Multi-resolution-Tract CNN with Hybrid Pretrained and Skin-Lesion Trained Layers

Jeremy Kawahara, and Ghassan Hamarneh

Medical Image Analysis Lab, Simon Fraser University, Burnaby, Canada
{jkawahar,hamarneh}@sfu.ca

Abstract. Correctly classifying a skin lesion is one of the first steps towards treatment. We propose a novel convolutional neural network (CNN) architecture for skin lesion classification designed to learn based on information from multiple image resolutions while leveraging pretrained CNNs. While traditional CNNs are generally trained on a single resolution image, our CNN is composed of multiple tracts, where each tract analyzes the image at a different resolution *simultaneously* and learns interactions across multiple image resolutions using the *same* field-of-view. We convert a CNN, pretrained on a single resolution, to work for multi-resolution input. The entire network is fine-tuned in a fully learned end-to-end optimization with auxiliary loss functions. We show how our proposed novel multi-tract network yields higher classification accuracy, outperforming state-of-the-art multi-scale approaches when compared over a public skin lesion dataset.

1 Introduction

The World Health Organization estimates that globally each year, between two and three million nonmelanoma skin cancers are diagnosed, and 130,000 melanoma skin cancers occur [16]. Classifying different types of skin lesions is needed to determine appropriate treatment, and computerized systems that classify skin lesions from skin images may serve as an important screening or second opinion tool. While considerable research has focused on computerized diagnosis of melanoma skin lesions [9], less work has focused on the more common nonmelanoma skin cancers and on the general multi-class classification of skin lesions. In this work, we focus on predicting multiple types of skin lesions that includes both melanoma and nonmelanoma types of cancers.

To classify skin lesions, Ballerini et al. [1] performed 5-class classification, and Leo et al. [12] performed 10-class classification over images that included nonmelanoma and melanoma skin lesions. They segmented lesions, extracted color and texture based features, and used K-nearest neighbours to classify the images. To perform 5- and 10-class skin lesion classification of nonmelanoma and melanoma skin lesions, Kawahara et al. [8] performed a two step process: first, using a Convolutional Neural Network (CNN) pretrained over ImageNet [13], they extracted image features at two different image resolutions, and second, these features were concatenated and used to train a linear classifier.

CNNs generally learn based on an image of a single fixed resolution (e.g., Krizhevsky et al. [10]). However, this single resolution may not be optimal and depends on the scale of the objects within the image. Information from multiple image resolutions may be critical in capturing fine details, especially in the domain of medical images (e.g., to discriminate pathology). As such, other works have proposed different multi-scale approaches. During testing, Sermanet et al. [14] used a fully convolutional neural network to extract predictions over multiple image resolutions and spatial locations and aggregated the predictions using a spatial max and averaging of scales. This simple aggregation approach, however, does not *learn* interactions across different resolutions (i.e. multi-resolution only applied during testing, not training). He et al. [5] proposed a spatial pyramid pooling layer applied after the last convolutional layer to produce fixed-sized responses regardless of the image size. The CNN is trained on images of multiple resolutions *sequentially*, causing the CNN to learn parameters that generalize across image resolutions. However, each prediction is based only on a single input resolution, and interactions across multiple input image resolutions are not considered. Bao et al. [2] proposed a multi-scale CNN trained and tested on image patches of different sizes (i.e. *different* field-of-view) simultaneously for segmentation, but did not explore multi-resolution input (i.e. *same* field-of-view at different resolutions) for whole image classification. Kamnitsan et al. [7] proposed a multi-scale dual-path 3D CNN for brain segmentation that, like the prior approach, considers *different* field-of-views. While the previously mentioned approach by Kawahara et al. [8] does consider multiple image resolutions, only a final linear classifier learns interactions across different image resolutions, and the CNN itself does not learn based on the input images.

In this work, we propose a CNN for skin lesion classification that *learns interactions across multiple image resolutions* of the same image simultaneously through multiple network tracts. Unlike prior multi-scale architectures [2,7], our network keeps the *same* field-of-view for image classification, uses auxiliary loss functions, and leverages parameters from existing pretrained CNNs. Leveraging pretrained CNN parameters (i.e. transfer learning) is especially useful with limited training images, and has resulted in consistent improvements in other medical image analysis tasks when compared to starting from random initialization [15]. Thus a key contribution of our work is to extend pretrained CNNs for multiple image resolutions, optimized end-to-end with a single objective function. We demonstrate that our proposed multi-tract CNN outperforms competing approaches over a public skin dataset.

2 Methods

We design a CNN to predict the true lesion class label y, given a skin lesion image x. Our CNN is composed of multiple tracts where each tract considers the same image at a different resolution using the same field-of-view. An end layer combines the responses from multiple resolutions into a single layer. We attach a *supervised loss layer* (i.e. layer with a loss function that compares predicted with

true class labels) to these combined responses, thus making the final prediction a learned function of multiple resolutions of the same image. This loss is back-propagated through all tracts causing the entire network to be optimized with respect to multiple image resolutions. We add auxiliary supervised loss layers to each tract, motivated by the work of Lee et al. [11], who found that adding additional "companion"/"auxiliary" supervised layers regularize the responses learned. In this work, auxiliary losses cause each tract to learn parameters that classify well at that particular resolution. At test time, we ignore the auxiliary classifiers and only use the final end classifier.

Converting a pretrained CNN to multi-tract CNN In order to train large CNNs with a limited skin dataset, we use a hybrid of the pretrained AlexNet [4,10] architecture and parameters θ_p, (omitting the 1000-d ImageNet-specific output layer) for early network layers, and additional untrained layers for later network layers that learn only from skin images. To pass images of different resolutions through all the layers pretrained on a single resolution, we convert (keeping the trained parameters) fully-connected layers to convolutional layers, as convolutional layers allow for variable sized inputs [14].

For practical considerations (e.g., limited GPU memory), we limit our discussion and experiments to two tracts, although this approach is applicable to additional tracts/resolutions. We refer to the two tracts as the *upper tract*, which takes in a low-resolution image, and the *lower tract*, which takes in a high-resolution image. Our full proposed network is shown in Fig. 1.

We pass an image $x^{(0)}$, of the same image resolution that the pretrained network (AlexNet) was trained on to the *upper tract* of our network. This produces responses of size $1\times1\times4096$. We add an additional convolutional layer with untrained (i.e. randomly initialized) parameters $\theta_t^{(0)}$, which produces responses of lower dimensionality $1\times1\times256$. To the *lower tract*, we pass an image $x^{(1)}$ with an image resolution greater than that of $x^{(0)}$. After being convolved with the pretrained parameters, the lower tract produces responses of $m \times m \times 4096$ (Fig. 1 *lower green box*). Other works have reduced this $m \times m$ dimensionality through pooling [5,8,14], but in this work, we add additional untrained $1\times1\times4096$ and $m \times m \times 64$ convolutional filters, $\theta_t^{(1)}$, that *learn* to reduce the dimensionality to $1\times1\times256$. Using these two layers instead of a single fully-connected layer, we significantly reduce the amount of needed parameters. Auxiliary supervised loss layers with untrained parameters, $\theta_l^{(0)}, \theta_l^{(1)}$, are added to the upper and lower tract responses.

An untrained convolutional layer takes as input the 256-dimensional responses from both tracts. We add a supervised loss layer to these combined responses making the final prediction a function of the image taken at two different resolutions. In order to reduce the total number of independent parameters in our model, we are inspired by the work on Siamese nets [3] to share the AlexNet weights θ_p, across the upper and lower tracts. This means that updates to θ_p will be based on both image resolutions. Finally, rather than storing separate image resolutions of the same image, we only store the highest desired resolu-

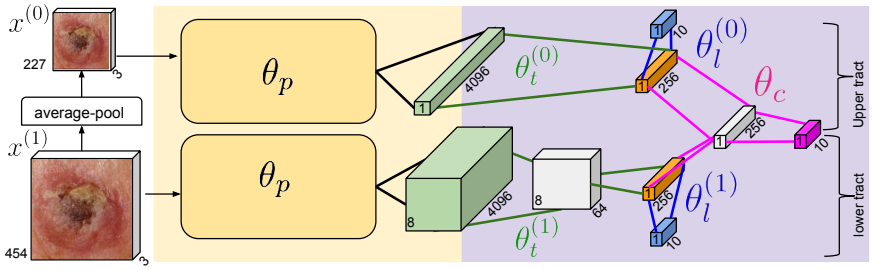

Fig. 1: The proposed two-tract fully convolutional multi-resolution neural network. The highest resolution image $x^{(1)}$, is averaged-pooled to create a low-resolution image $x^{(0)}$, as input to the *upper tract*. $x^{(1)}$ is fed to the *lower tract* to extract responses at a finer scale. As all layers are convolutional layers, a larger input produces larger responses (*green lower box*). After the layers with pretrained parameters θ_p, additional layers with unshared trainable parameters θ_t, are added. Each tract has a supervised auxiliary loss layer (*blue box*). The responses from both image resolutions are combined and an output layer makes the final prediction (*pink box*). Spatial dimensions (e.g., 8 mean 8×8) are given inside each box, and the number of channels are shown alongside each box.

tion. Within the network architecture itself, we average-pool the high-resolution image to the desired low-resolution scale, allowing for more efficient storage.

Multi-tract loss and optimization Our network has a supervised data loss that considers the combined high and low resolution images, as well as auxiliary data losses that each only considers the responses from a single image resolution. These equally weighted losses are averaged over n mini-batches of training instances along with a regularization over the parameters,

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = \frac{\lambda}{n} \sum_{i}^{n} \left(\ell(x_i, y_i; \theta_p, \theta_t, \theta_c) + \sum_{j=0}^{n_{\text{aux}}} \ell(x_i^{(j)}, y_i; \theta_p, \theta_t^{(j)}, \theta_l^{(j)}) \right) + \gamma ||\boldsymbol{\theta}|| \tag{1}$$

where $\ell(.)$ is the cross-entropy loss using a softmax activation function; the ith image is transformed into the jth resolution $x_i^{(j)}$; y_i is the ground truth class label of x_i; the parameters $\boldsymbol{\theta} = \{\theta_p, \theta_t, \theta_c, \theta_l\}$ are composed of the shared pretrained parameters θ_p, the unshared tract parameters $\theta_t = \{\theta_t^{(j)}\}$ where j indicates the tract, the parameters connecting the j auxiliary loss $\theta_l = \{\theta_l^{(j)}\}$, and the parameters connecting the tracts together θ_c; $||\boldsymbol{\theta}||$ is the L2 regularization over the parameters; and, λ, γ weight the terms where $\lambda = \frac{1}{n_{\text{aux}}+1}$ and n_{aux} is the number of auxiliary supervised layers in the network (e.g., 2).

We update our network parameters $\boldsymbol{\theta}$ using stochastic gradient descent with mini-batches. Thus, for the k+1 iteration, we compute $\boldsymbol{\theta}_{(k+1)}$, from the previous

k iteration parameters $\boldsymbol{\theta}_{(k)}$ and parameter updates $\boldsymbol{U}_{(k)}$ as,

$$\boldsymbol{U}_{(k+1)} = \mu \boldsymbol{U}_{(k)} - \alpha \nabla \mathcal{L}(\boldsymbol{\theta}_{(k)}) \qquad \text{and} \qquad \boldsymbol{\theta}_{(k+1)} = \boldsymbol{\theta}_{(k)} + \boldsymbol{U}_{(k+1)}, \tag{2}$$

using a low learning rate $\alpha = 10^{-4}$, as much of the CNN is pretrained; $\nabla \mathcal{L}(\boldsymbol{\theta}_{(k)})$ are the gradients of Eq. 1; and μ is a momentum parameter. We use Caffe [6] to implement our architecture and optimize Eq. 1 with mini-batches of size $n = 15$ (lowers GPU memory to allow for multiple tracts). As common in the literature [10], we set $\mu = 0.9$ and $\gamma = 0.0005$.

3 Results

We used the Dermofit Image Library[1] to test our proposed method. This dataset contains 1300 skin lesion images from 10 classes: Actinic Keratosis (AK), Basal Cell Carcinoma (BCC), Melanocytic Nevus/Mole (ML), Squamous Cell Carcinoma (SCC), Seborrhoeic Keratosis (SK), Intraepithelial Carcinoma (IEC), Pyogenic Granuloma (PYO), Haemangioma (VSC), Dermatofibroma (DF), and Malignant Melanoma (MEL). We randomly divided the dataset into three subsets of approximately the same number of images and class distributions. One subset is used to train (i.e. optimize Eq. 1), validate (e.g., test design decisions), and test. We resized the $x^{(0)}$ image to 227×227 and $x^{(1)}$ to 454×454. Each image was normalized by subtracting the per-image-mean intensity as in [8].

For our first experiments (Table 1 rows *a-c*), we implemented the two-step approach of Kawahara et al. [8] by extracting responses from the sixth layer (FC6) of the pretrained AlexNet for images $x^{(0)}$ and $x^{(1)}$, and max-pool the spatial responses of $x^{(1)}$. As in [8], these extracted responses are used to train a logistic regression classifier. We report the accuracy for classifying $x^{(0)}$ and $x^{(1)}$ individually, and on the concatenated responses from the two image resolutions (note this experimental setup only uses half of the training images that [8] did).

Our next experiments (rows *d,e*) show that our hybrid use of pretrained and additional skin-lesion trained layers improved classification accuracy. We split the two-tract network into upper and lower tracts and train each separately on a single resolution. For a fair comparison, we doubled the number of nodes in the layer before the auxiliary loss layer (i.e. Fig. 1 *orange layer*) to closely match the number of independent parameters within the two-tract model. The accuracy of the one-tract single-resolution model (rows *d,e*) improved over rows *a,b*, but is less than our proposed model (row *i*), indicating that considering multiple resolutions within our two-tract architecture improves accuracy.

Row *f* details the results of applying the classification approach of Sermanet et al. [14] to aggregate the CNN responses from multiple image resolutions. To implement their classification approach, we pass high-resolution $x^{(1)}$ images through the one-tract model (row *d*) trained on low-resolution $x^{(0)}$ images to produce class responses with spatial dimensions. We take the maximum spatial response and average it with the class responses computed from the low-resolution image to compute a class unnormalized probability vector.

[1] https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html

Table 1: Experimental results. *image res.* shows the image resolution in the train/test phase (e.g., 227/454 means image size 227×227 and 454×454). We report the averaged classification accuracy for the *valid* and *test* datasets. Rows *a-i* use multi-resolution versions of an image spanning the same field-of-view. Rows *j,k* use augmented image views, where row *k* combines the multi-resolution approach with augmented views.

	method	image res.	valid	test	
(a)	FC6+LogReg	227	0.674	0.705	single view
(b)	FC6+LogReg	454	0.649	0.700	
(c)	Kawahara et al. 2015 [8]	227/454	0.684	0.741	
(d)	1-tract (*ours*)	227	0.733	0.741	
(e)	1-tract (*ours*)	454	0.737	0.759	
(f)	1-tract + Sermanet et al. 2014 [14]	227/454(test)	0.719	0.748	
(g)	He et al. 2014 [5] (SPP)	224/448	0.688	0.711	
(h)	2-tract 0-aux-losses (*ours*)	227/454	0.723	0.755	
(i)	2-tract 2-aux-losses (*ours*)	227/454	**0.751**	**0.773**	
(j)	1-tract (*ours*)	454	0.760	0.775	aug. view
(k)	2-tract 2-aux-losses (*ours*)	227/454	**0.781**	**0.795**	

Row *g* uses He et al. [5]'s Spatial Pyramid Pooling (SPP) approach, which learns CNN parameters from multiple image resolutions. To implement, we use the pretrained Zeiler-Fergus (ZF) SPP network He et al. [5] provided (similar architecture to AlexNet) and replace their final output layer with our own. We train over ≈11 epochs before switching between 224×224 and 448×448 image resolutions, repeating 20 times for 9000 iterations (more iterations did not improve results). Each image resolution is fine-tuned for 1000 iterations. During testing, we averaged the CNN's output class responses from both resolutions.

Rows *h,i* show results using our two-tract multi-resolution architecture. Without auxiliary losses (row *h*), the two-tract model performs worse than the single tract (row *e*), highlighting the need to include the auxiliary loss functions (Eq. 1) to achieve the highest accuracy (row *i*). Note that we outperform [8], which was shown to outperform [1,12], and that [1,12] were non-CNN based approaches specifically designed for this dataset. The confusion matrix over the test data is shown in Fig. 2 (*right*). We ran additional experiments to cross-validate over the two other folds and obtained a statistically significant difference between the baseline of [8] (using the approach from row *c*) and our two-tract approach (row *i*) with a mid-p McNemar's test, p=0.0155.

We compare the accuracy of the final output classifier with the accuracy of the auxiliary classifiers (Fig. 2 (*left*)). Generally, the final classifier has a higher accuracy, indicating that this classifier (which considers the same responses as each auxiliary classifier) has learnt to combine responses from multiple image resolutions, and that this improves classification accuracy. This plot also highlights the advantage of pretrained parameters, as high accuracy occurs within 5000 iterations (1 hour of training), using a low number (430) of training images.

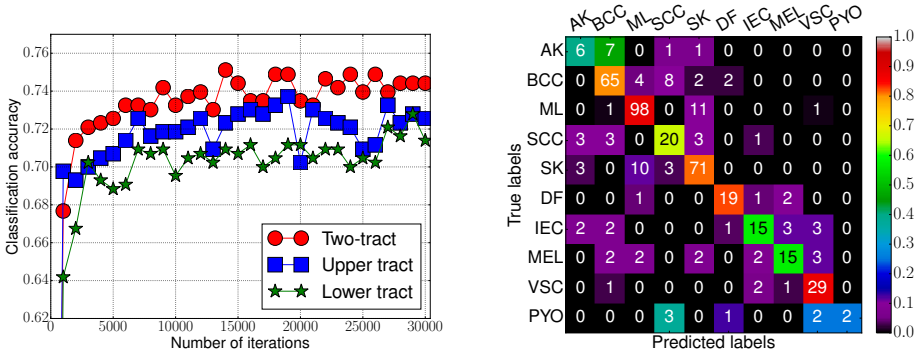

Fig. 2: (*left*) A comparison of the classification accuracy of the individual upper and lower single resolution-tracts with the two-tracts. Integrating multiple image resolutions yields higher accuracy. By using pretrained parameters, we reach a high accuracy within a short number of iterations. (*right*) The confusion matrix over the 10-classes from our test data using our proposed multi-tract CNN (heatmap indicates class-specific classification accuracy normalized across rows).

In order to focus on the effects of our proposed architecture and multi-resolution input, the experiments in rows *a*-*i* did not use data augmentation. Our final set of experiments demonstrates that our multi-resolution approach is *complementary* to the commonly used approach of training using different image views. We augment the training images with left-right flips, and rotations. Row *k* combines augmented image views with multi-resolution input, resulting in further accuracy improvements when compared to using only augmented views (row *j*) and using only multiple resolutions (row *i*), highlighting that the proposed multi-resolution input complements existing image augmentation approaches.

We did not compare to [2,7] as their approach was designed for 3D segmentation, and while their approach of taking as input different amount of spatial context is well motivated for patch-based segmentation of 3D volumes, it is less applicable to whole image classification. Further contributions we make that differ with their work include: pretrained CNNs for multiple resolutions, the use of auxiliary losses, and multi-resolution input.

Finally, we discuss possible reasons why successful approaches used in *computer vision datasets* (e.g., ImageNet [13] where images are captured at widely different scales), were found less effective for our skin diagnosis application (where dermatology images are captured at a similar scale). When the scale of objects widely differs, the SPP approach [5] to learn parameters that *generalize* over multiple scales, and the approach to *aggregate* responses over different scales [14], are desirable. However, in our case, where the objects' scale are roughly fixed, the different CNN-tracts learn to respond to characteristics that are *specific* to that resolution. This highlights how our proposed architecture is well designed for skin images captured at relatively fixed scales.

4 Conclusions

We presented a novel multi-tract CNN that extends pretrained CNNs for multi-resolution skin lesion classification using a hybrid of pretrained and skin-lesion trained parameters. Our approach captures interactions across multiple image resolutions simultaneously in a fully learned end-to-end optimization, and outperforms related competing approaches over a public skin lesion dataset.

Acknowledgments. Thanks to the Natural Sciences and Engineering Research Council (NSERC) of Canada for funding and to the NVIDIA Corporation for the donation of a Titan X GPU used in this research.

References

1. Ballerini, L., Fisher, R.B., Aldridge, B., Rees, J.: A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Celebi, M.E., Schaefer, G. (eds.) Color Medical Image Analysis. vol. 6, pp. 63–86. Springer Netherlands (2013)
2. Bao, S., Chung, A.C.S.: Multi-scale structured CNN with label consistency for brain MR image segmentation. CMBBE: Imaging & Visualization. pp. 1–5 (2016)
3. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similiarty metric discriminatively, with application to face verification. In: IEEE CVPR. pp. 349–356 (2005)
4. Donahue, J., et al.: DeCAF: A deep convolutional activation feature for generic visual recognition. In: ICML. vol. 32, pp. 647–655 (2014)
5. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV. vol. 8691, pp. 346–361. Springer (2014)
6. Jia, Y., et al.: Caffe: Convolutional Architecture for Fast Feature Embedding. In: ACM Conference on Multimedia. pp. 675–678 (2014)
7. Kamnitsas, K., et al.: Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. In: ISLES Challenge (2015)
8. Kawahara, J., BenTaieb, A., Hamarneh, G.: Deep features to classify skin lesions. In: IEEE ISBI. pp. 1397–1400 (2016)
9. Korotkov, K., Garcia, R.: Computerized analysis of pigmented skin lesions: A review. Artificial Intelligence in Medicine 56(2), 69–90 (2012)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS. pp. 1097–1105 (2012)
11. Lee, C.Y., Xie, S., Gallagher, P.W., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: AISTATS. vol. 38, pp. 562–570 (2015)
12. Leo, C.D., et al.: Hierarchical classification of ten skin lesion classes. In: Proc. SICSA Dundee Medical Image Analysis Workshop (2015)
13. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. IJCV 115(3), 211–252 (2015)
14. Sermanet, P., et al.: OverFeat: Integrated recognition, localization and detection using convolutional networks. In: ICLR (2014)
15. Shin, H.C., et al.: Deep convolutional neural networks for computer-aided detection. IEEE TMI 35(5), 1285–1298 (2016)
16. World Health Organization: INTERSUN: the global UV project. http://who.int/uv/publications/en/Intersunguide.pdf (2003), accessed: 2016-02-13