Focal Rank Loss Function With Encoder-decoder Network For Skin Lesion Segmentation

Lingdu Wang¹, Lei Zhang²*, and Xin Shu³

¹College of Computer Science, Sichuan university, Chengdu, Sichuan, 610065, China
²College of Computer Science, Sichuan university, Chengdu, Sichuan, 610065, China
³College of Computer Science, Sichuan university, Chengdu, Sichuan, 610065, China
*Corresponding author’s e-mail: leizhang@scu.edu.cn

Abstract. Skin cancer is a major health problem as melanoma is one of the deadliest types of skin cancer and causes thousands of deaths worldwide each year. Automatic segmentation of skin lesions is considered an essential step in computer-aided diagnosis (CAD) systems for melanoma detection, which help the specialists better examine pigmented skin lesions. However, the segmentation is a challenging task due to the low contrast between each lesion and its surrounding skin tissue which makes the boundary pixels hard to classify. Besides, the class imbalance problem exists in skin lesion segmentation problem. Thus, a new hybrid loss function based on focal loss and rank loss is proposed to alleviate the class imbalance problem and the hard-easy classified pixels problem in skin lesion segmentation problem. The proposed method is evaluated on the ISIC 2017 dataset using the proposed encoder-decoder network, and the result shows the loss outperforms most methods in terms of the dice coefficient and especially the sensitivity.

1. Introduction
Skin cancer is one of the most prevalent cancer worldwide, and melanoma is the deadliest form of skin cancer. It is estimated to have 106,110 new cases and 7180 deaths from melanoma in 2021 in the United States (US) alone [1]. Early diagnosis is rather important in the case of skin cancer as shown by the study, where the survival rate was as high as 90% for melanoma with early detection [2].

Automatically segmenting melanoma is an essential step which detects the locations and boundaries of lesion in computerized analysis of dermoscopic images. One of the difficulties of segmentation tasks can be summarized as the low contrast between each lesion and its surrounding skin tissue results in irregular and fuzzy lesion boundaries and it would make the boundary pixels hard to classify. It is obviously important to correctly classify more pixels for better segmentation performance. Thus, For the problem of boundary pixels (e.g., hard pixels) being difficult to classify which caused by irregular borders and blur, this paper [3] combines the dice loss [7] and rank loss to tackle it, which enables the network to pay more attention to those hard pixels and thus learn more discriminative information. Besides, the problem of class imbalance between foreground and background pixels exists in skin lesion segmentation. This inefficiency is a classic problem that is typically addressed via techniques such as bootstrapping [15] or hard example mining [16]. And Tsung-Yi Lin et al. [6] proposed the focal loss function as a more effective alternative to previous approaches for dealing with class imbalance which can automatically down-weight the contribution of easy examples during training and rapidly focus the model on hard examples.
The contributions of this work are two-fold. First, a new hybrid loss function based on the focal loss and the rank loss is proposed to address the issue of class imbalance and hard-easy classified pixels imbalance problem. The proposed loss makes the model pay more attention to the foreground pixels and enforces the predicted probability of boundary pixels closer to its ground truth. Second, we develop a U-Net [5] implementation which is the most popular network for biomedical image segmentation based on the concept of Fully Convolutional Networks (FCN) [4] re-using layers of ResNet50 [8] network as the encoder and change the skip connection form between the encoder and decoder to better concatenate the feature maps of different encoder layers to corresponding decoder layers to combine the global and local information.

2. Methodology

2.1. Network architecture

The network contains two-fold architecture: encoder and decoder. The encoder is a contracting path to extract different level features and the decoder is an expansive path to restore the image with contextual information extracted. And skip connections are employed to combine high-resolution local information with low-resolution global information to get better performance. In this study, we let the encoder reuses the main layers of ResNet50 with the average pooling layer and the fully connected layer removed. The first layer is a layer with a 7 × 7 (stride = 2) convolution kernel followed by max pooling (stride = 2) that generates 64 feature maps. Following four layers are used to extract 256, 512, 1024, 2048 feature maps with the resolution reduced by half each layer. Thus, the encoder output feature map is 1/32 of the input image. For the decoder, we perform upsampling by nearest interpolation with a factor of 2 and convolutional layer to reconstruct segmentation mask from the encoder output. Figure 1 shows the architecture of the network we used.

2.2. Loss Function

Many deep learning methods for segmentation usually employ the cross-entropy loss or dice loss as a loss function to optimize the segmentation network. The skin lesion mostly is a small part of the image which means the dataset has the class imbalance problem between foreground and background pixels. Thus, the cross-entropy loss tends to be dominated by the background part which means easily classified negatives comprise the majority of the loss and dominate the gradient. The balanced cross-entropy loss introduces a weighting factor to balance the importance of positive and negative pixels. However, it cannot differentiate between easy and hard pixels so the focal loss reshapes the loss function to down-weight easy pixels which makes the training concentrate on the hard pixels. The focal loss introduces a modulating factor \((1 - p)^\gamma\) to the balanced cross-entropy loss and the form is defined as equation (1).
\[ y \in \{\pm 1\} \text{ specifies the label of the pixel and } p \text{ is the probability predicted by the model for the class with label } y = 1. \]

\[ L_{\text{focal}}(p, y) = \begin{cases} -(1 - p) \alpha \log(p) & \text{if } y = 1 \\ -(1 - \alpha)p \log(1 - p) & \text{otherwise} \end{cases} \]  

(1)

Besides, it is generally acknowledged that the pixels that can be easily classified which means non-boundary pixels located in internal lesion and background, contribute little to the gradient optimization. The focal loss makes the loss smaller when the pixel’s predicted probability is closer to its label which means the hard pixels would get relatively larger loss and can be more important at optimizing process. Specially, to tackle this issue more precisely and efficiently, rank loss ranks the pixels of skin lesion and background respectively by the error after the propagation of each batch. The top k pixels that have the largest error in lesion or background are chosen to be the hardest pixels in its own area. Define \( p_{ni} \) and \( p_{nj} \) as the prediction values of the \( i \)th hard pixel of background and the \( j \)th hard pixel of skin lesion for the \( n \)th input image. The equation (2) describes the form of rank loss. It enables the model to pay more attention to the chosen hardest pixels because it would enforce \( p_{nj} > p_{ni} + \text{margin} \) in the training process.

\[ L_{\text{rank}}(X_n, Y_n) = \frac{1}{k^2} \sum_{i=1}^{k} \sum_{j=1}^{k} \max \{0, p_{ni}(X_n, Y_n) - p_{nj}(X_n, Y_n) + \text{margin} \} \]  

(2)

Thus, to alleviate the class imbalance problem between skin lesion pixels and the background pixels and the hard classified boundary pixels problem, we combine the focal loss and the rank loss as the focal rank loss which introduces a weighting factor \( \lambda \) to control the contribution of \( L_{\text{rank}} \).

\[ L = L_{\text{focal}} + \lambda L_{\text{rank}} \]  

(3)

3. Experiments

3.1. Dataset and Data preprocessing

The 2017 International Skin Imaging Collaboration (ISIC) skin lesion segmentation challenge dataset [8] is used for this study. The dataset contains 2000 training, 150 validation, 600 test dermoscopic images and corresponding lesion masks annotated by an expert clinician. The images in ISIC-2017 dataset have various resolutions ranging from 540 × 722 to 4499 × 6748 pixels. For this dataset, we use a simple pre-processing procedure to alleviate overfitting of the system and improve the robustness of the learning process. We first resize all the images to 256 × 256 using bicubic interpolation to reduce the computational cost. And we augmented the dataset by using the combination of multiple image processing steps, e.g., random scale, rotation, shifts and flips as the image augmentation is very important for building a powerful model with limited data.

3.2. Evaluation metrics

We employ five evaluation metrics, including pixel-wise accuracy (ACC), sensitivity (SE) and specificity (SP), jaccard index (JA), dice coefficient (DI). All the evaluation metrics mentioned above are computed in relation with the elements of the confusion matrix as follows. The TP, FP, FN, and TN symbolize the true positive, false positive, false negative and true negative, respectively.

\[ ACC = \frac{tp + tn}{tp + fn + tn + fp} \]  

(4)

\[ SE = \frac{tp}{tp + fn} \]  

(5)

\[ SP = \frac{tn}{tn + fp} \]  

(6)

\[ JA = \frac{tp}{tp + fn + fp} \]  

(7)
\[ DI = \frac{2 * tp}{2 * tp + fp + fn} \] 

\hspace{1cm} (8)\\

3.3. Implementation

All training and inference were performed on an Ubuntu system with NVIDIA Tesla K40 GPUs and Inter Xeon E5-1650 CPUs. The models were constructed on the Pytorch deep learning platform. We use the Adam as an optimizer with an initial learning rate set to \(3e^{-4}\). And an exponential LR learning rate schedule with the power set to 0.9 was implemented. All the models are trained for 200 epochs with the batchsize set to 32. For transfer learning, we used the pre-training model weights by the Resnet50 layers learning from the ImageNet dataset. We choose the hyper-parameters in our loss as \(\gamma = 1.0, \alpha = 0.8, \lambda = 0.002, k = 30, \text{margin} = 0.25\) after experiments.

3.4. Results

We first compared the proposed method to several published methods recently in Table 1. From the table, one can see that our method achieves the best performance on the sensitivity and dice coefficient. It shows that some methods like [9] and [10] have a rather high specificity with a not very good performance for dice coefficient. [10] has a higher specificity and a lower sensitivity than [9] and finally a lower dice coefficient which indicates that the sensitivity is more important in skin lesion segmentation problem. Thus, the proposed loss pays more attention to the foreground pixels and gets a 0.900 of sensitivity and surpass the second-best method by 2.6%, which indicates the proposed method obtains a better ability to segment the pixels of lesion areas.

Table 1. Comparison of skin lesion segmentation results on ISIC 2017 test dataset.

| Method          | ACC  | SE   | SP   | JA   | DI   |
|-----------------|------|------|------|------|------|
| Ours            | 0.942| 0.900| 0.958| 0.804| 0.879|
| Berseth et al. [9] | 0.932| 0.820| 0.978| 0.762| 0.847|
| Bi et al. [10]  | 0.934| 0.802| 0.985| 0.760| 0.844|
| SLSDeep [11]    | 0.936| 0.816| 0.983| 0.782| 0.878|
| RECOD [12]      | 0.931| 0.817| 0.970| 0.754| 0.839|
| MB-DCNN [3]     | **0.947**| 0.874| 0.968| **0.804**| 0.878|
| Yuan et al. [13]| 0.934| 0.825| 0.975| 0.765| 0.849|
| FRCN [14]       | 0.940| 0.854| 0.967| 0.771| 0.871|

Besides, to present a fair evaluation of the proposed loss function, we compared it with several common loss functions trained on the same network. Table 2 indicates that the proposed loss function can get better performance than these loss functions, especially in sensitivity. The high score of sensitivity shows that the proposed loss does get the model more focused on foreground pixels compared to other losses without the network effect.

Table 2. Comparison of the performance between the proposed loss and some other loss functions with the proposed network on ISIC 2017 test dataset.

| Loss Function | ACC  | SE   | SP   | JA   | DI   |
|---------------|------|------|------|------|------|
| Our loss      | **0.942**| **0.900**| 0.958| **0.804**| 0.879|
| CE loss       | 0.938| 0.849| 0.973| 0.787| 0.867|
| Dice loss     | 0.938| 0.836| **0.974**| 0.784| 0.863|
| Focal loss    | **0.942**| 0.886| 0.962| 0.798| 0.873|
| Dice Rank Loss| 0.937| 0.865| 0.969| 0.789| 0.867|
The result in Table 3 shows that the pure network has a better performance than other basic networks without the proposed loss function. It indicates that the network has a good ability to extract useful features and to combine the local information and global information.

Table 3. Segmentation performance of the used network with ce loss compared to FCN, U-Net, and SegNet for ISIC 2017 test dataset.

| Model  | ACC  | SE   | SP   | JA   | DI   |
|--------|------|------|------|------|------|
| Ours   | 0.938| 0.849| 0.973| 0.787| 0.867|
| FCN    | 0.927| 0.800| 0.967| 0.722| 0.839|
| U-Net  | 0.900| 0.672| 0.972| 0.616| 0.763|
| SegNet | 0.918| 0.801| 0.954| 0.696| 0.821|

4. Conclusions

The proposed focal rank loss function is proved to be useful to alleviate the class imbalance problem between lesion pixels and the background pixels and the hard-easy classified pixels imbalance problem. It gets a good performance on sensitivity and dice coefficient which means this loss does pay more attention to the relatively important foreground part. The reason is that the loss enforces the predicted probability to be closer to its ground truth especially for the foreground pixels so the experiment achieved a rather high sensitivity. Besides, a deep learning skin lesion segmentation model based on the encoder-decoder network architecture is used in this study. The encoder employs the ResNet layers to extract the features of the input when downsampling, in turn decoder layers for upsampling are used to restore the resolution of the input. In our future work, tuning of the hyper-parameters and the introduction of various augmentations may yield better segmentation performances. And we plan to investigate the potential of the method for other semantic segmentation tasks and try to improve performance from the perspective of network structure.

References

[1] Siegel R L, Miller K D, Fuchs H E, et al. Cancer Statistics, 2021[J]. CA: a Cancer Journal for Clinicians, 2021, 71(1): 7-33.
[2] Ge Z, Demyanov S, Chakravorty R, et al. Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images[C]/International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2017: 250-258.
[3] Xie Y, Zhang J, Xia Y, et al. A mutual bootstrapping model for automated skin lesion segmentation
and classification[J]. IEEE transactions on medical imaging, 2020, 39(7): 2482-2493.

[4] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[5] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[6] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[7] Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 fourth international conference on 3D vision (3DV). IEEE, 2016: 565-571.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[9] Berseth M. ISIC 2017-Skin Lesion Analysis Towards Melanoma Detection[J]. arXiv preprint arXiv:1703.00523, 2017.

[10] Bi L, Kim J, Ahn E, et al. Automatic skin lesion analysis using large-scale dermoscopy images and deep residual networks[J]. arXiv preprint arXiv:1703.04197, 2017.

[11] Sarker M M K, Rashwan H A, Akram F, et al. SLSDeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2018: 21-29

[12] Menegola A, Tavares J, Fornaciali M, et al. RECOD titans at ISIC challenge 2017[J]. arXiv preprint arXiv:1703.04819, 2017.

[13] Yuan Y. Automatic skin lesion segmentation with fully convolutional-deconvolutional networks[J]. arXiv preprint arXiv:1703.05165, 2017.

[14] Al-Masni M A, Al-Antari M A, Choi M T, et al. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks[J]. Computer methods and programs in biomedicine, 2018, 162: 221-231

[15] Sung K K. Learning and example selection for object and pattern detection[J]. 1996.

[16] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. IEEE, 2001, 1: I-I.