, semantic segmentation minimizes the computational complexity in fMRI applications [2]. For both diagnosis and treatment strategies, accurate classification and segmentation of the tissues are necessary and primordial, such as the quantitative volume measurement of different brain structures in brain fMRI. DWI datasets such as ImageNet [12], it is hard to collect an fMRI dataset for specific tissues of the human body and specific pathology features. Regular semantic segmentation task for most of the benchmarks.

Convolutional networks have been applied to a variety of recognition and perception tasks recently, e.g. image classification [4], object detection [5] and semantic segmentation [6]. The great success of Convolutional Neural Network (CNN) is attributed to the outstanding performance of trained hierarchical structure in data representations. The weights of CNN are typically calculated by back-propagation and gradient descent. This self-driven method has been keeping showing its unlimited potential in other conventional robotic fields [7] [8] as well.

Pixel labeling, which is also called semantic segmentation, is much more challenging compared with image recognition and object detection. The scale of perception changes from a whole image to pixel-wise elements. For image classification, the coarse distribution is taken as the output. But for semantic segmentation, the output is a dense prediction at every pixel location. Not like the traditional recognition and segmentation methods [9], [10], [11], where mutual information is used manually, fully convolutional networks (FCN) [6] is a CNN variant. It achieves the state-of-art effect in semantic segmentation task for most of the benchmarks.

However, not like computer vision tasks with various datasets such as ImageNet [12], it is hard to collect an fMRI dataset for specific tissues of the human body and specific pathology features. Regular semantic segmentation dataset is comprised of 3-channel RGB images, while the parameters of fMRI are not constant based on the application and usage. Realizing such a segmentation in a small dataset with multi-channel fMRI means massive work load for both supervised expertise and medical diagnosis. The multi-
channel infrastructure brings more co-related information, but also increases the data complexity. Training from small datasets like particular pathology and disease features can help build an expert system to recognize the same feature in the scene, which will bring great convenience for doctors. Another obstacle for pixel-wise segmentation in fMRI is the unclear boundary. Not like the typical RGB images by digital cameras, junction area between different tissues in fMRI are hardly recognizable even for experienced doctors. It also adds the complexity of the labeling process.

Motivated by the facts mentioned above, we stress the following contributions and features of this work:

- We propose a PCA-aided FCN structure to process 31-channel fMRI and to achieve pixel-wise semantic segmentation for primary tissues. Through PCA, the dimension of the origin data will be reduced to three and be further taken as the input of the fully convolutional network.
- The fine-tuning is implemented on a small-sized dataset as transfer learning. The experiment results show outperforming efficiency compared with conventional methods for fMRI feature learning.
- A revised labeling strategy is proposed, which ignores the junction areas between tissues, which significantly decreases the complexity of semantic labeling for segmentation.

The rest of paper is organized as follows. We present related works in fMRI and pixel labeling in Section II. In Section III, we describe our pre-processing configurations of the dataset with PCA and normalization. The configuration of the whole FCN structure is also introduced in this section. The detail of the training and evaluation on specific datasets are then presented in Section IV. At the end, Section V concludes the paper.

II. RELATED WORK

A. fMRI Segmentation

Most of conventional brain fMRI segmentation tasks are probabilistic, based on clustering methods like k-means. Usually, the MR images are represented as a collection \( X = \{x_1, x_2, \cdots, x_n\} \) with every element representing a pixel of the image. The clustering algorithm returns \( C = \{c_1, c_2, \cdots, c_m\} \) as the cluster centers. \( C \) are commonly initialized in random.

Barrad et al. [1] used a revised fuzzy clustering algorithm to segment fMRI. Standard fuzzy c-means considers the clustering an optimization problem. The objective function is defined as:

\[
\arg \min_C \sum_i^n \sum_j^m w_{ij} \|x_i - c_j\|^2
\]

where \( w_{ij} \) is the degree to which element \( x_i \) belongs to cluster \( c_j \) and \( h \) is the fuzzifier to determine the level of cluster fuzziness. Euclidean distance is used to evaluate the element and cluster center directly. After that, \( w_{ij} \) and \( c_j \) can be estimated iteratively, as

\[
w_{ij} = \frac{1}{\sum_{k=1}^m \left( \frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{-h}}
\]

\[
c_j = \frac{\sum_{i=1}^n w_{ij} x_j}{\sum_{i=1}^n w_{ij}^h}
\]

In [1], the nearby area of the element in an image was considered as well. Gaussian mixture model was also used to address brain MR images clustering [3]. Compared with fuzzy clustering mentioned above, the Gaussian kernel takes the place of norm functions between elements and cluster centers.

There are some challenges about ventricle fMRI segmentation, such as Statistical Atlases and Computational Modeling of the Heart (STACOM) [13] and Medical Image Computing and Computer Assisted Intervention (MICCAI) [14]. Petitjean supplied a review of conventional methods for left ventricle segmentation[15], which consisted of graph cuts and deformable models, etc.

All of these methods provided a reasonable segmentation result for fMRI datasets. However, these statistical-based or feature-based method were limited, because they tend to overfit on specific dataset like ventricle. Their utility in a generic dataset was extremely limited as well.

B. ConvNet in Semantic Segmentation

The convolutional neural network is leading the artificial intelligence right now. Related research [16] just beat the best human player in Go game which was once regarded as the most difficult task for artificial intelligence. In image classification, deep residual network [17] with hundreds of convolutional layers drove the network to learn identity feature and improve the accuracy of classification tasks in ImageNet [12]. Faster-rcnn [5] addressed all of the tasks in object detection including region proposal, bounding box regression, instance location, and classification. CNN related methods take the state-of-the-art place in almost every computer vision tasks. Not only in vision fields, the successes of RNN and LSTM in natural language processing [18] and speech recognition [19] are all based on revised convolutional neural network. In terms of robot control, motivated by the game controller developed by Google [20], deep reinforcement learning is implemented to teach the robot be adapted to the unfamiliar environment and achieve obstacle avoidance ability [7] [8]. We adopt most of these concepts in this work.

Among various vision tasks, a high-level prediction like semantic segmentation is much more complicated compared with simply predicting image context as single output. While for pixel-wise classification, the computation source is thousands times more than the image classification.

Long et al. [6] re-architected and fine-tuned classification nets to achieve dense prediction of semantic segmentation directly. They improve the efficiency of convets in semantic segmentation dramatically because FCN extended a
convolutional network to arbitrary-sized inputs [21]. Deep Convolutional Neural Networks (DCNN) were not sufficient to achieve semantic segmentation separately because of the poor localization property. Chen [22] combined the feature map at the final layer of DCNNs with a fully connected Conditional Random Field (CRF). The main breakthrough in [22] was the intersection of convolutional network and probabilistic graphical models. Notice that, CRF and FCN are the state-of-the-art semantic segmentation methods nowadays. [23] applied FCN processing cardiac segmentation in the ventricle dataset mentioned in Section II-A, which also focused on a unique dataset but not generic multi-channel fMRI.

III. PCA-aided Fully Convolutional Networks

Our approach benefits a lot from the recent popular deep net-related works including image classification [4], semantic segmentation [6] and transfer learning [24]. After PCA, the processed fMRI data are taken as the input of FCN.

A. Data Gathering and Preparation

This study was approved by the local institutional ethic review board. Written informed consent was obtained from all volunteers. Six young and healthy volunteers (age range, 40 years; median age, 38 years; three males and three females) were recruited in this study. In addition to the standard criteria for exclusion from fMRI, further exclusion criteria were applied using the absence of cervical spine related pain and disorders.

The dataset focuses on the vertebral area of human being. Except the background, there are five primary tissues including bones, spinal fluid, lumbar disc, vertebral body and cerebrospinal fluid shown in Fig. 1.

fMRI examinations were carried out on 3T clinical MR scanner (Achieva, Philips Healthcare, Best, The Netherlands) with standard abdomen coil. Mid-sagittal multi-echo T2-weighted images were acquired using a multi-shot spin-echo sequence with echo-planar readout and following parameters shown in Table I. Here TE means echo time. The intensity of image under different echo time reflects the attenuation of tissues. With 32 different echo times, 32 channels representing the different intensity of various tissues. The first echo of the sequence was excluded from further analysis to avoid the effect of the stimulated echo [25]. After cutting the black background edge, the dimension of every sample in the dataset turns to be $31 \times 256 \times 154$.

B. PCA Processing

It is evident that the 31-channel fMRI provide more information in detail about the pathology features. But because 31-channel images increase the calculation complexity dramatically, a dimension reduction method is necessary. Besides, state-of-the-art CNN structures were usually implemented on vision tasks which take raw RGB images with three channels as input. Preliminaries experiments also tell that if we attach the images with 31 channels as input layer to the network, the weights of the network is very difficult to converge. These difficulties are to be moderated.

In PCA analysis, we vectorize every channel of the sample to a vector. The sample is transformed to a matrix $X$ as $31 \times 39424$. PCA is a variant matrix transformation as

$$
\hat{X}_k = X - \sum_{s=1}^{k-1} X w_{(s)} w_{(s)}^T
$$

Here $w_{(s)}$ means the related eigenvector of $XX^T$'s sorted eigenvalues. And $w_{(s)}$ is also a united vector. It can be calculated like

$$
w_{(1)} = \text{arg} \max \left( \frac{w^T X^T X w}{w^T w} \right)
$$

The left part of Fig. 2 shows a sample of the origin 31-channel images and the transformation result after PCA and normalization processing. Fig. 3 shows all of the 31 singular values of $X$ which is also the square root of eigenvalues of $XX^T$ from all of the six samples of the datasets. The first three singular values take the weights of all the singular values in more than $99.9\%$. That means the transformation is more efficient and meaningful. The largest three singular values of $X$ can almost represent all of the origin information in the raw sample equivalently.

The fMRI datasets have a broad range of pixel intensities. The different ranges significantly influence the accuracy of segmentation. Even after PCA processing, the ranges of the three channels are still quite irregular. To locate the maximal and minimal elements among the three channels, we normalize the three channels to a range of $(0,255)$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOV</td>
<td>$240 \times 155 \times 6(FH\times AP\times RL)$</td>
<td>mm$^3$</td>
</tr>
<tr>
<td>matrix</td>
<td>$120 \times 72$</td>
<td>-</td>
</tr>
<tr>
<td>reconstruction</td>
<td>$32$</td>
<td>-</td>
</tr>
<tr>
<td>slice thickness</td>
<td>$5.0$</td>
<td>mm</td>
</tr>
<tr>
<td>No. of slice</td>
<td>$1$</td>
<td>-</td>
</tr>
<tr>
<td>No. of TEs</td>
<td>$32$</td>
<td>-</td>
</tr>
<tr>
<td>TE/TR</td>
<td>$8.0 \times 32/1202$</td>
<td>ms</td>
</tr>
<tr>
<td>NA</td>
<td>$8$</td>
<td>-</td>
</tr>
<tr>
<td>fat suppression</td>
<td>SPIR</td>
<td>-</td>
</tr>
<tr>
<td>scan time</td>
<td>$235.5$</td>
<td>s</td>
</tr>
</tbody>
</table>

TABLE I

fMRI GATHERING PARAMETERS
C. FCN Implementation

Fig. 4 illustrates our proposed FCN architecture. In the convolutional networks, each layer is a three-dimensional array of size $h \times w \times d$, where $d$ is the channel dimension and $h, w$ are the spatial sizes. Here the dimension of the input data is following the pre-processed fMRI data. The size of input array is thus $256 \times 154 \times 31$.

1) Dense Prediction: The feature learning of FCN uses the structure of VGG-16 net [4] which won ILSVRC14. As introduced in [6], all of the fully connected layers of VGG-16 are converted to convolutions. Finally, there are 15 convolutional (Conv) layers for feature learning. To increase non-linearity, every Conv layer is followed by a Rectified Linear Unit (ReLU) activation function layer. They are represented as red cubes in Fig. 4. The number under the Conv+ReLU cubes in Fig. 4 is the number of channels of the output data related to this cube. A $1 \times 1$ convolution layer with channel dimension six is appended to predict scores for each of the tissue classes including the background. As shown in Fig. 4, actually the Prediction cube is also a convolution layer. All of the output channel numbers of prediction convolution layers is six as well. The Deconv layer will upsample the coarse output to the similar size as input ($256 \times 154$).
TABLE II
TRAINING PARAMETERS

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>batch size</td>
<td>1</td>
</tr>
<tr>
<td>learning rate</td>
<td>$10^{-14}$</td>
</tr>
<tr>
<td>momentum</td>
<td>0.99</td>
</tr>
<tr>
<td>weight decay</td>
<td>$0.0005$</td>
</tr>
</tbody>
</table>

2) Hierarchy Combination: Max pooling layers, represented by blue cubes, are implemented after several Conv+ReLU cubes to spatially reduce the data size. We use the same strategy to combine layers of the feature hierarchy as introduced in [6]. The output of Pool3, Pool4, Pool5 are to be unsampled to with different strides. As the sequence illustrated in Fig. 4, Deconv1 and Predict2 will be firstly added conjugatedly and unsampled to Deconv2. Then Deconv2 and Predict3 will be added in sequence and unsampled to Deconv3. The dimension of Deconv3 is $6 \times 256 \times 154$. At last, a Softmax layer is appended to output the final prediction.

IV. EXPERIMENTS AND RESULTS

A. Training and Test

The whole structure is trained end-to-end and the initial weights of network are transferred from the training result of VOC dataset [26] in voc-fcn8s \(^1\) structure which was pre-trained with the parameters in [6]. In the fine-tuning procedure, we use Stochastic Gradient Descent (SGD) method with momentum. The latest research [27] showed that setting the batch size to one is beneficial for semantic segmentation. We tried different learning rate from $10^{-5}$ to $10^{-14}$, and smallest learning rate results in the best performance. One possible explanation is that the initialization from ImageNet-trained weights is important, so the tuning that makes it adaptive to the new datasets should be tiny and slow. Training parameters is described in detail in II.

We use Caffe [28] python interface as the deep learning framework to train and test all of the data and models. A unit NVIDIA Tesla K40c is used to support the computation.

There are six MR images in the vertebral datasets as illustrated in Fig. 3. We separate one of them as the test sample. Fig. 5 shows the training and test loss decreasing with the iteration number increasing. The two strategies introduced in Section III-B are all implemented and tested. All of the losses experience substantial reduction at first 5000 steps. After that, because of the existence of ignored area, the loss of Ignore-Bound strategy is obviously less than the one with the Fully-BP strategy. The test losses of both strategies are converged stably after 15000 iterations.

B. Results

A naive K-Nearest-Neighbor (KNN) method, which is commonly used nowadays, is implemented as comparison as well.

\[^1\]http://dl.caffe.berkeleyvision.org/fcn8s-heavy-pascal.caffemodel

In this experiment, there are five samples in the training dataset as mentioned in Section III-A. From the training dataset, we got $5 \times 256 \times 154$ pixels for each scan. After PCA and normalization process, there is a 3-dimensions vector for every pixel location. We calculate the median 3-dimensional vectors for pixels belong to same class and denote it by $y_j$ for class $j$. For all of the six classes introduced in Section III-A, let $Y = \{y_1, y_2, \ldots, y_6\}$ be the medians set for all of six classes. For a pixel in the test sample after PCA and normalization process, let the 3-dimensional pixel vector be $x$. We can locate the least normalized inner product between $x$ and the element in the median sets $Y$ as:

$$\arg \min_j x^T y_j / \|x\| \|y_j\|$$

Then $j$ is the prediction result of pixel $x$. By scanning the whole image, we can get the prediction for every pixel in the test data.

TABLE III
PARAMETER REPRESENTATIONS FOR CRITERIA

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i, j$</td>
<td>specific class number</td>
</tr>
<tr>
<td>$n_{ij}$</td>
<td>pixels of class $i$ predicted to belong to class $j$</td>
</tr>
<tr>
<td>$n_c$</td>
<td>number of classes (6)</td>
</tr>
<tr>
<td>$s_i$</td>
<td>pixels labeled with class $i$ ($\sum_j n_{ij}$)</td>
</tr>
</tbody>
</table>

We use several common semantic segmentation and scene parsing methods to compare the experiment results, using criteria like pixel accuracy and region intersection over union (IU). Based on the parameters introduced in Table III, we compute:

- mean IU:

$$\frac{1}{n_c} \sum_i n_{ii} / (s_i + \sum_j n_{ji} - n_{ii})$$

- frequency weighted IU (fw IU):
The evolution of test results and label maps are described in Fig. 1. As the result of loss reduction, the shapes of the segmentation results progressively tended to the same as the ground-truth at the first several thousands steps. After 15000 iterations, there is hardly any revision anymore. By comparison, we find that the Fully-BP strategy prefers to predict the junction area as background. On the other hand, the Ignore-Bound strategy is prone to regard the junction area as tissues especially for bone area. The confusion matrices of the test result after 15000 iterations for both strategies which reflect the difficulties of prediction for bones are illustrated in Fig. 6. An probable explanation is that the bone area is the most difficult one to label with unclear boundary mostly as shown in Fig. 2. There may be too many labeling mistakes in this area with the Fully-BP strategy. For the precision of different classes, the Ignore-Bound strategy substantially improves the precision of all classes except the background part. However, it takes the cost to increase the false positive counts for primary tissues as well.

Comparison among various methods and FCN methods against different training periods are listed in Table II. FCN methods show great improvement compared with other methods. Another fact is that FCN only needs 90 milliseconds to calculate the forward prediction, and the conventional method takes 1.4 seconds to finish the traversal analyzing for every pixel in the whole image. From Table II, the Fully-BP strategy shows high performance in mean-IU and frequency-
weighted-IU. The Ignore-Bound strategy has increased accuracy especially for primary tissues. For the five primary classes of tissues except the background, the differences in mean IU and fw IU are also reduced compared with the Fully-BP strategy. And the best performance of the Ignore-Bound strategy appeared after around 10000 iterations. It takes less time to achieve the satisfied result compared with the Fully-BP strategy which need 30000 iterations. Generally speaking, the Ignore-BP strategy shows benefits in converging speed and accuracy.

V. CONCLUSION

This papers demonstrated the effectiveness to apply FCN deep-learning methods on semantic segmentation of multi-channel fMRI. The utility of a PCA-aided fully convolutional network model is proved under efficient end-to-end training using a small-sized dataset. It indicates that the transfer learning and Ignore-Bound label strategy are evidentially sufficient when collecting small datasets, labeling them and realizing the pixel-wise classification in near real-time. The proposed framework outperforms other state-of-the-art methods in precision.

A few limitations to our approach remain. Most notably, as with most CNN approaches, the training parameters is decided after a lot of attempts. Fine-tuning from other pre-trained models is also time consuming. The labeling process of Ignore-Bound strategy need an appropriate threshold to distinguish the junction area.

In the future, the labeling strategy can be further improved considering the limitations in parts of the experimental results. The combination with CRF introduced in [22] should also be implemented and tested. Eventually, we will not only test on the tissue segmentation but also apply the segmentation method on pathology related areas directly.

REFERENCES


