Machine Learning Engineer Nanodegree

Capstone Proposal

Zhenxiong Yang
November 8, 2018

Proposal

Domain Background

Pneumonia is a big killer in Healthcare area for human-beings. Pneumonia ranks third in twenty leading principal hospital discharge groups in United States in 2015, only next to heart disease and chest pain[1]. There are about 544,000 visits in hospital taking up 4.4% distribution[1]. And it causes more than 50,000 deaths which nearly half of them are 85 years and over[2]. X-ray is the best way to diagnose pneumonia[3]. However, usually diagnosing pneumonia precisely needs experienced doctors checking carefully. So it is vital to confirm the pneumonia accurately and quickly through X-ray photographs. This can help doctors to make effective analysis and decisions, and also save patients' waiting time in hospital. This a Kaggle project.

Problem Statement

Pneumonia detection can be divided into two parts. The first is to classify whether it shows pneumonia or non-pneumonia. This is binary classification problem. The second is to localize which regions of pictures show pneumonia if it is pneumonia. Every region is presented in rectangle described by a left-top coordinate and width and height. There may be more than one region in a picture. This is the object detection problem.

Datasets and Inputs

The data is from Kaggle Data and consists of several files. The train images and test images files show patients' X-ray original photographs and also attach the personal information, like patientId, sex, age and so on.

The detailed_class_info.csv maps patientId to result class. There are three classes here, "No Lung Opacity/Not Normal", "Normal", "Lung Opacity". Both "Normal" and "No Lung Opacity/Not Normal" images are non-pneumonia targeted as 0, and "Lung Opacity" images are pneumonia targeted as 1.

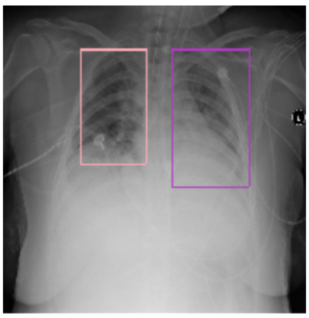

Figure 1: pneumonia

The train_labels.csv illustrate patients' results, including 'patientId', 'x-min', 'y-min', 'width', 'height', 'target'. The 'x-min' and 'y-min' is the coordinate of the region. 'width' and 'height' describe the region size. The 'target' is 1 if it is pneumonia and 0 if not. They are consistent with the classes in detailed_class_info.csv. It illustrates the pneumonia region clearly in Figure 1. I will use train pictures as X_data, and the train_labels as y_label according to corresponding 'patientId'.

Solution Statement

The sample_submission.csv file specifies the submission format. This is also the solution format. Every row consists of 'patientId', 'confidence' 'x-min' 'y-min' 'width' 'height'. The final output prediction is in this format. If a picture shows non-pneumonia, only 'patientId' is outputted. If an image has multiple regions showing pneumonia, all the related coordinates and width and weight are outputted. If detection object exists, the 'confidence' is IoU(defined in Evaluation Metrics part). Otherwise, 'confidence' is zero.

Benchmark Model

As a project from Kaggle, the top 10% in Private Leaderboard is considered as the benchmark model. The private leaderboard is calculated with approximately 99% of the test data.

Evaluation Metrics

The Evaluation Metrics is according to Kaggle Evaludation. The score is evaluated on the mean average precision at different intersection over union (IoU) thresholds. The IoU of a set of predicted bounding boxes and ground truth bounding boxes is calculated as:

$$IoU(A, B) = \frac{A \bigcap B}{A \bigcup B}$$

The metric sweeps over a range of IoU thresholds, at each point calculating an average precision value. The threshold values range from 0.4 to 0.75 with a step size of 0.05: (0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75). At each threshold value t, a precision value is calculated based on the number of true positives (TP), false negatives (FN), and false positives (FP) resulting from comparing the predicted object to all ground truth objects: $\frac{TP(t)}{TP(t)+FP(t)+FN(t)}$. The average precision of a single image is calculated as the mean of the above precision values at each IoU threshold. In nearly all cases confidence will have no impact on scoring.

Project Design

As is discussed above, this problem contains two sections, classification and detection. The loss function is mainly on this two parts. There are several excellent models to solve this kinds of problems, like Faster-RCNN[4] or Mask-RCNN[5], Yolo[6] and SSD[7]. Fast-RCNN processes accurately but slow(7 FPS with mAP 73.2%). Yolo is faster but less accurate(45 FPS with mAP 63.4%).

SSD has better accuracy and speed(59 FPS with mAP 74.3% on VOC2007 test). But SSD is not good at predicting small regions compared with Fast-RCNN. So I would like to try SSD first to see the result. If other one is better than SSD I may adjust this model.

SSD predicts both confidence and the center point, width, height of a region. The loss function is roughly like this: $L = \frac{1}{N}(L_{conf} + \alpha L_{loc})$. It computes the confidence loss and location loss. So if we set α very high, it is close to the evaluation metric above that is only considering the location.

Data preprocessing

The coordinate format in a region is 'x-min', 'y-min'. But SSD requires center coordinate format. So I need to preprocess the input data format to 'x-center', 'y-center' and postprocess the predict data back to 'x-min', 'y-min'. Maybe it is better to take the background out and only leave the pneumonia regions for a pneumonia image. I could try this way.

On the other side, for the train_labels data, if an image has two pneumonia regions, it has two pieces of data with the same patientId. I need to merge these data together.

Library

Python 2, numpy, pandas, pydicom, cv2, tqdm, glob, scikit-learn, Keras, TensorFlow, Caffe. Other libraries will be added if necessary.

References

[1] Rui P, Kang K, *National Ambulatory Medical Care Survey: 2015 Emergency Department Summary Tables.* Table 27. www.cdc.gov/nchs/data/nhamcs/web_tables/2015_ed_web_tables.pdf

[2] *Deaths: Final Data for 2015. Supplemental Tables.* Tables I-21, I-22. www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_06_tables.pdf

[3] *How Is Pneumonia Diagnosed?.* NHLBI. 1 March 2011. https://www.nhlbi.nih.gov/health-topics/pneumonia

[4] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.* https://arXiv:1506.01497v3

[5] Kaiming He Georgia Gkioxari Piotr Dollar Ross Girshick, *Mask R-CNN.* https://1703.06870v3

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, *You Only Look Once: Unified, Real-Time Object Detection.* https://arXiv:1506.02640v5

[7] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg, *SSD: Single Shot MultiBox Detector.* https://arXiv:1512.02325v5