SIIM 2017 Scientific Session
Analytics & Deep Learning Part 3
Friday, June 2 | 1:15 pm – 2:45 pm

The Impact of Segmentation on the Accuracy and Sensitivity of a Melanoma Classifier Based on Skin Lesion Images

Jack Burdick, Florida Atlantic University; Oge Marques, PhD (Presenter); Adrià Romero López;
Xavier Giró-i-Nieto; Janet Weinthal

Hypothesis

The accuracy and sensitivity of a convolutional neural network binary classifier which detects melanoma in dermoscopic images improves with the use of segmented images (images which contain only the lesions and exclude the data surrounding the lesion).

Introduction

In the United States alone, there were an estimated 76,380 new cases of melanoma and an estimated 6,750 deaths due to melanoma in 2016 (Siegel, Miller, & Jemal, 2016). Early screening can increase life expectancy (Freedberg et al., 1999), but melanoma left undiagnosed can be fatal. Dermatologists use many heuristic classification methods to diagnose melanoma (Argenziano et al., 1998; Nachbar et al., 1994), but to limited success with only 65 - 80% accuracy (Argenziano & Soyer, 2001). A tool capable of aiding physicians to classify skin lesions could potentially save numerous lives each year.

Typically, segmentation is used as a preprocessing method in the classification process to remove potentially irrelevant information (outside the region of interest within an image) from the input images (Li et al., 2009). Segmentation eliminates the pixel values outside of the lesion, so that only the pixel values of the lesion are the only data considered when training and testing the network.

Previous studies have successfully used image segmentation to produce promising results (Jaworek-Korjakowska & Kleczek 2016), but other studies have produced promising results without segmentation (Kawahara, BenTaieb, & Hamarneh, 2016; Esteva et. al, 2017; Lopez, et al, 2017; Codella et. al, 2016). Segmenting the skin lesion removes ostensibly nonessential data surrounding the lesion, including hair and other lesions. This surrounding data eliminated by segmentation may obfuscate vital information, but may also provide greater context for the convolutional neural network (CNN). Better understanding the effects of image segmentation may help to refine the classification process. To investigate the effects of segmentation, we compare the performance metrics of CNNs trained on perfectly segmented images against the same metrics obtained using unsegmented images. We then further investigate the effects of segmentation by classifying partially segmented images, in which the perfectly segmented region is dilated to include surrounding non-lesion pixels.

Methods

The ISIC dataset is publicly available and contains 1279 images of skin lesions and corresponding hand-labeled binary masks of the lesion. Each image is labeled as either benign or malignant. The ISIC dataset is pre-partitioned into 900 training images and 379 testing images (International Skin Imaging Collaboration).

Rather than train a CNN model from scratch using our small dataset, a transfer learning approach, which extracts layers from a neural network previously trained on a different, larger dataset (Yosinski et al., 2014), was implemented. VGGNet (Simonyan & Zisserman, 2014), a CNN trained on the ImageNet dataset, which contains over 14 million natural images (Russakovsky et al., 2015), was selected. The VGGNet became popular after achieving excellent results in the ILSVRC-2014 competition (Russakovsky et al., 2015). Specifically, the VGG-16 architecture (Figure 1), which has fewer parameters than the other VGG ConvNet configurations, was selected because it has been shown to generalize well to other datasets (Simonyan & Zisserman, 2014). Our VGG-16 architecture consists of the same five convolutional blocks and a slightly modified final fully connected block consisting of one 256, relu activated, node layer and lastly a single sigmoidal node. We froze the first four convolutional blocks, initialized the fully connected layers with uniform distribution (a process known as Uniform Initialization), and then trained the CNN with our images from the ISIC dataset.

Figure 1

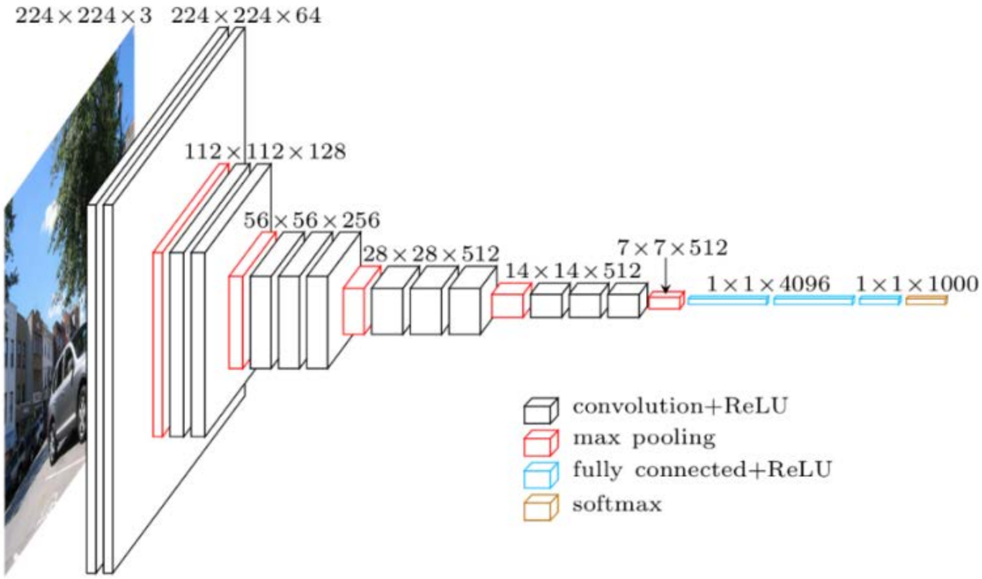

$224 \times 224 \times 3$
$224 \times 224 \times 64$
$112 \times 112 \times 128$
$56 \times 56 \times 256$
$28 \times 28 \times 512$
$14 \times 14 \times 512$
$7 \times 7 \times 512$
$1 \times 1 \times 4096$
$1 \times 1 \times 1000$
convolution+ReLU
max pooling
fully connected+ReLU
softmax

Figure 1. Original VGG16 architecture (Vachet 2016)

To equalize the number of images in each class and balance the dataset, the training and test sets were reduced through downsampling. The dataset was divided into a 70-30% training/testing split, with 230, 116, and 150 images in the training, validation, and test sets respectively. Input images for the three cases - full segmentation, partial segmentation, and no segmentation - are created from the dataset. The unsegmented images were used in their unaltered form from the dataset. The perfectly segmented images were generated by performing a bitwise AND operation on the unaltered images and its corresponding binary mask provided by the ISIC dataset as a reference image. The resulting image is a perfectly segmented skin lesion where the values outside the lesion are converted to a zero value. A similar method was used to create the partially segmented images, except the original binary masks were first morphologically dilated with a disk-shaped structuring element (50 pixel radius). Additional preprocessing methods including resizing and normalization were performed to match the input size expected by the VGG16 architecture. The images were resized to 224×224 pixels and normalized by performing a channel-wise rescaling to [-1, 1] (by subtracting the mean and dividing by the result of the max minus minimum values). Figure 2 shows the the three input sets and their corresponding segmentation masks.

Figure 2

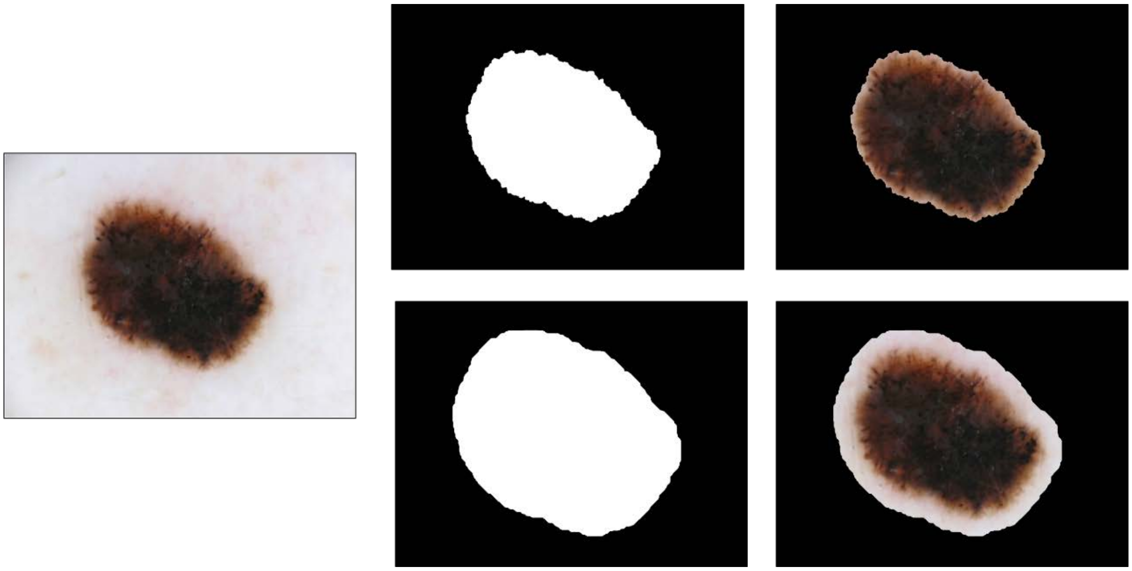


Figure 2. Representative sample of the three input datasets (ISIC Archive). The unaltered image (left), and the segmented images (right) with corresponding mask (center).

Results

To ensure the architecture produces the same results for each epoch, the random number generator was seeded with an arbitrary value that remained constant throughout testing. The input images were the only parameter that changed between sets.

Stochastic gradient descent (learning rate=1e-6, momentum=0.9) and binary cross entropy (Keras Objectives) were implemented as the optimizing and loss functions. The fully connected layers included a 0.5 value for the dropout optimization. A batch size of 26 was used and 60 epochs were performed.

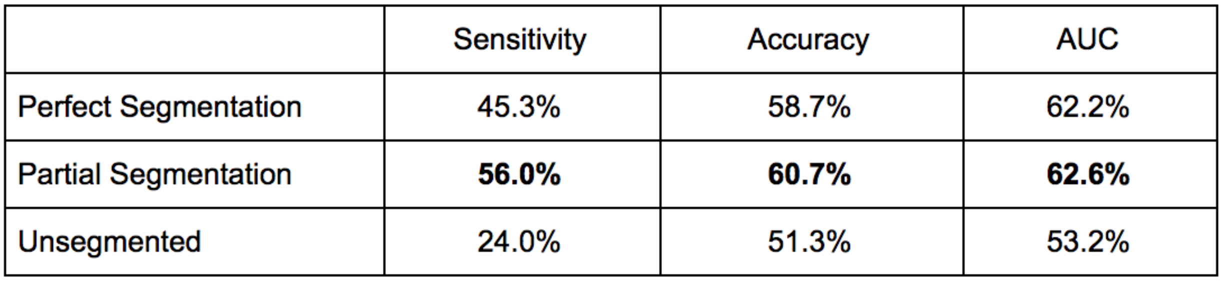
Balancing through downsampling was performed by removing images from the training set by selectively obtaining the first 173 images from the directory of each class (alphabetically listed), for a total of 346 images. In the test dataset, images from the majority class were removed randomly to produce 75 benign and 75 malignant images.

The model performance evaluation was performed using the created balanced testing dataset with the metrics listed below:

- Accuracy: the number of correct predictions divided by the total number of predictions made.
- Sensitivity: the fraction of true positives that are correctly identified.
- AUC (Area Under the Curve): the area under an ROC (Receiver Operating Characteristic) curve plotting the true positive rate vs. the false positive rate.

Examining the results for the classification of unsegmented and segmented images revealed improved metrics for all cases when perfect segmentation was applied; an accuracy of 51.3% and 58.7%, sensitivity of 24.0% and 45.3%, and AUC of 53.2% and 62.2%, respectively. When the partially segmented results are compared these results, all metrics further improve. The Accuracy increased to 60.7%, sensitivity to 56.0%, and AUC to 62.6% (Figure 3). The corresponding confusion matrices for all three cases are shown in Figure 4. Figure 5 includes sample images corresponding to the confusion matrix for the unsegmented case.

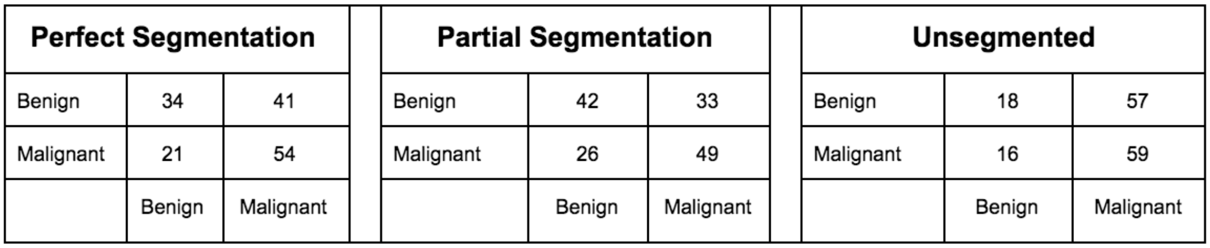
Figure 3



	Sensitivity	Accuracy	AUC
Perfect Segmentation	45.3%	58.7%	62.2%
Partial Segmentation	**56.0%**	**60.7%**	**62.6%**
Unsegmented	24.0%	51.3%	53.2%

Figure 3. Accuracy, sensitivity, and AUC of lesion classification for unsegmented, partially segmented, and perfectly segmented datasets.

Figure 4



Perfect Segmentation		
Benign	34	41
Malignant	21	54
	Benign	Malignant

Partial Segmentation		
Benign	42	33
Malignant	26	49
	Benign	Malignant

Unsegmented		
Benign	18	57
Malignant	16	59
	Benign	Malignant

Figure 4. Confusion matrices of lesion classification for unsegmented, partially segmented, and perfectly segmented datasets

Figure 5

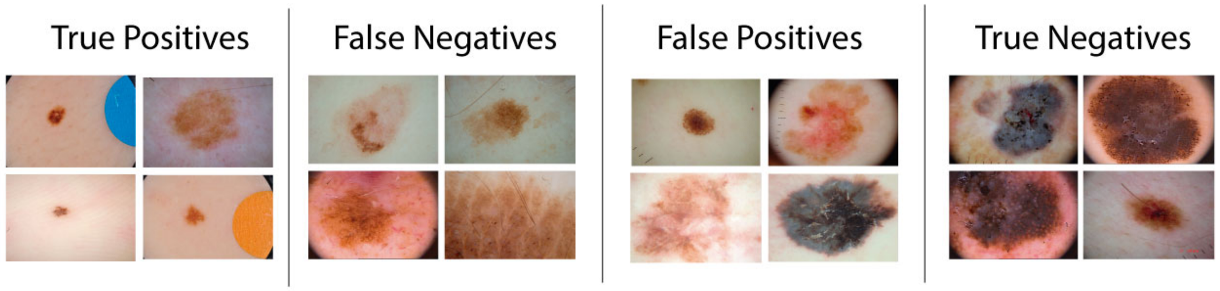


Figure 5. Sample classification results for correct and incorrectly predicted unaltered images.

Discussion

Skin lesion images often contain background noise that may be detrimental to image classification. Sample images including a distracting background are shown in figure 6. Ostensibly, the background values serve as distractions that negatively affect lesion classification and must be segmented.

Figure 6

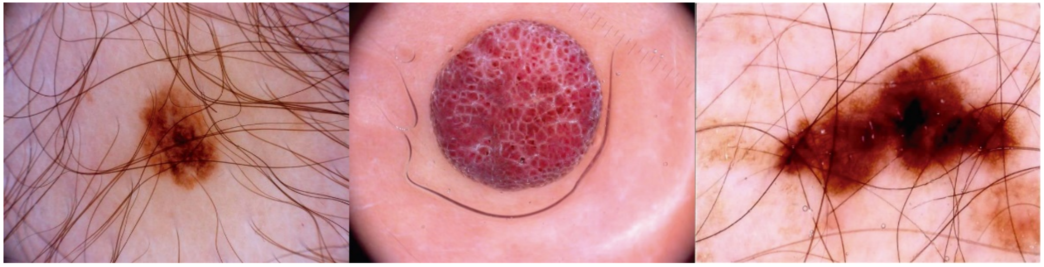


Figure 6. Classification difficulties caused by background noise.

Accuracy, sensitivity, and AUC all improved when perfect segmentation was applied to the images from the ISIC dataset. Of the three performance metrics, sensitivity was the most improved, increasing from 24.0% to 45.3%. In medical applications, sensitivity is often considered to be one of the most important metrics (Lalkhen and McCluskey, 2008). Sensitivity measures the efficacy of the CNN in identifying melanoma. A diagnostic test should be optimized to correctly recognize as many cases of melanoma as possible, even at the cost of false positives (Lalkhen and McCluskey, 2008). Ultimately, misidentifying benign lesions as melanomas will cost patients their time, but misidentifying melanomas as benign lesions will cost patients their lives.

Dilation of the segmentation mask beyond the border of the lesion further improves performance metrics. This improvement in performance suggests that despite common classification pipelines, entirely isolating the lesion within the image may be less effective than segmentation and subsequent dilation to include additional surrounding values. Future work should develop effective segmentation methods and explore the effects of the inclusion of data surrounding the lesion during classification.

Conclusion

We investigated the hypothesis that segmentation improves classification performance of skin lesions using convolutional neural networks. Examination of classification using different degrees of segmentation on images from the ISIC Archive dataset with the VGG-16 architecture revealed unexpected results. Early results indicate that the highest classification performance occurs when a pre-classification segmentation and an unconventional subsequent dilation of the segmentation region is performed.

Future work could expand on these findings and further investigate the degree to which dilating a segmentation mask to include surrounding data improves classification performance. Our results suggest that classification performance would increase, reach an optimal value, and then decrease, as additional contextual values are included in the dataset. This apparent trend should be examined across multiple datasets and classifiers. Furthermore, the effects of transfer learning on this trend should be investigated by comparing the results of architectures that do not use transfer learning.

References

1. Argenziano G, Soyer HP: Dermoscopy of pigmented skin lesions–a valuable tool for early diagnosis of melanoma. The Lancet Oncology 2(7):443-9, 2001.
2. 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. Archives of Dermatology 134(12):1563-70, 1998.
3. Codella N, Nguyen, QB Pankanti S., Gutman D, Helba B, Halpern A, Smith JR: Deep learning ensembles for melanoma recognition in dermoscopy images. arXiv preprint arXiv:1610.04662, 2016.
4. 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): 115-8, 2017.
5. Freedberg KA, Geller AC, Miller DR, Lew RA, Koh HK: Screening for malignant melanoma: a cost-effectiveness analysis. Journal of the American Academy of Dermatology 41(5):738-45, 1999.
6. International Skin Imaging Collaboration: Melanoma Project Website. Available: https://isic-archive.com/
7. Jaworek-Korjakowska J, Kleczek P: Automatic classification of specific melanocytic lesions using artificial intelligence. BioMed Research International 2016, 2016.
8. Kawahara J, BenTaieb A., Hamarneh G.: Deep features to classify skin lesions. IEEE International Symposium on Biomedical Imaging (IEEE ISBI), 1397-1400, 2016.
9. Keras Objectives. Website. Available: https://keras.io/objectives/

10. Lalkhen AG, McCluskey A: Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia, Critical Care & Pain 8(6): 221-223, 2008.
11. Li X, Aldridge B, Ballerini L, Fisher R, Rees J: Depth data improves skin lesion segmentation. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2009, 1100-11007, 2009.
12. Lopez AR, Giro-i-Nieto X, Burdick J, Marques O: Skin lesion classification from dermoscopic images using deep learning techniques. In Biomedical Engineering (BioMed), 2017 13th IASTED International Conference on, 49-54, 2017.
13. 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. Journal of the American Academy of Dermatology. 30(4):551-9, 1994 .
14. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC: ImageNet large scale visual recognition challenge. International Journal of Computer Vision 115(3):211-52, 2015.
15. Simonyan K, Zisserman A: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
16. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2016. CA: A Cancer Journal for Clinicians 66(1):7-30, 2016.
17. Uniform initialization. Keras Framework. Website. Available: https://keras.io/initializations/
18. Vachet A: A Brief Report of the Heuritech Deep Learning Meetup #5. Heuritech - Le Blog. Website. Available: https://blog.heuritech.com/2016/02/29/a-brief-report-of-the-heuritech-deep-learning-meetup-5/ N, 2016.
19. Yosinski J, Clune J, Bengio Y, Lipson H: How transferable are features in deep neural networks? Advances in Neural Information Processing Systems, 3320-3328, 2014.

Keywords

skin lesion, segmentation, medical image analysis, deep learning, machine learning, medical decision support, convolutional neural networks, transfer learning