J Digit Imaging (2018) 31:435–440
https://doi.org/10.1007/s10278-017-0026-y

Rethinking Skin Lesion Segmentation in a Convolutional Classifier

Jack Burdick[1] · **Oge Marques**[1] · **Janet Weinthal**[1] · **Borko Furht**[1]

Published online: 18 October 2017

Abstract Melanoma is a fatal form of skin cancer when left undiagnosed. Computer-aided diagnosis systems powered by convolutional neural networks (CNNs) can improve diagnostic accuracy and save lives. CNNs have been successfully used in both skin lesion segmentation and classification. For reasons heretofore unclear, previous works have found image segmentation to be, conflictingly, both detrimental and beneficial to skin lesion classification. We investigate the effect of expanding the segmentation border to include pixels surrounding the target lesion. Ostensibly, segmenting a target skin lesion will remove inessential information, non-lesion skin, and artifacts to aid in classification. Our results indicate that segmentation border enlargement produces, to a certain degree, better results across all metrics of interest when using a convolutional based classifier built using the transfer learning paradigm. Consequently, preprocessing methods which produce borders larger than the actual lesion can potentially improve classifier performance, more than both perfect segmentation, using dermatologist created ground truth masks, and no segmentation altogether.

✉ Oge Marques
omarques@fau.edu

Jack Burdick
jburdick2015@fau.edu

Janet Weinthal
jweinthal2012@fau.edu

Borko Furht
bfurht@fau.edu

[1] Florida Atlantic University, Boca Raton, FL, USA

Keywords Medical decision support systems · Deep learning · Medical image analysis · Convolutional neural networks · Skin lesions · Machine learning

Introduction

One in 33 men and one in 52 women in the USA will develop melanoma, a deadly form of skin cancer, in their lifetimes [1]. In 2016, Melanoma is estimated to have killed over 10,000 people in the USA alone [1]. 98% of patients survive when melanoma is diagnosed early, however only 17% of patients survive when melanoma is left undiagnosed until its distant stage [1]. Melanoma treatment costs over $3.3 million dollars in the USA annually [2], while the indirect cost of melanoma due to premature mortality is estimated to be over $3 billion [3]. Early detection is key to minimizing economic costs and saving lives.

Dermatologists have developed many methods to diagnose melanoma, including the ABCD (Asymmetry, Border, Color, and Differential structure) Rule [4], the 7-point Checklist [5], and the CASH (Color, Architecture, Symmetry, and Homogeneity) algorithm [6]. However, studies have shown that the success of these heuristic algorithms is limited [7–9]. This has provided a strong motivation for the use of computer aided diagnosis (CADx) systems powered by deep learning architectures, which might improve the accuracy and sensitivity of melanoma detection methods and potentially outperform medical professionals working on the same task [10].

This paper focuses on the role, importance, and impact of skin lesion segmentation prior to classification on the eventual results of classification architectures. More specifically, we evaluate the performance of two state-of -the-art

convolutional classification architectures after applying varying levels of a post-segmentation morphological dilation [11]. Although excellent classification results have been achieved without any image preprocessing whatsoever, given the current nature of readily available clinical data sets, image segmentation is necessary to ensure both accurate CNN training and classification results. Many images contain background noise and/or artifacts (Fig. 1), which can potentially lead to incorrect classification results due to the network either learning incorrect features while training or the already trained classifier using invalid image data. As such, without perfect skin lesion images, skin lesion image segmentation methods [12] are necessary to produce reliable results.

By performing comparisons between augmented segmentation and prior methods in a systematic and reproducible manner, we demonstrate empirically that a certain amount of segmentation border enlargement can improve image classification results. This departure from conventional methods, which traditionally use no segmentation or aim for dermatologist-like segmentation, could improve classification performance.

Related Work

Previous works have developed deep learning solutions which classify melanoma with accuracies ranging between 70 and 95% [10, 13–17]. While segmentation is often a key step in image classification [18], recent works have produced excellent results without segmentation [10, 16] and have even found segmentation to be detrimental to accuracy in skin lesion classification [13, 15]. However, in [13], authors found that sensitivity increases despite the decrease in accuracy when classification is performed on unaltered—instead of perfectly segmented—skin lesion images.

Methods

This section explains the methods used to investigate the effects of including additional pixel data that lie outside the target lesion in an image. Conceptually, our goal is to demonstrate empirically that extending borders beyond the lesion to include background pixels may improve the performance of a two-class (melanoma vs. benign) skin lesion classifier using a CNN. To demonstrate, we use morphological dilation to produce progressively more dilated versions of the manually generated ground-truth masks that precisely outline the skin lesion of interest, producing a series of masks which increasingly expand the segmented area beyond the lesion (Fig. 2).

Input Image Preparation

Dilated masks were created, prior to the subsequent resizing, through a morphological dilation, using the original raw images and binary masks provided by the ISIC dataset (Section "Dataset") as reference images and disk-shaped structuring elements of various pixel sizes (25, 50, 75 and 100 pixels). For the few images containing artifacts (such as markers or stickers, see Fig. 1), the dilated versions of the masks do not include any marker information.

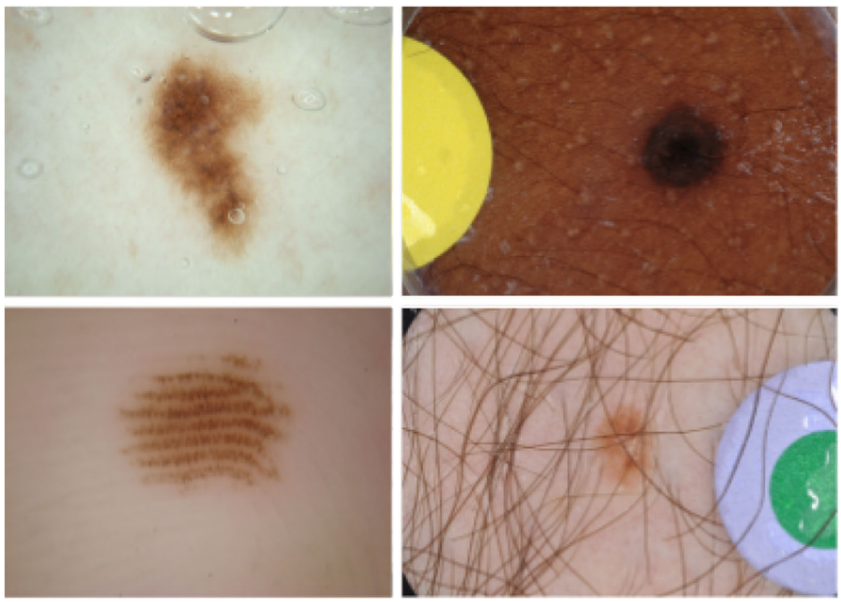

Fig. 1 Sample input images

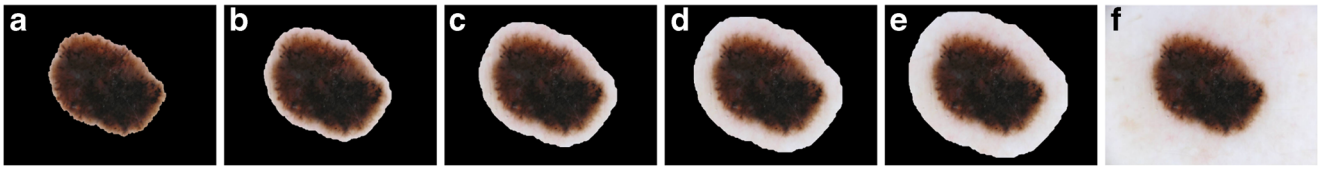

Fig. 2 Input images: a perfectly segmented image (**a**), progressively larger imperfectly segmented images (obtained with dilation using disk structuring elements with radii 25 (**b**), 50 (**c**), 75 (**d**), and 100 (**e**) pixels, and an unsegmented image (**f**)

Classifier Architecture

CNN Architecture Two deep convolutional classifier architectures were implemented: VGG-Net [19] and InceptionV3 [20]. VGG-16 is a well-known CNN, made popular due to its excellent performance achieved on the ImageNet dataset [21]. The VGG-16 architecture has been shown to generalize well to a variety of datasets [19] and—in previous works—has produced promising results on the ISIC dataset [13, 16]. InceptionV3, the third version of the Inception architecture [22], is popular due to its lower computational cost compared to VGG and other widely used CNN architectures [20].

Design Considerations Since the input layer of the VGG-16 architecture expects a 224×224 pixel RGB image and the InceptionV3 architecture expects a 299×299 pixel RGB image, the input images must be resized accordingly. After an input image has been loaded, it is passed through the layers as designated by the given architecture. The networks are concluded with a classifier block consisting of fully connected (FC) layers.

A few modifications were made to the original VGG-16 and InceptionV3 architectures:

- The final fully connected output layer performs a binary classification (melanoma vs. benign), not 1,000 classes as previously designed for the ImageNet dataset.
- In VGG-16 three, not four, fully connected layers are implemented as the final classifying block.
- In InceptionV3, a global average pooling layer followed by two fully connected layers is implemented as the final classifying block.
- The activation function in the modified final layer has been changed from Softmax to Sigmoidal.

Transfer Learning

Rather than attempting to train an entire CNN from scratch using a small dataset, a transfer learning approach was utilized. Transfer learning makes use of features previously learned from a different, larger dataset [23]. Using the principles of transfer learning, we initialized the architecture by loading weights from the networks pre-trained on the ImageNet dataset [24]. ImageNet contains 1.2 million images labeled with 1,000 classes, none of which include skin lesions.

In this work, the initial layers of the network were frozen (prevented from being modified during training) because they contain more generic features which can be transferred to any natural image. By training only the final layers of the network on the ISIC dataset, the network learns finer features specific to skin lesion images.

All convolutional layers in the original VGG-16 architecture are initialized with weights from the ImageNet dataset. The first four convolutional blocks were frozen while the fifth, and final, convolutional block was initialized with weights saved from the pretrained architecture and left unfrozen for training. A similar principle was applied to the InceptionV3 architecture, where all layers in the base model were initialized with weights from training the model on the ImageNet dataset and subsequently frozen.

Implementation Considerations

Python [25] was selected as the main programming language. In some specific tasks, such as creating the dilated masks, MATLAB [26] was also used. Keras [27], a deep learning framework that provides a layer of abstraction on top of lower level deep learning libraries was used as the main deep learning based classification framework. During experimentation, Theano [28] was the underlying framework implemented for the VGG-16 architecture and Tensorflow [29] was the underlying framework implemented for the InceptionV3 architecture. Notable additional libraries and dependencies include a general purpose machine learning and image processing library—Scikit-learn library [30] and PIL [31], respectively.

GPUs (Graphics Processing Units) were used to meet the computational demands of training a CNN. CuDNN and CUDA libraries [32], required for programming Nvida GPUs, were utilized.

Experiments and Results

This section presents the results of the experiments using the proposed methods and the selected implementation.

Dataset

The ISIC Archive dataset [33], as used in the ISBI 2016 Challenge [34], was adapted for this work. The ISIC Archive dataset is publicly available and contains 1279 RGB images that are pre-partitioned into 900 training images and 379 testing images. All images are labeled (benign or malignant) and include a corresponding binary image mask. The training dataset was manually split 70–30% to create a validation dataset while training. As the original dataset is unbalanced, containing a majority of benign images, the training, validation, and testing subsets were balanced through down sampling to produce an equivalent number of images for each class. The final training, validation, and test sets contain 115, 58, and 75 images, from each class, respectively.

Parameters

Information on classifier related parameters can be found in the Keras documentation [27]. In VGG-16, stochastic gradient descent was used as the optimizing function with a learning rate of 10^{-6} and a momentum value of 0.9. An adaptive optimizer, RMSProp [35], was used for InceptionV3. The binary cross entropy loss function was selected as the loss function for both architectures. All layers in the fully connected classifier block, with the exception of the final output single node which was sigmoidal, were specified to have a ReLU activation function. In VGG-16, the second to last fully connected layer (256 nodes) was initialized with a normalized distribution of random values and included a dropout value of 0.5. The final classifier block in InceptionV3 is three layers deep; the first layer, a global average pooling layer, is followed by two fully connected layers (64 nodes and 1 node, respectively). To ensure consistency, the random number generator was seeded with a constant value for all methods.

Results

The model evaluation was performed using the created balanced testing dataset.

The main metrics used in the evaluation of this work are listed below:

- *Accuracy*, the number of correct predictions divided by the total number of samples.
- *Sensitivity*, the fraction of true positives that are correctly identified.
- *Precision*, the fraction of positives that are relevant.
- *AUC (Area Under the Curve)*, the area under an ROC curve plotting the true positive rate vs. the false positive rate.

Table 1 Test results for VGG-16

Dilation	Accuracy	Sensitivity	Precision	AUC
None	0.587	0.453	0.568	0.622
25	*0.613*	0.533	*0.598*	*0.642*
50	0.607	0.560	*0.598*	0.626
75	0.593	*0.573*	0.590	0.608
100	0.553	0.347	0.538	0.579
N/A	0.513	0.240	0.509	0.532

In VGG-16, all methods and inputs were trained for a consistent number of epochs—selected based on the behavior of the loss and accuracy reported during training. 60 epochs with a batch size of 26 were performed in VGG-16 and 25 epochs with a batch size of 32 were performed in InceptionV3.

Testing results are shown for different amounts of dilation ("none," 25, 50, 75, and 100-pixel radius, plus a baseline case—where no segmentation mask is used—denoted as "N/A") in Tables 1 and 2, with the best values highlighted (in italics). Sample results are shown for the unsegmented set in Fig. 3.

Discussion

We found that dilating the original binary masks to include a border of skin pixels surrounding the actual lesion consistently improved every classification metric—accuracy, sensitivity, precision, and AUC—when any amount of morphological dilation was applied. More interestingly, expanding the border of the segmented region at first improved and then deteriorated classification performance, suggesting that there is an ideal amount of contextual information that is beneficial to the classifier. For the given dataset, classifiers, and methodology, all metrics performed best when the raw image was segmented and subsequently dilated to include roughly 75 additional pixels in each direction.

Table 2 Test results for InceptionV3

Dilation	Accuracy	Sensitivity	Precision	AUC
None	0.573	0.667	0.590	0.643
25	0.627	0.667	0.638	0.696
50	0.613	0.680	0.631	0.700
75	*0.693*	*0.760*	*0.723*	*0.739*
100	0.653	0.733	0.683	0.738
N/A	0.633	0.613	0.628	0.680

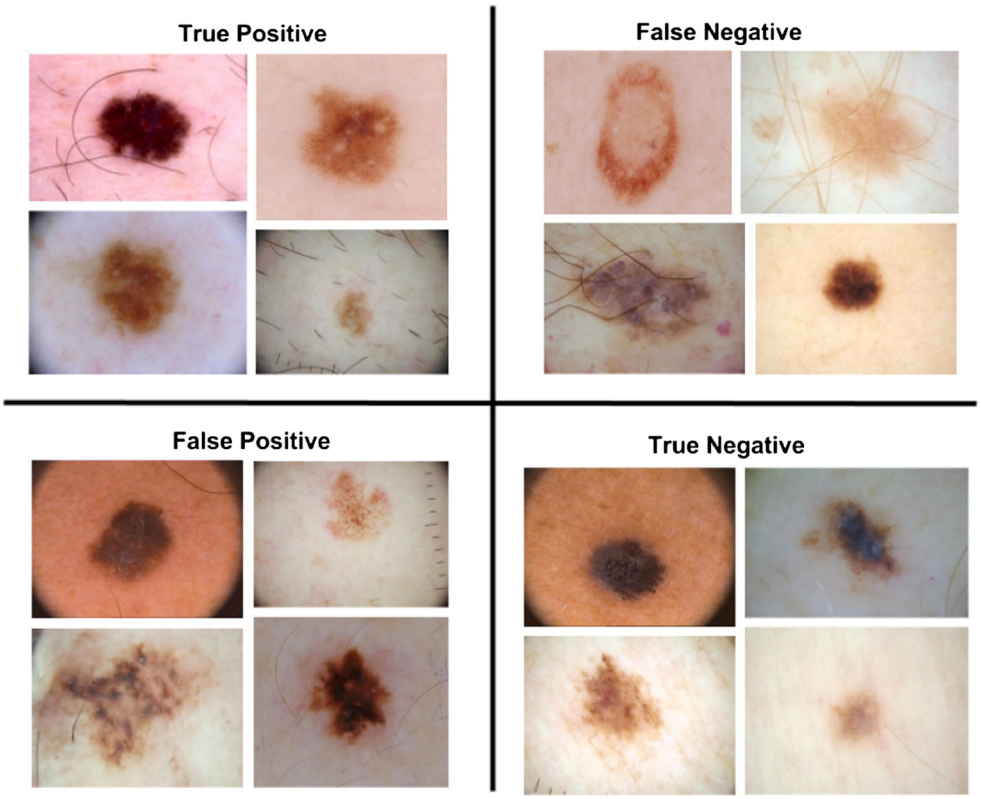

Fig. 3 Confusion matrix example images from unsegmented inputs classified on the VGG-16 classifier

The process of segmenting and subsequently expanding the segmentation border to include values beyond the target lesion has produced better results across all metrics of interest when using a deep learning based classifier built upon the transfer learning paradigm. The results, though collected on a small dataset, reveal a clear range within which the dilation of the segmentation border improves classification results. It should be noted that the balanced test split used for evaluation does not accurately represent the balance a classifier would face in a clinical setting, where the balance may be shifted considerably in favor of benign lesions.

Limitations

The main limitations of the proposed approach are as follows:

- Resizing the images for each network's specified input dimensions may adversely affect the classifier's performance. It is possible that fine structure information, such as texture, globules, vessels, etc., which may provide information relevant to classification, is lost when the images are downsized.
- Balancing the dataset may also adversely affect the classifier performance by decreasing the total number of samples available for training and validation.

Conclusions

Our work investigated how the inclusion of a non-lesion border around a lesion of interest impacts classification results when classifying skin lesions with a convolutional neural network.

Results indicate that the skin surrounding the lesion may provide contextually relevant information to a deep learning classifier that aids in skin lesion classification. Segmentation with enlarged, rather than dermatologist-like, masks achieves higher performance in image classification. There appears to be a "sweet spot," in which the degree to which the surrounding skin included is neither too great nor too small, providing a "just right" amount of context. Specifically, experimental results for the given dataset, classifier, setup, and methodology indicate that implementing a segmentation method that captures a border of roughly 75 pixels surrounding the lesion in all directions as a preprocessing step results in better classification performance in the indicated metrics of interest.

Potential avenues for future work include (i) using larger and/or different datasets, (ii) using a dilation size that is proportional to the target lesion rather than absolute, and (iii) investigating transfer learning's role in these results.

Acknowledgements The authors gratefully acknowledge funding from NSF Award No. 1464537, Industry/University Cooperative

Research Center, Phase II under NSF 13-542. We are also thankful to the 33 corporations that are the members of the Center for their active participation and funding.

References

1. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2016. CA: Cancer J Clin 66(1):7–30, 2016
2. Guy GP, Machlin SR, Ekwueme DU, Yabroff KR: Prevalence and costs of skin cancer treatment in the US, 2002- 2006 and 2007-2011. Am J Prev Med 48(2):183–187, 2015
3. Guy GP Jr, Ekwueme DU: Years of potential life lost and indirect costs of melanoma and non-melanoma skin cancer. Pharmacoeconomics 29(10):863–874, 2011
4. Nachbar F, Stolz W, Merkle T, Cognetta AB, Vogt T, Landthaler M, Bilek P, Braun Falco O, Plewig G: The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. J Am Acad Dermatol 30(4):551–559, 1994
5. Argenziano G, Fabbrocini G, Carli P, De Giorgi V, Sammarco E, Delfino M: Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Arch Dermatol 134(12):1563–1570, 1998
6. Henning JS, Dusza SW, Wang SQ, Marghoob AA, Rabinovitz HS, Polsky D, Kopf AW: The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy. J Am Acad Dermatol 56(1):45–52, 2007
7. Argenziano G, Soyer HP: Dermoscopy of pigmented skin lesions–a valuable tool for early diagnosis of melanoma. Lancet Oncol 2(7):443–449, 2001
8. Kittler H, Pehamberger H, Wolff K, Binder M: Diagnostic accuracy of dermoscopy. Lancet Oncol 3(3):159–165, 2002
9. Vestergaard M, Macaskill P, Holt P, Menzies S: Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. Brit J Dermatol 159(3):669–676, 2008
10. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(7639):115–118, 2017
11. Marques O: Morphological image processing. In: Practical image and video processing using MATLAB. Wiley - IEEE, chap. 13, 2011, pp 299–334
12. Oliveira RB, Mercedes Filho E, Ma Z, Papa JP, Pereira AS, Tavares JMR: Computational methods for the image segmentation of pigmented skin lesions: a review. Comput Methods Programs Biomed 131:127–141, 2016
13. Burdick J, Marques O, Romero Lopez A, Giró Nieto X, Weinthal J: The impact of segmentation on the accuracy and sensitivity of a melanoma classifier based on skin lesion images. In: SIIM (Society for Imaging Informatics in Medicine) 2017 annual meeting. Pittsburgh, 2017, pp 1–6
14. Codella N, Nguyen QB, Pankanti S, Gutman D, Helba B, Halpern A, Smith JR: Deep learning ensembles for melanoma recognition in dermoscopy images, arXiv:1610.04662, 2016
15. Kawahara J, BenTaieb A, Hamarneh G: Deep features to classify skin lesions. In: IEEE International symposium on biomedical imaging (IEEE ISBI). Prague, 2016, pp 1397–1400
16. Lopez AR, Giro-i Nieto X, Burdick J, Marques O: Skin lesion classification from dermoscopic images using deep learning techniques. In: 13th IASTED International conference on biomedical engineering (BioMed). IEEE, Innsbruck, 2017, pp 49–54
17. Yu L, Chen H, Dou Q, Qin J, Heng PA: Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging 36(4):994–1004, 2017
18. Jaworek Korjakowska J, Kłeczek P: Automatic classification of specific melanocytic lesions using artificial intelligence. BioMed Res Int, 2016, p 17
19. Simonyan K, Zisserman A Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014
20. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z: Rethinking theInception architecture for computer vision. arXiv:1512.00567, 2016, pp 2818–2826
21. Deng J, Dong W, Socher R, Li LJ, Li K, Fei Fei L: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on computer vision and pattern recognition. IEEE, MIami, 2009, pp 248–255
22. Applications - InceptionV3. [Online]. https://keras.io/applications/#inceptionv3, 2016. Accessed 01 Sept 2017
23. Yosinski J, Clune J, Bengio Y, Lipson H: How transferable are features in deep neural networks? In: Advances in neural information processing systems. Curran Associates, Inc., New York, 2014, pp 3320–3328
24. Krizhevsky A, Sutskever I, Hinton GE: ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems, vol 25. Curran Associates, Inc., New York, 2012, pp 1097–1105
25. Python. [Online]. https://www.python.org/, 2016. Accessed 01 Sept 2017
26. MathWorks: MATLAB. [Online]. https://www.mathworks.com/products/matlab.html, 2016. Accessed 01 Sept 2017
27. Keras documentation. [Online]. https://keras.io/, 2016. Accessed 01 Sept 2017
28. Theano 0.8.2. documentation. [Online]. http://deeplearning.net/software/theano/, 2016. Accessed 01 Sept 2017
29. TensorFlow. [Online]. https://www.tensorflow.org, 2017. Accessed 01 Sept 2017
30. scikit-learn: machine learning in python. [Online]. http://scikit-learn.org/, 2016. Accessed 2017-09-01
31. Python Imaging Library (PIL). [Online]. http://www.pythonware.com/products/pil/. Accessed 01 Sept 2017
32. CUDAm, Nvidia. [Online]. http://www.nvidia.com/object/cuda_home_new.html. Accessed 01 Sept 2017
33. ISIC Archive - International skin imaging collaboration: melanoma project. [Online]. https://isic-archive.com/, 2016. Accessed 01 Sept 2017
34. ISIC: ISBI 2016: Skin lesion analysis towards melanoma detection. [Online]. https://goo.gl/2A1913, Accessed 2016. 31 Aug 2017
35. tf.train.RMSPropOptimizer. [Online]. https://goo.gl/ULzaug, 2017. Accessed 01 Sept 2017