Skin Lesion Analysis Towards Melanoma
Detection via End-to-end Deep Learning of
Convolutional Neural Networks

Katherine M. Li and Evelyn C. Li

Abstract. This article presents the design, experiments and results of our solution submitted to the 2018 ISIC challenge: Skin Lesion Analysis
Towards Melanoma Detection. We design a pipeline using state-of-the-art Convolutional Neural Network (CNN) models for a Lesion Boundary
Segmentation task and a Lesion Diagnosis task.

Keywords: Convolutional Neural Network, Medical Image Analysis,
Skin Lesion Analysis, Melanoma Detection

Recently, state-of-the-art research in Artificial Intelligence (AI) has rapidly transitioned from research laboratories into production [1]. This development is
driven by a number of forces in the industry: 1) widespread availability of a

driven by a number of forces in the industry: 1) widespread availability of a large amount of data, in the form of images (e.g., ImageNet), audio, video, text and user logs; 2) significant increase in affordable computation power (e.g., NVidia GeForce GTX 1080Ti cards, which offer 11.3 TFLOPs of computation, have 11GB GDDR5X memory at 484GB/sec memory bandwidth, and cost only \$699); and 3) accessiliby of state-of-the-art machine learning platforms. A unique aspect of the recent deep learning revolution is that all major deep learning platforms, whether TensorFlow [2], PyTorch [3], Caffe [4], CNTK [5], or MxNet [6], are all built as open-source software. This enables rapid cross-pollination of ideas among researchers and engineers. It also allows big organization and individuals alike to use state-of-the-art machine learning tools without incurring (traditionally significant) license fees, or the need to develop technology from scratch.

The rapid development of AI systems has the opportunity to significantly impact our lives, aiding humans in making mission-critical or life-and-death decisions. We believe that the task of skin lesion analysis toward melanoma detection organized by the International Skin Image Collaboration (ISIC) is one such image classification task that will benefit from the use of state-of-the-art machine learning technology. As discussed in [7], skin cancer is the most prevalent form of cancer in the United States. Melanoma, the most dangerous form of skin cancer, leads to over 9,000 deaths a year. Most melanomas are first identified visually, but unaided visual inspection only has a diagnostic accuracy of roughly 60% [8]. Recently, medical professionals have used dermoscopy, a new technology of visual inspection, that both magnifies the skin and eliminates surface reflection.

With proper training, it is shown that a human expert can achieve a diagnostic accurcy of 75%-84% [8,9]. Moreover, a growing shortage in dermatologists per capita [10] leads to interest in using AI techniques for automated assessment of dermoscopic images.

The rest of the paper will discuss our submission for the ISIC 2018 challenge on lesion segmentation and disease classification. We will briefly discuss our observation of the dataset and task in Section. 2. The methodology and results will be discussed in Section. 3.

2 Dataset and Task

The HAM10000 Dataset [11] is used in our training pipeline. For task 1, Lesion Boundary Segmentation, we have 2594 images and corresponding ground truth response masks. For task 3, Lesion Diagnosis, we have 10015 images used for training and validation. There are 7 classes in the HAM10000 Dataset, including 1) melanoma, 2) Melanocytic nevus, 3) basal cell carcinoma, 4) Actinic keratosis/Bowens disease (intraepithelial carcinoma), 5) benign keratosis (solar lentigo/seborrheic keratosis/lichen planus-like keratosis), 6) dermatofibroma and 7) vascular lesion. The number of images in each category is extremely imbalanced. For example, the largest category, Melanocytic nevus, has 6705 images while the smallest category, dermatofibroma, only has 115 images.

3 Method, Experiment Design and Results

When training human dermoscopic students, procedural algorithms, such as the "3-point checklist", the "ABCDE rule", and the "7-point checklist" were developed [12]. For example, the "ABCDE" rule suggests checking for suspicious moles using the following signs: 1) Asymmetry (one half of the mole doesn't match the other), 2) Border irregularity, 3) Color that is not uniform, 4) Diameter greater than 6 mm (about the size of a pencil eraser), and 5) Evolving size, shape or color. There has been debate in the medical community about whether this procedural guideline or relying on personal analysis is more beneficial. Similarly, when designing an AI system for dermoscopic image analysis, there can also be two different approaches. First, there is the feature based approach, in which multiple sub-AI systems are developed, each of which will be trained to identify one feature (e.g., one AI sub-system each will be trained to measure Asymmetry, Border irregularity, Color uniformity, Diameter and Evolving size), and then an aggregated AI system will be used to combine the individual features and make disease diagnostics. An alternative is the end-to-end approach, in which an AI system, in particular a deep convolutional neural network (CNN) will be trained directly from the image datasets. We apply the end-to-end training approach, mainly because the current dataset lacks individual features (such as the "ABCDE" feature discussed above), and as a result, training multiple individual feature regression/classification subsystems may lead to added development

cost. If the ISIC organizer later providers metadata with individual features, especially features related to the $\underline{\mathbf{E}}$ volving size, shape, or color and the $\underline{\mathbf{D}}$ iameter, both of which cannot be reliably estimated from the image by the end-to-end CNN approach, we will be interested to investigate if the use of those individual features can improve disease detection accuracy.

In the ISIC 2018 challenge, it is required to only use images for lesion boundary segmentation and lesion diagnosis tasks. In Sec. 3.1, we describe the method we used in Task 1 and in Sec. 3.2, we describe the method we used in Task 3.

3.1 Task 1: Lesion Boundary Segmentation

We start this task by using the state-of-the-art object segmentation model Mask RCNN [13]. We use ResNet50 as backbone network. The pre-trained weights on ImageNet dataset are used to initialize the backbone network. In this task, we do not separate the 7 categories in the training set. Instead, we treat all the lesion regions as one class, and add a background class during the training. We train the model on a single NVIDIA Titan X GPU. The network proposes 200 ROI candidates per image and classifies if each ROI contains a lesion region. We use a detection confidence threshold of 0.9 in our experiment to eliminate most of the false positive detections. We also randomly select 20 images from the training set as validation set. Furthermore, we use the Jaccard index metric to measure our model. The model was trained for 40 epochs and we get 0.783 in our validation set.

We notice that the initialize backbone network weights was trained on the ImageNet dataset, which has 1000 classes of common visual objects. The ImageNet dataset is widely used in many computer vision applications, but its visual domain is quite different from the lesion image dataset. The first few layers of a CNN can be considered as feature extractor, and the ResNet50 model was trained on the ImageNet dataset, so the convolutional filters of the ResNet50 are tuned to extract features from the common objects that appeared in the ImageNet dataset. This causes the ResNet50 to sometimes be unable to locate the lesion region in the HAM10000 dataset. Thus, we create an improved solution: we first train the ResNet50 model using the data from the Lesion Diagnosis task, and after the model is well tuned and achieves a reasonable overall classification accuracy, we use the weights from this model to initialize the backbone network in our Mask RCNN model. With this method, we improve our validation performance to 0.818.

3.2 Task 3: Lesion Diagnosis

We start our baseline experiment by training the multi-class classification model on a variety of CNN models, including ResNet, DenseNet and Inception. The dataset is split into 80% training and 20% validation set. We use a similar setup in each model. A global average pooling layer is added after the base network and followed by a fully connected layer for prediction. We use softmax as the activation function in the prediction layer and multi-class cross entropy as the

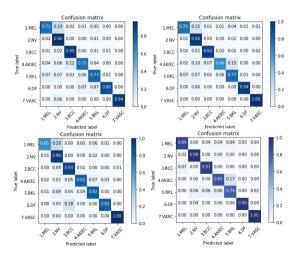


Fig. 1. Confusion matrix from DenseNet201, ResNet152, Inception_V4 and DenseNet201 with cropped images

loss function. Each model is trained for 300 epochs, and the loss and validation loss are converged after around 240 epochs. As we mentioned in Section 2, the classes of the HAM10000 dataset are extremely imbalanced. We use class weights to balance the dataset. The higher weights will be given to the samples in the class with small size and the lower weights will be given to the samples in the class with large size. We test our models in the validation set and the confusion matrix of each model is shown in Fig. 1.

Training on Cropped Lesion Images We notice that we may improve the classification performance in the future by removing the background from the lesion images and letting the classifier only focus on the lesion region. Based on the segmentation model we get from Task 1, we can first perform lesion boundary segmentation on the dataset. The lesion region is cropped based on the segmentation results and used as new training/validation data. We achieve about 2% improvement on normalized multi-class accuracy by using cropped images for training and validation. It is important to point out that this improvement highly depends on the performance of the lesion boundary segmentation model. The classification results may suffer if the performance of the segmentation model is not good enough.

4 Conclusions

This article summarized the methods we used in the ISIC 2018 challenge. We proposed several improvements to address the problem of Lesion Boundary Segmentation and Lesion Diagnosis.

References

- Stoica, I., Song, D., Popa, R.A., Patterson, D.A., Mahoney, M.W., Katz, R.H., Joseph, A.D., Jordan, M., Hellerstein, J.M., Gonzalez, J., Goldberg, K., Ghodsi, A., Culler, D.E., Abbeel, P.: A berkeley view of systems challenges for ai. Technical Report UCB/EECS-2017-159, EECS Department, University of California, Berkeley (Oct 2017)
- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X.: Tensorflow: A system for large-scale machine learning. In: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16). (2016) 265–283
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. (2017)
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22Nd ACM International Conference on Multimedia. MM '14. New York, NY, USA, ACM (2014) 675–678
- Yu, D., Eversole, A., Seltzer, M., Yao, K., Guenter, B., Kuchaiev, O., Seide, F., Wang, H., Droppo, J., Huang, Z., Zweig, G., Rossbach, C., Currey, J.: An introduction to computational networks and the computational network toolkit. Microsoft Technical Report (2014)
- Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., Zhang, Z.: Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. CoRR abs/1512.01274 (2015)
- Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N.K., Kittler, H., Halpern, A.: 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). CoRR abs/1710.05006 (2017)
- Kittler, H., Pehamberger, H., Wolff, K., Binder, M.: Diagnostic accuracy of dermoscopy. The lancet oncology 3(3) (2002) 159–165
- 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. British Journal of Dermatology 159(3) (2008) 669–676
- Kimball, A.B., Resneck Jr, J.S.: The us dermatology workforce: a specialty remains in shortage. Journal of the American Academy of Dermatology 59(5) (2008) 741– 745
- 11. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. arXiv preprint arXiv:1803.10417 (2018)
- Argenziano, G., Soyer, H.P., Chimenti, S., Talamini, R., Corona, R., Sera, F., Binder, M., Cerroni, L., De Rosa, G., Ferrara, G., et al.: Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet. Journal of the American Academy of Dermatology 48(5) (2003) 679–693
- 13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Computer Vision (ICCV), 2017 IEEE International Conference on, IEEE (2017) 2980–2988