

Rapid tumor detection by YOLO

Po-Nan Li

liponan@stanford.edu

Introduction

In the first lecture of BIOMEDIN260, Dr. Rubin presented several radiological images to students and very few of which could point out the location of an abnormality. Indeed, a tumor or an abnormality that is obvious to a trained doctor or radiologist might be obscure to other people, such as a computer scientist. Such a gap between the two communities might explain why many cancer imaging routines have not benefited from recent advancements in machine learning and deep learning. In other words, the cost and the inaccessibility to annotated data might prevent the deep learning-enabled progress of cancer imaging that could have made.

To encourage the creation of publicly accessible datasets, The Cancer Imaging Archive (TCIA) [1] invited the attendees at the Radiological Society of North America (RSNA) 2017 annual meeting to help annotate some of the CT scans from 352 subjects. For a CT scan shown on a web-based interface, a participant would be prompted to simply draw a line that best describes the longest diameter of a tumor he/she saw. Such simplism meant to encourage participants to label as many images as possible and ended up generating annotations for 2,345 CT images [2].

Although the bounding boxes resulted from the labels by radiological professionals are not as informative as polygon labels for segmentation tasks, they are useful for training an object detector, such as the YOLO (you only look once) models [3-6], which are considered the best object detectors in the one-stage category [6-8]. In this project, I will combine these two components, which are less studied in the context of cancer imaging, crowd-sourced simple labels and the YOLO detector, exploring the potential from this intriguing combination.

Methods

Data

The dataset used in this project include CT scans from TCGA-LUAD, TCGA-KIRC, TCGA-LIHC and TCGA-OV collections from the TCIA website [1]. Annotations are crowd-sourced by the volunteering professionals attending the RSNA 2017 annual meeting [2], where 2,345 samples were collected, 2,302 of which were successfully matched with a scan in the image collections and those that could not be matched were discarded. Out of the 2,302 image-annotation pairs, 2,102 are used for training; 100 (4.3%) are used for online validation, parameter tuning and model selection, and the other 100 (4.3%) were kept for final testing, which I had not conduct until the stage

when I prepared for the final write-up. Table 1 displays the statistics of the subsets. The annotation dataset reveals that roughly 60% of the records were labeled by trained radiologists and others by other professionals. Additionally, a majority of labeled images are anatomically renal.

Table 1. Statistics of data subsets, categorized by labeler status or anatomy.

Subset	n	Radiologist?		Anatomy			
		Yes	No	Renal	Ovarian	Lung	Liver
Train	2,102	1,262	840	1,156	397	315	234
Val	100	63	37	57	20	18	5
Test	100	62	38	51	19	18	12

I use the original image resolution, 512x512 for training and inference. For training stability, values in each image are offset and quantized to $[0, 255]$. To take advantage of the pre-trained model on ImageNet, all images are converted to 3-channel “color” images.

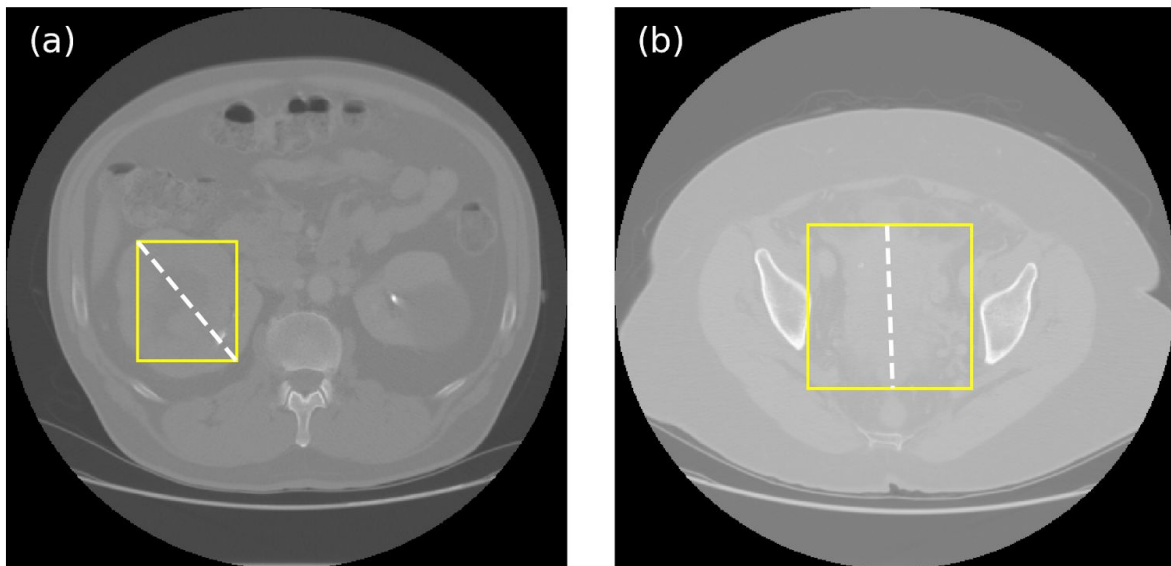


Figure 1. Representative (a) renal and (b) ovarian images and annotations from the validation set. (a) shows an example where the crowd-source volunteer drew the label in a diagonal fashion; (b) is an example that the labeller drew a nearly vertical line to label the tumor, in which case a square bounding box is automatically used.

Each image has exactly one label, which consists of a pair of coordinates $(x_{\text{start}}, y_{\text{start}})$ and $(x_{\text{end}}, y_{\text{end}})$. The pair is converted to a tuple of four parameters describing a bounding box: (x, y, w, h) , where x and y are the coordinates of the box centroid and w and h are the width and height, respectively. Figure 1 shows two examples of image and annotation pairs from the validation set. Note that the conversion from raw drawing to bounding boxes is not always straight-forward. The drawing shown in Figure 1(a), for example, is preferred as the two vertices can be easily converted to a rectangle. The drawing in Figure 1(b), on the other hand, depicts the height of the tumor but carries no information about the width. As a result, the following is in place to address this: if the aspect ratio between $|x_{\text{end}} - x_{\text{start}}|$ and $|y_{\text{end}} - y_{\text{start}}|$ is larger than 3, a square bounding box with side $h = w = \max(|x_{\text{end}} - x_{\text{start}}|, |y_{\text{end}} - y_{\text{start}}|)$ will be used. Here I only consider a single class: tumor. Benignity is not considered as it is not available from the annotation set.

Models

In this project I consider three generations of the YOLO architecture, YOLOv2, YOLOv3 and YOLOv4. YOLOv1 is not used as its detection mechanism is largely different from its successors and it is less used by the community nowadays. I use officially recommended hyper-parameters and default architecture for each of the models considered, except that the output layers are modified for single-class detection in lieu of multi-class. Table 2 compares YOLO's 4 generations.

For each of the YOLO models, I consider three different training schemes: use pre-trained weights and fine-tune all weights, use pre-trained weights but only update output layers (sometimes known as transfer learning), or train all weights from scratch. Note that the terms fine tuning and transfer learning are sometimes used interchangeably or in an opposite way, causing confusions. In this report, fine tuning is defined as the whole set of pretrained weights are updated with new dataset; transfer learning is defined as only the output layers are graded, back-propagated and updated while other upstream weights remain untouched.

Table 2. YOLO architectures and training schemes considered in this project. BFLOPs: Billion floating point operations. Note: YOLOv1 is listed here for comparison but is not used in this project.

Architecture	YOLOv1	YOLOv2	YOLOv3	YOLOv4
Year	2015	2016	2018	2020
# Conv. layers	24	23	75	94
BFLOPs	52.448	44.441	98.923	90.226
Training scheme	N/A	Pretrain (P), transfer learning (T), from scratch (S)		

Experiments

The workflow is carried out as follows: the validation dataset is used to choose the “best model,” and then the best model is run to infer the test dataset. Each of the 3 YOLO models, YOLOv2, v3 and v4, is independently trained with 3 different training schemes. Altogether, 9 different models are considered as candidates. Each model is trained for 10,000 iterations with batch size of 64. To prevent overfitting, each model is saved and validated every 1,000 iterations and the snapshot that has the highest metric is selected, by the process detailed below.

Model selection

Figure 2 shows the validation metrics of 3 different pretrained (P) models during training (fine tuning the pretrained model). For other two training schemes, i.e. transfer learning (T) and training from scratch (S), see Figures S1 and S2 in the Appendix. Four metrics are used to evaluate the model performance: precision, defined as number of true positive over number of predicted positive with confidence score > 0.25 ; TP rate, known as recall, defined as number of true positive (score > 0.25) over number of true positive; IOU, or intersection over union, is the average IOU of true positive with score > 0.25 ; mAP or mean average precision, is defined as the area under precision-recall curve (AUC) with IOU threshold 0.50.

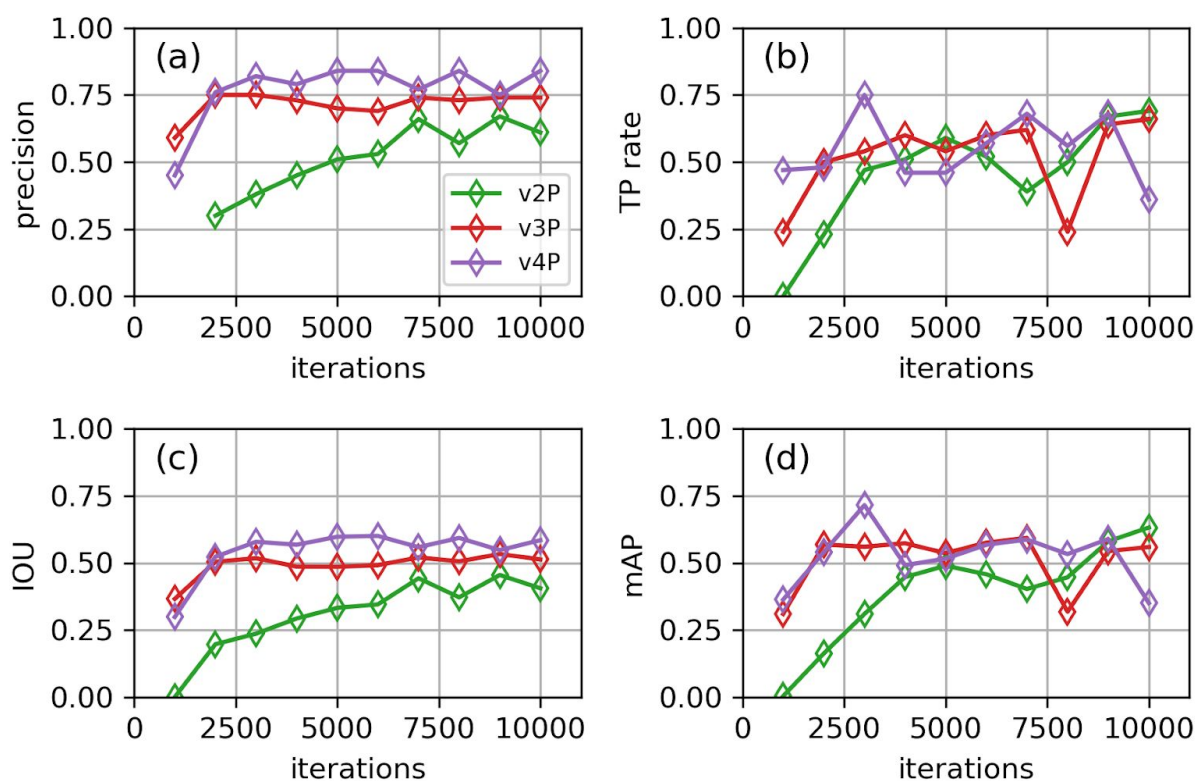


Figure 2. Validation metrics vs. training iterations with various YOLO models pretrained on the ImageNet dataset. P suffix indicates models are pretrained.

Although it is usually practical to use IOU to evaluate the performance of an object detection system as it emphasizes the geometrical correctness of the predicted bounding boxes, it should be recognized that the labels from the RSNA 2017 dataset are not perfect nor should be considered as oracles. On the other hand, it is widely known that only looking at precision or recall (TP rate) would lose a greater picture as there is trade-off between them. I therefore use the widely accepted metric, mAP, as the criterion to pick the best model. As Table 3 can show, the pretrained YOLOv4 model (v4P) outperforms v2P and v3P models by all criteria. To choose the best snapshots of v4P, I follow the discussion above and pick the one with the highest mAP.

Table 3. Best 3 models, in terms of 4 different criteria. IOU: intersection over union. TP: true positive. mAP: mean average precision. P: pretrained model; T: transfer learning model; S: trained from scratch model.

Criterion	Precision		TP rate		Average IOU		mAP	
No. 1	v4P	0.84	v4P	0.75	v4P	0.601	v4P	0.715
No. 2	v3S	0.77	v2P	0.69	v3P	0.533	v2P	0.632
No. 3	v4T	0.76	v3P	0.66	v3T	0.524	v3T	0.623

Figure 3 shows four representative images ((a)-(d) are renal, ovarian, lung, liver, respectively) from the validation dataset and the corresponding detection results from 3 YOLO pretrained models after fine tuning. The v2P model (green boxes) successfully detected the tumors in all 4 example images, although there is a redundant box in Figure 3(c). The v3P model (red boxes) fails to recognize the tumors in Figure 3(b) and (d). The v4P model (purple boxes) performs similarly as v2P and has a false positive box in Figure 3(d). Bounding boxes by all models are apparently much larger than the labels, despite their centroids being typically very close to the labels’.

Best model on the test set

Once the best model is chosen, I test it on the test dataset, which was kept untouched until now. Table 4 shows the best model’s performance on the validation and test sets. Surprisingly, although both the validation and test sets were unseen by the model and have the same distribution, because the best model is picked based on the validation result, there is a non negligible gap between the validation and test result (also known as variance), suggesting that the model selection process might have over-fit to the validation data.

Figure 4 shows four representative images ((a)-(d) are renal, ovarian, lung, liver, respectively) from the test dataset and the corresponding detection results from the best model. Again, the predicted bounding boxes are generally much larger than the label

boxes and the centroids are slightly off but relatively acceptable. See Table S1 for metrics categorized by labeler status or anatomy.

Detection time

All experiments were performed on a NVIDIA Tesla T4 GPU. The average detection time for a single 512x512 image with the “best model” is 22 ms, or 45.5 frames per second.

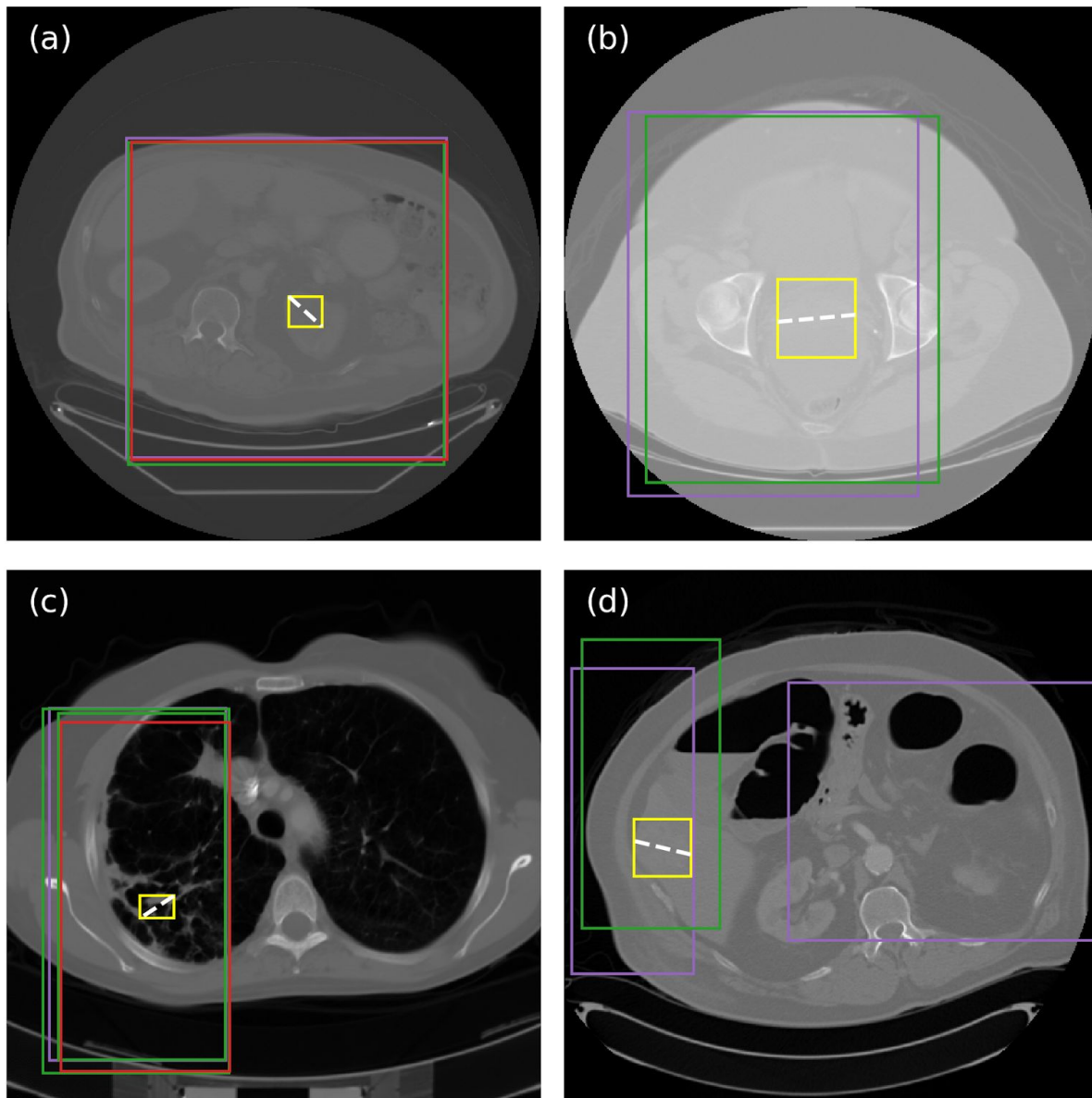


Figure 3. Representative (a) renal, (b) ovarian, (c) lung and (d) liver images from the validation dataset and predicted bound boxes from the 3 pretrained models after fine tuning: v2P (green), v3P (red) and v4P (purple). Dashed white lines and yellow boxes are the raw drawing by the RSNA 2017 participants and the converted label boxes, respectively.

Table 4. Performance of the chosen “best model.”

Dataset	Precision	TP rate	Average IOU	mAP
Val	0.82	0.75	0.579	0.715
Test	0.73	0.69	0.514	0.613

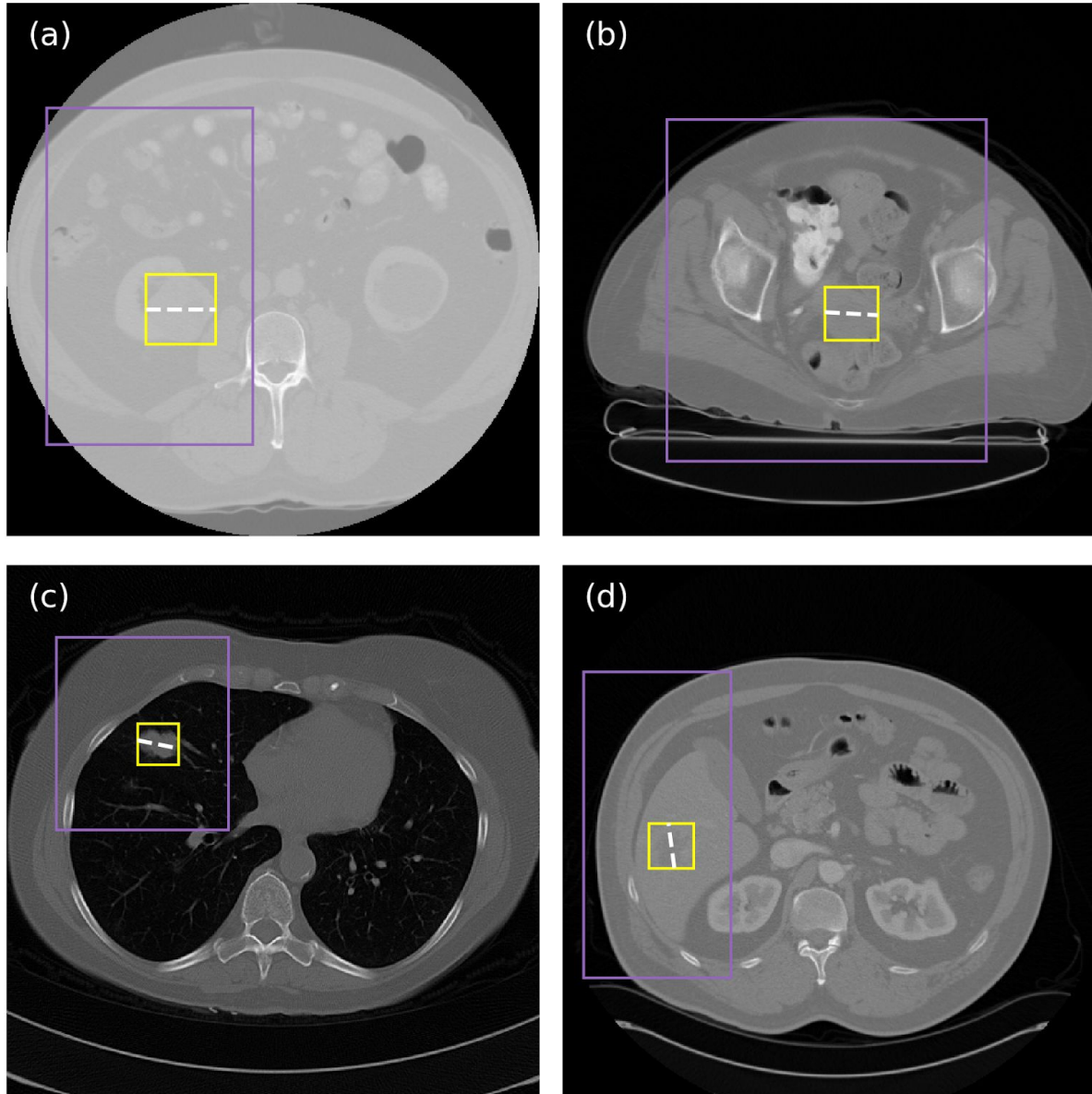


Figure 4. Representative (a) renal, (b) ovarian, (c) lung and (d) liver images from the test dataset and predicted bound boxes from the pretrained v4P model after fine tuning, namely the best model. Dashed white lines and yellow boxes are the raw drawing by the RSNA 2017 participants and the converted label boxes, respectively.

Discussion

This project aims to capitalize on the crowd-sourced annotations and understand whether these simple drawings can be utilized as labels for training a tumor detection system. In this study I have learned that, despite the good quality, these drawings do not faithfully represent the true dimension of tumors in a two-dimensional space. Even though in the data processing pipeline, a workaround is implemented to address drawings with extreme aspect ratios, not having a precise bounding box might ultimately prevent the model from properly learning and recognizing the boundary of a tumor, which can be crucial in medical imaging contexts.

It should be also pointed that despite that medical images have variance attributed from different instruments, different manufacturers, different institutions, as Dr. Rubin has discussed in his lectures, the task of tumor detection is not necessarily as hard as an image challenge, like the ImageNet contest, because the object sizes (i.e. tumor sizes) do not vary significantly over cases and images, and the pixel values are typically well calibrated. The use of a model with complex architecture, e.g. YOLOv4, might have been an overkill. It might have been a better strategy to employ a simple model and carefully tune for its hyper-parameter. Nonetheless, it should be emphasized that the rationale of considering several off-the-shelf models is to study how easy it is to repurpose an ImageNet-pretrained model for medical imaging applications and to understand the capacity of a generic model.

Conclusion

This project demonstrates a potential use of a CT dataset with crowd-sourced annotations, where tumors are crudely labeled with a straight line. Several YOLO models are trained to detect the tumor in the images and one of which is ultimately selected for the final testing. The main finding of this project is that all YOLO models considered (v2, v3, v4) are capable of detecting the tumor in a CT image but the bounding box sizes are generally overestimated, which can be attributable to the raw annotations and thus the imprecise labels.

Code and data availability

Python code and data are available at <https://github.com/leeneil/bmi260>.

Acknowledgment

P.-N.L. thanks Mars Huang for his time and efforts on helping set-up the Google Cloud virtual instance, which made this project possible.

References

- [1] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F, “The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository,” *Journal of Digital Imaging* **26**, 1045–1057 (2013).
- [2] Jayashree Kalpathy-Cramer, Andrew Beers, Artem Mamonov, Erik Ziegler, Rob Lewis, Andre Botelho Almeida, Gordon Harris, Steve Pieper, David Clunie, Ashish Sharma, Lawrence Tarbox, Jeff Tobler, Fred Prior, Adam Flanders, Jamie Dulkowski, Brenda Fevrier-Sullivan, Carl Jaffe, John Freymann, and Justin Kirby, “Crowds Cure Cancer: Data collected at the RSNA 2017 annual meeting,” The Cancer Imaging Archive. DOI: 10.7937/K9/TCIA.2018.OW73VLO2
- [3] Joseph Redmon and Santosh Divvala and Ross Girshick and Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” arXiv:1506.02640 (2015).
- [4] Joseph Redmon and Ali Farhadi, “YOLO9000: Better, Faster, Stronger,” arXiv:1612.08242 (2016).
- [5] Joseph Redmon and Ali Farhadi, “YOLOv3: An Incremental Improvement,” arXiv:1804.02767 (2018).
- [6] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection,” arXiv:2004.10934 (2020).
- [7] Zhong-Qiu Zhao, Peng Zheng, Shou-Tao Xu, and Xindong Wu, “Object detection with deep learning: a review,” *IEEE Transactions on Neural Networks and Learning Systems* **30**, 3212–3232 (2019).
- [8] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, Matti Pietikäinen, “Deep Learning for Generic Object Detection: A Survey,” *International Journal of Computer Vision* **128**, 261–318 (2020).

Appendix

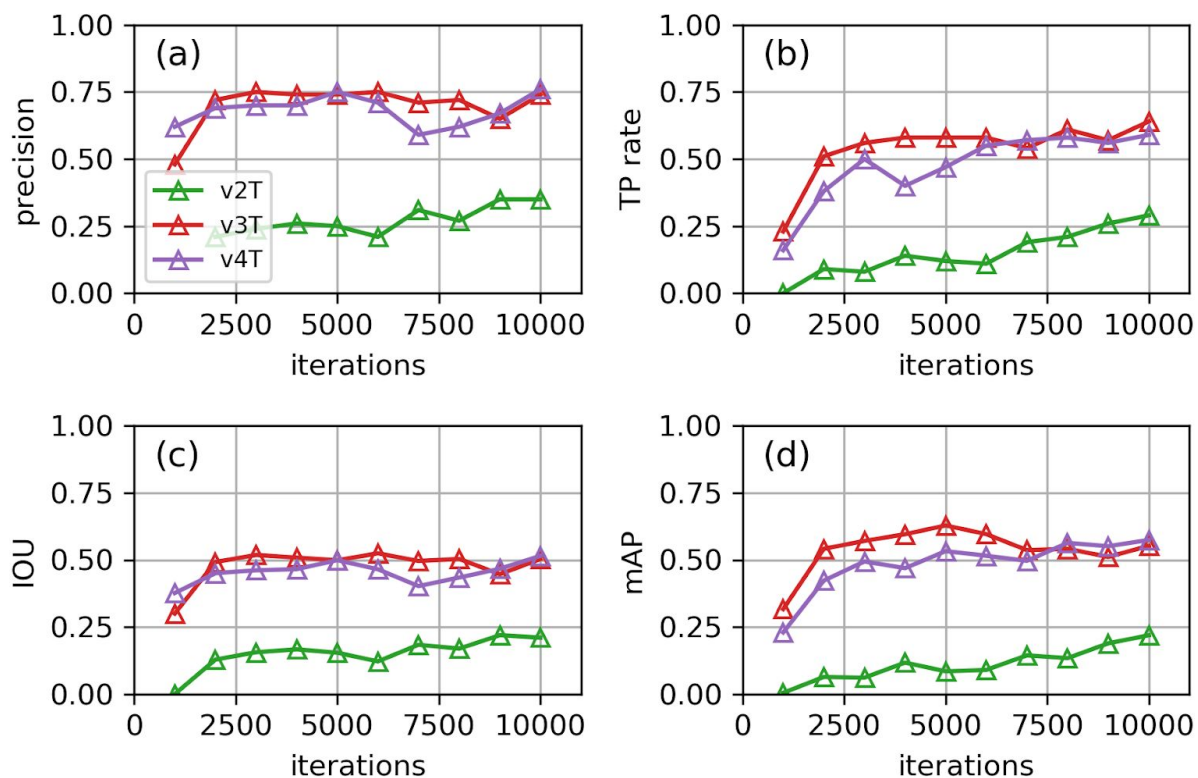


Figure S1. Validation metrics vs. training iterations when transfer learning is used, i.e. only the output layers are back-propagated and updated. T means transfer learning.

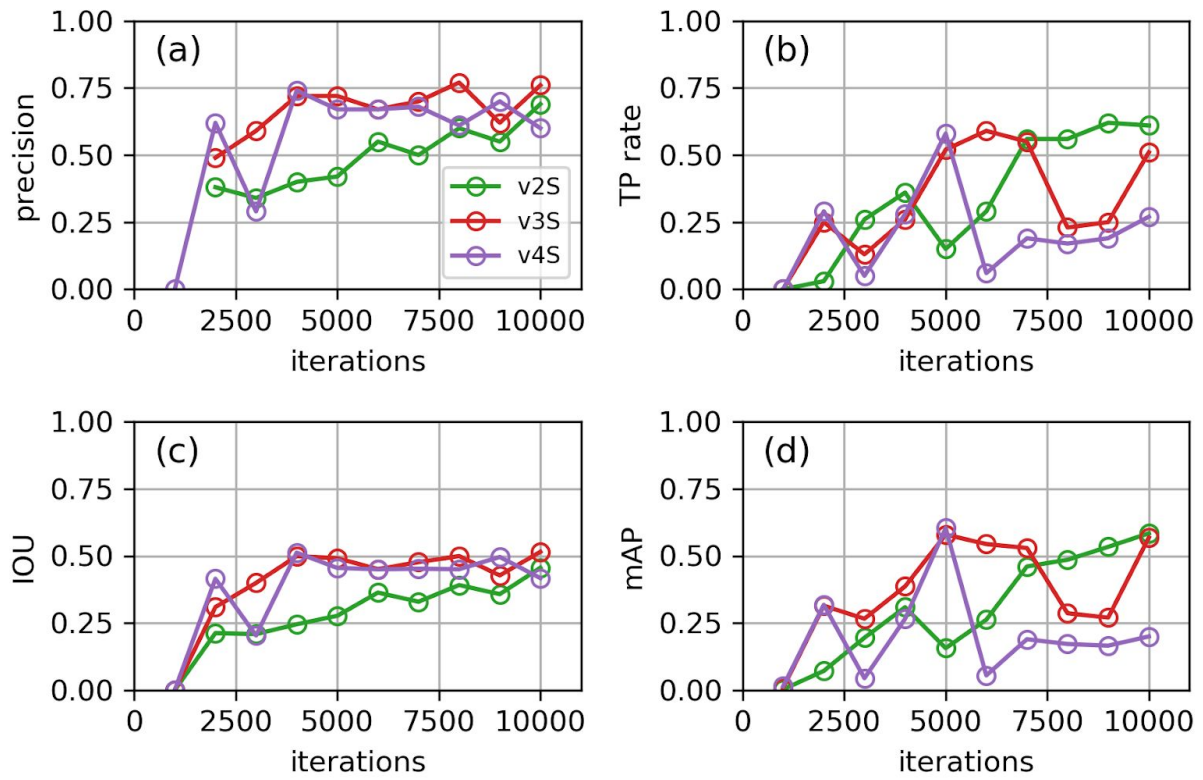


Figure S2. Validation metrics vs. training iterations when training from scratch. S indicates training from scratch.

Table S1. Performance of the chosen “best model” on categorized test data subsets.

Subset		<i>n</i>	Precision	TP rate	Average IOU	mAP
Radiologist?	Yes	62	0.68	0.66	0.473	0.547
	No	38	0.82	0.74	0.585	0.722
Anatomy	Renal	51	0.80	0.76	0.559	0.689
	Ovarian	19	0.61	0.58	0.440	0.486
	Lung	18	0.71	0.67	0.487	0.590
	Liver	12	0.70	0.58	0.468	0.447
Overall		100	0.73	0.69	0.514	0.613