Skin lesions: HAM 10000 dataset. BIODS220 Project.

Anastasia Butskova **anastasb@stanford.edu**

Abstract

We introduce the following techniques to tackle the problem of skin lesion diagnosis with deep learning: mixUp data augmentation, 2 types of self-supervised learning for pre-training the data, 2 stage EfficientNet and focal loss optimization as a proxy for the best average recall. We find the pre-trained EfficientNet with 2 prediction stages and the focal loss with $\gamma = 1$ to be the most effective set up for maximizing the average recall between the 7 classes in this problem.

1 Introduction

Skin cancer initial diagnosis is mostly performed visually by dermatologists, and combining it with an automated diagnosis produced by computer vision algorithms is therefore a logical step to make this diagnostics more efficient and potentially more reliable. Moreover, it would be highly desirable if there existed FDA approved mobile applications, which allowed skin lesion image uploads and preliminary diagnostics with the alerts of when one needs to see the doctor immediately.

This project aims at tackling the ISIC 2018 Challenge Part 3 dedicated to the Lesion Diagnosis prediction. This is a multi-label classification problem with 7 skin condition classes, including the cancerous and non-cancerous ones. While this problem has been tackled before, the individual recall scores are not yet maximized across all the classes for this particular data set (HAM 10000), with the less popular classes being more disadvantaged in this sense. Different schemes for pre-processing (random crop with various sizes, downsampling, multi-crop, ordered crop, hair removal from the images if any) and transfer learning with the classic architectures, such as GoogleNet Inception v3, DenseNet, ResNet, MobileNet, and also model ensembling have been used before, with quite promising results, sometimes beating the certified dermatologists' consensus.

In this project, we explored several more techniques for maximizing the average recall score between the seven classes. First, we experimented with the pre-processing techniques, mixUp in particular. Second, we used focal loss for optimization purposes, for which cross entropy is a subset of, and experimented with its hyperparameters γ and α-s to solve the imbalanced data problem, which is common in medical datasets, this one included. Moreover, we used a 2 stage EfficientNet pretrained network where we first performed a binary classification of the most popular class versus everything else, and then sequentially classified among the six classes left in case the image was not predicted to be the most popular class at the first stage. We also experimented with the two self-supervised approaches as a potential substitution for the pre-trained networks: 1) rotating the images by a certain angle (randomly chosen from 6 angle sizes) and predicting the angle of rotation (classification problem using the EfficientNet); 2) distorting the image by randomly swapping several patches and restoring it using a U-Net (Ronneberger, et al [15]) like architecture.

The best model in terms of the average recall and the overall accuracy is a 2 stage pretrained EfficientNet with $\gamma = 1$ used in the Focal Loss, while self-supervised approaches did not perform as well as the above.

Finally, we also applied Grad-CAM to the most efficient model to make the results more interpretable.

Preprint. Under review.

2 Related work

For this project we made use of the papers dedicated to the medical imaging, and skin lesion diagnosis in particular, and also non-medical methodology papers, which we deemed relevant for the problem being tackled.

Skin lesion diagnosis. There has been a number of deep learning papers in recent years dedicated specifically to the dermatology related imaging. Esteva, et al [2] use a large data set of nearly 130K images to apply a pre-trained Google Inception v3 architecture and then fine-tune it to perform two binary classifications of cancerous versus non-cancerous cases, and achieve very high results, as measured by AUC comparable to those from the tested expert dermatologists. Some papers (Gessert, et al [3]) experiment with pre-processing strategies aiming at capturing local information by using several patches per image and then applying their designed attention mechanism to also grasp the global information about the image. Patch selection techniques include random sampling and deterministic selection at particular coordinates (corners and the center).

Self-supervised learning for medical imaging. Chen et al [9] introduce a self-supervised approach for the model pre-training stage, where the original image is randomly distorted and then restored. The fine-tuning stage performs the classification (or regression task) with the initial parameters learnt during the image restoration. The authors use this technique to utilize the unlabelled data at the restoration stage; we will try using the above for labelled data only, in a hope that the model will be able to learn the image features better this way. We will also be applying the above approach (with some modifications) to the current dataset.

MixUp. MixUp (Zhang, et al [4]) is a useful augmentation technique that trains the network on the linear combinations of its images and their labels, which is also regularizing. We note that it is also applicable in the multi-class case, where we combine the one-hot-encoded pairs of labels and calculate the cross entropy loss as the weighted loss for the two targets.

Imbalanced data and the Focal Loss. While the above paper tackles the imbalanced data problem, which is relevant for our dataset, by using oversampling and loss weighting, we find the Focal loss (Lin, et al [5]) particularly useful for these purposes. The formula for the loss is $-\alpha_t(1-p_t)^\gamma \log p_t$, where α_t and γ are its hyperparameters, and p_t is a predicted probability for the true class. We note that the loss is equivalent to the cross entropy loss when $\gamma = 0$ and $\alpha_t = 1$ for all the classes. When $\gamma > 0$, examples which are confidently and correctly predicted, have a smaller loss impact, as compared to the cross entropy loss, while those classified incorrectly, have a higher weight in the loss. α_t role is standard, as it represents weights directly set by a user to artificially increase the less popular classes, or decrease the more popular ones.

EfficientNet. When experimenting with various architectures, we found EfficientNet (Tan, et al [6]) achieving a particularly high accuracy. This network successfully optimizes the compound effect of depth/width/resolution using a highly effective compound coefficient and builds up on top of MobileNets (Howard, et al [7]) and ResNet (He, et al [8]). Not only does it increase the accuracy, but also reduces the complexity by using MobileNet's depthwise separable convolutions. The technique applies filters across the depth dimension separately rather than simultaneously as in other CNN-s, and then uses 1x1 convolutions to establish the depth dimension for the next layer.

Model Ensembling. Finally, model ensembling has proven to be successful for many skin diagnosis imaging tasks, and the paper by Gessert, et al ([10]), which uses an ensemble of deep learning models including EfficientNets, SENet (Hu, et al [11]), and ResNeXt WSL (Xie, et al [12]), selected by a search strategy, achieves AUC results above 90% across all the eight classes they are predicting.

Grad-CAM. For neural network interpretation purposes, we find Grad-CAM (Selvaraju, et al [13]) to be of a particular interest, given its flexibility (can be applied to any layer in contrast to it predecessor CAM (Zhou et al [14])) and its self-supervised nature (no segmentation masks needed to obtain Grad-CAM heatmaps). This allows the medical models in particular to be less of a black box for physicians and patients by showing which areas of an image are the most important ones for a network to make a prediction.

3 Data

The data set contains 10,015 labeled images of size 450x600 (HAM 10000 data set) of seven skin decease classes: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma and vascular lesion (examples in Figure 1). It also contains information about the sex and age of the patients, which we are not likely to incorporate into the model at this point. The data is highly imbalanced, with more than 60 % of the images being of the NV: "Melanocytic nevus" class and some classes being extremely rare (less than 2%) (class frequencies presented in Figure 2). The goal of the project is to experiment with various approaches and achieve a higher result, as measured by the normalized multi-class accuracy metric.

Figure 1: Image examples

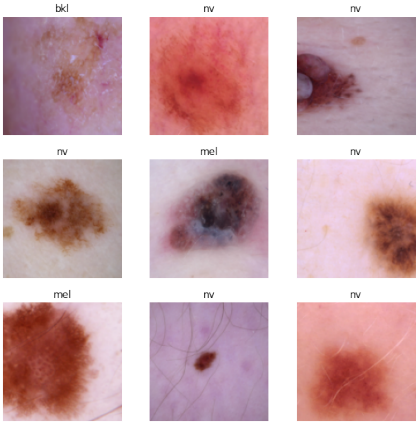
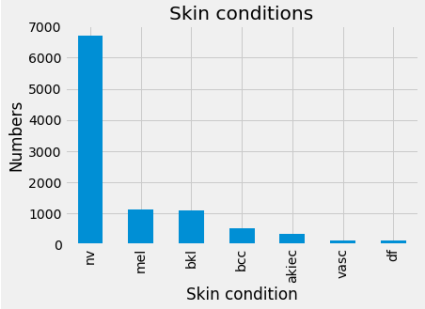


Figure 2: Class frequency



4 Approach

Base model. For the base model we pre-processed the images by random cropping to the 128x128 size (we also tried 256x256 size), and tried ResNet50 with unpretrained weights and focal loss gamma equal to 0 (equivalent to the cross entropy loss). The result obtained for the average recall was not high enough on the validation set (only 42%).

Further experiments. We then experimented with γ-s ranging from 0 to 5. We also added alphas, which are inversely proportional to the class frequency, to the focal losses, but did not see much improvement from the above. We tried downsampling and then added mixUp. This technique can be viewed as a regularization one, and allows for a better generalizaion of the trained model, as evidenced by the paper by Zhang, et al [4].

Then, we utilized transfer learning by using ResNet50 and EfficientNet b7 pretrained on Imagenette and using several Dense layers with ReLU activations on top of them. EfficientNet has significantly increased both the overall accuracy and the average recall.

Confusion matrices from EfficientNet made us realize that classes are mostly confused with the most popular class labelled 'nv'. Therefore, we created a modified 2-stage EfficientNet, which performs the binary classification of 'nv' versus all first, and then performs the classification between the rest of the classes based on the first stage result. This approach and also adding a discount α factor to the most popular class 'nv' allowed to improve both the accuracy and the average recall.

Self-supervised learning. We tried two self-supervised approaches with the intention to replace the pre-trained models with the self-supervised learning stage. The first approach was to first predict the rotation angle from the six angles (random rotations from the choice of siz angles were applied to all the images in the dataset for training). Cross entropy loss was used at this stage and EfficientNet was used as an architecture. We then performed the final classification (dense layer was added on top of the existing structure to classify between the seven classes) with the focal loss and $\gamma = 1$. We tried both sequential training of the two objectives and simultaneous training. The results for this self-supervised approaches were not high in terms of the average recall, and this is likely due to the fact that the images are mostly rotation invariant, i.e. it is a hard task to learn a rotation angle from the rotated skin lesion image, and it is probably not a good enough self-supervised task for such images.

The second self-supervised approach was using image context restoration block instead of a pre-trained model. The input to the model was the corrupted image, and the corruption was done as follows: we split each image into the non-intersecting patches of size 8x8 (we tried size 16x16 too) and randomly pick 2 of them, and swap them. We applied the below algorithm (Chen et al [9]), which does the above procedure, 20 times, and got a distorted image:

Algorithm 1: Image context disordering

Input: original image x_i

Output: image with disordered context $\tilde{x_i}$

for iter = 1,2..20 do:

randomly select a patch $p_1 \in x_i$

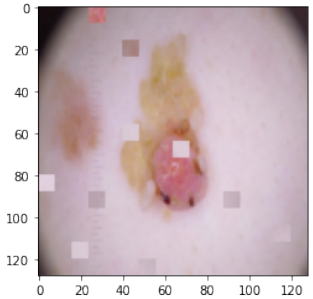
randomly select a patch $p_2 \in x_i$

$p_1 \cap p_2 = \emptyset$

swap p_1 and p_2

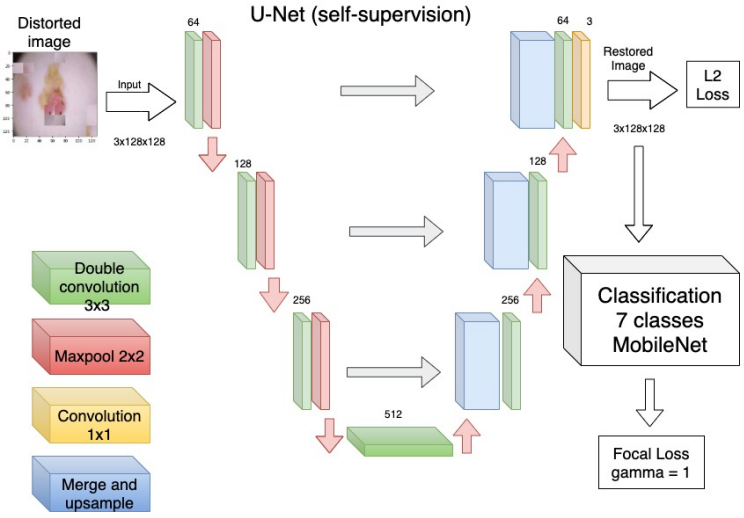
The first part of the training network was restoring the original image with the architecture similar to U-Net (shown below in Figure 4), while the second part of the network was predicting the skin lesion label using the MobileNet. There were two losses to minimize here: the L2 loss for the first part, which was minimizing the differences between the original image and the model-restored distorted image. Focal loss with $\gamma = 1$ was used for the second part of the network, aimed at the image labeling. The idea behind such a network is to allow the model to learn the context based information from the image restoration before performing the classification task. The idea was inspired by Chen's et al paper [9].

Figure 3: Distorted image example.



The results of the second self-supervised approach were better than for the first one, but still worse than that of the 2 stage pre-trained EfficientNet, and it is likely due to the fact that we did not train

Figure 4: Self-supervised learning: second approach.



the self-supervised part long enough, but were limited by the Kaggle kernel constraints. We also note that Chen et al [9] used unlabelled data to train the self-supervised part, but we only included the labelled data from the original dataset when training the model.

For interpretation purposes, we implemented Grad-CAM for our best model in order to have a better idea of which areas of an image are being used for making a prediction. We noticed that for the top losses (misclassified examples) the network was oftentimes ignoring the useful areas of the skin abnormality itself.

5 Experiments

All the below results are calculated on the validation set, which is 20% of the dataset and was chosen using stratified random sampling across the 7 classes. Training is run for 20 epochs with the Adam optimizer and a learning rate of 0.001. It was run for 30 epochs for the self-supervised parts in both of the self-supervised approaches when run sequentially. Test set result (on 1511 images) was obtained for the best model, which is the 2 stage EfficientNet, and has an overall test accuracy of 0.891 and the average recall of 0.784 between the 7 classes on the test set.

We start with the unpretrained models, where the base model is ResNet-50 with $\gamma = 0$ for the focal loss, which is in this case equivalent to the cross entropy loss. We experiment with different levels of γ, which is an additional penalty for the examples non-confidently classified, as compared to the ones classified confidently, and see how the average recall increases with the gamma being positive. However, average recall is still not high enough. Moreover, increasing the γ from 0 deteriorates the overall accuracy, as a lot of the dominant benign images become misclassified in favor of the less popular classes. We note that we also experimented with the size 256x256, but did not see a substantial improvement, as compared to the 128x128 image sizes. We also experimented with the alpha parameters of the focal loss and kept them as ones for all the unpretrained models (equal class weights).

We therefore proceeded with the pretrained models with the focal loss and $\gamma = 1$ (focal loss with such γ demonstrated the highest average recall in the unpretrained case). We note that we applied the mixUp augmentation technique when training both of the EfficientNet based models below. An example of mixUp, which is the linear combination of the two images, is shown below in Figure 5. For this picture, the two images are equally weighted for illustration purposes. We used weights 0.7 and 0.3 for the model though, as these weights are in the range of the most effective ones, as suggested by the relevant paper [4]. We combine the first two images to obtain the artificial image on the right. The loss for the mixed up images 1 and 2, with weights 0.7 and 0.3 respectively, is a

weighted loss of the mixed prediction with the target of image 1 and the target of image 2, so we can still use cross entropy or focal losses and the one-hot-representation for the targets.

Figure 5: mixUp example.

We compared the confusion matrices for the EfficientNet model and the 2 stage EfficientNet model. We saw that in the first case there is a lot of confusion happening between the class 'nv' and everything else. 2 stage EfficientNet improves this result, raising the average recall to 0.806.

Unpretrained (128x128 size):

Model name	Accuracy	Average Recall
ResNet-50, $\gamma = 0$	0.748	0.422
ResNet-50, $\gamma = 1$	0.540	0.561
ResNet-50, $\gamma = 1.5$	0.549	0.560
ResNet-50, $\gamma = 2$	0.534	0.531
ResNet-50, $\gamma = 5$	0.521	0.525

Figure 6: 2 stage EfficientNet Confusion Matrix.

	akiec	bcc	bkl	df	mel	nv	vasc
akiec	45	3	6	1	3	4	3
bcc	4	81	3	1	5	8	1
bkl	7	1	181	0	14	10	7
df	3	3	0	16	0	1	0
mel	3	1	7	0	158	47	7
nv	1	2	11	0	18	1304	5
vasc	0	0	0	0	0	1	27

Pretrained (128x128 size):

Model name	Accuracy	Average Recall
ResNet-50, $\gamma = 0$	0.862	0.753
Resnet-50, $\gamma = 1$	0.843	0.727
EfficientNet, $\gamma = 1$	0.901	0.800
2 stage EfficientNet, $\gamma = 1$	0.905	0.806

Self-supervised learning before performing the classification:

Model name	Accuracy	Average Recall
Approach 1	0.753	0.452
Approach 2	0.782	0.541

Figure 7: 2 Stage EfficientNet Train and Validation Losses

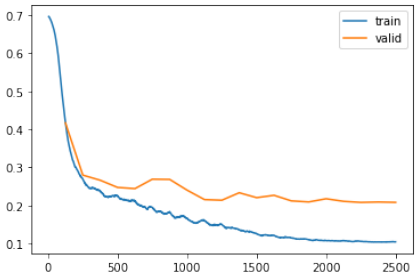
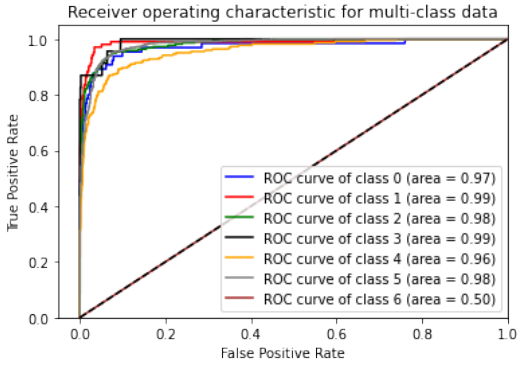


Figure 8: 2 Stage EfficientNet ROC curves for all the classes



We can see that the 2 Stage EfficientNet achieves the highest accuracy and average recall on the validation set. Its training process and ROC curves for all the classes are shown above (Figures 7 8). The seventh class has the lowest AUC and we can see from the confusion matrix below that while its recall is high, its false positive rate is also relatively high.

We show Grad-CAM examples for the EfficientNet below (Figure 9). The images on the right show the two of the top losses examples, where we can see that the areas impacting the network prediction (yellow) are not on the main tissue itself. For the correctly and confidently classified examples on the left, the yellow areas are covering the skin condition.

Figure 9: Grad-CAM examples.

Conclusion

We have experimented with several new approaches and found mixUp and focal loss's positive gamma hyperparameter beneficial for improving the average recall across the seven classes. We also found that if the dominant class is present, it is an efficient technique to first train the model to classify this class versus all, and then proceed with further classification, depending on the first

binary outcome. As for the self-supervised approaches, the second one with the image distortion and restoration performed better than the first one with the rotation angle prediction, as expected, as these images are mostly rotation invariant and the rotation angle is thus hard to predict. However, self-supervised learning for this labelled dataset did not beat the pretrained models. As for the further steps, we would like to try training the self-supervised part of the second approach more (for 90-100 epochs) and see if it can improve the result. We would also like to ensemble several models to improve the results in the future, and also extend our training database by adding more skin lesion images, labeled and unlabeled (unlabeled images to be used for the self-supervised part).

Acknowledgements

I was using some part of the code from the fastai website: `https://github.com/fastai/fastbook`. I was also using the efficientnet implementation in pytorch found on github. I was partially using this colab file: `https://colab.research.google.com/drive/1DS6WJPiH7rc2AsIwmGpz1y-t8zLBGN56?usp=sharing#scrollTo=UlNLS_swIzch`.

The rest of my code can be found on Canvas.

References

[1] `https://challenge2018.isic-archive.com/task3/`

[2] Esteva, A., Kuprel, B., Novoa, R., Ko, J., Swetter, S., Blau, H. & Thrun, S. (2017) *Dermatologist-level classification of skin cancer with deep neural networks.* Nature 542, 115–118.

[3] Gessert, N., Sentker, T., Madesta, F., Schmitz, R., Kniep, H., Baltruschat, I., Werner, R., & Schlaefer, A. (2019) *Skin Lesion Classification Using CNNs with Patch-Based Attention and Diagnosis-Guided Loss Weighting.* arXiv: 1905.02793.

[4] Zhang, H., Cisse, M., Dauphin, D. & Lopez-Paz, D. (2018) *mixup: BEYOND EMPIRICAL RISK MINIMIZATION.* arXiv:1710.09412.

[5] Lin, T-Y., Goyal, P., Girshick, R., He, K. & Dollar, P. (2018) *Focal Loss for Dense Object Detection.* Facebook AI Research (FAIR).

[6] Tan, M. & Le, Q. (2019) *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.* arXiv:1905.11946v5.

[7] Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017) *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.* arXiv:1704.04861v1.

[8] He, K., Zhang, X., Ren, S. & Sun, J. (2015) *Deep Residual Learning for Image Recognition.* arXiv:1512.03385v1.

[9] Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M. Rueckert, D. (2019) *Self-supervised learning for medical image analysis using image context restoration.* Medical Image Analysis, Volume 58, 101539.

[10] Gessert, N., Nielsen, M., Shaikh, M., Werner, R. Schlaefer, A. (2020) *Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data.* MethodsX Volume 7, 100864.

[11] Hu, J., Shen, L., Albanie, S., Sun, G. Wu, E. (2017) *Squeeze-and-Excitation Networks.* arXiv:1709.01507v4.

[12] Xie, S., Girshick, R., Dollar, P., Tu, Z. He, K. (2017) *Aggregated Residual Transformations for Deep Neural Networks.* arXiv:1611.05431v2.

[13] Selvaraju, R., Cogswell, Das, A., Vedantam, R., Parikh, D., Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. arXiv:1610.02391.

[14] Zhou, B., Khosla, A., Oliva, A., Torralba, A (2016). Learning Deep Features for Discriminative Localization. arXiv:1512.04150v1.

[15] Ronneberger, O., Fischer, P., Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597.