Novel Knowledge Distillation Techniques for Visual Skin Disease Detection

Rohan Davidi
Stanford University
Department of Computer Science
rohand25@stanford.edu

Rachel Park
Stanford University
Department of Computer Science
rachpark@stanford.edu

Abstract

This paper explores the application of novel knowledge distillation techniques to improve the performance of lightweight convolutional neural network (CNN) models for skin lesion classification. Motivated by the substantial clinical potential of smaller, more efficient skin lesion classifiers, we evaluate the performance of knowledge distillation methods to transfer knowledge from a high-capacity EfficientNet-B7 teacher model to a smaller MobileNetV2 student model on the highly imbalanced ISIC 2019 dataset. We focus on exploring logit-based distillation techniques, specifically KL divergence loss and the correlation-based DIST loss, as well as curriculum learning as a promising augmentative distillation method. We evaluate and compare different configurations of these frameworks on the task of skin lesion classification by assessing the accuracy of their class predictions. Our experiments show that certain configurations of logit-based knowledge distillation, particularly those using KL divergence loss, improve skin lesion classification performance of our baseline student model. Moreover, we found that a curriculum learning distillation approach, which involves re-ordering the input data to the student model based on the teacher's prediction confidence, generally degrades the performance of our baseline student model. Our findings suggest that logit-based knowledge distillation can enhance model performance, and the choice of distillation technique and loss weights are critical to achieving optimal results for this task.

1. Introduction

Skin diseases constitute the fourth leading non-fatal burden worldwide, impacting approximately 1.9 billion individuals [12]. Early detection is essential to preventing complications and improving outcomes, especially as many of these diseases may progress to the most life-threatening and least predictable forms of skin cancer. However, a significant challenge in achieving timely diagnosis and treatment is the global shortage of healthcare professionals. Moreover, the diversity and similar symptomatology of skin diseases makes the task of characterizing disease complex and time-consuming for even experienced dermatologists.

In response to these challenges, there is a growing demand for computational tools that can accurately diagnose skin diseases and improve access to vital medical care. Deep neural networks have shown considerable promise, with remarkable achievements being shown for this task with complex, large-scale CNNs such as ResNet [9] and EfficientNet [23] [3] [6] [7] [21]. These research works also indicate that sophisticated CNNs with more layers and blocks tend to have better predictive performance with sufficient training data. While these large models can allow us to obtain better prediction results, they also require significant computing resources for both training and inference. This is impractical for resource-constrained clinical environments, and an ideal world where mobile devices extends the reach of dermatologists beyond the clinic. The development of light-weight skin lesion classification models that can be deployed on mobile devices thus holds significant clinical potential. These models could provide more accessible and efficient diagnostic capabilities, expanding the reach of dermatologists and improving patient outcomes in underserved areas.

Knowledge distillation (KD) is a promising strategy for model compression and acceleration, achieved by transferring knowledge from a large, complex ”teacher” model to a smaller, more efficient ”student” model. Formally popularized in 2015 by Hinton et al.[10], KD is motivated by the idea that large-scale teacher models are able to understand a richer representation of the training dataset than the lighter student models, and the student model can derive some extent of the teacher's representation by being trained on a combination of the original dataset and the soft labels of the teacher model. In practice, KD involves minimizing a loss function that balances the traditional cross-entropy loss on the true labels with a distillation loss on the teacher's soft labels. This enables the student model to achieve performance comparable to the teacher model while being significantly smaller and faster, making it suitable for deployment in resource-constrained environments.

In this paper, we intend to achieve performance improvements on skin disease classification with lightweight models designed to be trained and inferenced on mobile processors or other low-resource technology. In particular, we explore the use of state-of-the-art knowledge distillation techniques and evaluate a diverse set of model types and hyperparameters. We aim to find an optimal ensemble of techniques that improves the performance of lightweight models for accurately classifying the skin disease present in skin lesion image inputs.

2. Related Work

Our work can be seen as an extension of work done using CNNs for the task of skin lesion classification. Estava et al. [6] used transfer learning with a pre-trained CNN to classify benign skin lesions from malignant melanomas, outperforming dermatologist discrimination rates. In particular, deep, largle-scale CNNs such as the EfficientNet models have been shown to perform well as skin lesion classifiers on the ISIC 2018 and 2019 datasets [3] [7] [21]. For instance, Gessert et al. [7] put together a multi-resolution ensemble of EfficientNet models with loss balancing, achieving a balanced multiclass accuracy rate of 63.4% on the ISIC 2019 leaderboard. Sun et al. [21] applies data augmentation with a similar model ensemble as Gessert's to achieve an accuracy rate of 66.2% on the 2019 leaderboard.

While the above research clearly demonstrates that deep learning and CNNs are the preferred technique for skin lesion image classification and investigates the development of large-scale models for the task, how to optimize the task performance with lighter weight models is a challenge for real-world deployment that remains comparatively understudied. There is a growing body of literature that investigates and proposes novel techniques for knowledge distillation as a model compression method. The seminal paper by Hinton et al. [10] suggests a logit-based knowledge distillation method that trains the smaller model by exactly aligning its logits with the teacher's, i.e. minimizing the Kullback-Leiber (KL) divergence between their logits. Huang et al. [11] investigates optimizing this loss function during the knowledge distillation process, demonstrating improved performance with a correlation-based loss dubbed DIST loss on popular image recognition benchmarks including ImageNet [5] and COCO [14]. Other papers have also investigated improving the knowledge distillation data transferred to the student using another machine learning technique called curriculum learning, in which a model is taught by using easy samples firstly and gradually adding more difficult ones as opposed to random order [1]. Researchers have investigated various ”difficulty” proxies in curriculum learning during the training process, with some using methods as simple as sorting the inputs by magnitude [1] and others going as far as to train a separate discriminator model [26]. For instance, Zhu et al. [26] demonstrates that integrating curriculum learning into the knowledge distillation framework improves performance by using an adversarial trained discriminator for measurement of difficulty with respect to the original classification task loss. To the best of our knowledge, it remains unexplored how effectively curriculum learning can be used for lightweight models in skin lesion classification. As such, we focus on comparing various state-of-the-art knowledge distillation methods for this task.

3. Data

We train our models on the International Skin Imaging Collaboration Challenge Dataset of 2019 (ISIC2019), a publicly available repository comprising of 25,331 dermascopic images. The ISIC2019 training dataset consists of several dermoscopic image databases: BCN_20000 [4], HAM100000 [24], and the MSK dataset [3]. Its aim is to improve the diagnosis and treatment of melanoma among nine different diagnostic categories: melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (solar lentigo/seborrheic keratosis/lichen planus-like keratosis) (BKL), dermatofibroma (DF), vascular lesions (VASC), and squamous cell carcinoma (SCC). Examples from each of these different categories can be seen in Figure 1.

The dataset was then split into a training set, validation set, and testing set with a ratio of 70:20:10 so that the each split contained 17,728 images, 5,062 images, 2,541 images respectively. We resized each of the images to a dimension of 224x224 pixels. We then normalized the dataset based on the means and standard deviations of the ImageNet dataset [5], which our pre-trained models were originally trained on, to help stabilize and speed up the training process. Upon analysis, we determined that this dataset presents a unique challenge for our task because it has a severely imbalanced distribution across the different classes (Fig. 2). To mitigate the effects of the imbalance on our model performance, we used a weighted random sampler during training to ensure that each batch contains a balanced representation of classes.

4. Methods

4.1. Baseline Models

4.1.1 Student Model

In order to assess and employ the use of knowledge distillation as a means to provide high performance under low-resource settings, we chose to use a MobileNetV2 base model pretrained on the ImageNet training data for image classification [20]. To this base we attach a trainable

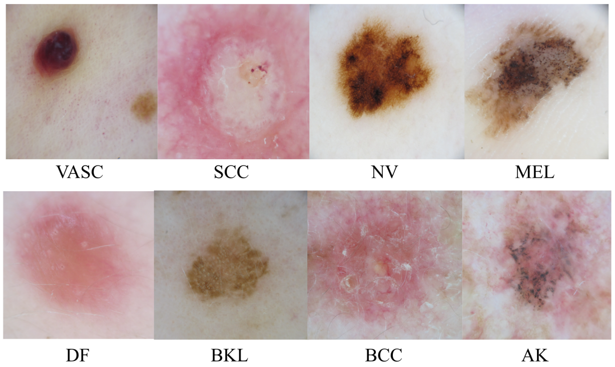

Figure 1. Examples of different skin lesions

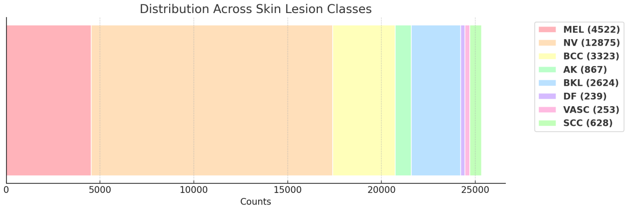

Figure 2. Distribution of training dataset across skin lesion classes

dense layer from MobileNetV2-ImageNet's 1000-size output layer to the desired class size of 8. The final model architecture contains 2,234,120 trainable parameters.

4.1.2 Teacher Model

After trialing the use of various large CNN model architectures (ResNet50V2, ResNet101V2, and EfficientNet-B7) via cross-validation, we chose to use a PyTorch implementation of EfficientNet-B7 pretrained on ImageNet image classification [17] [23]. Identical to the student model, to this base we attach a trainable dense layer to the desired class size of 8. The final model architecture contains 63,807,448 trainable parameters. We chose to restrict our search for a teacher architecutre to CNNs in order to maintain structural similarity to the MobileNetV2-based student model. We highlight the size of the models trialed to emphasize fit with the problem statement for applying techniques for small model improvement.

4.2. Logit Distillation

The specific form of knowledge distillation adopted is logit distillation. Logit distillation leverages the pre-softmax logits as representations of knowledge from a fully-trained teacher model. These logits are used in the training process to encourage the lighterweight student model's learning of similar output distributions.

Model	Trainable Parameter Count	Total Size (MB)
MobileNetV2 Student	2.2M	8.52
ResNet50V2 Teacher	24.8M	94.64
ResNet101V2 Teacher	42.8M	163.32
EfficientNetB07 Teacher	63.8M	243.41

Table 1. Sizes of Considered Model Architectures

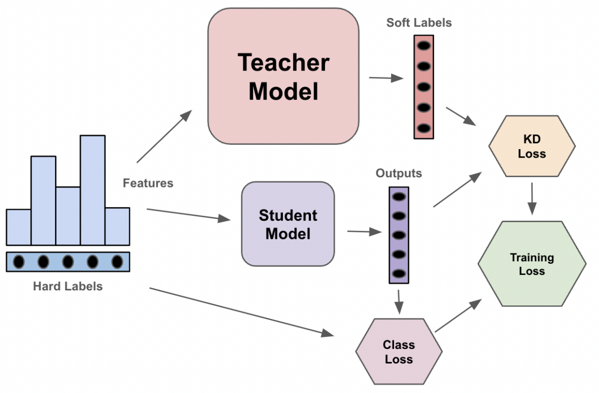

Figure 3. Logit Distillation Method

The specific approach we employ was to first train the teacher model to convergence on the training dataset using the standard cross-entropy loss for classification tasks. Subsequently, as seen in Figure 3, the student model's parameters were optimized by minimizing a hybrid objective function comprised of two components: the conventional classification loss computed using ground truth labels L_{CLS}, and an auxiliary knowledge distillation term L_{KD} quantifying the discrepancy between the student's output logits and the logits outputted by the trained teacher model. This distillation term imposes a regularization effect, encouraging the student to emulate the teacher's output distribution over class labels.

An alternative source of knowledge transfer is the use of feature distillation. This depends on each model sharing some apparent embedded feature representation that is comparabale such that the discrepancy of these as opposed to the logits could be used for the KD loss term. As our chosen architectures does not give way to this, we proceed with investigating the use of logits.

In order to rigorously assess the capabilities of logit distillation, we chose to explore the use of both of the two most commonly employed state-of-the-art distillation loss techniques Kullback-Leibler Divergence Loss and DIST Loss. With either choice of knowledge distillation loss, the final loss of the student is computed over each batch using a weighting of α for KD loss and $1 - \alpha$ for the classification loss as described in Figure 4. To ensure a rigorous survey of logit distillation for our given task, we experiment across a wide range of α values for evaluating the use of both loss KD loss functions mentioned.

$$L = \alpha \cdot L_{KD} + (1 - \alpha) \cdot L_{CLS}$$

Figure 4. Logit Distillation Loss Calculation

4.2.1 Kullback-Leibler Divergence Loss

Our first approach to logit distillation puts forward a Kullback-Leibler Divergence loss function to represent discrepancy between student and teacher knowledge. Kullback-Leibler (KL) Divergence is a statistic derived from information theory used to quantify the divergence between two probability distributions [13]. Utilizing the softmax function, the student and teacher logits can easily be converted and interpreted as a probability distribution (often done in classification tasks) making KL Divergence a natural fit. Specifically, in training the student model, we softmax the teacher's outputs for he given batch producing soft targets or a "true distribution" Q. We then apply softmax, again, to the student's outputs for the batch and produce the model distribution (P). After doing so, we are able to evaluate function described in Figure 5 which computes the logarithm of the outputs for each of these classes and sums their differences weighted by the classes student probability (P) producing our knowledge distillation loss L_{KD}.

4.2.2 DIST Loss

Our next approach is a correlation-based metric DIST loss introduced as a direct remedy for potential shortcomings of loss functions that are unbounded in their measure of student-teacher discrepancy [11]. More specifically, loss functions such as KL Divergence can produce potentially irreversably large loss values as it is an unbounded function weighted by the logits themselves. When using a potentially significantly stronger teacher, a term such as $\log\left(\frac{P(x)}{Q(x)}\right) = \log P(x) - \log Q(x)$ can potentially demand a magnitude of similarity that is too high to assume. Thus, DIST loss is an alternative for such an outcome that sacrifices this finer-grain discrepancy analysis for each class by opting to preserve the relational similarity of the distribution through use of interclass and intra-class correlation. Particularly, the function,, calculates the interclass loss L_{inter} as 1 minus the mean Pearson correlation between the student and teacher outputs. Similarly, the function encourages similarity to the teacher output's relations L_{intra} between classes by using 1 minus the mean Pearson correlation of the transposes of these outputs. These components are summed as seen in Figure 6.

$$L_{KD} = D_{KL}(P \| Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

Figure 5. KL Divergence-Based Loss

$$L_{KD} = \beta \cdot L_{inter} + \gamma \cdot L_{intra}$$

Figure 6. DIST-Based Loss

As suggested by the experimentation in DIST's founding paper, for CNN architectures assessed in image classification, the weightings of these two components (β, γ) are initialized to 1.

4.3. Teacher-Ranked Curriculum Learning

In addition to our survey of logit distillation, we propose the enhancement of knowledge distillation with a teacher-involved curriculum learning procedure. In order to mitigate the potentially complex information produced by the incorporation of transferring teacher knowledge, we motivate training in an order of increasing difficulty. Specifically, we examine use of the teacher's logit-derived confidence as a measure of difficulty.

To do so, the training data is unloaded from the loader directly, to account for the use of any custom sampling in the data loader, at which point the trained teacher model computes logits and its softmax scores for each example. Measuring the model "confidence" in a prediction as the probability of the predicted class we sort the training data and reload it for student training. This setup is visually depicted in Figure 7 in which the connection from the teacher to the student's input data in red marks our confidence sorting.

We view this as a potential for further leveraging the teacher's understanding capabilities. As the logit distillation procedure alters the loss function, we consider this representation of the complexity of matching teacher knowledge as opposed to typical curriculum learning techniques focusing on optimizing for measures of classification loss convergence difficulty.

Accounting for change in behavior dependent on the loss function, this method is applied in experimentation with a range of α weighting values and both loss functions to ensure thoroughness.

5. Experiments

5.1. Model Setup

All experiments were trained using the student model architecture detailed in methods: MobileNetV2 with a dense

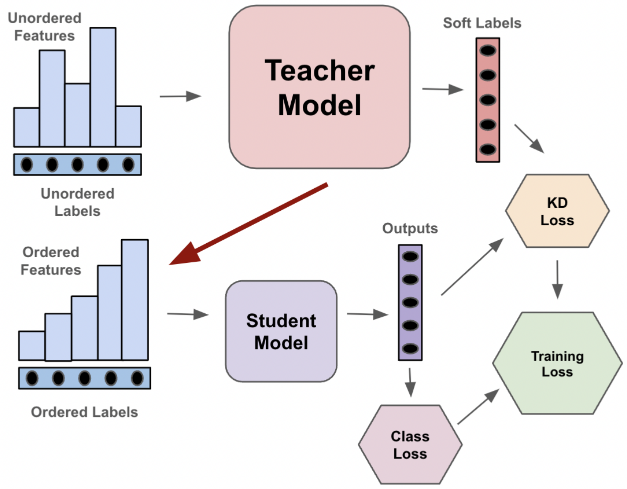

Figure 7. Curriculum Learning using Teacher-Based Reordering

output layer. By training on our preprocessed ISIC training data and cross validating with the validation split (5 folds) of the dataset, we experimented with the use of various training setups and combinations of trainable/untrainable base layers. In doing so, we noticed the need for substantial unfreezing of the base model for adequate fitting as well as a reduced learning rate with weight decay regularization to avoid overfitting. Ultimately, we found a fully unfrozen base model, learning rate of 1e-5, and Adam optimizer with 0.01 weight decay to produce the most promising configuration for all experiments. We found the same configuration to be appropriate for the training of the EfficientNet-B07 teacher model as well.

5.2. Experimental Setup

Ensuring consistency, all experiments are run using an Adam optimizer, batch size of 32, and train for a maximum of 25 epochs. Early stopping is implemented with a patience of 5 corresponding to the maximum number of consecutive epochs without a new lowest validation accuracy. We employ a classification loss of categorical cross entropy for all experiments as it is the standard for multi-class classification and this particular dataset.

5.3. Evaluation Metrics

Our quantitative evaluation consists of two primary metrics. The first one is accuracy on the classification task represented as a raw percentage (correct classifications / total examples). Serves as straightforward assessment of quality of model when used in application. However, accuracy is not invariant to class imbalance and can often be inaccurate as a representation of the quality of model understanding.

The second, which allows for a deeper and more balanced measurement, is the weighted F1 score. Use of F1 score provides robustness to class imbalance as it incorporates both reliability in a certain positive classification (precision) and ability to capture all relevant classifications (recall) in balance with one another. For each class, precision is calculated as the proportion of True Positive classified examples out of the total classifications (True Positives + False Positives). Recall is calculated, in each class, as the proportion of correctly identified classifications (True Positives) out of the total possible correct classifications (True Positives + False Negatives).

The F1 score for each class is calculated as the harmonic mean of these two quantities scaled by 100. We calculate the weighted average (by class counts in test data) of class-wise F1 scores to combine these results while accounting for the class imbalance which we know to exist in the data.

5.4. Results

In terms of F1 score and accuracy accuracy, based on Table 1, we find the best performing model is the EfficientNet-b07 Teacher Baseline. Amongst the MobileNetV2 student models, we observe the use of KL Divergence loss with a KD-loss weight of 0.1 produces the highest F1 and accuracy scores. Across all distillation weightages α, KL Divergence loss without Curriculum Learning outperforms all other model types including DIST distilled models and is the only method which outperforms the MobileNetV2 Student baseline (by 0.85 F1 and 0.93% accuracy). Amongst DIST Distilled models, we notice the model performance is largely invariant to the α_{DIST} weight. On the other hand, KL Divergence distillation notices a decrease in performance as the value of α_{KL} is increased dropping by more than 1.5% from $\alpha_{KL} = 0.1$ to $\alpha_{KL} = 0.9$. Taking a closer look at the performance of KL Divergence distillation in Table 3, we notice KL distillation performance reaches a local maxima around $\alpha_{KL} = 0.1$ with lowered performance for lower and higher values even observed in a finer grain level. Similar trend can be seen when Curriculum Learning also applied as the best performance is reached with $\alpha_{KL} = 0.1$.

From Figure 8, we also note a significant drop in performance with the use of curriculum learning for both DIST and KL Divergence loss methods. Amongst the two, for all loss weightage levels other than $\alpha = 0.5$, the respective performance of the distillation models follow the F1 performance order from the best to worst of KL Distillation without CL, DIST Distillation without CL, DIST Distillation with CL, followed by DIST Distillation with CL.

By examining the class-wise F1 scores in Figure 9, it becomes apparent there still exists fairly significant variance in performance based off of class as the two largest subclasses of the dataset (NV and BCC) are the classes on which all of the method's best performing models produce the highest F1 score.

Model	F1	Accuracy
EfficientNet Teacher Baseline	**77.646**	**77.33**
MobileNetV2 Student Baseline	72.659	71.84
KL Distilled ($\alpha_{KL} = 0.1$)	**73.508**	**72.77**
KL Distilled ($\alpha_{KL} = 0.3$)	72.310	71.55
KL Distilled ($\alpha_{KL} = 0.5$)	72.009	71.15
KL Distilled ($\alpha_{KL} = 0.7$)	71.551	70.76
KL Distilled ($\alpha_{KL} = 0.9$)	71.986	71.19
DIST Distilled ($\alpha_{DIST} = 0.1$)	71.292	71.43
DIST Distilled ($\alpha_{DIST} = 0.3$)	71.351	71.47
DIST Distilled ($\alpha_{DIST} = 0.5$)	71.129	71.09
DIST Distilled ($\alpha_{DIST} = 0.7$)	71.292	71.07
DIST Distilled ($\alpha_{DIST} = 0.9$)	70.665	70.05
KL Distilled + CL ($\alpha_{KL} = 0.1$)	68.508	67.18
KL Distilled + CL ($\alpha_{KL} = 0.3$)	66.728	66.71
KL Distilled + CL ($\alpha_{KL} = 0.5$)	66.801	66.63
KL Distilled + CL ($\alpha_{KL} = 0.7$)	66.896	66.39
KL Distilled + CL ($\alpha_{KL} = 0.9$)	66.840	66.31
DIST Distilled + CL ($\alpha_{DIST} = 0.1$)	68.368	67.18
DIST Distilled + CL ($\alpha_{DIST} = 0.3$)	68.907	67.77
DIST Distilled + CL ($\alpha_{DIST} = 0.5$)	66.140	65.80
DIST Distilled + CL ($\alpha_{DIST} = 0.7$)	69.266	68.36
DIST Distilled + CL ($\alpha_{DIST} = 0.9$)	69.041	68.12

Table 2. F1 scores and Accuracies comparing best performing KL Distilled and DIST Distilled student models with and without Curriculum Learning (CL) in comparison with EfficientNet-B07 Teacher Baseline and MobileNetV2 Student Baseline. Additional comparison of KL Distilled and DIST Distilled student models with varying α_{KL} and α_{DIST} values.

Model	F1	Accuracy
KL Distilled ($\alpha_{KL} = 0.05$)	73.322	72.49
KL Distilled ($\alpha_{KL} = 0.075$)	73.452	72.77
KL Distilled ($\alpha_{KL} = 0.1$)	**73.508**	**72.77**
KL Distilled ($\alpha_{KL} = 0.125$)	73.270	72.45
KL Distilled ($\alpha_{KL} = 0.15$)	73.111	72.33

Table 3. F1 scores and Accuracies for KL Distilled student models with more fine-grained α_{KL} values.

5.5. Discussion

5.5.1 Quantitative Analysis

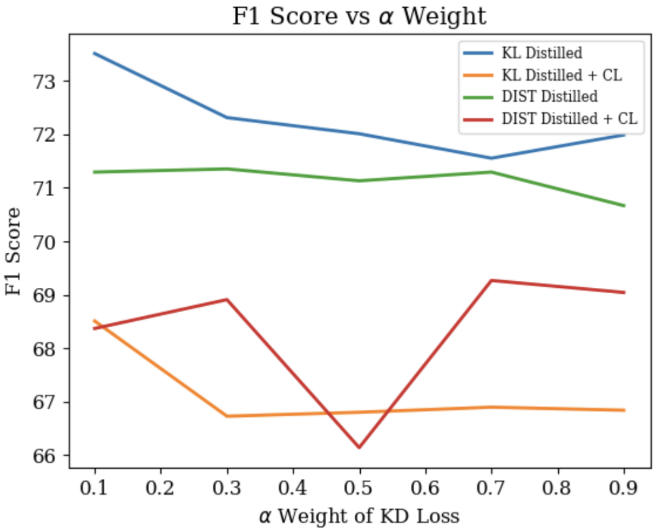

Figure 8. F1 scores across weightages of the KD loss term for KL Divergence Loss and DIST Loss with and without Curriculum Learning.

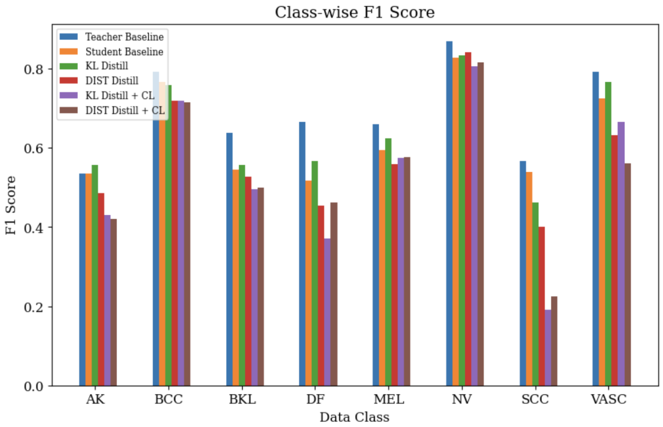

Figure 9. F1 scores by class for best performing model of each distillation method and use of CL compared with MobileNetV2 Student Baseline and EfficientNet-B07 Teacher Baseleine.

From our evaluation metrics, we understand it is possible to improve the F1 and accuracy performance of a light-weight model (MobileNetV2) with the use of knowledge distillation for skin lesion classification. We find that the combination of methods and parameters that produces this improvement is the use of KL Divergence loss weighted $\alpha_{KL} = 0.1$ without the use of curriculum learning. We find, though, that this added performance is not extremely significant as it produces an increased F1 of 0.85. However, as this is still not a statistically insignificant result, the use of KL Divergence loss in knowledge distillation holds potential promise for our said task.

Similarly, we find that no model with integration of DIST-based logit distillation produce F1 or Accuracy performance better than the base MobileNetV2 Student model. This in combination with the relative invariance to the loss weighting leads us to believe that for our given task, the model capabilities of the MobileNetV2 might have neared its capacity for the skin lesion task and dataset.

On the other hand, the consistent F1 out performance of DIST loss-based methods by KL Divergence suggests the direct comparison of logits is more effective in matching the teacher's output than the inter/intraclass relations for the different classes of skin lesions. This is understandable as intraclass DIST-loss weights the maintenance of intraclass relations which may not be prevalent in skin lesions as it is unlikely there exists a strong hierarchy of proximity of these classes with one another. In comparison a task such as identifying a living creature would have a clear relation in knowledge proximity to other objects.

Another explanation is that distillation losses have a scale different to the classification loss making the linear combination with this set of weight parameters unmoving. However, upon observation of component loss curves this is not the case.

Regarding Teacher-ordered Curriculum Learning, we find that there exists consistent deprecation of F1 and accuracy performance with the use of this preprocessing strategy. Upon observing the variance that exists in the Class-Wise F1 scores, we anticipate it is the case that the Teacher model exhibits imbalances in its prediction probabilities across examples. In other words, the ordering based on the confidence of the Teacher may have produced one that starts by teaching the student batches consisting mainly of a subset of the total classes. This leads to low generalizability in the early stages of training which would explain a deprecation of performance. We find this consistent with the inheritance of a higher class-wise F1 variance in models using curriculum learning as seen in the results.

5.5.2 Qualitative Analysis

Across our experiments, we identified two common themes of classification errors to analyze. A common error across all the models was the misclassification of AK as BCC or BKL. This makes sense given that the three classes of skin lesions are pretty visually similar (Fig. 1), and the AK class has very low representation in the training dataset relative to both the BCC and BKL classes. We also observed more confusion and lower true positives in the classes SCC, VASC, and DF, indicating that these models may need more data for these classes. This makes sense given that these are the three classes with the lowest representation in our training dataset. These behaviors all imply that even with weighted sampling, our student model did not have the capacity to build a rich and complex enough representation for the classes that were severely underrepresented in the training data. Most of our classification errors fell into this category, highlighting the persistent challenge with training a lightweight model on an extremely imbalanced dataset.

We also noticed that our best student model trained with KL Divergence Loss had the same most frequent misclassification as our teacher model—classifying NV as DF—with 1015 and 1077 errors respectively (Fig. 10). This particular misclassification is not prominent in our baseline student model and our best student model trained with DIST loss (Fig. 11 in Appendix). This would make sense because KL Divergence loss demands a perfect match of the logits between the student and teacher models that leads to the student model inheriting a representation very similar to the teacher model, whereas DIST loss has a more cumulative and softer loss function that allows the student to factor in intra- and inter-class relations along with the representation learned by the teacher model. Overall, this analysis highlights areas in which characteristics of our training setup, as well as our knowledge distillation techniques, may introduce inherent limitations to our model performance.

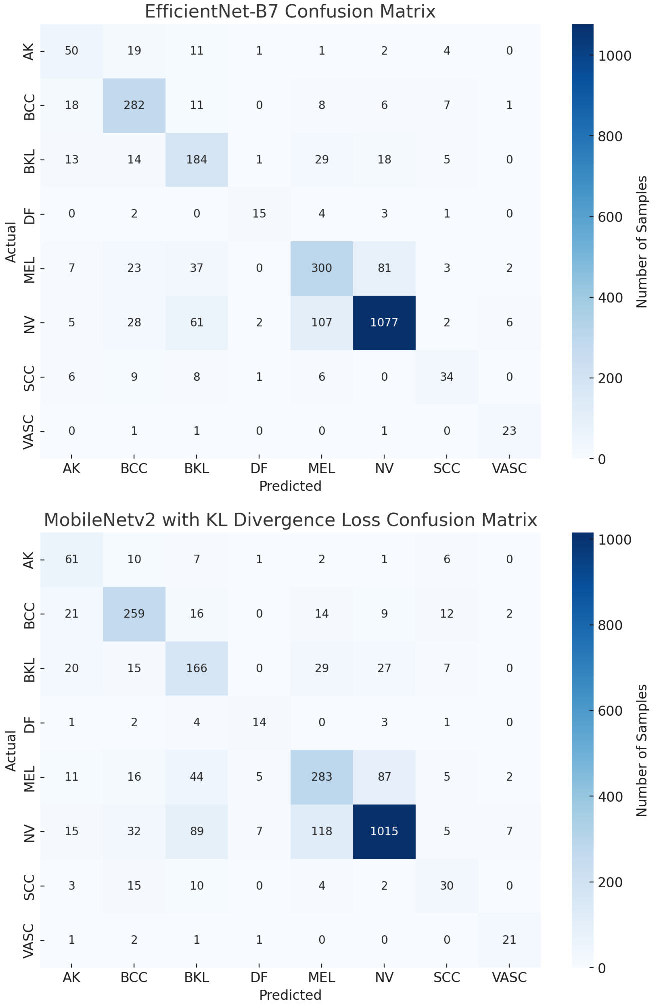

Figure 10. Confusion matrices for the EfficientNet-B7 Teacher Baseline and the best performing model of the knowledge distillation method with KL Divergence Loss.

6. Conclusions

This project rigorously examines the use of knowledge distillation for skin lesion classification. This investigation results in determining there exists configurations of Logit Distillation, particularly, the use of KL Divergence in linear combination with Categorical Cross Entropy Loss to improve performance of small scale models. We are able to show the particular weighting of these losses that produces accuracy and F1 higher than their base models is a 0.1 to 0.9 weighting of L_{KL} and L_{CLS}. In comparison, the use of DIST-based logit distillation is determined to not produce any delectably better knowledge for this task across various weightings and configurations. Similarly, we find that the use of Curriculum Learning based on the use of teacher model's logit confidence leads to deteriorating performance in this task.

We anticipate strategies that assist in making the task more so within the student model's capacity such as random cropping of the image given the nature of skin lesions could produce more pronounced benefit. Further, we would suggest the exploration of feature distillation through use of alternative teacher-student model architectures such that the underlying embedded features are measured for discrepancy as opposed to outputs. Furthermore, the use of an alternate ordering method such as an adversarial-trained discriminator model could be of use to improve these method performances with curriculum learning.

7. Appendix

8. Contributions and Acknowledgements

R.D. designed the experiments, implemented the knowledge distillation and curriculum learning methods, and conducted training evaluations. R.P. conducted the literature review, implemented the training pipeline, and performed the experiments. R.D. and R.P. analyzed the results and wrote the paper.

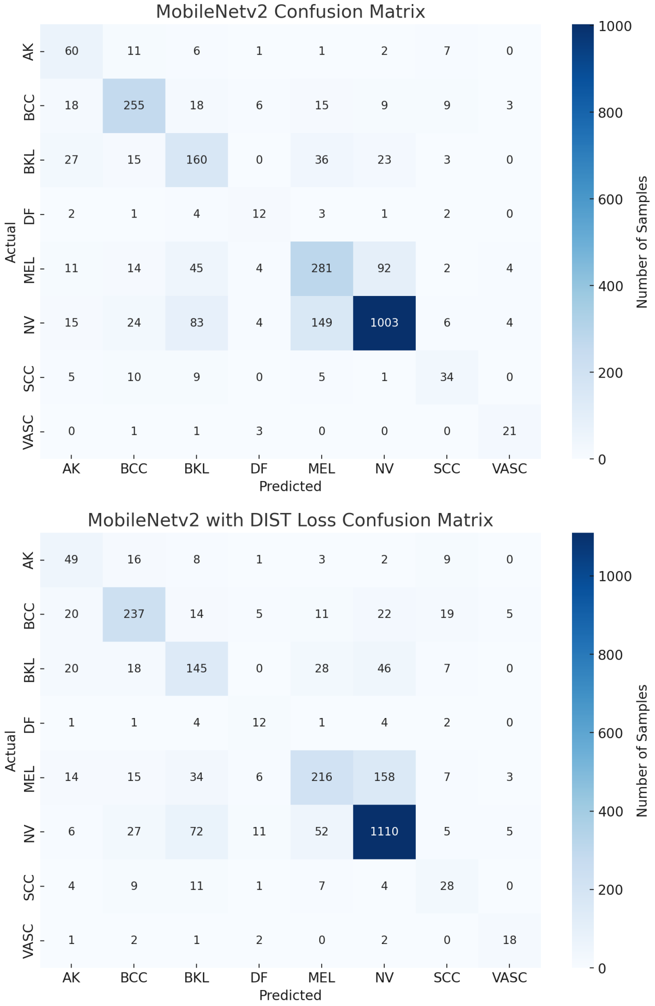

Figure 11. Confusion matrices for the MobileNetv2 Student Baseline and the best performing model of the knowledge distillation method with DIST Loss.

References

[1] Yoshua Bengio et al. "Curriculum learning". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM. 2009, pp. 41–48.

[2] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[3] Noel C. F. Codella et al. *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)*. 2018. arXiv: 1710.05006 [cs.CV].

[4] Marc Combalia et al. "BCN20000: Dermoscopic Lesions in the Wild". In: *arXiv preprint arXiv:1908.02288* (2019). URL: https://arxiv.org/abs/1908.02288.

[5] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[6] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542 (2017), pp. 115–118. DOI: 10.1038/nature21056.

[7] Nils Gessert et al. "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data". In: *MethodsX* 7 (2020), p. 100864. DOI: 10.1016/j.mex.2020.100864.

[8] David Gutman et al. "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)". In: *arXiv preprint arXiv:1605.01397* (2016).

[9] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: `1512.03385 [cs.CV]`.

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: `1503.02531 [stat.ML]`.

[11] Tao Huang et al. *Knowledge Distillation from A Stronger Teacher*. 2022. arXiv: `2205.10536 [cs.CV]`.

[12] C. Karimkhani et al. "Global Skin Disease Morbidity and Mortality: An Update From the Global Burden of Disease Study 2013". In: *JAMA Dermatology* 153.5 (May 2017), pp. 406–412. DOI: `10.1001/jamadermatol.2016.5538`.

[13] S. Kullback and R. A. Leibler. *On Information and Sufficiency*. 1951.

[14] Tsung-Yi Lin et al. *Microsoft COCO: Common Objects in Context*. 2015. arXiv: `1405.0312 [cs.CV]`.

[15] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: `https://www.tensorflow.org/`.

[16] Luke Melas-Kyriazi. *EfficientNet-PyTorch: A PyTorch implementation of EfficientNet*. `https://github.com/lukemelas/EfficientNet-PyTorch`. 2019.

[17] Luke Melas-Kyriazi. *Lukemelas/EfficientNet-PyTorch: A pytorch implementation of EfficientNet*. `https://github.com/lukemelas/EfficientNet-PyTorch`. 2019.

[18] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[19] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[20] Mark Sandler et al. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 2019. arXiv: `1801.04381 [cs.CV]`.

[21] Qian Sun et al. "Skin Lesion Classification Using Additional Patient Information". In: *Biomed Res Int* 2021 (2021), p. 6673852. DOI: `10.1155/2021/6673852`.

[22] Shangquan Sun et al. *Logit Standardization in Knowledge Distillation*. 2024. arXiv: `2403.01427 [cs.CV]`.

[23] Mingxing Tan and Quoc V. Le. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. 2020. arXiv: `1905.11946 [cs.LG]`.

[24] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific Data* 5 (2018), p. 180161. DOI: `10.1038/sdata.2018.161`.

[25] Yuzhu Wang et al. *Improving Knowledge Distillation via Regularizing Feature Norm and Direction*. 2023. arXiv: `2305.17007 [cs.CV]`.

[26] Qingqing Zhu et al. "Combining Curriculum Learning and Knowledge Distillation for Dialogue Generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by Marie-Francine Moens et al. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1284–1295. DOI: `10.18653/v1/2021.findings-emnlp.111`. URL: `https://aclanthology.org/2021.findings-emnlp.111`.