Skin Lesion Segmentation and Classification Using Deep learning

Nithish Gurram*
Computer Engineering
University of Houston-Clear Lake
Houston, Texas
GurramN8099@Uhcl.edu

Unal Sakoglu
Computer Engineering
University of Houston-Clear Lake
Houston, Texas
Sakoglu@Uhcl.edu

Liwen Shih
Computer Engineering
University of Houston-Clear Lake
Houston, Texas
Shih@Uhcl.edu

ABSTRACT

This work describes a supervised deep-learning-based skin lesion classification system that classifies eight categories of cancerous skin tissues, where images were provided by International Skin Cancer Imaging Collaboration (ISIC) in 2019. We trained and tuned parameters of an AlexNet deep learning neural network architecture to learn about eight categories based on a training set provided by ISIC, cross-validated it on that set, and submitted results to ISIC website to have the trained algorithm evaluated. We report that, by employing data augmentation, which includes rotated, cropped, and noisy versions of the original images, the cross-validation accuracy improved to around 79% and the area under the curve (AUC) of the test results improved to around 78%, from 50%

CCS CONCEPTS

• Computing Methodologies → Machine Learning

KEYWORDS

Deep Learning, Convolutional Neural Network, AlexNet, Skin Lesion Segmentation and Classification.

ACM Reference format:
Nithish Gurram, Unal Sakoglu and Liwen Shih. 2020. Skin Lesion Segmentation and Classification using Deep Learning. In Practice and Experience in Advanced Research Computing 2020, Portland Marriott Downtown Waterfront Hotel, Portland, OR, United States, July 26-30, 2020, *ACM, New York, NY, USA, 4 pages.* https://doi.org/10.1145/1234567890

1 Introduction

Skin cancer is one of the most recurrent types of cancers, with melanoma being the deadliest form [1]. Since skin cancer occurs on the surface of the skin, its lesions can be evaluated by visual inspection. Dermoscopy is a noninvasive skin imaging modality which permits visualizing more profound levels of the skin as its surface reflection is removed. Dermoscopy has demonstrated improvement for diagnosis of skin cancer compared to unaided visual inspection [1]. However, clinicians should receive adequate training for those improvements to be realized. In order to make expertise more widely available, the International Skin Imaging Collaboration (ISIC) has developed the ISIC Archive, an international repository of dermoscopic images, for both the purposes of clinical training, and for supporting technical research toward automated algorithmic analysis by hosting the ISIC Challenges [2].

Most attempts in utilizing Deep Learning for skin lesion segmentation have focused solely on the training or increasing the accuracy of image classification. Many studies in image classification have proposed artificially increasing training dataset size by creating modified versions of data, called data augmentation, in conjunction with applying Deep Learning algorithms. To achieve this objective in this work, we artificially created augmented data by using image processing techniques.

2 Background

The International Skin Imaging Collaboration (ISIC) is an international effort to improve melanoma diagnosis, sponsored by the International Society for Digital Imaging of the Skin (ISDIS). The ISIC Archive contains the largest publicly available collection of quality controlled dermoscopic images of skin lesions [2]. Presently, the ISIC Archive contains over 13,000 dermoscopic images, which were collected from leading clinical centers internationally and acquired from a variety of devices within each center. Broad and international participation in image contribution is designed to insure a representative clinically relevant sample. All incoming images to the ISIC Archive are screened for both privacy and quality assurance. Most images have associated clinical metadata, which has been vetted by recognized melanoma experts. [5-7] A subset of the images has undergone annotation and markup by recognized skin cancer experts. These markups include dermoscopic features (i.e., global, and focal morphologic elements in the image known to discriminate between types of skin lesions).

Dermoscopy is an imaging technique that eliminates the surface reflection of skin. By removing surface reflection, visualization of deeper levels of skin is enhanced. Prior research has shown that when used by expert dermatologists, Dermoscopy provides improved diagnostic accuracy, in comparison to standard photography. As inexpensive consumer dermoscopic attachments for smart phones are beginning to reach the market, the opportunity for automated dermoscopic assessment algorithms to positively influence patient care increases [1].

2.1 Dataset

The training dataset has 25,331 color (RGB) JPEG images of skin lesions, metadata entries of age, sex, general anatomic site, and common lesion identifier, training label data “Ground truth.csv” has 25,331 entries of gold standard lesion diagnosis and matching

metadata with various information such as age, sex, and general anatomic site. The test dataset has 8,238 color JPEG images of skin and corresponding metadata. In this work, any metadata information was not used for training the algorithms. Most of the images were of size 1022 x 767 or similar. For each color component, JPEG images had 8-bits (levels 0 through 255). The images had to be rescaled down to appropriate size for our algorithms. A subset of the images (scaled and cropped) are shown in Figure 1. [2]

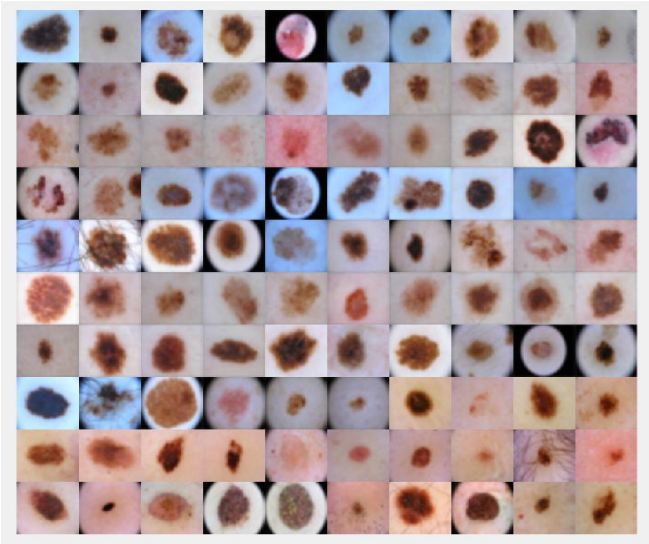

Figure 1: A random selection of ISIC database images

In the training data set, there were eight categories of different types of skin cancer images. The list is as follows: melanoma (MEL), melanocytic nevus (NEV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (solar lentigo /seborrheic keratosis / lichen planus-like keratosis) (BKL), dermatofibroma (DF), vascular lesion (VASC), squamous cell carcinoma (SCC). In addition, in the test dataset, there was an additional unknown category of skin cancer images (UNK).

3 Methods

3.1 Data Augmentation

We expanded the training image dataset by artificially creating rotated, noisy, and cropped versions of the images. Images in this dataset were first scaled down to 64x64 from their original sizes before the training. For rotations, we added rotated versions of each image rotated with 10°, 15°, 30°, 45°, 75°, 90°, 120°, 180°, 270°, 275°. For noise, we added two noisy versions of each image by adding two different noise levels: a) uniformly distributed random noise of intensity values between 0 and 5, b) uniformly distributed random noise of intensity values between 0 and 10. For random cropping, we first scaled the original images down to 128x128. We then cropped a random 64x64 from the larger 128x128 images, where the cropping offset in x- and y-dimensions was a uniformly distributed random integer between 0 and 32.

Overall, the augmentation has increased the number of training images 13x, from 25,331 to 329,303 images.

3.2 AlexNet Architecture

In this section we explain the methodology employed in this paper, comprising the factors present in our experimental design, the hardware/software infrastructure used to run the project, the statistical design of the experiment and the AlexNet deep learning neural network algorithm.

The architecture of AlexNet consists of 25 layers, eight of which are convolutional layers: five convolutional layers and three fully connected layers. AlexNet uses Rectified Linear Units (ReLU) instead of the tanh function, which was standard previously [3]. ReLU's advantage is in training time; a CNN using ReLU was able to reach a 25% error on the CIFAR-10 dataset six times faster than a CNN using tanh [3].

Recently, computational power of graphical processing units (GPUs) have increased tremendously. AlexNet allows for multi-GPU training by splitting the model's neurons on these GPUs [3]. This means that a bigger model can be trained, and it also cuts down on the training time. CNNs traditionally "pool" outputs of neighboring groups of neurons with no overlapping. However, when the authors introduced overlap, they saw a reduction in error by about 0.5% and found that models with overlapping pooling generally find it harder to overfit.

For our work, we used the following AlexNet parameters: First convolutional layer filter number = 144, zero-center normalization in layer 1, 96 11x11x3 convolutions with stride [4 4] and padding [0 0 0 0] in layer 2, reluLayer (rectifier linear units) in layer 3, cross-channel normalization with 5 channels per element in layer 4, 3x3 max pooling with stride [2 2] and padding [0 0 0 0] in layer 5, 2 groups of 128 5x5x48 in layer 6, reluLayer in layer 7, cross channel normalization with 5 channels per element in layer 8, 3x3 max pooling with stride [2 2] and padding [0 0 0 0] in layer 9, 384 3x3x256 convolutions with stride [1 1] and padding [1 1 1 1] in layer 10, reluLayer in layer 11, two groups of 192 3x3x192 convolutions with stride [1 1] and padding [1 1 1 1] in layer 12, reluLayer in layer 13, two groups of 128 3x3x192 convolutions with stride [1 1] and padding [1 1 1 1] in layer 14, reluLayer in layer 15, 2x2 max-pooling with stride [2 2] and padding [0 0 0 0] in layer 16, fully connected layer of 256 in layer 17, reluLayer in layer 18, 50% drop-out in layer 19, fully connected layer of 256 in layer 20, reluLayer in layer 21, 50% drop-out in layer 22, fully connected layer of 8 in layer 23, softmax function in layer 24, and classification of the output in the final layer 25.

3.3 Training, Cross-Validation and Testing

32,930, or approximately 10% of the training images were set aside for cross-validation; therefore, the remaining 296,373 or 90% of the images were used for training. The training parameters were as follows: momentum of 0.9, initial learning rate of 0.001, learning rate schedule as piecewise, learning rate drop factor of 0.1, learning rate drop period of 8, L2-regularization with parameter 0.004, maximum epochs of 100, mini batch size of 128. After the algorithm's cross-validation accuracy reached a plateau, training

was stopped and weights and category scores were saved, and then applied to the test dataset, and was uploaded to ISIC's website for evaluation.

There was an additional unknown (UNK) category in the test image dataset. Since we did not have any member of the UNK category in the training image set, in our classification, we classified a test image as belonging to UNK category, only if the maximum category score among the eight known categories were less than twice (threshold of 2) the average of the remaining category scores. This constituted a confidence-based categorization of whether the new test image belonged to a different category which was not seen by the algorithm before. The threshold of 2 was set empirically and somewhat arbitrarily. A lower threshold would decrease the probability of assigning a new image to UNK category. Classification of the UNK category correctly was not a focus of this paper.

3.3 Computational Details

We implemented, trained and tested the algorithm using MATLAB scientific computing software [8] Deep Learning toolbox on a Dell Inspiron Workstation with Intel Core i7 processor and 16 GB of RAM with standard GPU. We set the training algorithm to have 100 epochs, each with 2493 iterations. The cross-validation frequency was 50 iterations. After running just over 14,000 minutes, which was shortly after 10 epochs were complete, the cross-validation accuracy seemed to have reached a plateau.

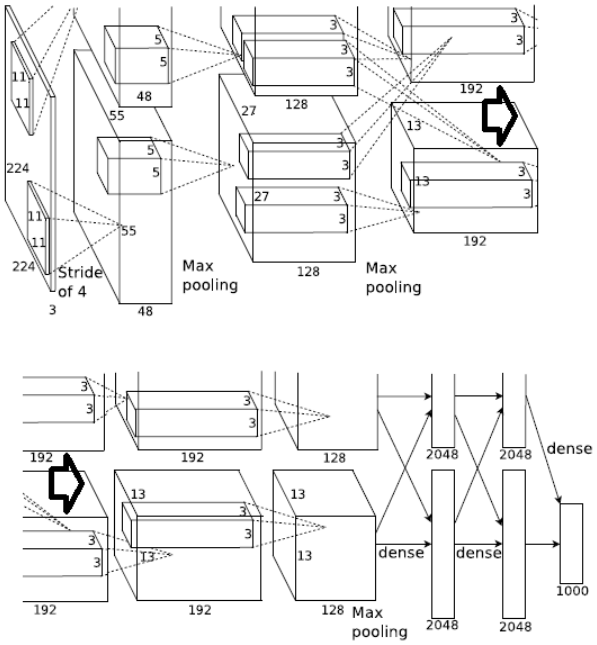

Figure 2: The Original Alex Net Architecture [3]

4 Results

After 10 epochs and about 14,000 minutes, the training run reached 79% cross-validation (CV) accuracy (Fig. 3). Due to time restrictions, we decided to stop the training, saved weights and scores, and applied classification to test data. Figure 4 summarizes the results of classification on the test data, whereas Figure 5 presents the ROC curve (TPR vs FPR) for each of the nine categories, including the UNK category. Average area under the curve (AUC), including the UNK's AUC, was 0.784. When the UNK category is excluded, average AUC is approximately 0.820. With no data augmentation, CV accuracy was 0.72 and AUC was around 0.68 (with UNK) and 0.70 (without UNK), therefore, augmentation improved the classification performance.

Figure 3: Training and Cross-Validation.

	Diagnosis Category										
	MEL	NEV	BCC	AK	BKL	DF	VAS	SCC	UNK	**Mean**	**Mea2**
AUC	0.842	0.900	0.808	0.783	0.728	0.830	0.822	0.410	0.501	**0.784**	**0.820**
AUC 80	0.660	0.803	0.647	0.589	0.463	0.555	0.58	0.670	0.181	**0.581**	**0.631**
Avg. Prec.	0.606	0.825	0.319	0.132	0.245	0.086	0.322	0.116	0.193	**0.316**	**0.331**
Acc.	0.859	0.833	0.805	0.941	0.899	0.988	0.987	0.977	0.807	**0.900**	**0.912**
Sens	0.503	0.779	0.534	0.0428	0.165	0.000	0.284	0.026	0.004	**0.260**	**0.292**
Spec	0.931	0.859	0.845	0.989	0.969	1.000	0.997	0.998	0.990	**0.954**	**0.950**
DC	0.545	0.751	0.417	0.0684	0.222	0.000	0.377	0.048	0.006	**0.270**	**0.303**
PPV	0.595	0.725	0.342	0.170	0.337	1.000	0.547	0.235	0.429	**0.487**	**0.494**
NPV	0.903	0.890	0.924	0.951	0.824	0.980	0.979	0.806	0.547	**0.928**	**0.976**

Figure 4 Display of Mean values after Data Augmentation. AUC: area under the ROC curve; AUC80: AUC where sensitivity > 80% only; Avg. Prec: Average precision; Acc: Accuracy; Sens: Sensitivity, Spec: Specificity; DC: Dice Coefficient; PPV: positive predictive value; NPV: negative predictive value. Categories: MEL: melanoma, NEV: melanocytic nevus; BCC: basal cell carcinoma; AK: actinic keratosis; BKL: benign keratosis variants; DF: dermatofibroma; VAS: vascular lesion; SCC: squamous cell

carcinoma; UNK: unknown (in the test set only). Mean: Mean including the UNK; Mea2: Mean excluding the UNK.

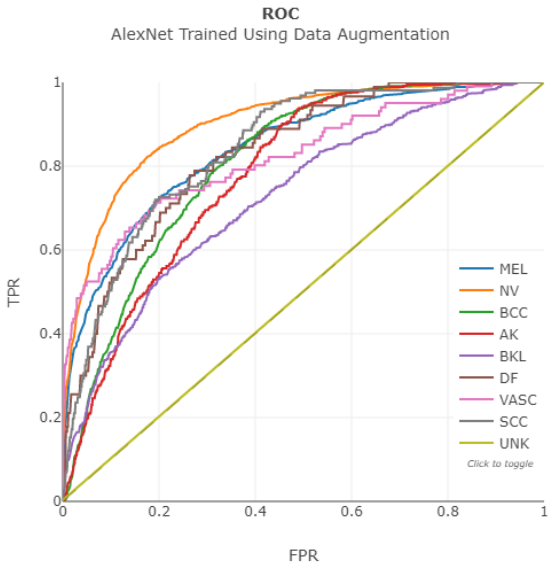

Figure 5: Display of ROC Curve Characteristics of each Lesion type

5 Conclusions

In this work we modified and trained an AlexNet deep-learning neural network for skin lesion classification system that classifies eight categories of cancerous skin tissues, images of which were provided by International Skin Cancer Imaging Collaboration (ISIC) in 2019. We trained and tuned parameters of the AlexNet architecture to learn about eight categories of skin cancer based on a training set provided by ISIC, cross-validated it on that set, and submitted results to ISIC website to have the trained algorithm evaluated on a test set. By employing data augmentation, which includes rotated, noisy and cropped versions of images, the cross-validation accuracy was improved to around 79%, and the area under the curve (AUC) of the classification of the test dataset was around 78% from 50%(82% when the novel unknown category results in the test dataset are excluded). These metrics represent 9-10% improvement over classification metrics with no augmentation; therefore, augmentation certainly improves the classification performance significantly, albeit costing much longer training time. Future work includes coming up with a better confidence strategy and metric in order to classify novel / unknown category or categories in the test images, more augmentation techniques which can include sheared, scaled, color-processed and morphology-processed versions of images, and use of supercomputing resources for speeding-up training of the algorithm, and further optimizing neural network algorithm parameters.

ACKNOWLEGEMENTS

Student Nithish Gurram thanks Dr. Unal Sakoglu and Dr. Liwen Shih for their supervision throughout this project; and he thanks the Computer Engineering Program at University of Houston – Clear Lake, for supporting his graduate studies.

REFERENCES

[1] Lowell BA, et al. “Dermatology in primary care: Prevalence and patient disposition" In: Journal of the American Academy of Dermatology (JAAD), vol. 45, no. 2. 2001.
[2] https://workshop2020.isic-archive.com/
[3] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, 2012. ImageNet Classification with Deep Convolutional Neural Networks *Communications of the ACM*. **60** (6): 84–90. doi:10.1145/3065386. ISSN 0001-0782
[4] Automated Skin Lesion Classification Using Ensemble of Deep Neural Networks in ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection Challenge Md Ashraful Alam Milton https://arxiv.org/abs/1901.10802
[5] Deep-Learning Ensembles for Skin-Lesion Segmentation, Analysis, Classification: RECOD Titans at ISIC Challenge 2018 https://arxiv.org/abs/1808.08480
[6] A Novel Multi-task Deep Learning Model for Skin Lesion Segmentation and Classification Xulei Yang, Zeng Zeng, Si Yong Yeo, Colin Tan, Hong LiangTey, Yi Su. https://arxiv.org/abs/1703.01025
[7] Segmentation of Skin Lesions from Digital Images Using Joint Statistical Texture Distinctiveness https://ieeexplore.ieee.org/abstract/document/6701329/
[8] https://www.mathworks.com/help/deeplearning/index.html?s_tid=srchtitle