The 7th IEEE International Conference on E-Health and Bioengineering - EHB 2019
Grigore T. Popa University of Medicine and Pharmacy, Iasi, Romania, November 21-23, 2019

Automatic Detection of Melanoma with Yolo Deep Convolutional Neural Networks

Yali Nie[1], Paolo Sommella[2], Mattias O'Nils[1], Consolatina Liguori[2], Jan Lundgren[1]
[1]dept. of Electronics Engineering, Sundsvall, Sweden
yali.nie@miun.se, mattias.onils@miun.se, Jan.Lundgren@miun.se
[2]dept. of Industrial Engineer, Salerno, University of Salerno, Italy
psommella@unisa.it, tliguori@unisa.it

***Abstract*—In the past three years, deep convolutional neural networks (DCNNs) have achieved promising results in detecting skin cancer. However, improving the accuracy and efficiency of the automatic detection of melanoma is still urgent due to the visual similarity of benign and malignant dermoscopic images. There is also a need for fast and computationally effective systems for mobile applications targeting caregivers and homes. This paper presents the You Only Look Once (Yolo) algorithms, which are based on DCNNs applied to the detection of melanoma. The Yolo algorithms comprise YoloV1, YoloV2, and YoloV3, whose methodology first resets the input image size and then divides the image into several cells. According to the position of the detected object in the cell, the network will try to predict the bounding box of the object and the class confidence score. Our test results indicate that the mean average precision (mAP) of Yolo can exceed 0.82 with a training set of only 200 images, proving that this method has great advantages for detecting melanoma in lightweight system applications.**

***Keywords*—Image processing; Melanoma; Yolo; Object Detection.**

I. INTRODUCTION

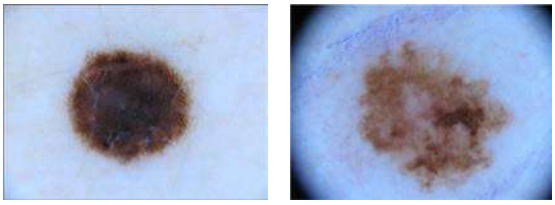
Skin diseases remain a major cause of disability worldwide, and skin conditions pose a significant threat to patient well-being. Skin conditions accounted for 1.79% of the global burden, measured in disability-adjusted life years (DALYs), of 306 diseases and injuries in 2013 [1]. Both geographic and age-related factors affect the skin disease burden, with melanoma being the single most common diagnosis in resource-rich regions such as Australia, New Zealand, Northern Europe, and North America. Malignant melanoma is a highly aggressive cancer that tends to spread to other parts of the body, and may be fatal if not treated early. If melanoma is detected at an early stage, it can usually be completely removed with surgery. Recently, deep learning has played a vital role in the early detection of cancer. In this paper, we use deep convolutional neural networks (CNNs) to detect melanoma. Furthermore, mobile applications in e-healthcare is a recent trend and future direction that is explored here. As an example, a cross-sectional survey of HIV-positive patients in Botswana found that 91% felt that a mobile teledermatology visit would provide the same level of care as a face-to-face visit [2].

For object detection based on deep learning, three main detector families have been encountered by researchers: R-CNN, SSD and Yolo. The region-CNN (R-CNN) family [3, 4, 5], the single shot detector (SSD) [6], and the You Only Look Once (Yolo) series [7, 8, 9]. All these algorithms treat object detection as a regression problem, taking a given image and simultaneously learning bounding box (BBox) coordinates and the corresponding probabilities of class labels. R-CNN [3] and its variants comprise the original R-CNN, Fast R-CNN [4], and Faster R-CNN [5], which tend to be very accurate but incredibly slow, mainly due to being two-stage detectors. SSD [6] offers a good tradeoff between speed and accuracy. However, we are aiming to develop a mobile melanoma detector in the future, and speed is paramount for this application. For this purpose, a Yolo algorithm is the best option as it is a one-stage detector. Yolo is less accurate but significantly faster than R-CNN. Undoubtedly, Yolo has some drawbacks and limitations; for example, it cannot handle small objects well and does not deal well with objects grouped close together. In melanoma detection, however, these phenomena are rarely encountered.

II. RELATED WORK

Recent years have seen considerable research into melanoma detection from dermoscopic images. Earlier techniques for detecting skin cancer include the ABCDE method [10] and the seven-point checklist [11, 12]. These methods [10, 11, 12, 13] focus on extracting low-level visual features, such as color, edge, and texture descriptors. Some, such as ANN [14], SVM [15], and kNN [16], use machine learning techniques. Segmentation methods have also been reported by Ganster et al. [17]. Compared with conventional methods that extract low-level handcrafted features, methods based on deep learning can extract deeper and more generic features. In recent research, various approaches based on deep learning have been proposed [18, 19, 20, 21]. Deep learning, especially deep convolutional neural networks (CNNs), can automatically categorize input datasets of, for example, audio and images. CNNs can directly learn from the combined raw pixel data and class labels through end-to-end learning. In this paper we apply the Yolo series of state-of-the-art, real-time object detection systems [7, 8, 9], which have proven good

competitors to Fast R-CNNs and SSDs in terms of both detection and speed. The International Skin Imaging Collaboration (ISIC) dataset [21] is one of the largest collections of dermoscopic images. In our paper, we chose 200 images from the ISIC dataset, with half of the images being benign and half malignant (see example images in Figure 1).



(a) (b)

Fig. 1. Images of (a) benign lesion and (b) melanoma.

III. Detection via the YOLO DCNN

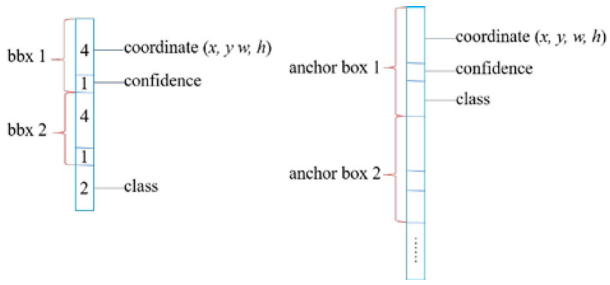
Yolo is essentially a unified detection model without a complex processing pipeline that uses the whole image as the network input, which will be divided into an $s \times s$ grid. After the selection from the network, the model directly outputs the position of the object border and the corresponding category in the output layer as shown in Figure 2. However, Yolo is not effective in detecting close objects and small populations. In response to these problems in the original YoloV1 network, YoloV2 and YoloV3 borrowed some ideas from SSD [6] and Faster R-CNN [5]. These combined approaches yield big improvements in terms of higher detection accuracy and speed. YoloV2 and YoloV3 are improved models of YoloV1 built without complex networks, but incorporating a wide variety of ideas from other deep learning models and their applying methods, such as applying batch normalization, anchor boxes, and fine-tuning with high-resolution classifiers.

Fig. 2. The structure of Yolo detection with s × s grid input image, after Yolo selects, it find the final location of the melanoma.

A. YoloV1

YoloV1 uses a single convolutional neural network to transform the target detection problem into a regression problem that extracts BBoxes and class probabilities directly from the image. The entire inspection process is divided into three phases: zooming the image to 448 × 448; detection and classification using grid cell; and Non-Maximum Suppression (NMS) [22], which is used to filter redundant detection candidates.

Specifically, Yolo first divides the image into $s \times s$ grid cells. If the center of a target falls into the grid, the grid is responsible for detecting the target. Each grid cell predicts two BBoxes, one box confidence score, and two conditional class probabilities, as shown in Figure 3(a). Each BBox contains five values: coordinates x, y, w, h and the confidence value of the target. The x and y coordinates are offsets of the corresponding cell. The w and h coordinates are the BBox width and height, respectively. Yolo has 24 convolutional layers followed by two fully connected layers.



(a) (b)

Fig. 3. The final predictions of YoloV1 (a) and improved YoloV2 (b).

B. YoloV2

The YoloV2 deep learning framework, the improved version of YoloV1, is proposed for modeling the architecture and training the model.

1) The YoloV2 detection network extracts features based on the Darknet-19 model, whose last convolution layer is changed to three 3 × 3 convolution layers and 1024 channels, adding a 1 × 1 convolution layer after each convolution to compress the features and speed up operation. Since convolution does not require reshaping, the spatial information is well preserved.

2) YoloV2 changes the grid behind the multi-layer convolution and pools operations from the original 7 × 7 grid to 13 × 13, where a large grid increases the size of the network's feature map. The larger grid makes the grid denser and increases the density of the BBox. When the image contains multiple objects, especially small objects, the enlarged network can increase the number of objects extracted, thereby improving the detection accuracy. However, the complexity and computation of the model will also increase. To establish an acceptable tradeoff between accuracy and detection speed, it is therefore crucial to select the appropriate grid size.

3) YoloV2 draws on the idea of using an anchor [5], and using K-means to cluster the BBoxes of self-made datasets by dimension. The sizes and numbers of anchors affect the speed of detection and the accuracy of the BBox location in the model. The anchor box predicts the type of object and its coordinates. There are five anchor boxes in YoloV2. The final prediction is shown in Figure 3(b).

4) In YoloV2, a route layer is used as a pass-through layer to extract feature maps from shallow layers. The reorganization layer can combine shallow information with deep feature extraction information.

C. YoloV3

YoloV3, based on the idea of ResNet [23], incorporates Darknet-19 from YoloV2 to propose a new deeper and wider feature extraction network called Darknet-53. YoloV3 has nine anchor boxes. The main improvement is the application of a

feature pyramid network (FPN) [24] that can generate a higher-quality feature graph pyramid.

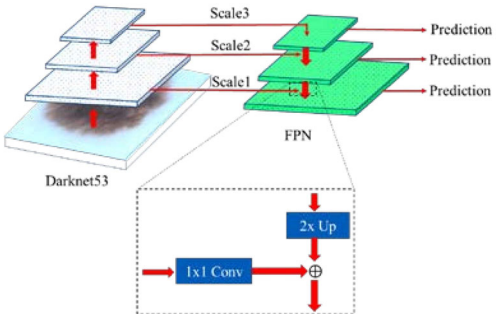


Fig. 4. Darknet-53 with Yolo layers.

As shown in Figure 4, the input image passes through the backbone network, Darknet-53, which mainly adopts blocks as its basic components. Every residual block consists of a 3 × 3 and a 1 × 1 convolutional layer paired with a shortcut connection. After Darknet-53, the input image will be sampled three times, 32 times, 16 times, and eight times, successively, to obtain the multi-scale feature map. After 32× downsampling, the feature map is too small, so YoloV3 uses upsampling with stride 2 to double the size of the resulting feature map, resulting in 16× downsampling. Similarly, the feature map sampled at 16 times is sampled with a step size of two, and a feature map sampled at eight times is obtained. There are shortcut connections that concatenate the intermediate layers of Darknet-53 to the right layer after upsampling. The pyramid structure shown in Figure 4 will be referred to as Yolo layers. Every layer has sub-layers of convolution, batch normalization, and leaky ReLU activation. The three Yolo layers have grids of different resolutions and three anchors with different scale shapes. These three Yolo layers are capable of capturing different scales of objects.

IV. Results and Evaluation

Our experiment was run on a server equipped with a Ubuntu 16.04 system. The development environment was Python 3.6. From YoloV1 to YoloV3, we applied three different frameworks: 24 convolutional layers [7], Darknet-19, and Darknet-53. To compare all the mean average precision (mAP) values of the Yolo versions under the same conditions, we set both the batch and the subdivision to 32, since other numbers would lead to out-of-memory errors. The image size was 448 × 448. The other parameters of YoloV1 were as follows: momentum was 0.9; decay was 0.0005; learning rate was 0.0005; policy was in the steps 200, 400, 600, 20000, and 30000; scales took 2.5, 2, 2, 0.1, and 0.1; and the maximum number of batches was set to 40000.

A. Dataset description

The melanoma detection dataset in this paper was self-made. The main data sources were selected from the ISIC dataset, a readily available skin disease database. Since the Yolo detector is a supervised deep learning algorithm, it needs manually labeled data. We used the PASCAL VOC2007 dataset format as a reference, and all the positions and sizes of the melanomas were saved as a formatted .xml file. Simulating a classification application where the amount of incoming malignant images can be small, only 200 images were selected, half benign and half malignant, to keep balanced categories. There were 160 images for training, 20 for validation, and the remaining 20 for testing. To increase the number of data in the training set and improve the generalization ability of the model, data augmentation was used with random scaling and translations to enhance the data.

B. Evaluation method

In object detection, there are two distinct tasks: one is determining whether an object exists in the image, and the other is to find the object's location. Furthermore, a typical dataset will have more than one class, whose distribution is non-uniform. Biasing will occur with a simple accuracy-based metric, since it is also important to assess the risk of misclassification. A model score detecting each BBox to assess the model at different levels of confidence was therefore proposed. In this paper, the mAP [25] was used as the evaluation criteria. The mAP score is determined by calculating the mAP over all classes and all Intersection over Union (IoU) [26] thresholds. In this experiment, the IoU threshold was considered to be 0.5 and the mAP was averaged over the two object classes.

C. Comparison and discussion

The performance of the three versions of Yolo is compared in Table I. The mAP of YoloV3 is 0.770, which is 0.064 lower than that of YoloV2; however, YoloV1's mAP is the lowest at 0.371. Hence, the proposed YoloV2 better balances the relationship between feature representation capacity and the network overfitting problem. The YoloV3 model is proposed for small object detection, as it has poor detection results on large targets.

TABLE I. Mean Average Precision (mAP) comparison of Yolo versions

Framework	Benign	Malignant	mAP
YoloV1	0.41	0.33	0.37
YoloV2	0.85	0.82	0.83
YoloV3	0.79	0.75	0.77

The quality of the Yolo results is shown in Figure 5. Figure 5(a) and (b) illustrate the correct detection of benign and malignant lesions, respectively. Figure 5(c) shows that lesions of more than one disease may appear in a single image. However, the classification task mostly judges one class in an image. Here, the detection task can solve more than one class in an image. During image processing, noise always occurs and will have a large impact on the accuracy of an experiment. Additionally, extra features such as hairs are present in Figure 5(d). If a handcrafted method is used, it will be difficult to remove all the hairs. With a deep learning method, however, the process of removing the hairs can be ignored. We can also obtain results like those in Figure 5(e), where even the machine

is confused about the melanoma classification. It is therefore very important to label correctly before training.

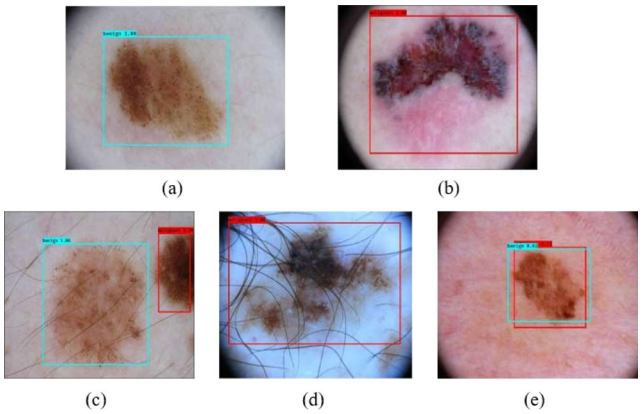


Fig. 5. Yolo detection results.

V. Conclusions

Our main contribution to the work is to apply the State-of-the-art yolo network to do the detection of melanoma. Even with a small dataset, the performance is promising. The presented comparison of three DCNN networks (Yolo) shows that YoloV2 has the highest accuracy, reaching approximately 83% classification accuracy. The main conclusion why YoloV2 performed better for Automatic Detection of Melanoma in this dataset is that the Darknet-19 feature extraction and the larger grid size is better suited for the images in the tested dataset. These results also shows that further efforts needs to include studies of how the DCNN network can be improved and in order to further develop the classification, there is a great need for larger datasets and more classes of melanoma.

References

[1] S. Divya, K. Cheldize, D. Brown, and E.E. Freeman, 2017. Global burden of skin disease: Inequities and innovations. Current Dermatology Reports, 6(3), pp. 204–210.

[2] R. S. Azfar, J.L. Weinberg, G. Cavric, Ivy A. Lee-Keltner, W.B. Bilker, J.M. Gelfand, and C.L. Kovarik, 2011. HIV-positive patients in Botswana state that mobile teledermatology is an acceptable method for receiving dermatology care. Journal of Telemedicine and Telecare, 17(6), pp. 338–340.

[3] G. Ross, J. Donahue, T. Darrell, and J. Malik, 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 580–587).

[4] G. Ross, 2015. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1440–1448).

[5] R. Shaoqing, K. He, R. Girshick, and J. Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems (pp. 91–99).

[6] L. Wei, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Fu, and A.C. Berg, 2016, October. Ssd: Single shot multibox detector. In European Conference on Computer Vision (pp. 21–37). Springer, Cham.

[7] R. Joseph, S. Divvala, R. Girshick, and A. Farhadi, 2016. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779–788).

[8] R. Joseph, and A. Farhadi, 2017. YOLO9000: better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7263–7271).

[9] R.Joseph, and A. Farhadi, 2018. YOLOv3: An incremental improvement. arXiv preprint, arXiv:1804.02767.

[10] W. Stolz, 1994. ABCD rule of dermatoscopy: A new practical method for early recognition of malignant melanoma. European Journal of Dermatology, 4(7), pp. 521–527.

[11] M. Scott W., C. Ingvar, K.A. Crotty, and W.H. McCarthy, 1996. Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features. Archives of Dermatology, 132(10), pp. 1178–1182.

[12] G. Di Leo, A. Paolillo, A. Pietrosanto, P. Sommella, G. Fabbrocini and S. Cacciapuoti, 2015, May. A distributed measurement system for dermoscopic analysis of pigmented skin lesions. In 2015 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings (pp. 1646–1651). IEEE.

[13] G. Rahil, M. Aldeen, and J. Bailey, 2012. Computer-aided diagnosis of melanoma using border- and wavelet-based texture analysis. IEEE Transactions on Information Technology in Biomedicine, 16(6), pp. 1239–1252.

[14] A.H. AlAsadi and B.M. ALsafy, 2016. Early detection and classification of melanoma skin cancer. Riga, Latvia: Lambert Academic Publishing.

[15] Y. Xiaojing, Z. Yang, G. Zouridakis, and N. Mullani, 2006, August. SVM-based texture classification and application to early melanoma detection. In 2006 International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 4775–4778). IEEE.

[16] B. Lucia, R.B. Fisher, B. Aldridge, and J. Rees, 2013. A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In Color Medical Image Analysis (pp. 63–86). Springer, Dordrecht.

[17] H. Ganster, , P. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, 2001. Automated melanoma recognition. IEEE Transactions on Medical Imaging, 20(3), pp. 233–239.

[18] Y. Lequan, H. Chen, Q. Dou, J. Qin, and P. Heng, 2016. Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Transactions on Medical Imaging, 36(4), pp. 994–1004.

[19] E.Z. Chen, X. Dong, J. Wu, H. Jiang, X. Li, and R. Rong, 2018. Lesion attributes segmentation for melanoma detection with deep learning. bioRxiv, https://doi.org/10.1101/381855.

[20] X. Fengying, H. Fan, Y. Li, Z. Jiang, R. Meng, and A. Bovik, 2016. Melanoma classification on dermoscopy images using a neural network ensemble model. IEEE Transactions on Medical Imaging, 36(3), pp. 849–858.

[21] N.C.F. Codella, D. Gutman, M.E. Celebi, B. Helba, M.A. Marchetti, S.W. Dusza, A. Kalloo et al, 2018, April. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). In 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018) (pp. 168–172). IEEE.

[22] R. Rasmus, M. Guillaumin, and L.V. Gool, 2014, November. Non-maximum suppression for object detection by passing messages between windows. In Asian Conference on Computer Vision (pp. 290–306). Springer, Cham.

[23] H. Kaiming, X. Zhang, S. Ren, and J. Sun, 2016. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770–778).

[24] L. Tsung-Yi, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, 2017. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2117–2125).

[25] E. Mark, and J. Winn, 2011. The PASCAL visual object classes challenge 2012 (VOC2012) development kit. Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep.

[26] R. Adrian, 2016. “Intersection over Union (IoU) for object detection.” http://www.pyimagesearch.com/2016/11/07/intersection-overunion-iou-for-objectdetection.

[27] Z. Zhong-Qiu, P. Zheng, S. Xu, and X. Wu, Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems. 2019 Jan 28.