Lung Cancer Detection from CT Scan Images Using Deep Convolutional Neural Networks

Yuqing Zhang[1], Rahul Bazaz[2], Howard Fan[1], Maulik Shah[3]

[1]Graduate Program in Bioinformatics [2]Undergraduate Program in Computer Science & Business Admin
[3]Graduate Program in Computer Science, Boston University, Boston MA 02215

Overview

➢Objective

Lung cancer strikes millions of people every year, and features a high mortality rate. Early detection of lung cancer is critical for patient recovery and survival. We participated in the Data Science Bowl 2017 (DSB-17) competition, which was held by the Kaggle community for developing diagnostic algorithms for lung cancer, using CT scan images.

➢Methods

We are given CT images from 1595 patients with high risk of lung cancer, including 1397 for training and 198 for evaluation. Each patient is associated with a 3D CT image and a cancer/normal binary label. In addition, we used public data from the LUNA-16 challenge, which contains CT images of 601 patients, and the locations of lung nodules in each patient.
We developed a predictive model based on deep convolutional neural networks (DCNNs). We first trained a U-Net segmentation model on the LUNA-16 data and predicted top nodule candidates in the DSB-17 dataset. Then, we built a DCNN model, which used the U-Net predictions together with the original CT images of the nodule candidates for the cancer/normal classification.

➢Results

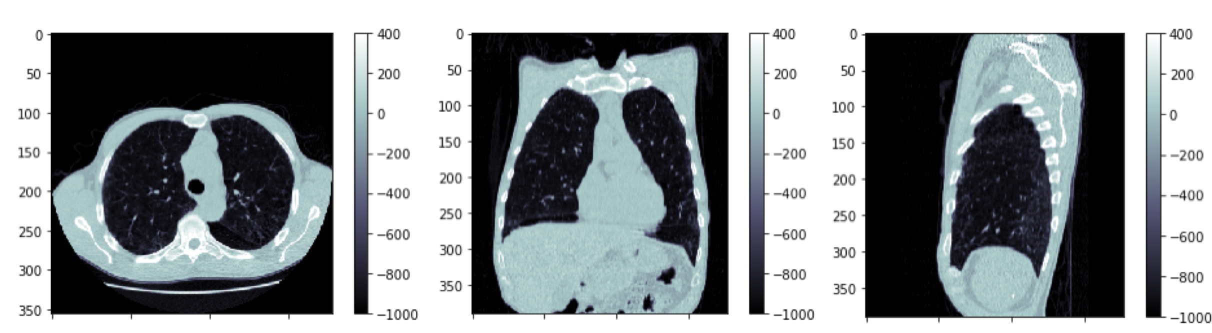
Our model achieved 0.55 loss and 76.7% accuracy on the DSB-17 evaluation set, which was comparable to the top-50 performances among all (around 2000) participating teams.

Lung Cancer Detection from CT Scan

➢Lung cancer detection task

Input: 3D CT-scan image of the patient's lung (e.g. right figure)

Output: binary classification of whether the patient has cancer



an example CT-scan image in DSB-17 dataset, along z, x, y axis

➢Datasets for lung cancer detection

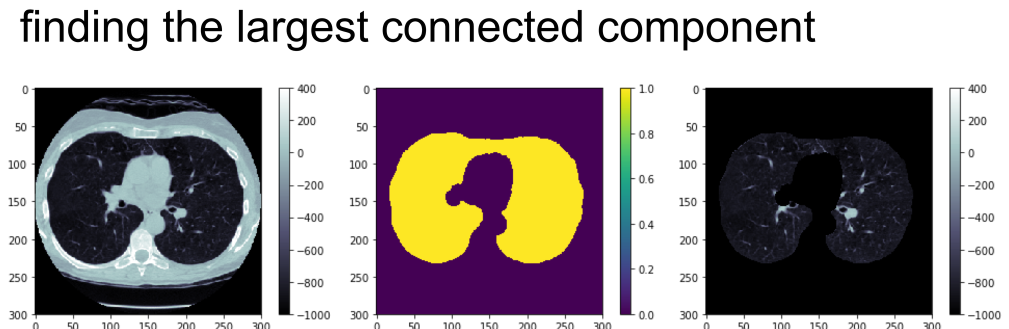
1. **Data Science Bowl 2017 (DSB-17) dataset** [1]:
   - contains 1595 CT scan images
   - 1397 for training (362 cancer and 1035 non-cancer)
   - 198 for test/evaluation (57 cancer and 141 non-cancer).
   - only a binary cancer/non-cancer label is associated with each CT scan image
2. **LUNA-16 dataset** [3]:
   - contains 601 CT scan image
   - the sizes, (X, Y, Z) coordinates, and diameters of *nodules* in each CT scan image

➢Summary of our method (based on deep neural networks)

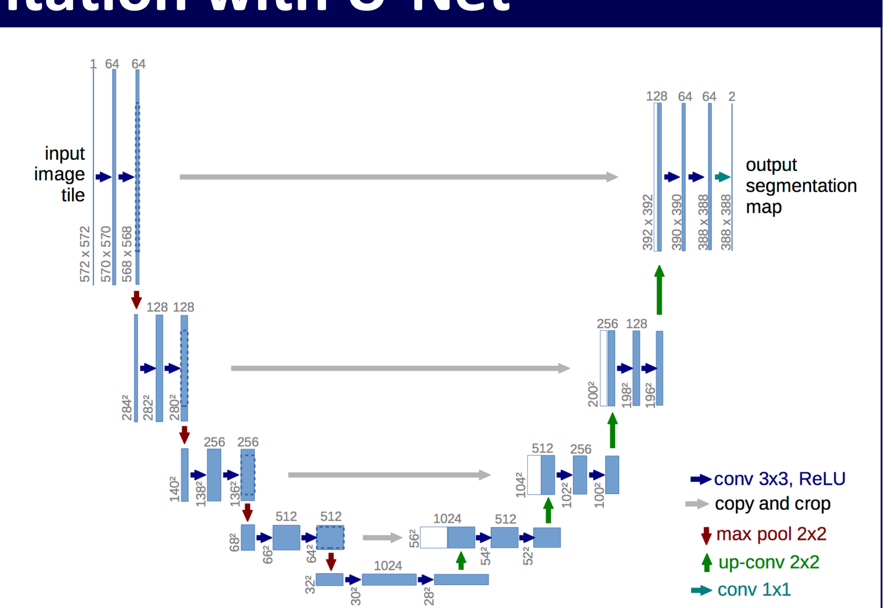
- **Nodule segmentation with U-Net (trained on LUNA-16)**
- **Identification of nodule candidates via clustering**
- **Cancer classification based on nodule candidates (trained on DSB-17)**

Methods - Nodule Segmentation with U-Net

Data preprocessing: segment the lung region by finding the largest connected component



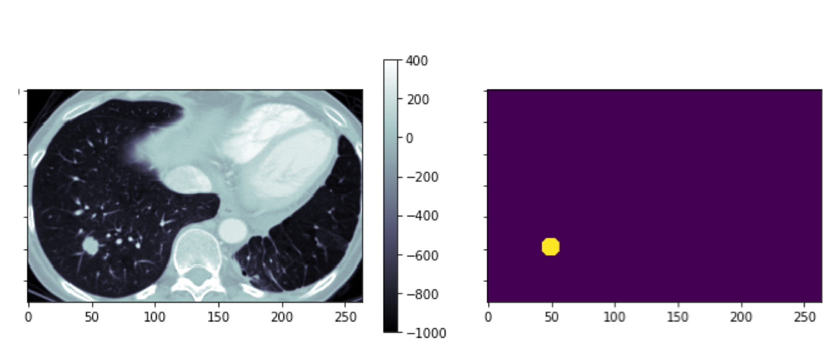
original CT-scan *lung mask* *segmented lung*



the U-Net [2] model for nodule segmentation

➢Segment nodules from CT-scan with *U-Net* [2]

- ***U-Net***: a fully convolutional neural network to predict pixel-wise probability from image
- ***Slice to probability***: get nodule probability in every location from the 2D CT-scan slices
  - Using the nodule mask in LUNA-16
- ***3D prediction***: 3 U-Nets trained on *2D slices* along x, y, z axis to predict nodule probabilities over the slices in the 3 axes.
  - average 3 axes for final nodule probability

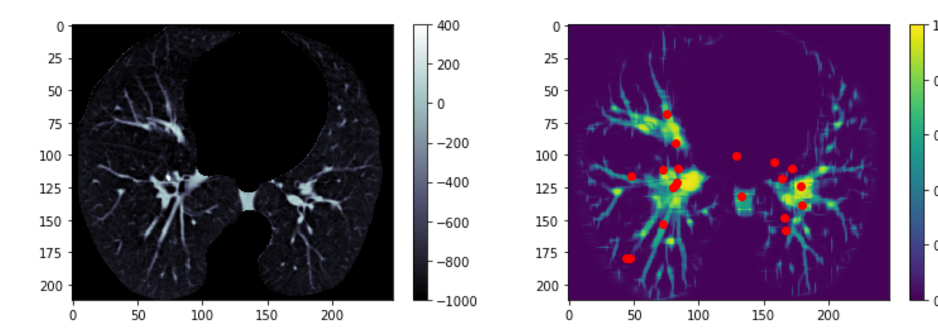


a slice of CT scan along z-axis in LUNA-16 *corresponding nodule mask (yellow circular region) for this slice.*

Methods - Cancer Classification over Nodule Clusters

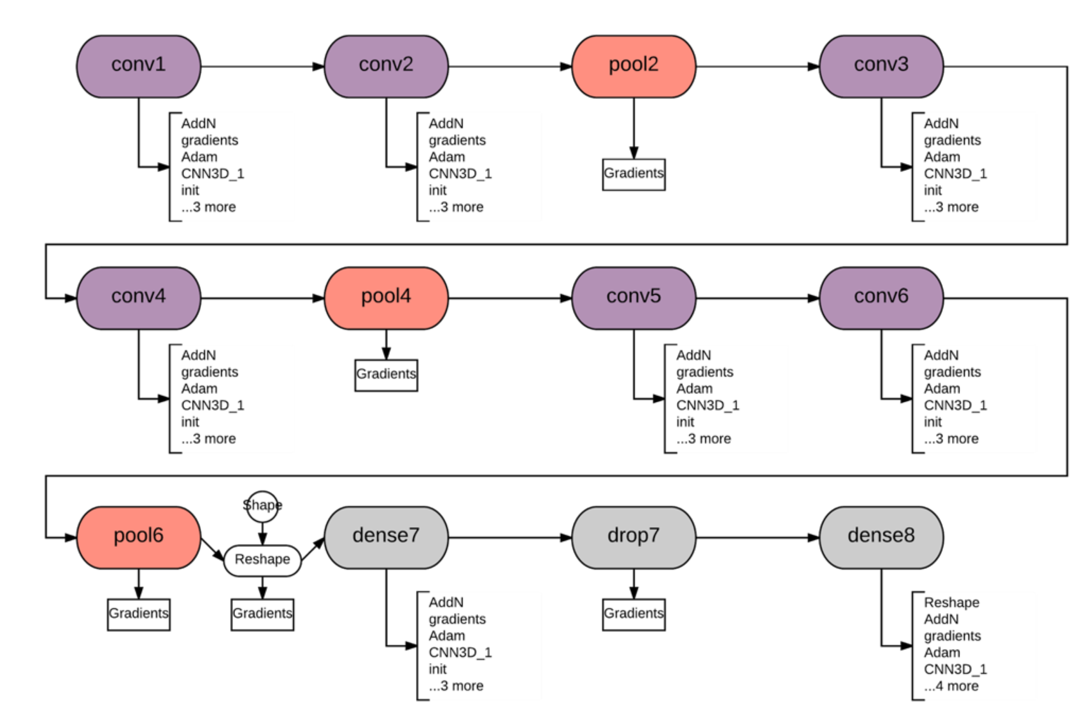
➢Identify top nodule candidates

- In the 3D probability map, include locations with probability of being a nodule larger than a threshold (0.8)
- Find connected components in the filtered locations
- Nodule candidates: top 20 components with maximum sum of probabilities
- Locate centroids in the nodule candidates
- Select a 32 x 32 x 32 cube centered at the centroids



input CT-scan image, showing a slice along the z-axis *predicted nodule probabilities with Identified top nodule candidates marked in red*

➢Classification over nodule candidates



structure of the classification 3D CNN

- Two channel input: the input size is (20, 32, 32, 32, **2**) containing both the original image and the nodule probability map from U-Net
- Apply 3D CNN architecture on each of them to get 20 scores
- Take the maximum of the 20 scores as each patient score
  - Rationale: the patient has cancer if one of the nodules is malignant, so take the max risk.
- Apply sigmoid cross-entropy loss on the maximum score to train the 3D neural network.

Model Performance

➢Training

- Log loss: 0.4780; Accuracy: 78.88%

➢Test

- Log loss: **0.5525**; Accuracy: **76.77%** (~14 epoch)

➢Comparable to **Top 50** (among ~1900 teams) in Data Science Bowl Leaderboard.

Conclusions and Future Work

➢Conclusions

- In this project, we addressed the challenging lung cancer detection problem with deep convolutional neural networks.
- We first trained U-Net models on LUNA-16 data for nodule segmentation and predicted top nodule candidates on the DSB-17 dataset, then built a 3D convolutional neural network and used the nodule probability together with the original CT images of the nodule candidates for cancer/non-cancer classification.
- Our final model achieves reasonable performance on the DSB-17 dataset.

➢Future Work

- Our model used all nodules during the candidate selection, regardless of whether the nodules are benign or malignant. This information is provided in the LUNA-16 dataset, but not used in our model. Building predictions with only malignant nodules may further improve the prediction accuracy.

References

[1] Booz Allen. Data Science Bowl 2017. https://www.kaggle.com/c/data-science-bowl-2017, 2017.
[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 234–241. Springer, 2015.
[3] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas de Bel, Moira SN Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. arXiv preprint arXiv:1612.08012, 2016.