CS230

Augmented Bladder Tumor Detection using Deep Learning

Mark Laurie, Colt McNealy, Victor Yin
Department of Computer Science
Stanford University
markl21@stanford.edu, vhsy@stanford.edu, cmcnealy@stanford.edu

Abstract

The recurrence rate for patients diagnosed with bladder cancer is 70%. That is why they are required to be constantly checked upon by medical personnel. One way of checking is White light cystoscopy (WLC), but it often misses 20% of lesions resulting in an unnecessary increase in patient mortality. Therefore, we propose two object detection frameworks for bladder tumor detection that may supersede the currently used YOLOv3 model for WLC: Faster-RCNN and EfficientDet. While unable to acheive a working Faster-RCNN model, we acheived superior recall with the EfficientDet architecture. Although precision was inferior, we note that we trained on a much smaller dataset and did not use transfer learning. We therefore predict that, with additional data and transfer learning, the EfficientDet architecture will demonstrate both improved precision and recall compared to YOLOv3, leading to fewer missed lesions and patient mortality.

1 Introduction

With approximately 80,470 diagnoses in 2019, bladder cancer is the sixth most common malignancy in the United States [1]. Patients diagnosed with bladder cancer experience recurrence rates nearing 70%, meaning that they require endoscopic surveillance from diagnosis to death; consequently, bladder cancer is the most expensive cancer to treat [2]. Currently, white-light cystoscopy (WLC) is the standard-of-care for bladder cancer diagnosis and surveillence; suspicious lesions observed via WLC are removed endoscopically via transurethral resection of bladder tumor (TURBT). Patients undergo repeat WLCs until either remission or death [3]. Unfortunately, 20% of lesions, particularly those categorized as either non-papillary or multi-focal, are missed by WLC, resulting in increased disease morbidity and mortality [4].

Recently, neural networks have proved to be useful tools for medical image analysis, specifically convolutional neural networks (CNNs). So far, the Liao and Xing labs have developed CystoNet, which has demonstrated initial promise in the real-time detection of suspicious papillary lesions during WLC and TURBT, using just a small subset of our video database [5]. The development of a real-time (RT), deep-learning-based paradigm for image-based bladder cancer diagnosis, localization, and surveillance during WLC may reduce the probability that suspicious lesions are missed, which may significantly improve patient outcomes.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

2.1 Previous Implementation: YOLO

The Liao and Xing labs used the YOLO object detection framework as CystoNet's first backbone architecture [5]. Their architecture consists of 5 convolutional blocks each separated by four max-pooling layers followed by two fully-connected layers. Their original test set consisted of 7542 frames of 44 lesions originating from WLCs and TURBTs from 2016-2019. On this test set, the per-frame and per-tumor (correct flagging of histologically confirmed bladder cancer in at least one frame) sensitivities were 90.9% and 95.5% in the tumor cohort, respectively, while the per-frame specificity was 98.6% in the normal frame cohort [5].

Using a different test set, one from which our test set originates, per-frame recall dropped to 58.1% while per-frame specificity was 87.4% (see contributions). Per-frame precision was 98.8% while per-frame negative predictive value was 10.6%. Its F1 score was 0.731. All metrics did not have a minimum required IoU threshold. We hypothesize that the distribution of the original test set does not match that of this new test set, which is more representative of RT WLC and TURBT images. Therefore, we will compare the results of our models with these metrics.

2.2 Object Detection

Our problem is complicated due to the monochromatic appearence of bladder endothelia and the subtle differences between normal and non-papillary tumor endothelia. We first looked at Sinon Nrea's approach which used five convolutional layers with relu activation layers, batch normalization, maxpooling layer and dropout layers [6]. Our baseline model was based off this architecture and had an F1 score of 0.715 for lesion identification.

We then planned to implement an object detection framework that was efficient enough to detect tumors in real time.We found one paper by Mingxing Tan that introduced us to the EfficientDet model architecture [7]. This model built upon previous scaling neural networks such as EfficientNet and incorporated a novel Bi-directional feature network along with other scaling rules. The new bi-directional feature network is used to enable information to flow in top-down and bottom-up directions. More importantly, it does this using regular and efficient connections, which distinguishes it from the NAS-FPN architecture that also achieves the same effect but is not generalizable. For our purposes, we hypothesize that the the Bi-directional feature network approach will perform better. See below for our initial results.

In addition to the EfficientDet architecture, we experimented with the recent Faster-RCNN architecture. The Faster-RCNN architecture is the latest of the RCNN family of object detection models. RCNN combines the region proposal network (RPN) and the convolutional neural network (CNN) of the previous Fast-RCNN model into one network, thus improving accuracy and reducing training time. While we did not achieve success with the Faster-RCNN architecture, we still believe it has promise as a method for bladder tumor detection.

3 The Task: Real-time Bladder Tumor Detection

Given an image originating from the cystoscope positioned inside the bladder, our goal is to not only classify identified tumors with respect to type but also flag the region of suspicion with a bounding box. Even though CystoNet's YOLO-based architecture's F1 score is respectable, it exibibits difficulty in detecting non-papillary lesions. While CystoNet exceeds expectations regarding papillary lesion detection, this task is already straightforward as lesions with this morphology are difficult to miss. Detecting non-papillary lesions and *carcinoma in situ* (CIS), however, is more difficult, as 20% are missed during WLC [8]. CystoNet's utility, therein, lies in its ability to flag non-papillary lesions out during WLC and then allow for the urologist decide whether to proceed with transurethral resection of bladder tumor (TURBT) if the flagged region appears suspicious. Therefore, our primary goal is to improve the per-frame sensitivity for non-papillary lesions during WLC.

A

#Cases	Training	Test	Total
Normal	2	2	4
Papillary	9	4	13
Non-papillary	4	3	7
Total	15	9	24

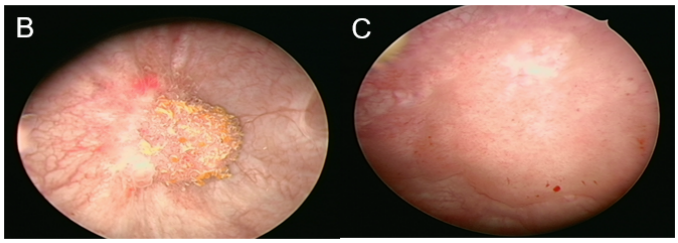

Figure 1: Our cystoscopic dataset. **A:** Training and test set splits for CystoNet by case. **B:** Image of papillary lesion used for model training. **C:** Image of non-papillary lesion used for model training.

4 Dataset and Preprocessing

With institutional review board approval, cystoscopy videos were collected from consenting patients between 2016 and 2020, with one video corresponding to one patient. For our efficientDet implementation, we significantly increased the size of our dataset relative to that used for the baseline and revised our dataset split. We trained the efficientDet architecture using 15 cases and evaluated with 9 cases. In the training set, 2 videos captured unconfirmed normal bladder endothelia, 9 vidoes captured pathologically confirmed papillary carcinoma, and 4 videos captured 1 or more cases of non-papillary carcinoma. In the test set, 2 videos captured unconfirmed normal bladder endothelia; 4 videos captured pathologically confirmed papillary carcinoma, and 3 videos captured pathologically confirmed cases of non-papillary carcinoma. Our model was trained using approximately 29,000 frames from the training set and approximately 15,000 frames from the test set. We decided to forgo an additional validation set due to our dataset's small number of cases and note that, although the number of total frames is sufficient, many of these frames are nearly identical as the cystoscope can be stationary for several seconds at a time.

Our dataset was annotated in adjunct with urologists and clinical coordinators of the Liao Lab. To ensure that annotations are as close to ground truth as possible, We sent sample annotated frames to urologists for review and revision before annotating the entirety of each case. We used Computer Vision Annotation Tool (CVAT) to annotate all cases [10]. Significant pre-processing was required to parse, convert, and combine individual datasets.

4.1 Screen Tone Removal

Before feeding the images into our model, we first performed screentone removal (STR) on the raw videos. Qualitatively, STR highlights the edges and texture of an image, which facilitates the detection of tumors by our model. The STR algorithm uses a Laplacian Gaussian filter to detect and preserve edges and lines, and a screentone mask. The two masks are pixel-wise OR'ed together and the resulting mask is finally applied to the image.

5 Methods

We tried two different model architectures: EfficientDet and MaskRCNN. Furthermore, we compared the performance of both models with and without screentone removal as a pre-processing step.

5.1 EfficientDet

Efficientdet was introduced at CVPR in 2020 as a new family of scalable and efficient object detectors. Overall, EfficientDet is smaller and uses much less computation compared to other state of the art detectors. The key difference is that Efficientdet does not use a top-down feature pyramid network that previous detectors use. Compared to other models that offer an additional bottom-up flow (such as PANet), Efficientdet does so at the cost of less computation as well. This is done with a bi-directional feature network.

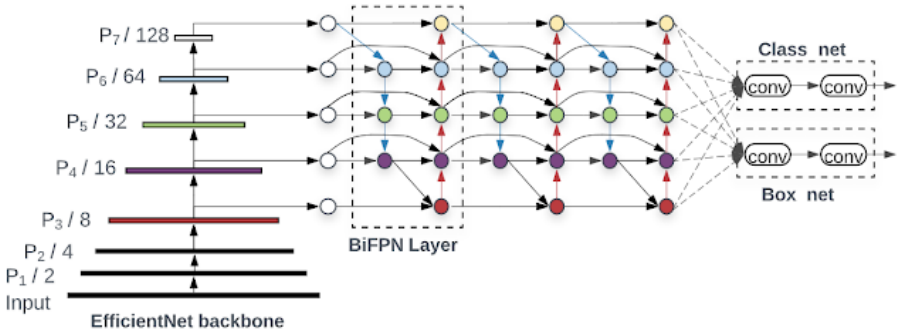

Figure 2: EfficientDet architecture. Notice the bi-directional feature network.

The bi-directional feature network takes the EfficientNet backbone network and repeatedly applies bidirectional feature fusion. To further increase efficiency, the paper also suggested a new fast normalized fusion technique. At the same time, another suggestion was to make sure to replace regular convolutions with less expensive depth wise separable convolusions. This was what we followed. EfficientDet also works well on the COCO dataset, exceeding prior state-of-the-art models using 4x less parameters and 9.4 less computation. For our project we built upon prebuilt architectures that guided us on figuring out the connections in the bi-directional feature network. This was really important for implementing Efficientdet and our forward propagation step. Our code was also built on top of the backbone net the EfficientNet model which we accessed by loading a pretrained model. Finally, we included a classifier and regressor.

Our model was trained using a batch size of 4, early stopping patience of 2, and a learning rate of $1 \cdot 10^{-4}$. We used a batch size of 4 because CUDA failed to allocate memory for batch sizes of greater than 4. We did not want to train via stochastic gradient descent as we believed this would decrease training speed. We used early stopping in an attempt to minimize overfitting. We also used the Adam Optimizer to reduce overfitting. Our model stopped training after four epochs.

5.2 FasterRCNN

In their paper "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", Ren *et al* propose an improvement over Girshick's *FastRCNN*, which is the second model of the RCNN family [14]. The standard RCNN architecture uses the selective search algorithm to propose a series of regions in which an object may be found. Then, a pre-trained CNN is used on each of those regions to attempt to detect the presence of certain objects. The largest downside to this approach is the computational intensity of the selective search as well as evaluating the CNN on all of the proposed regions.

In the paper "Fast R-CNN", Ross Girshick improves the computational cost and performance of the RCNN architecture [13]. The key insight of Girshick's work is that it is possible to share computation when evaluating the CNN on proposed regions by replacing the last pooling layer with a ROI pooling layer, which converts an input of any size into a fixed-size feature vector. The last fully-connected layer of the CNN is replaced with a softmax layer (for classes) and a fully-connected layer for regression on bounding boxes.

Finally, the Ren *et al* paper improves upon "Fast R-CNN" by integrating the region proposal network and the CNN into one network. The architecture of the Faster-RCNN model is shown in the Appendix.

6 Experiments/Results/Discussion

6.1 Results with FasterRCNN

Unfortunately, we were unable to successfully train a working implementation of FasterRCNN in the time allotted. Because we began training a FasterRCNN model after the EfficientDet model, we did

Model	IoU (P > 0.5)	AP	AR	F1
Baseline	**0.00**	**99.1**	**56.9**	**72.29**
YOLO	**0.00**	**98.8**	**58.1**	**73.10**
EfficientDet	**0.00**	**76.4**	**84.4**	**80.20**
EfficientDet	0.05	58.3	72.9	64.79
EfficientDet	0.10	37.9	57.9	45.81
EfficientDet	0.15	30.3	50.9	37.99
EfficientDet	0.20	26.3	46.8	33.68
EfficientDet	0.25	22.3	42.7	29.30
EfficientDet	0.30	18.9	39.3	25.52
EfficientDet	0.35	15.4	35.7	21.52
EfficientDet	0.40	13.4	33.2	19.09
EfficientDet	0.45	11.5	30.8	16.75
EfficientDet	**0.50**	**9.8**	**28.4**	**14.57**
EfficientDet	0.75	0.7	7.3	1.28
EfficientDet	**0.00:0.50**	**29.3**	**47.6**	**36.27**
EfficientDet	0.50:0.95	2.9	11.5	4.63

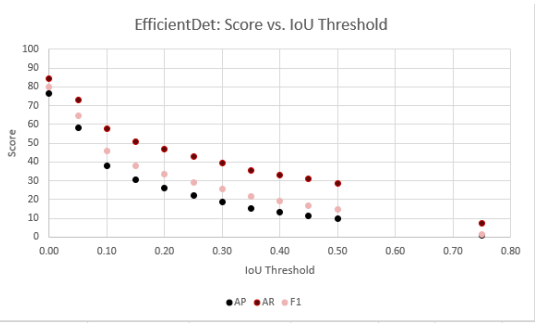

Figure 3: EfficientDet, YOLOv3, and baseline results. **LEFT:** Table displaying average precision, average recall, and F1 scores at differing IoU thresholds with minimum confidence score of 0.5. **RIGHT:** Plot of scores vs. IoU threshold.

not have enough time to perform an expansive hyperparameter search and as such did not achieve a model that converged. However, we still believe that there is potential for a FasterRCNN model to be effective at our task, and will consider such a model in our future work.

6.2 Results with EfficientDet

We present the results for our implementation of the EfficientDet architecture. With a minimum required confidence score of 0.5 and a minimum IoU threshold of 0.0, our EfficientDet model demonstrated improved recall but reduced precision when compared to the YOLOv3 and baseline architectures (0.844 and 0.764, respectively). As the required minimum IoU threshold for correct prediction increased, precision and recall subsequently decreased. Our mAP average with range 0.0 to 0.5 was 0.293 while mAR with the same range was 0.476. AP and AR at the standard required minimum IoU of 0.5 were 0.098 and 0.284, respectively. AP, AR, and F1 scores are plotted against each IoU threshold in **Figure 3**. Examples of correctly detected papillary and non-papillary in both WLC and BLC along with mislableed and poorly annotated predictions are shown in **Figure 4**.

6.3 Github Repository Link

https://github.com/markl21/CystoNet2

7 Conclusion/Future Work

Given our results, we conclude that EfficientDet is a promising object detection framework that exhibits improved recall when compared to YOLOv2. Although precision is decreased, we note that recall is the more important metric to consider as the cost of false negatives is higher than the cost of false positives. In the clinic, falsely flagged areas of suspicion are less morbid for the patient than are missed malignant lesions. Even though performance suffered at higher required IoUs, we predict that EfficientDet will be able to replace YOLOv3 as the object detection algorithm of choice.

We also note that EfficientDet could make predictions to images approximately four times faster than that of YOLOv3 on the same local machine (2.8GHz intel i7-7700HQ Octacore CPU, Nvidia 1050 Ti GPU with 4GB Memory, 16GB RAM). While YOLOv3 made predictions at approximately 5 fps, EfficientDet made predictions at approximately 20 fps. EfficientDet's optimized model architecture consequenltly may confer increased accessibility to socioeconomically disadvantaged communities that may not be able to afford a computer with strong GPU capability.

7.1 Future Work

Although our EfficientDet implementation was successful, we believe we could achieve greater precision by training on a larger dataset. Due to limited availability of data and computational power, we were only able to train our model on fifteen different tumors. However, the laboratory

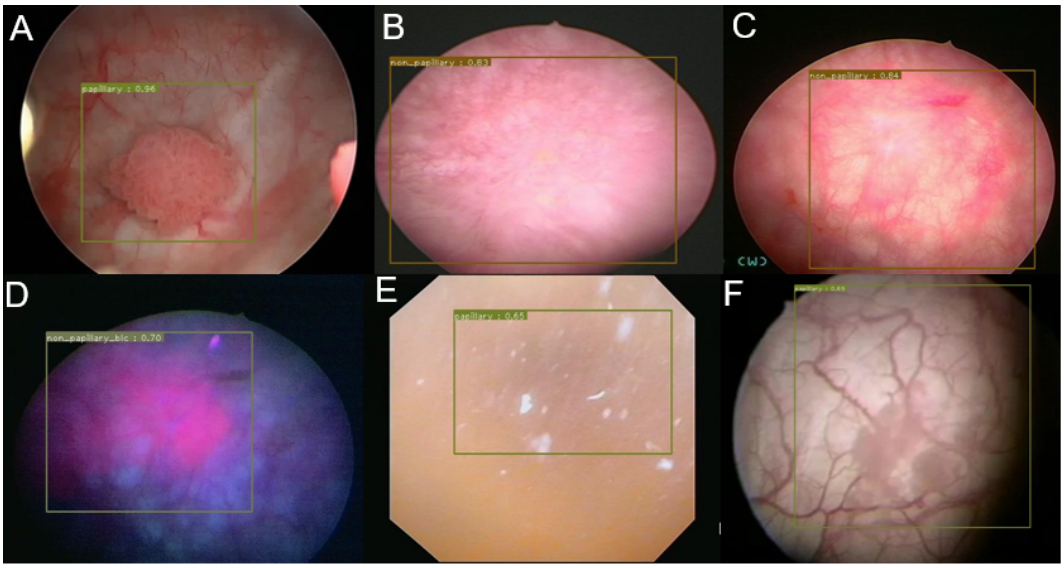

Figure 4: Visualization of EfficientDet performance. **A:** Successful detection of papillary lesion during WL TURBT. **B:** Successful detection of non-papillary CIS during WLC. **C:** Successful detection of non-papillary CIS during WLC from unrelated patient. **D:** Successful detection and classification of non-papillary CIS during BLC. **E:** False positive flagging of post-BCG debris during WLC of patient with no discovered areas of suspicion. **F:** Successful classification yet suboptimal box prediction of papillary lesion with IoU of 0.41.

that provided our training set will soon have dozens more examples of expert-annotated tumors for training, which we believe would increase the accuracy of our model. We also believe that an examination of other state-of-the-art object detection architectures is warranted. Although we were unable to achieve results, we believe that Faster-RCNN could show promise.

Lastly, we believe that further research can be done into tumor segmentation rather than simple tumor detection, as determining tumor borders is just as, if not more, important as detecting the tumor clinically.

8 Contributions and Acknowledgements

Mark Laurie, Colt McNealy, and Victor Yin annotated roughly equal proportions of the cystoscopy dataset. Mark performed significant preprocessing of raw data and significantly modified the forked EfficientDet repository to adapt to our needs. Mark, Victor, and Colt contributed equally to the writeups and video report. Victor researched and implemented screentone removal. Colt implemented the initial baseline and started research on the FastRCNN implementation.

We thank members of the Liao lab for collecting RT performance metrics using the YOLO-based implementation of CystoNet. and for annotating a large minority of our database. We also thank Dr. Lei Xing and Xiaomeng Li for their technical mentorship.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69(1):7-34. Epub 2019/01/09. doi: 10.3322/caac.21551. PubMed PMID: 30620402.

[2] Avritscher EB, Cooksley CD, Grossman HB, Sabichi AL, Hamblin L, Dinney CP, Elting LS. Clinical model of lifetime cost of treating bladder cancer and associated complications. Urology. 2006;68(3):549-53. Epub 2006/09/19. doi: 10.1016/j.urology.2006.03.062. PubMed PMID: 16979735.

[3] Cheung G, Sahai A, Billia M, Dasgupta P, Khan MS. Recent advances in the diagnosis and treatment of bladder cancer. BMC Med. 2013;11:13. Epub 2013/01/19. doi: 10.1186/1741-7015-11-13. PubMed PMID: 23327481; PMCID: 3566975.

[4] Stenzl A, Burger M, Fradet Y, Mynderse LA, Soloway MS, Witjes JA, Kriegmair M, Karl A, Shen Y, Grossman HB. Hexaminolevulinate guided fluorescence cystoscopy reduces recurrence in patients with nonmuscle invasive bladder cancer. The Journal of urology. 2010;184(5):1907-13. Epub 2010/09/21. doi: 10.1016/j.juro.2010.06.148. PubMed PMID: 20850152.

[5] Augmented Bladder Tumor Detection Using Deep Learning Shkolyar, Eugene et al. European Urology, Volume 76, Issue 6, 714 - 718

[6] Simon, and Nrea. "Breast Cancer Detection Using Convolutional Neural Networks." arXiv.org, March 19, 2020. https://arxiv.org/abs/2003.07911

[7] "EfficientDet: Towards Scalable and Efficient Object Detection." Google AI Blog, April 15, 2020. https://ai.googleblog.com/2020/04/efficientdet-towards-scalable-and.html

[8] Soubra, A., Risk, M. C. (2015). Diagnostics techniques in nonmuscle invasive bladder cancer. Indian journal of urology : IJU : journal of the Urological Society of India, 31(4), 283–288. https://doi.org/10.4103/0970-1591.166449

[9] Classification of Lesions in Breast Ultrasound Images Using Neural Networks, n.d. https://github.com/mungujn/machine-learning-detect-cancer/blob/master/Undergraduate dissertation, Lesion classification using machine learning, Nickson Mungujakisa, github.commungujn.pdf.

[10] Sekachev BMNA, Z. . Computer Vision Annotation Tool: A Universal Approach to Data An-notation. Retrieved 2019 [cited 2020 10/03/2020]. Available from: https://software.intel.com/en-us/articles/computer- vision-annotation-tool-a-universal-approach-to-data-annotation.

[11] Sainath, T. N., Vinyals, O., Senior, A., Sak, H. (2015). Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). doi: 10.1109/icassp.2015.7178838

[12] Feichtenhofer, Christoph, et al. "Slowfast networks for video recognition." Proceedings of theIEEE International Conference on Computer Vision. 2019.

[13] Girshick, Ross. "Fast R-CNN." 2015.

[14] Ren, Shaoqing et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." 2015.

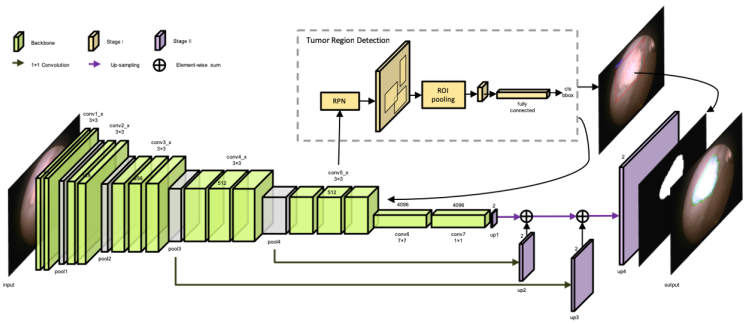

Figure 5: YOLO architecture used in prior implementation of CystoNet.

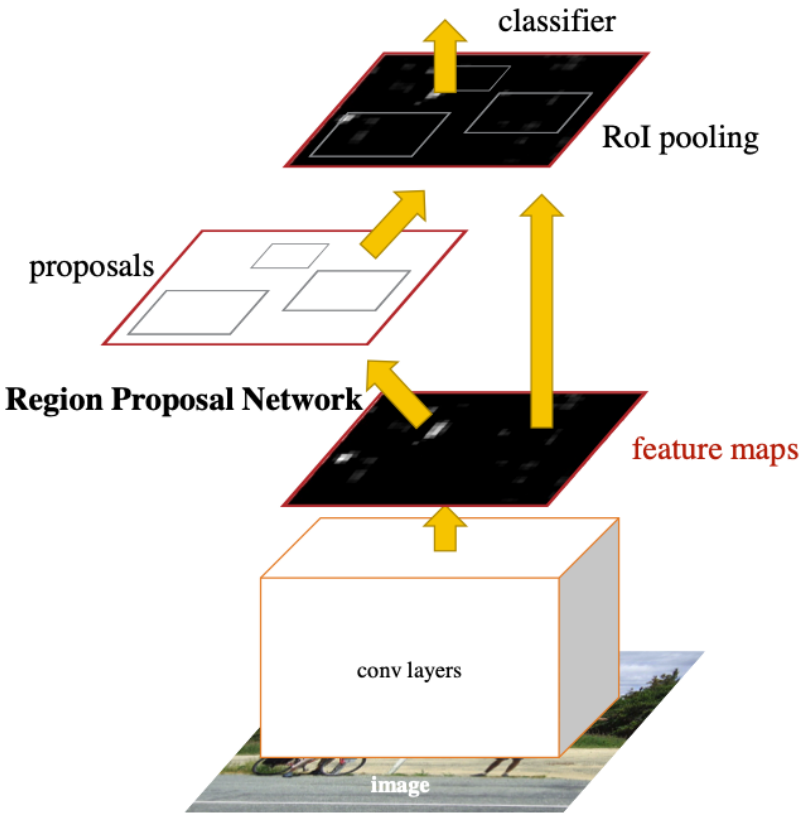

Figure 6: Faster-RCNN Architecture.

Appendix Figures