A Hierarchical Approach to Skin Lesion Classification

Rahul Kulhalli, Chinmay Savadikar, Bhushan Garware

Abstract

This manuscript describes the approaches and methodologies used by our team, SkinLegion, for undertaking the ISIC 2018 Task 3 challenge - Lesion Diagnosis.
After several combinations and architectural approaches, we conclude that the 5-stage hierarchical classification pipeline yields the best results on the online validation set. We also conclude that despite the heavy class imbalance, data augmentation can significantly aid a Convolutional Neural Network extract a good set of features for the minority classes.
All our models and scripts are available on our GitHub repository - https://github.com/rahulkulhalli/SkinLegion

1 Introduction

Deep learning has recently generated a lot of interest in the medical domain. Convolutional Neural Networks are rapidly approaching human expertise levels and achieving state-of-the-art performance. Greater compute is resulting in the discovery of newer, more efficient architectures.

2 Literature Survey

i. **ISIC 2017 (RECOD Titans)**
We derived our base approach for the ISIC2018 challenge from a team which competed in last year's image classification - the RECOD Titans [1]. Their one-vs-all approach towards Skin Lesion detection inspired us to create our own set of hierarchical classifiers so as to address the class imbalance problem. By trial-and-error, we found that the minority classes (namely DF and VASC) consistently performed well even with a low amount of augmentations. This piece of knowledge was critical for our motivation in restructuring the classifier hierarchy.

ii. **Deep Learning in Medical Image Analysis**
We followed some very valuable insights brought into the spotlight by Geert et. all [2] and Razzak et. all [3]. It conforms with the notion that Dermoscopic image classification is an area where Convolutional Neural Networks are rapidly gaining ground.
For this specific domain, Esteva et. all [4] have performed ground-breaking research in classifying skin cancer with the help of deep learning and concluded that the final performance of a fully-trained Convolutional Neural Network is at par with that of expert dermatologists. Furthermore, they also conclude that with more data, these models might even outperform human experts. Esteva et. all used the state-of-the-art InceptionV3 CNN architecture and performed transfer learning to help the model learn complex representations about the various skin cancer lesions. Saliency map visualizations indicate that the CNN learns to accurately distinguish between the different types of skin cancer lesions.

iii. **Addressing Class imbalance**
Class imbalance is detrimental to the performance of Deep Learning. An extensive study of class imbalance problem in deep learning has been performed by Buda et. al [5]. Primary methods to address this issue include data level methods like oversampling and undersampling, and classifier level methods like thresholding, cost sensitive learning and $one-class$ classification. Due to its simplicity and the use of data augmentation, we performed cost sensitive learning, wherein a higher cost is assigned to misclassification of minority classes. The cost is calculated as class weights given by the formula:

$$weight_i = \frac{number\ of\ samples}{(number\ of\ classes * number\ of\ samples\ of\ class\ i)}$$

To reduce the training time, we initialized the bias of the final decision unit of the minority with a value slightly greater than or less than the standard initialization of 0, depending on whether the minority class is classified as 0 or 1 in binary or the index of the minority class in categorical settings.

3 Methodology

i. **Data**

As the majority of our financial budget was focused on computational resources, we could not leverage the use of external datasets which required a fee for their usage. However, the data provided for this year′s challenge was several scales larger than that provided for the previous year′s challenge. This alleviated the need for use of external data. Our data was extracted from the ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection grand challenge datasets [6]. The dataset comprises of 10,015 images belonging to 7 classes, namely MEL, NV, BCC, AKIEC, BKL, DF, and VASC, obtained from the HAM10000 dataset . Figure 1 shows the class distribution of the dataset.

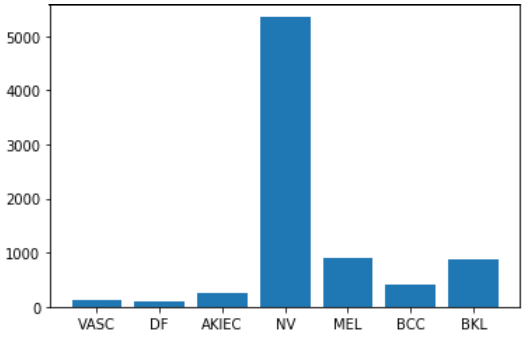

Figure 1: Class distribution of the dataset.

As can be seen in the figure, the class balance of the data provided is highly skewed. Our models were trained by only using the data provided as the training set and did not make use of the auxiliary data. Going by the standard norm, we split the dataset into 3 sets with the same class distribution - a *Training Set*, a *Validation Set*, and a *Test Set*. The Test Set was constructed using 10% of the images from each class, so as to keep the class distribution same. The Validation Set was constructed by using 10% of the remaining 90% of the images from each class. Although the provided data is larger than previous years challenge, the number of images of the minority classes is still less as compared to modern deep learning standards. For this reason, we made use of augmentations on the Training Set. Table 1 describes the nature of augmentations made. We used the Augmentor library for performing augmentations [7].

Type of Augmentation	**Probability**	**Details**
Zoom	0.9	min:1, max:1.25
Flip Left - Right	1.0	-
Flip Top - Bottom	1.0	-
Shear	0.9	Max Left:15, Max Right:15
Rotate	1.0	Max Left:20, Max Right:20
Random Distortion	0.9	Grid Width: 4, grid height: 4, Magnitude: 8
Skew	0.9	Magnitude: 0.4

Table 1: Augmentation details

4 Approaches

i. **7-way classifier**

Transfer learning [8] is the practice of fine-tuning the weights of a model trained on a very large corpus of data so as to help the model ’adapt’ to a different domain. After several combinatorial trials, we concluded that the InceptionV3 architecture gave us the most robust results. The bottleneck layer consists of 40 hidden units, followed by a Batch Normalization layer, a Dropout layer with a dropout intensity of 0.4, and the final decision layer. The baseline F1-score on our held-out test set was around 0.76. Our base benchmark, after submitting the validation results, was 66.8%. We believe that we can improve this further by fine-tuning the number of hidden units, learning rate, etc.

ii. **Hierarchical classifier**
This approach has been derived from the ISIC 2017 'RECOD Titans'. We hypothesize that a hierarchical classification approach may have some advantages over conventional end-to-end training - each model in the architecture may become a 'specialist' module, and may act as a discriminator for a certain class. Albeit subject to fine-tuning, we are hopeful that this method will yield better results as compared to the baseline benchmark.

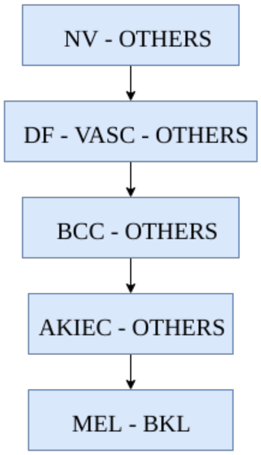

Figure 2: The 5-stage hierarchical approach

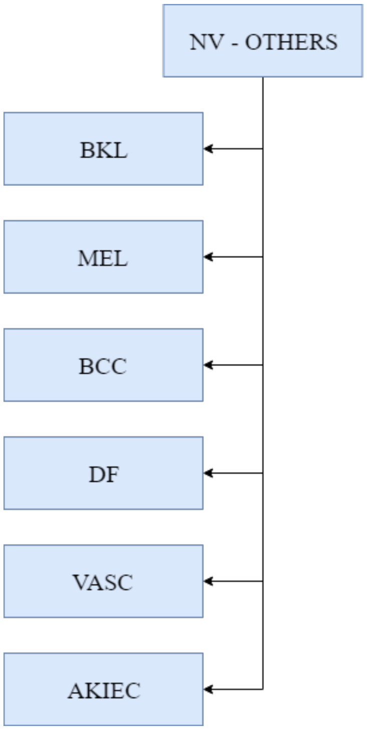

Figure 3: The 2-stage hierarchical approach

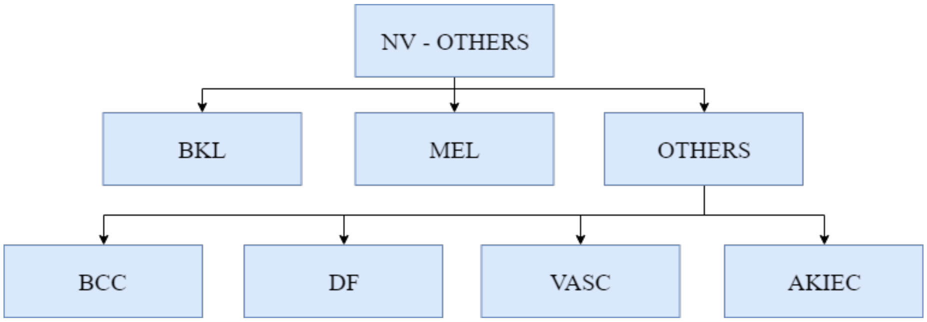

Figure 4: The 3-stage hierarchical approach

A pipelining script connects the individually-trained components together and creates a consolidated result.
After evaluating all the three aforementioned approaches, we chose to continue with the 5-stage classifier because it showed us the most promising results. Our first submission using this approach yielded an

accuracy of 79.2%.
One may also question the obviously-interchanged levels of the classifiers. We found that the MEL and BKL classes, when compared to other classes, yielded bad results but performed very well when placed against each other; thus the decision to add them as the leaves of the classifier tree.

iii. **End-to-end custom InceptionV3**

To address the problem of class imbalance, we hypothesized the creation of 'branches' at premature locations in the InceptionV3 network. We believe that this architecture, in addition to avoiding overfitting, will allow the minority classes gain important information from the majority classes.
All the 'branched' outputs are concatenated into a softmax output at the end of the model. This way, information will travel back directly to the branched output, allowing for richer representations.

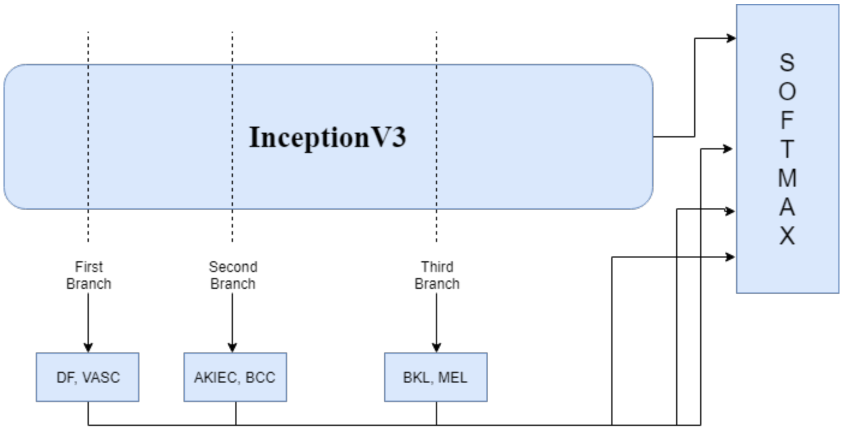

Figure 5: The proposed branched approach for InceptionV3

References

[1] Afonso Menegola, Julia Tavares, Michel Fornaciali, Lin Tzy Li, Sandra Eliza Fontes de Avila, and Eduardo Valle. RECOD titans at ISIC challenge 2017. *CoRR*, abs/1703.04819, 2017.

[2] Geert J. S. Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *CoRR*, abs/1702.05747, 2017.

[3] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. Deep learning for medical image processing: Overview, challenges and future. *CoRR*, abs/1704.06825, 2017.

[4] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[5] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *CoRR*, abs/1710.05381, 2017.

[6] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data*, 5:180161, 2018.

[7] Marcus D. Bloice, Christof Stocker, and Andreas Holzinger. Augmentor: An image augmentation library for machine learning. *CoRR*, abs/1708.04680, 2017.

[8] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning.