COMPARISON OF APPROACHES TO SEMANTIC SEGMENTATION OF PULMONARY NODULES IN COMPUTED TOMOGRAPHY SCANS

Anthony Liu, Dani Sim, Bruce Chen

Brandeis University, Department of Computer Science

ABSTRACT

The standard practice for visualizing a three-dimensional profile of the human thorax is a computed tomography (CT) scan. For every interval of slice thickness, axial slice images are taken by x-rays and ultimately concatenated to capture a patient's lungs and surrounding tissue. The heterogeneity across pulmonary CT scans, similarity in densities of pulmonary structures, and poor resolution pose an issue for segmentation, whether that be a task conducted by a radiologist or via a computer-aided detection (CAD) program. A relevant method for the latter to conduct such a task is semantic segmentation, a pixel-wise dense classification. Adapted from National Data Science Bowl 2017 finalist, Julian de Wit, a U-Net style fully convolutional network (FCN) was implemented to segment out the regions of interest in pulmonary computed tomography images. Subsequently, a comparison and an investigation of ROI parsing techniques was conducted. Our neural networks were trained on subsets from the LUNA16 dataset, available at `https://luna16.grand-challenge.org/data/`. The LUNA16 dataset is not a balanced dataset with the number of negative samples greatly overwhelming the number of positive, cancerous samples. The greatest dice similarity coefficient (DSC) achieved for our FCN implementation was 0.326.

1. INTRODUCTION

This report covers our work to implement a U-Net FCN to semantically segment two-dimensional slices of pulmonary CT scans, compare it to methods that achieved better accuracy, and evaluate the shortcomings of our implementation. We conducted this research as a part of the course "COSI 177A - Scientific Data Processing in Matlab". In addition to trying to solve the overall problem of semantic segmentation, this project also served as our efforts to learn and research more on machine learning as a field. The task of segmentation is inherently difficult due to the poor resolution and narrow range of intensities afforded by CT scans. In addition, the LUNA16 dataset is not balanced, an issue that can be dealt with during preprocessing and preparation of the training holdout files. Each pixel was categorized as belong to one of two classes: nodule and not-a-nodule. The performance was assessed by calculating the Dice Similarity Coefficient (DSC).

2. DATA

The CT scan images and segmentation annotations used for ground truth in cross-validation were sourced from the LUNA16 challenge. CT scan images were stored in metaimage format as paired mhd and raw files. It should be noted that there is a heavy bias for negative samples in the dataset with only 1186 positive samples of cancerous xyz coordinates across all ten subsets. The resolution also varies; the smallest slice thickness is 0.8 mm with the greatest value being 2.5 mm. The annotations csv file is formatted such that there are columns with patient id, xyz world coordinates, nodule diameter, and cancerous (1)/not cancerous (0).

3. NETWORK ARCHITECTURE

The network architecture is illustrated in Figure 1. It consists of a contracting path (left side) and an expanding path (right side). The contracting path disputes semantics of the pixel, taking context into consideration, and the expanding path localizes the pixel. Both the contracting and expanding paths are nearly identical to the original architecture of U-Net except for some slight modifications. The contracting path is made up of repeating convolutional layers consisting of 3x3 unpadded convolutions with a ReLU activation layer following and after every two convolutional layers follows a 2x2 max pooling operation with a stride of 2. At each downsampling step the number of channels are doubled, beginning with 32 and ending with 512. One of the two modifications made consists of a batch normalization layer following the max pooling operation. A batch normalization layer was added with the reasoning that by reducing the covariate shift from earlier layers, we would reduce the redistribution of weights from the earlier layers. This allows the model train faster, allows the application of a higher learning rate and reduces overfitting. Every step in the upsampling path consists of an upsampling of the feature map via merging the previously layer with one from the downsampling path. This allows merging of both coarse and fine features, after which the filters are halved and two 3x3 convolutions are applied. The

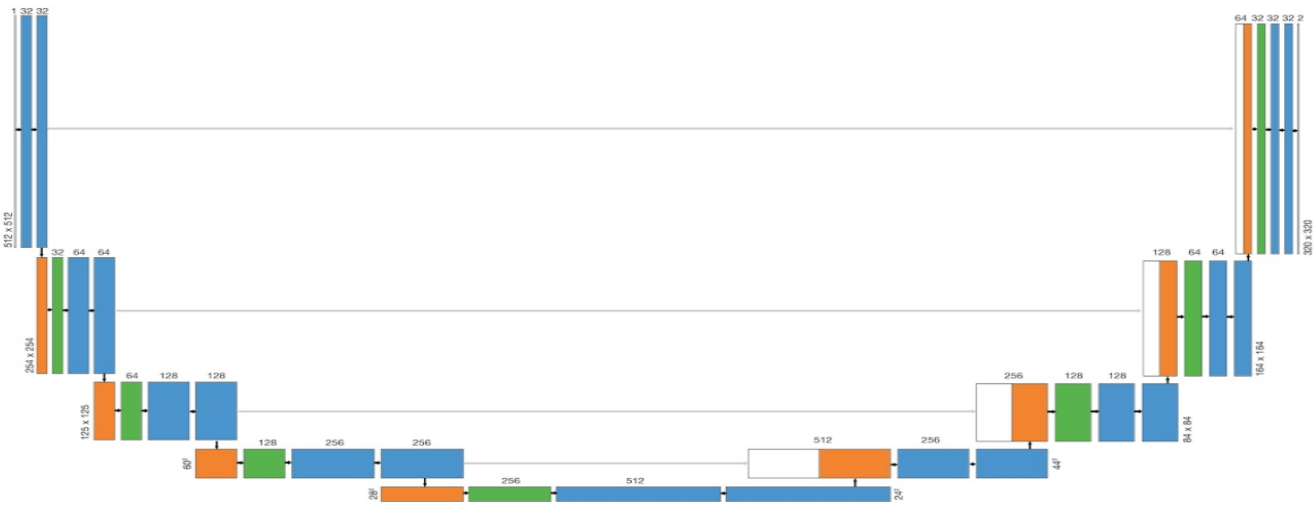


Fig. 1. Schematic of implemented U-Net Fully Convolutional Neural Network.

final layer is a 1x1 convolution that is used to map each of the 32-component feature vectors to one of two classes.

The second modification made was the use of the Momentum and Nesterov optimizer combination which was used in place of the default ADAM optimizer. This decision was due to previous papers citing that this optimizer hybrid was conducive to faster convergence to the absolute minima and reduced oscillation.

4. DATA AUGMENTATION

The LUNA16 dataset contains annotations for 551,065 candidates across all the data subsets with only 1186 cancerous nodules. With a dataset that has three orders of magnitude more negative samples than positive samples, the train FCN's capacity to distinguish between cancerous and non-cancerous regions will be skewed, leading to problematic false negatives. Data augmentation can be implemented to balance a dataset without throwing out negative samples. Random transpositions can be applied to the positive samples to augment the number of cancerous nodules for the FCN to train on. The Keras wrapper allows for the implementation of data augmentation with IMAGEDATAGENERATOR(). Data augmentation is only applied during training on-the-fly.

5. TRAINING

The input images and their corresponding segmentation masks were used to train the network with the stochastic gradient descent implementation in Keras. Due to the unpadded input, the output images were smaller than the input images by a constant border width. The images of size 512*512 were fed in mini-batches of size 4 and trained over 100 epochs, taking around 3 hours to complete in a subsection of the larger dataset (206,650 candidates) running on one NVIDIA GTX-Titan GPU. As mentioned previously, the loss function used was the dice loss coefficient which is defined as:

$$DSC(A,B) = \frac{2 * (A \cap B)}{(A + B)}$$

where A is the classification result and B is the ground truth. The DSC is a measure of how accurate the neural network performs, measuring not only how many positives the network generates, but also penalizes for the false positives that it cannot find.

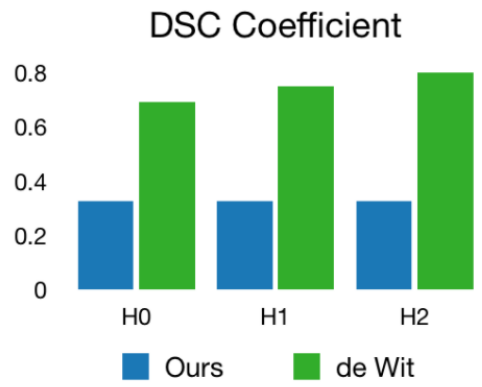
6. RESULTS

After training, our network scored a DSC of 0.326 while the network created by de Wit scored a maximum score of 0.810. While our model only used one cohort, de Wit's implementation utilized a mixture of models in order to maximize their results. Figure 3 depicts some generated plots that indicate the deficits in our trained U-Net FCN. Only a small part of the cancerous annotated region is predicted to be cancerous in the predicted mask (lower left plot in Fig. 3). This suggests a couple issues: More intensive preprocessing such as thresholding may help to detect more of the cancerous region. Additionally, due to the fact that our group's implementation did not take into consideration the negative bias of the dataset, there indeed seems to be an overt bias for detecting negative, non-cancerous regions.

7. DISCUSSION

The deficits in the accuracy we were able to achieve compared to that accomplished by Julian de Wit might be ex-

(a)

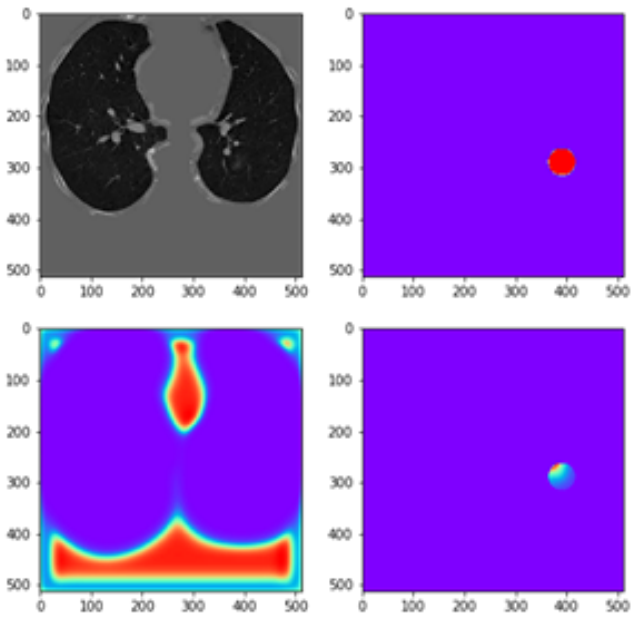


(b)

	h0	h1	h2
Ours	0.326	0.326	0.326
de Wit	0.691	0.746	5415

Fig. 2. (a) Comparison of DSC coefficients achieved by our FCN (blue) and that achieved by Julian de Wit and Daniel Hammack's FCN (green). (b) Table of DSC coefficients

Fig. 3. Predicted mask generated by our FCN implementation for one test axial image. (a) top left: original axial slice (b) top right: true mask from LUNA16 annotations (c) bottom left: predicted mask (d) bottom right: predicted mask overlaid with true mask



plained by multiple factors: De Wit's results were a combination of outputs from multiple models. One was trained on the full LUNA16 dataset, another was trained on thresholded nodules from LUNA16, and one was trained on the NDSB dataset and thresholded nodules from NDSB. Subsequently, De Wit boosted his classification results from his U-Net segmentation FCN with that of his collaborator, Daniel Hammack. De Wit conducted three-dimensional segmentation training up a 32x32x32 three-dimensional convolutional network. Our FCN implementation extracts two-dimensional axial slices to segment. It may have been beneficial to train three different networks, each one segmenting in different two-dimensional planes. In preprocessing, de Wit discriminated between positive candidates and negative candidates in the training datasets so as to use all positive samples available. Additionally, he implemented very heavy (i.e. lossless) augmentation. As noted previously, we trained our neural network on 206,650 candidates which was only a subset of the total number of candidates available in LUNA16. Of this smaller subset of candidates, we did not separate out positive samples from the negative ones nor did we augment the number of positive samples that did exist within the 206,650 candidates used. Both De Wit and Hammack made use of both LUNA16 and LIDC/NDSB's datasets. We spent a lot of time analyzing and making changes to the internal network architecture. However, upon additional research, the accuracy we were able to achieve would have increased more if we had spent time researching and accommodating for the shortcomings of the dataset in itself. With the timing constraints and initial lack of background knowledge, we hit a ceiling in regards to the scope of the project we were able to achieve. We were not able to experiment more with preprocessing tasks as we were predominantly occupied with changing parameters and methods used inside the U-Net architecture. A lot of our time was spent tackling with initial memory and dependency problems as well as a lack of knowledge about supervised learning later on.

8. ACKNOWLEDGEMENTS

This study was supported by Brandeis University's COSI 177A instructor, Antonella de Lillo, and teacher's assistant, Solomon Garber. Access to NVIDIA GTX-Titan GPUs was granted by Aaditya Prakash.

9. REFERENCES

[1] de Bel, T., van den Bogaard, C., Kotov, V., Scholten, L., Walasek, N.: LUNA16 Competition: False Positive Reduction (Project Report: Computer-Aided Diagnosis in Medical Imaging) (2016), Semantic Scholar

[2] Delougu, P., Cheran, S.C., De Mitri, I., Nunzio, G.D., Fantacci, M.E., Fauci, F., Gargano, G., Torres, E.L., Massafra, R., Oliva, P., Martinez, A.P., Raso, G., Retico, A., Stumbo, S., Tata, A.: Pre-processing methods for nodule detection in lung CT (2006), Internation Congress Series: 1281: 1099-1103

[3] Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation (2015), arXiv:1505.04597v1 [cs.CV]

[4] Setio, A.A.A., Ciompi, F., Litjens, G., Gerke, P., Jacobs, C., van Riel, S.J., Wille, M.M.W., Naqibullah, M., Sanchez, C.I., van Ginneken, B.: Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks (2016), IEEE: 35(5): 1160-1169

[5] Setio, A.A.A., Traverso, A., de Bel, T., Berens, M.S.N., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., van der Gugten, R., Heng, P.A., Jansen, B., de Kaste, M.M.J., Kotov, V., Lin, J.Y., Manders, J.T.M.C., Sonora-Mengana, A., Garcia-Naranjo, J.C., Papavasileiou, E., Prokop, M., Saletta, M., Schaefer-Prokop, C.M., Scholten, E.T., Scholten, L., Vandemeulebroucke, J., Walasek, N., Zuidhof, G.C.A., van Ginneken, B., Jacobs, C.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge (2017), arXiv:1612.08012v4 [cs.CV]

[6] Shelhamer, E., Long, J., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation (2016), arXiv: 1605:06211v1 [cs.CV]

[7] Zou, K.H., Warfield, S.K., Bharatha, A., Tempany, M.C., Kaus, M.R., Haker, S.J., Wells, W.M. III, Jolesz, F.A., Kikinis, R.: Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index (2004), Acad Radiol.: 11(2): 178-189