Semi-supervised multi-task learning for lung cancer diagnosis

Naji Khosravan and Ulas Bagci
Email: najikh@cs.ucf.edu, bagci@crcv.ucf.edu
Center for Resaerch in Computer Vision (CRCV), University of Central Florida (UCF), Orlando, FL.

***Abstract*— Early detection of lung nodules is of great importance in lung cancer screening. Existing research recognizes the critical role played by CAD systems in early detection and diagnosis of lung nodules. However, many CAD systems, which are used as cancer detection tools, produce a lot of false positives (FP) and require a further FP reduction step. Furthermore, guidelines for early diagnosis and treatment of lung cancer are consist of different shape and volume measurements of abnormalities. Segmentation is at the heart of our understanding of nodules morphology making it a major area of interest within the field of computer aided diagnosis systems. This study set out to test the hypothesis that joint learning of false positive (FP) nodule reduction and nodule segmentation can improve the computer aided diagnosis (CAD) systems' performance on both tasks. To support this hypothesis we propose a 3D deep multi-task CNN to tackle these two problems jointly. We tested our system on LUNA16 dataset and achieved an average dice similarity coefficient (DSC) of 91% as segmentation accuracy and a score of nearly 92% for FP reduction. As a proof of our hypothesis, we showed improvements of segmentation and FP reduction tasks over two baselines. Our results support that joint training of these two tasks through a multi-task learning approach improves system performance on both. We also showed that a semi-supervised approach can be used to overcome the limitation of lack of labeled data for the 3D segmentation task.**

I. Introduction

Lung cancer has the highest rate of mortality among the cancer related deaths [1]. Lung nodules are primary indicators of lung cancer and early diagnosis of them can increase survival rate considerably. Detecting these nodules along with different shape and size measurements enables radiologists to have an early diagnosis of their malignancy [2]. Toward this goal, many computer aided systems has been developed and shown to play a key role in early diagnosis of lung cancer [3], [4].

Existing automated lung nodule detection systems produce a lot of false positives (FP). Hence, there is an additional step needed to further reduce these FPs. This is a fundamental component of nearly all available CAD systems in the literature [3]. In the FP reduction step, candidates are being classified as *nodule* or *non-nodule* using discriminative features. Segmentation, on the other hand, is of interest as it is the first step toward quantification and different shape/size and volume measurements. In this study, we argue about the use of segmentation within the FP removal step. Since a good 3D segmentation of lung nodules leads to accurate volume/shape measurement analysis in cancer screening and treatment planning, it can be used as a discriminator information for FP identification. Although some studies used different nodule attributes in a multi-task manner with pretrained networks to do nodule characterization [5], till now, none of previous studies used segmentation within a FP reduction jointly.

This paper proposes a new methodology for addressing both *FP reduction* and *segmentation* problems, jointly. We propose a general model (Figure 1) that can perform both tasks with high accuracy through a multi-task learning (MTL) strategy. Our proposed model has a novel 3D deep encoder-decoder CNN architecture. We also exploit a semi supervised approach for training our model to avoid the need for large number of manual annotations for 3D segmentation masks. **Our contributions** can be specified as: **1)** This is the first study to propose joint segmentation and FP reduction of lung nodules through a fully 3D CNN, which is a critical step toward using CAD systems efficiently in clinical applications. **2)** Our work opens a door to possible improvements of CAD systems via a MTL approach. **3)** This work will generate fresh insight on how to tackle the problem of lack of available annotated medical image data through a semi-supervised learning method, which is more efficient if used along with MTL.

II. Method

The proposed 3D deep MTL algorithm is based on Convolutional Neural Network (CNN) and learns segmentation and FP reduction through some *shared and task specific layers*. The proposed architecture along with the training strategy is illustrated in Fig.1. In the rest of this section, we explain the proposed framework step by step.

A. Multi-Task Learning

MTL allows solving multiple learning tasks at the same time by optimizing multiple loss functions instead of one [6]. MTL can be beneficial in multiple senses: (1) *Generalization ability:* in MTL, a single model can be used to perform multiple tasks at the same time. Such as, in our case, it is desirable to have one general model, with the same accuracy if not better, instead of having multiple separate models. (2) *Highlighting underlying features:* depending on the selection of the tasks, features learned from one task can act discriminative for other tasks as well. These features might not always be easy to learn by a single task network due to their complexity or more discriminating effect of other features. However, learning multiple tasks jointly can strengthen the effect of these underlying features and boost the performance on one or all tasks. (3) *Dealing with lack of*

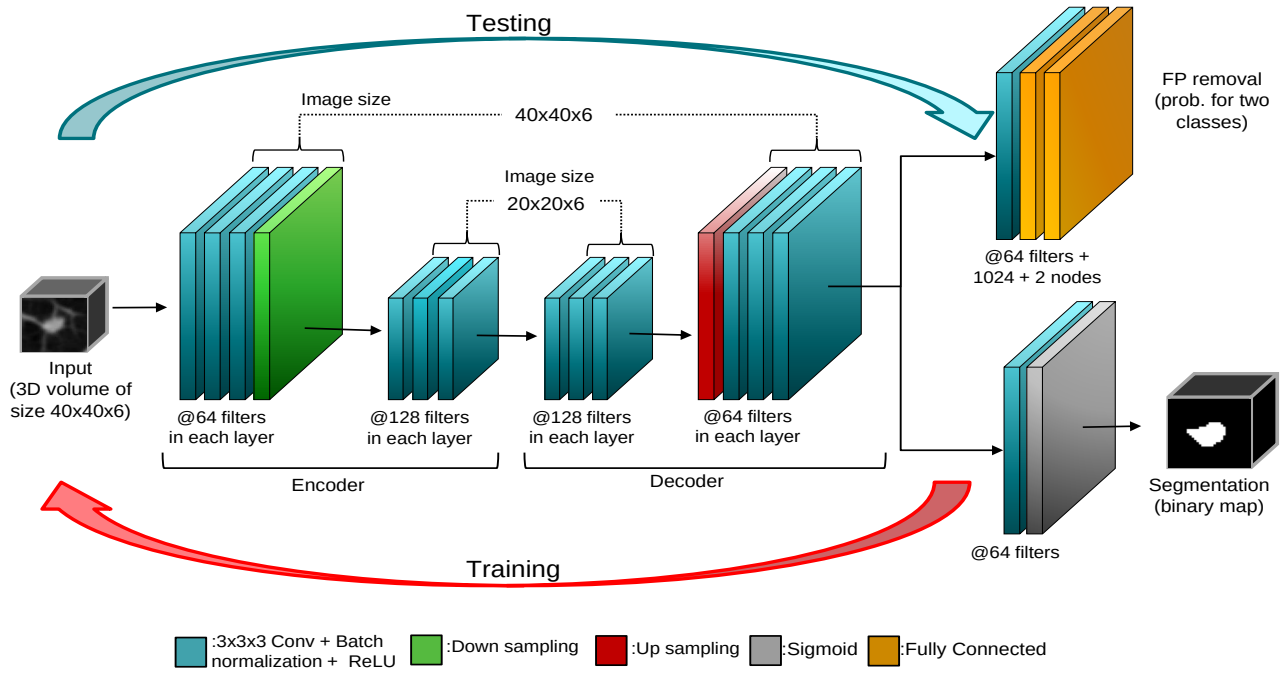

Fig. 1. The 3D deep multi-task CNN architecture. The size of all convolution kernels is set to $3 \times 3 \times 3$ with a stride of 1 in each dimension. The downsampling and upsampling operators are performed only in the xy plane. All convolution layers are $3D$. The network has 14 shared layers, 3 FP removal specific layers, and 2 segmentation specific layers. Red and blue arrows show the semi-supervised learning paradigm to train the proposed network.

data: in radiology field, it is not easy to gather large number of annotated data for training deep networks. An MTL model can benefit each task during training due to actively sharing features in relevant tasks.

The problem of jointly learning multiple tasks can be formulated as follows. Assume that we have N supervised tasks. The training set for each task can be considered as $D_n = (x_{in}, y_{in})$. In which $i = 1 : k_n$, where k_n is the number of training samples for the n_{th} task. With $x_{in} \in X^{(n)}$ and $y_{in} \in Y^{(n)}$ the problem of learning multiple tasks, jointly, can be narrowed down to the optimization problem of:

$$\min_{w} \sum_{n=1}^{N} L(Y^{(n)}, f(X^{(n)}) + \lambda \|f\|, \tag{1}$$

where $L : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^{+}$ is the loss function measuring the per-task prediction error, f is the multi-task model and w is the model's parameter set. In our study, we use cross entropy as loss function for both tasks. Cross entropy, also known as negative log likelihood, measures the similarity between two probability distributions and conventionally defined as:

$$L(Y^{(n)}, f(X^{(n)}) = \sum_{i=1}^{k_n} -y_i \log(f(x_i)), \tag{2}$$

where y_is are the true labels and $f(x_i)$s are the predictions for each task. To optimize equation 1, ADAM optimizer was used with an initially selected learning rate of 10^{-3}.

Since morphology (i.e., size, volume, and shape) information plays a key role in screening, diagnosis, and prognosis, we earlier postulate that this information can be effectively used for FP rejection, which is a significant challenge for most CADs. There is a strong need for reducing those findings (FPs) because it tremendously increases the workload of radiologists. We proved in the following that an MTL based CAD system can solve these two problems jointly: segmenting nodules while deciding whether they are FP or not. We believe that once the shape and appearance information can be highlighted in the shared layers of a network, other task specific layers can also learn if the nodule is a true nodule or not. In other words, features for classification and segmentation are combined through shared layers of the proposed network. To our best, this is the first study conducting this for both FP removal and segmentation.

B. Architecture

The inputs to our network are 3D volumes and the outputs are probabilities of each volume belonging to class of nodules or non-nodules. Our second output is a binary segmentation mask for those nodules. Our network has 19 layers: the first 14 layers are trained on both tasks, 5 task specific layers (2 for segmentation, 3 for classification) are trained only on one of the tasks. Each convolution layer in the architecture consists of a set of 3D convolution kernels (with size of $(3, 3, 3)$ and a stride of 1) following by a batch normalization (BN) and a rectified linear unit activation (ReLU). Number of kernels in each layer is depending on its location in the architecture. A max-pooling layer with the kernel size of $(2, 2)$ is used to perform down-sampling in the encoder. A bilinear interpolation is used for the up-sampling images in the decoder.

Our network *forks* after $14th$ layer into two branches (see Fig.1). Segmentation specific branch contains a convolution layer following by a sigmoid layer, which produces binary masks. FP reduction branch contains a convolution layer followed by two fully connected layer. The fully connected layers have 1024 and 2 nodes, respectively, and output the probability of each patch belonging to each one of classes (nodule vs. non-nodule).

C. Semi-supervised training

Due to the large number of parameters, deep CNNs need a large amount of annotated data to be trained efficiently. However, finding a large amount of such data is very challenging and expensive, specifically in the field of medical imaging. Semi-supervised learning methods are one way to address such issues. In semi-supervised methods, the model is initially trained on the part of data set which has labels. This model is then used to estimate labels for unlabeled data, which will be used to refine the model. The algorithm for semi-supervised learning strategy is illustrated in Algorithm 1. It can be argued that semi-supervised approach, if utilized naively, can lead to error propagation in the model and even cause worse performance. This problem, however, can be solved by iteratively performing prediction and training on small portions of unlabeled data and improving performance step by step. Constant improvements of results in our case supports that our algorithm perfectly handles error propagation and outperforms the baseline.

Algorithm 1: Semi-Supervised training algorithm

Input : labeled data: (X_l, Y_l), unlabeled data: X_u
Train model f on (X_l, Y_l)
for x *in* X_u **do**
  Predict on $x \in X_u$
  Add $(x, f(x))$ to labeled data
  Retrain model f
end
Return refined model f;

III. Results

Data: To evaluate our network we used Lung Nodule Analysis (LUNA16) Challenge dataset [7]. This dataset is gathered from the largest publically available LIDC-IDRI dataset. Scans with a slice thickness greater than 2.5 mm were excluded from the dataset leaving a total of 888 chest CT scans. The dataset contains the location of nodules accepted by at least 3 out of 4 radiologists leading to a total of 1186 nodule annotations. We performed our experiments on a total number of more than $500,000$ candidate locations provided by the dataset for the FP reduction task, which are a combination of outputs of candidate generation methods in the literature. This dataset is divided into 10 subsets by the provider. We performed 10-fold cross validation to evaluate our method. To handle the unbalance ratio between nodules and non-nodules we performed data augmentation on the nodules (shift in 6 directions). It should be mentioned that the number of segmentation masks available for this study was only **270** out of 1186 total nodule annotations and the masks for the rest (916 nodules) was created using the proposed semi-supervised strategy.

Segmentation: We used Dice Similarity Coefficient (DSC) as the metric to measure segmentation accuracy. To show the improvements, we compared the final model to 2 baselines of our model. Learning curves are plotted in Fig 2. In first baseline, we trained the model as a single task model using only the portion of annotated data which is available (depicted as single-manual ground truth (GT) in the plot-green). In second baseline, we trained the model jointly on both segmentation and FP reduction tasks as a MTL network with the same manual GT (depicted as joint-manual GT in the plot-pink). This multi-task model was used to generate annotations for the rest of the dataset. Next, we trained the model using the semi-supervised approach (depicted as joint-combined GT in the plot-blue). Note that we trained all models from scratch.

As shown, MTL based network outperforms single task based methods and semi-supervised approach improves results of MTL further. Our network reaches a DSC of **91%** compared to the baseline which does not go beyond **82%**.

FP reduction: To observe the effect of proposed semi-supervised MTL method on FP reduction performance, we compared the learning curves of three training strategies as follows. Single task network trained to only perform FP reduction (depicted as Single-green), Multi-Task using only manual GT available for the segmentation (depicted as Joint-Manual GT in the plot-pink) and Multi-Task using semi-supervised approach (depicted as Joint-Combined GT in the plot-blue). Figure 2 shows sensitivity through training epochs. As expected improvements are observed in the classification results (from **88%** to **98%**). Our results also show that, beside improving segmentation results, using a semi-supervised approach benefits FP reduction task as well (Joint-Combined GT in the plot-blue). This supports our rationale behind proposing a multi-task network strongly by showing that a better segmentation, which is highlighting shape and appearance information better, helps the other relevant task (FP reduction). Summary of the best performance on each task using different learning strategies is illustrated in Table I.

TABLE I
DICE SIMILARITY COEFFICIENT AND SENSITIVITY FOR THREE DIFFERENT LEARNING METHODS IS SHOWN.

Training strategy	DSC%	Sensitivity%
Single task	82%	88%
Multi task (manual GT)	86%	95%
Semi-Supervised multi task	**91%**	**98%**

Furthermore, to have a more accurate evaluation of our system, we used Free-Response Receiver Operating Characteristic (FROC) analysis [8]. Sensitivity at 7 FP/scan rates

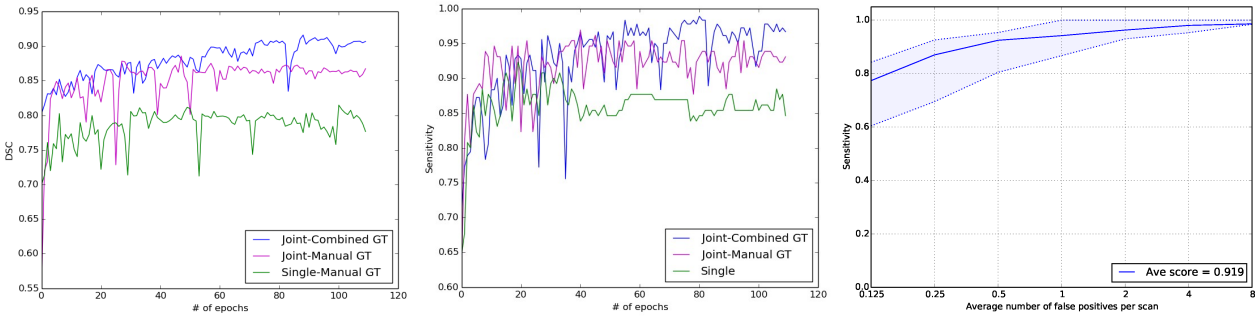

Fig. 2. Comparison of two baselines with proposed method. First baseline is single task network, second is semi-supervised MTL. *Left*: Dice similarity coefficient over first 100 learning epochs is shown. *middle*: Showing sensitivity for FP reduction task over the first 100 epochs. Improvement of segmentation through different training strategies are depicted. *Right*: is showing the FROC curve.

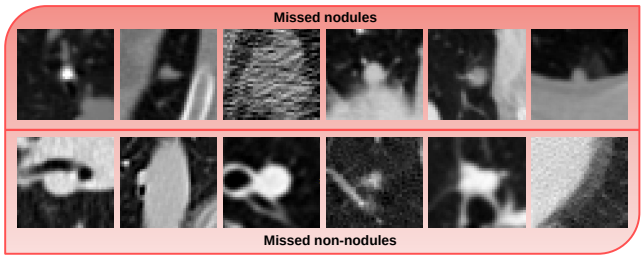

Fig. 3. Limitation of our system/failing cases: The first row shows 6 examples of missing nodules. The bottom row shows some examples of non-nodules which are mistakenly considered as nodules.

(i.e. $0.125, 0.25, 0.5, 1, 2, 4, 8$) is computed and the corresponding results are plotted in Fig. 2. The overall *score* of system is defined as the average sensitivity for these 7 FP/scan rates. Our network achieved an average score of $\sim$**92%** (see Table.II).

TABLE II

SYSTEM PERFORMANCE IN TERMS OF SENSITIVITY BASED ON NUMBER OF FPS/SCAN.

FPs/scan	0.125	0.25	0.5	1	2	4	8	Average
Sensitivity	0.773	0.870	0.924	0.941	0.962	0.980	0.986	**0.919**

IV. DISCUSSION AND CONCLUDING REMARKS

In this study, we proposed a 3D deep multi-task CNN for simultaneously performing segmentation and FP reduction. We showed that sharing some underlying features for these tasks and training a single model using shared features can improve the results for both tasks, which are critical for lung cancer screening. Furthermore, we showed that a semi-supervised approach can improve the results without the need for large number of labeled data in the training. It should be also note that there are some cases that our algorithm missed for FP reduction task. We illustrated some of those rarely seen examples of missing cases in Fig.3. One reason seems to be the small size of the missed nodule. Alternatively, very similar appearance of missing cases to other abnormalities and normal lung parenchyma. As an alternative direction to semi-supervised approach, one may use GAN to generate realistic data. One recent study created realistic nodules to support this idea [9].

REFERENCES

[1] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal, "Cancer statistics, 2017," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 1, pp. 7–30, 2017.

[2] Heber MacMahon, John HM Austin, Gordon Gamsu, Christian J Herold, James R Jett, David P Naidich, Edward F Patz Jr, and Stephen J Swensen, "Guidelines for management of small pulmonary nodules detected on ct scans: a statement from the fleischner society," *Radiology*, vol. 237, no. 2, pp. 395–400, 2005.

[3] Ingrid Sluimer, Arnold Schilham, Mathias Prokop, and Bram van Ginneken, "Computer analysis of computed tomography scans of the lung: a survey," *IEEE transactions on medical imaging*, vol. 25, no. 4, pp. 385–405, 2006.

[4] Carol E DeSantis, Chun Chieh Lin, Angela B Mariotto, Rebecca L Siegel, Kevin D Stein, Joan L Kramer, Rick Alteri, Anthony S Robbins, and Ahmedin Jemal, "Cancer treatment and survivorship statistics, 2014," *CA: a cancer journal for clinicians*, vol. 64, no. 4, pp. 252–271, 2014.

[5] Sarfaraz Hussein, Kunlin Cao, Qi Song, and Ulas Bagci, "Risk stratification of lung nodules using 3d cnn-based multi-task learning," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 249–260.

[6] Rich Caruana, "Multitask learning," in *Learning to learn*, pp. 95–133. Springer, 1998.

[7] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas de Bel, Moira S.N. Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, Robbert van der Gugten, Pheng Ann Heng, Bart Jansen, Michael M.J. de Kaste, Valentin Kotov, Jack Yu-Hung Lin, Jeroen T.M.C. Manders, Alexander Sóñora-Mengana, Juan Carlos García-Naranjo, Evgenia Papavasileiou, Mathias Prokop, Marco Saletta, Cornelia M Schaefer-Prokop, Ernst T. Scholten, Luuk Scholten, Miranda M. Snoeren, Ernesto Lopez Torres, Jef Vandemeulebroucke, Nicole Walasek, Guido C.A. Zuidhof, Bram van Ginneken, and Colin Jacobs, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The luna16 challenge," *Medical Image Analysis*, vol. 42, no. Supplement C, pp. 1 – 13, 2017.

[8] HL Kundel, KS Berbaum, DD Dorfman, D Gur, CE Metz, and RG Swensson, "Receiver operating characteristic analysis in medical imaging," *ICRU Report*, vol. 79, no. 8, pp. 1, 2008.

[9] Maria JM Chuquicusma, Sarfaraz Hussein, Jeremy Burt, and Ulas Bagci, "How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis," *arXiv preprint arXiv:1710.09762*, 2017.