Ensembling Convolutional Neural Networks for Skin Cancer Classification

Aleksey Nozdryn-Plotnicki, Jordan Yap, and William Yolland

I. Introduction

EARLY detection and routine monitoring play a crucial role decreasing the mortality rate of skin cancer. Survival rates decrease significantly if left to be treated in an advanced stage [1].

The ISIC challenges are split into 3 tasks. In this work we only focus on Task 3: Disease Classification. In previous ISIC challenges, classification tasks are comparably easy and less meaningful in a clinical setting. ISIC 2016 [2] required a binary decision of benign vs malignant. ISIC 2017 [3] presented 2 binary classification tasks. The first required a classification of melanoma vs. nevus and seborrheic keratosis, and the second required a classification of seborrheic keratosis vs. melanoma and nevus. The ISIC 2018 challenge [3] [4] demands a much more difficult 7 way classification task which is more representative of a real-world clinical scenario.

II. Dataset

In this work we combine multiple public and proprietary datasets shown in Table I. We map all suitable diagnoses from our additional datasets to the 7 diagnoses present in the HAM dataset: melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), and vascular lesion (VASC).

Each dataset is then split into 5 folds, stratified by diagnosis. For both the HAM10000 dataset [4] and one of our own proprietary datasets, lesion identifiers are available for each image. The lesion identifiers are considered during fold creation to ensure that no two images which belong to the same lesion are present in both training and validation folds. This prevents any data leakage in our validation set and ensures that classifier validation performance is more representative of test time performance.

TABLE I
Dataset Diagnosis Distribution

Dataset	Number of images
ISIC Archive	4163
HAM10000	10015
proprietary	33644

Tschandl et al. [4] state they perform manual changes to image histograms to correct for over/under exposure and undesired colour shifts. In our work we perform colour constancy on all of our additional images during training and testing as a preprocessing step using the Shades of Gray method proposed by Finlayson and Trezzi [5] with Minkowski norm p=6. We believe this preprocessing step is important to normalize images across separate datasets where methods of capture and age of images vary greatly [6].

III. Methods

We train a number of different networks with varying architectures separately as shown in Table III. All models have been initialized with weights obtained from pre-training on ImageNet [7]. The reported loss and balanced accuracy are from averaging the results of 5-fold cross validation.

We resize all images so the short side is $1.25\times$ larger than the input size. Next a random square crop with the size in [0.8, 1.0] of the resized image is taken and resized to the desired input size of the model. We perform random horizontal flips, random rotations of [0, 90, 180, 270] degrees and augment brightness, saturation and contrasts by a random factor in the range [0.9, 1.1]. All networks were trained in a similar manner to their original implementations, and we changed only the initial learning rate, the size of the last fully connected layer and the mean used for normalization.

We ensemble models with a stacking scheme [8]. For each model configuration and each of 5 cross-validation folds, we train a model and make out of sample predictions on the held out validation fold. For each model configuration and each training

A. Nozdryn-Plotnicki, J. Yap and W. Yolland are with MetaOptima Technology Inc., Vancouver, BC, Canada (Corresponding Author: jordan@metaoptima.com)

All authors contributed equally and are listed in alphabetical order

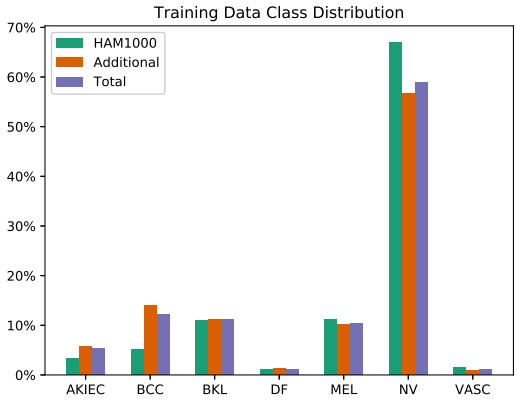
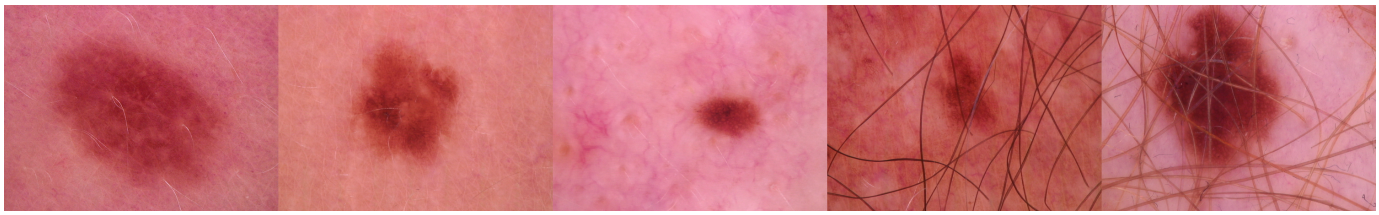
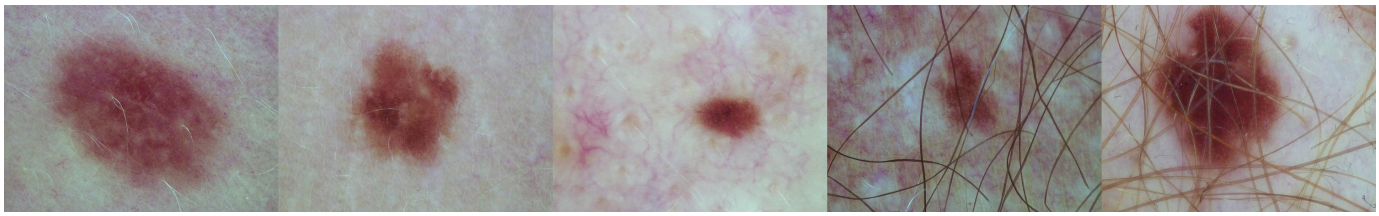


Fig. 1. Proportion of diagnosis in our dataset



(a) Before colour constancy



(b) After colour constancy

Fig. 2. Results of applying colour constancy to images found in the HAM10000 dataset.

image this yields a set of predicted probabilities for the 7 diagnoses in a relatively unbiased way. We treat these probabilities as features and train an XGBoost Classifier on them. Our final ensemble is comprised of our best performing models, along with a meta-model. The meta-model is a mean of predictions made by all models not selected for the ensemble weighted by the inverse of their cross validated log loss. We also include a one-hot encoded categorical feature which encodes the dataset of origin for each image. We re-calibrate the ensembler and change its prior. For test images, for each model configuration we average predictions made by each of the 5 models trained across the folds.

A. Metrics

The primary competition metric was "normalized multi-class accuracy metric (balanced across categories)" which we refer to as *balanced accuracy*. More precisely, this is the unweighted average of per-class accuracy values. Neural network classifiers estimate Bayesian a posteriori probabilities informed by a learned prior from the training distribution [9]. To maximize performance on the balanced accuracy metric, we believe that assuming a balanced prior is critical.

We address the issue of balanced priors in two ways. First, we experimented with training on balanced mini-batches (in expectation). Second, we trained networks with a weighted loss functions where weights are determined using the inverse frequency of classes in the training data. Both linear and logarithmic relationships among loss-weights were explored. Regardless of a models training scheme, we are able to adjust the final predicted probabilities with a change of prior derived using information from both the training datasets and the HAM10000 dataset [10].

Given a class c from a possible set C, we define a set of scores for each class as $s(C = c)$. For a given model, predictions for an input image X are taken as a set of class-wise probabilities $P_{CNN}\left(C = c \mid X\right)$. These probabilities are generated by a model which believes the distribution of data follows the probability distribution of the training set $P_{train}(C = c)$ for each class c. Given that the competition metric promotes a balanced evaluation among classes from C, regardless of prior dataset

TABLE II
MODELS USED IN ENSEMBLE

Model	Input Size	Loss	Balanced Accuracy
DPN-92(5k)	224×224	0.331	0.787
DPN-92(5k)	224×224	0.333	0.786
Resnet-152	224×224	0.333	0.770
Densenet-161	224×224	0.334	0.771
Inceptionv3	299×299	0.334	0.770
Inceptionv3*	299×299	0.359	0.757
seresneXt-50	224×224	0.345	0.774
ResNet-50	224×224	0.350	0.772
ResNet-34	224×224	0.356	0.762
ResNet-34	224×224	0.358	0.759
ResNet-50**	224×224	0.364	0.766
seresneXt-50†	224×224	0.366	0.793
seresneXt-50†	224×224	0.372	0.801
seresnet-50	224×224	0.393	0.720
ResNet-18	224×224	0.381	0.736
ResNet-50	224×224	0.437	0.721
ResNet-18‡	224×224	0.438	0.774
Squeezenet1.1	224×224	0.558	0.555
histogram	NA	0.797	0.323

* - froze first half
** - trained with partial proprietary data
† - weighted loss
‡ - trained on balanced data

distribution, we would like to correct the model intuition to maximize this metric by reweighting the prior. We do this using a balanced distribution $P_{balanced}(C = c)$ after prediction and are left with an adjusted set of scores $s_{adj}\left(C = c \mid X\right)$ for each image X.

$$s_{adj}\left(C = c \mid X\right) = \frac{P_{balanced}(C = c)}{P_{train}(C = c)} * P_{CNN}\left(C = c \mid X\right)$$

$$P_{adj}\left(C = c \mid X\right) = \frac{s_{adj}\left(C = c \mid X\right)}{\sum_{c \in C} s_{adj}\left(C \mid X\right)}$$

For a given model, we reweight predicted class probabilities by the ratio of a balanced frequency $P_{balanced}(C) = 1/count(C)$ to the training frequency $P_{train}(C)$. The reweighted probabilities are an updated score s_{adj} for each image, that when normalized act as a new set of class probabilities P_{adj}.

In addition to reweighted class probabilites, we perform significant test time augmentation. We center crop the resized image ($1.25\times$ larger than input as mentioned above) with proportional crops of [0.8, 0.9, 1.0]. For each crop we perform all 8 combinations of 90 degree rotations and horizontal flips leading to 24 network predictions per sample. We ensemble these 24 predictions with a class-wise mean. During validation stages in training, we simply take a center crop at 0.9 and resize to network input size.

TABLE III
TEST TIME AUGMENTATION (TTA) IMPACT

Model	Single Sample Loss	TTA Loss
ResNet-34	0.399	0.358
ResNet-152	0.359	0.333
DPN-92	0.362	0.331
Densenet-161	0.357	0.334
Inceptionv3	0.383	0.334

IV. EXPERIMENTS

All reported results are obtained from the mean over 5-fold cross validation on the HAM10000 training set unless stated otherwise. Our ensemble generates class probabilities for which we apply our change of prior and report. Throughout this paper when reporting single model performances we present un-adjusted scores unless stated otherwise.

Working purely with colour distribution was surprisingly effective. We were able to achieve a balanced accuracy of 0.511 with a change of prior using XGBoost with default hyper-parameters. The following colour features were used:

- Create a 16 bin colour histogram for each channel

- per-channel mean, median, min, max, variance, mode, skew, kurtosis
- 8 bin per-channel, 3D histogram
- 4 bin per-channel, 3D histogram
- min, max, min index, max index of each 3D histogram

V. Discussion

In this work we have shown how to obtain good performance for the task of skin lesion classification. It is important to note however that such a large ensemble of models is not at all practical for use in a real-world scenario. Every image has to be sent through all models to obtain predictions and the resulting predictions sent through the final ensemble which adds a significant amount of additional computation.

We would also like to highlight that obtaining a classifier that is close to or on par with the performance of dermatologists is only one part of a larger problem. Implementing and making such systems available for use by clinical practitioners in real-time also presents some interesting problems. For future challenges, adding constraints such as limiting memory usage or a limit on the number of FLOPS can make the work produced by participants more usable.

VI. Conclusion

The 2018 edition of the ISIC Task 3: Skin lesion classification challenge is more difficult and clinically relevant compared to the previous years. In this work we show how it is possible to obtain a high classification accuracy through ensembling multiple models trained for 7-class skin lesion classification. We note however that the approach taken in this work would not be suitable for a real-world use case.

References

[1] D. S. Rigel, J. Russak, and R. Friedman, "The evolution of melanoma diagnosis: 25 years beyond the abcds," *CA: a cancer journal for clinicians*, vol. 60, no. 5, pp. 301–316, 2010.

[2] D. Gutman, N. C. F. Codella, M. E. Celebi, B. Helba, M. A. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (ISIC)," *CoRR*, vol. abs/1605.01397, 2016.

[3] N. C. F. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. Mishra, H. Kittler, and A. Halpern, "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*, vol. abs/1710.05006, 2017.

[4] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, p. 180161, 2018.

[5] G. Finlayson and E. Trezzi, "Shades of gray and colour constancy," *IS&T/SID Twelfth Color Imaging Conference*, pp. 37–41, 2004.

[6] C. Barata, M. E. Celebi, and J. S. Marques, "Improving dermoscopy image classification using color constancy," *IEEE journal of biomedical and health informatics*, vol. 19, no. 3, pp. 1146–1152, 2015.

[7] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[8] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[9] M. D. Richard and R. P. Lippmann, "Neural network classifiers estimate bayesian a posteriori probabilities," *Neural computation*, vol. 3, no. 4, pp. 461–483, 1991.

[10] C. D. M. Saerens, P. Latinne, "Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure," *Neural Computation*, vol. 14, no. 1, pp. 21–41, Jan 2002.