Computerized Medical Imaging and Graphics 66 (2018) 90–99

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

An application of cascaded 3D fully convolutional networks for medical image segmentation

Holger R. Roth[a,*], Hirohisa Oda[a], Xiangrong Zhou[b], Natsuki Shimizu[a], Ying Yang[a], Yuichiro Hayashi[a], Masahiro Oda[a], Michitaka Fujiwara[c], Kazunari Misawa[d], Kensaku Mori[a,*]

[a] Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Japan
[b] Gifu University, Yanagido, Gifu, Japan
[c] Nagoya University Graduate School of Medicine, Nagoya, Japan
[d] Aichi Cancer Center, Kanokoden, Chikusa-ku, Nagoya, Japan

ARTICLE INFO

Keywords:
Fully convolutional networks
Deep learning
Medical imaging
Computed tomography
Multi-organ segmentation

ABSTRACT

Recent advances in 3D fully convolutional networks (FCN) have made it feasible to produce dense voxel-wise predictions of volumetric images. In this work, we show that a multi-class 3D FCN trained on manually labeled CT scans of several anatomical structures (ranging from the large organs to thin vessels) can achieve competitive segmentation results, while avoiding the need for handcrafting features or training class-specific models.

To this end, we propose a two-stage, coarse-to-fine approach that will first use a 3D FCN to roughly define a candidate region, which will then be used as input to a second 3D FCN. This reduces the number of voxels the second FCN has to classify to ~10% and allows it to focus on more detailed segmentation of the organs and vessels.

We utilize training and validation sets consisting of 331 clinical CT images and test our models on a completely unseen data collection acquired at a different hospital that includes 150 CT scans, targeting three anatomical organs (liver, spleen, and pancreas). In challenging organs such as the pancreas, our cascaded approach improves the mean Dice score from 68.5 to 82.2%, achieving the highest reported average score on this dataset. We compare with a 2D FCN method on a separate dataset of 240 CT scans with 18 classes and achieve a significantly higher performance in small organs and vessels. Furthermore, we explore fine-tuning our models to different datasets.

Our experiments illustrate the promise and robustness of current 3D FCN based semantic segmentation of medical images, achieving state-of-the-art results.[1]

1. Introduction

Recent advances in fully convolutional networks (FCN) have made it feasible to train models for pixel-wise segmentation in an end-to-end fashion (Long et al., 2015). Efficient implementations of 3D convolution and growing GPU memory have made it possible to extent these methods to 3D medical imaging and train networks on large amounts of annotated volumes. One such example is the recently proposed 3D U-Net (Çiçek et al., 2016), which applies a 3D FCN with skip connections to sparsely annotated biomedical images. Alternative architectures for processing volumetric images have also been successfully applied to 3D medical image segmentation (Milletari et al., 2016; Chen et al., 2016; Dou et al., 2017). In this work, we show that a 3D FCN, like 3D U-Net, trained on manually labeled data of several anatomical structures (ranging from the large organs to thin vessels) can also achieve competitive segmentation results on clinical CT images, very different from the original application of 3D U-Net using confocal microscopy images. We furthermore compare our approach to 2D FCNs applied to the same images.

Our approach applies 3D FCN architectures to problems of multi-organ and vessel segmentation in a cascaded fashion. A FCN can be trained on whole 3D CT scans. However, because of the high imbalance between background and foreground voxels (organs, vessels, etc.) the network will concentrate on differentiating the foreground from the background voxels in order to minimize the loss function used for training. While this enables the FCN to roughly segment the organs, it causes particularly smaller organs (like the pancreas or gallbladder) and vessels to suffer from inaccuracies around their boundaries.

* Corresponding authors.
E-mail addresses: rothhr@mori.m.is.nagoya-u.ac.jp (H.R. Roth), kensaku@is.nagoya-u.ac.jp (K. Mori).
[1] Our code and trained models are available for download: github.com/holgerroth/3Dunet_abdomen_cascade.

https://doi.org/10.1016/j.compmedimag.2018.03.001
Received 27 October 2017; Received in revised form 12 March 2018; Accepted 12 March 2018

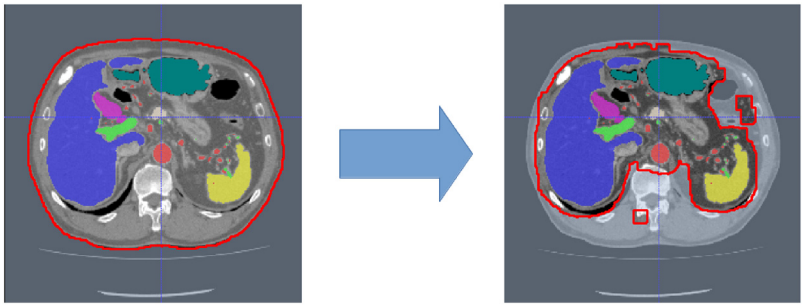

Fig. 1. Cascaded 3D fully convolutional networks in a coarse-to-fine approach: the first stage (left) learns the generation of a candidate region for training a second-stage FCN (right) for finer prediction. Outlined red area shows candidate region C_1 used in first stage and C_2 used in second stage. Colored regions denote ground truth annotations for training (best viewed in color).

To overcome this limitation, we learn a second-stage FCN in a cascaded manner that focuses more on the boundary regions. This is a coarse-to-fine approach in which the first-stage FCN sees around 40% of the voxels using only a simple automatically generated mask of the patient's body. In the second stage, the amount of the image's voxels is further reduced to around 10%. In effect, this step narrows down and simplifies the search space for the FCN to decide which voxels belong to the background or any of the foreground classes; this strategy has been successful in many computer vision problems (Viola and Jones, 2004; Li et al., 2016). Our approach is illustrated on a training example in Fig. 1.

1.1. Related work

Multi-organ segmentation has attracted considerable interest over the years. Classical approaches include statistical shape models (Cerrolaza et al., 2015; Okada et al., 2015), and/or employ techniques based on image registration. So called multi-atlas label fusion (Rohlfing et al., 2004; Wang et al., 2013; Iglesias and Sabuncu, 2015) has found wide application in clinical research and practice. Approaches that combine techniques from multi-atlas registration and machine learning are also common place and have been successfully applied to multi-organ segmentation in abdominal imaging (Tong et al., 2015; Oda et al., 2016). However, a fundamental disadvantage of image registration based methods is there extensive computational cost (Iglesias and Sabuncu, 2015). Typical methods need hours of computation time in order to complete on single desktop machines (Wolz et al., 2013).

The recent success of deep learning based classification and segmentation methods are now transitioning to applications of multi-class segmentation in medical imaging. Recent examples of deep learning applied to organ segmentation include (Roth et al., 2017; Zhou et al., 2016b,a; Christ et al., 2016). Many methods focus on the segmentation of single organs like prostate (Milletari et al., 2016), liver (Christ et al., 2016), or pancreas (Roth et al., 2015, 2016b). Multi-organ segmentation in abdominal CT has also been approached by works like (Hu et al., 2017; Gibson et al., 2017). Most methods are based on variants of FCNs (Long et al., 2015) that either employ 2D convolutional layers in a slice-by-slice fashion Roth et al. (2016b); Zhou et al. (2016b,a); Christ et al. (2016), 2D convolutions on orthogonal (2.5D) cross-sections (Roth et al., 2015; Prasoon et al., 2013), and 3D convolutional layers (Milletari et al., 2016; Chen et al., 2016; Dou et al., 2017; Kamnitsas et al., 2017). A common feature of these novel segmentation methods is that they are able to extract the features useful for image segmentation directly from the training imaging data, which is crucial for the success of deep learning (LeCun et al., 2015). This avoids the need for *hand-crafting* features that are suitable for detection of individual organs.

1.2. Contributions

Due to the automatic learning of image feature and in contrast to previous approaches of multi-organ segmentation where separate models have to be created for each organ (Oda et al., 2016; Tong et al., 2015), our proposed method allows us to use the same model to segment very different anatomical structures such as large abdominal organs (liver, spleen), but also vessels like arteries and veins. Furthermore, other recent FCN-based methods that applied in medical imaging in cascaded/iterative fashion were often constrained to using rectangular bounding boxes around single organs (Roth et al., 2017; Zhou et al., 2016b) and/or performing slice-wise processing in 2D (Christ et al., 2016; Zhou et al., 2016a).

2. Methods

Convolutional neural networks have the ability to solve challenging classification tasks in a data-driven manner. Given a training set of images and labels $\mathbf{S} = \{(I_n, L_n), n = 1, \ldots, N\}$, I_n denotes the raw CT images and L_n denotes the ground truth label images. Each L_n contains K class labels consisting of the manual segmentations of the foreground anatomy (e.g. artery, portal vein, lungs, liver, spleen, stomach, gallbladder, and pancreas) and the background for each voxel in the CT image. Our employed network architecture is the 3D extension by Çiçek et al. (2016) of the U-Net proposed by Ronneberger et al. (2015). U-Net, which is a type of fully convolutional network (FCN) (Long et al., 2015) was originally proposed for bio-medical image applications, utilizes deconvolution (Long et al., 2015) (or sometimes called up-convolutions (Çiçek et al., 2016)) to remap the lower resolution feature maps within the network to the denser space of the input images. This operation allows for denser voxel-to-voxel predictions in contrast to previously proposed sliding-window CNN methods where each voxel under the window is classified independently making such architecture inefficient for processing large 3D volumes. In 3D U-Net, operations such as 2D convolution, 2D max-pooling, and 2D deconvolution are replaced by their 3D counterparts (Çiçek et al., 2016). We use the open-source implementation of 3D U-Net[2] based on the Caffe deep learning library (Jia et al., 2014). The 3D U-Net architecture consists of analysis and synthesis paths with four resolution levels each. Each resolution level in the analysis path contains two $3 \times 3 \times 3$ convolutional layers, each followed by rectified linear units (ReLU) and a $2 \times 2 \times 2$ max pooling with strides of two in each dimension. In the synthesis path, the convolutional layers are replaced by deconvolutions of $2 \times 2 \times 2$ with strides of two in each dimension. These are followed by two $3 \times 3 \times 3$ convolutions, each of which has a ReLU. Furthermore, 3D U-Net employs shortcut (or skip) connections from layers of equal resolution in the analysis path to provide higher-resolution features to the synthesis path (Çiçek et al., 2016). The last layer contains a $1 \times 1 \times 1$ convolution that reduces the number of output channels to the number of class labels K. This architecture has over 19 million learnable parameters and can be trained to minimize a weighted voxel-wise cross-entropy loss (Çiçek et al., 2016). A schematic illustration of 3D U-Net is shown in Fig. 2.

2.1. Loss function: adjustments for multi-organ segmentation

The voxel-wise cross-entropy loss is defined as

[2] http://lmb.informatik.uni-freiburg.de/resources/opensource/unet.en.html.