Rapid and Robust Learning for Homogeneous Image Data

（同質画像に対する高速・堅牢な学習）

by

ミシュラ　ソーラヴ　Mishra, Sourav

48-177411

Submitted to the
Department of Information & Communication Engineering,
Faculty of Information Science & Technology
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

The University of Tokyo

December 2020

Rapid and Robust Learning for Homogeneous Image Data

同質画像に対する高速・堅牢な学習

by

Mishra, Sourav

Submitted to the
Department of Information & Communication Engineering,
Faculty of Information Science & Technology
on December 4, 2020, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The advent of deep learning has brought great successes in solving several difficult computer vision problems that were previously based on rule-based approaches. These advances have mostly relied on standardized data which are typically large, annotated, and well-investigated. However, the real-world use-cases involve much smaller long-tailed data corpus which is characterized by being small and homogeneous. Such data present a plethora of learning problems, which includes uneven performance and low confidence in model decisions. Using primarily homogeneous data corpora (such as those of dermatological images), we have attempted to show how classifiers can be trained rapidly to their optimum performance, discuss the rationale in such decisions and investigate points of failure, which could be scope for improvements in future learning methodology on non-standardized data corpus.

Acknowledgments

In completing this thesis, I am grateful to my advisor, Dr. Toshihiko Yamasaki. He has always encouraged me to think deeper and keeping investigating tough problems. He was there for me whenever I needed help. I am also grateful to Dr. Aizawa, who painstakingly brainstormed with me on several occasions to come up with ideas to investigate. I extend my gratitude to my committee members (Dr. Sato, Dr. Iba and Dr. Naemura) who have taken time out of their busy schedules to evaluate my doctoral work. Ms. Matsubayashi and Ms. Egawa lent me a helping hand whenever I needed to complete relevant documentation or inquire about procedures. My sincere thanks to them. I am grateful to Dr. Hideaki Imaizumi for providing me data, resources, and easy access to physicians for dermatological applications.

My fortuitous meeting with June three years ago changed my life and views. My doctoral research could not have finished without her constant support. I faced one of the most challenging phases in my life so far, trying to balance my health with pursuing research. My aunt, Dr. Pragati has been my doctor, friend, mentor, and above all, a spiritual compass.

This journey of a thousand miles began with small steps years ago. This thesis is for my parents. These few pages are just a chapter in their tireless efforts to make me who I have eventually become.

असतो मा सद्गमय।
तमसो मा ज्योतिर्गमय।
मृत्योर्माऽमृतं गमय॥

Asato mā sadgamaya,
tamaso mā jyotirgamaya,
mṛtyormā'mṛtaṃ gamaya

From falsehoods lead me to the truth,
From darkness lead me to light,
From death lead me to immortality.

Pavamānamantra
Bṛhadāraṇyaka Upaniṣhad (1.3.28.)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The dawn of deep learning has immensely improved discriminative tasks such as classification and detection in the computer vision domain. Earlier approaches, which were predominantly rule-based, have given way to neural network based decisions. These methods are constantly improving and showing us new possibilities.

Although several new applications have emerged from deep learning methods, most of them rely on training on a few standard dataset corpora such as ImageNet [24], Microsoft COCO [63] etc. But ImageNet was never designed keeping complex downstream tasks in mind. Real world tasks do not often translate to the few labels and annotations present in such commonly used datasets. The challenges posed by these tasks are harder. Consequently, several other task-specific datasets have emerged. But they are far smaller, long-tailed in their label distribution and involve visually complex categories. Learning on such small dataset and translating them to successful applications can be usually challenging.

Operating on fine-grained images has been a long-standing challenge in computer vision. These images are composed of a single super-category, with several distinct subcategories. They are representative of several new online conferences and challenges in recent years. We define *Homogeneous images* as a niche subset of fine-grained images, wherein the subcategories are not only part of a single meta-category, but also visually very alike. Figure 1-1 describes the visual complexity of dataset categories. The datasets progressively become harder towards right, often requiring expert opinion.

The higher visual complexity leads to computationally expensive and often unreliable learning in many homogeneous data corpora. This is due to absence of attributes such

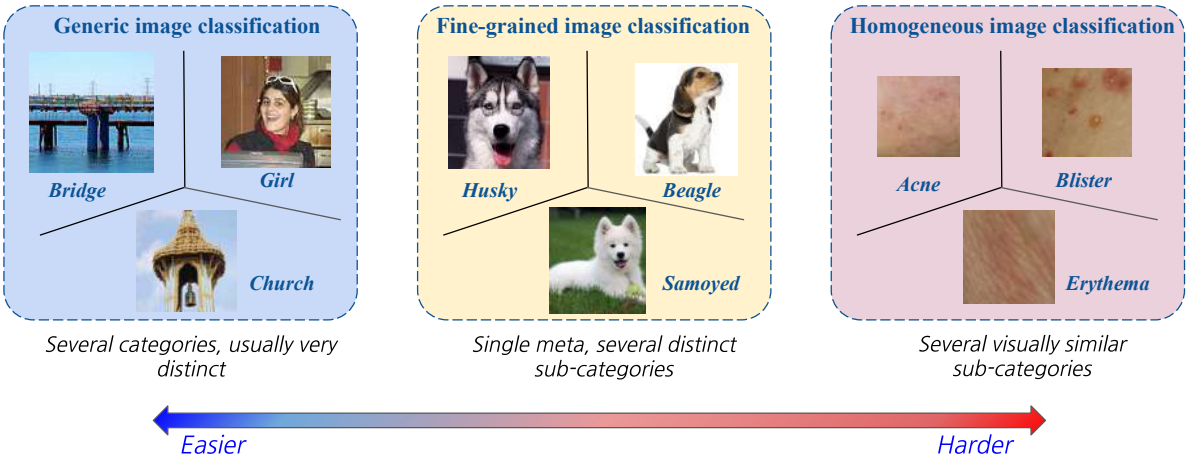


Figure 1-1: Hierarchy of datasets by visual complexity

as landmarks, texture, color etc., which are not prominent or consistent between the classes. Furthermore, this problem is compounded by photogrammetric factors such as image depth, presence of shadows, varying resolutions and significant inter-class & intra-class variations. Although human cognition is very capable of adapting to these factors, machine learning with such visual complexities is a challenging task, often with rewarding returns. Advances in fine-grained and homogeneous image analysis have directly impacted fields such as medical imaging, wildlife conservation, crop analysis, precision agriculture and remote sensing applications.

The majority of computer vision tasks rely on *Transfer learning*. This is the process of re-purposing pre-trained networks for downstream tasks. Such networks are usually trained with best parameter selection on large image datasets. During the process of transfer learning, the last few layers of the pre-trained network are nulled and the network trained with the custom datasets for desired predictions. This is possible because the data is continuously abstracted between layers of the network and the representations are hierarchically learnt in between succeeding layers. The fundamental, initial layers usually learn basic geometrical features which can be applicable to any vision task. The final few layers, which become specific to the task are given careful treatment. The ability of networks to be fine-tuned to downstream tasks, has made transfer learning the backbone of modern deep learning research. Figure 1-2 gives a schematic of this process.

Transfer learning is not optimal. This process works well if the downstream task has sufficient data, albeit a fraction of large corpus. Many a times, real challenges have a few dozen samples per category. This may not prove useful to reliably learn the downstream

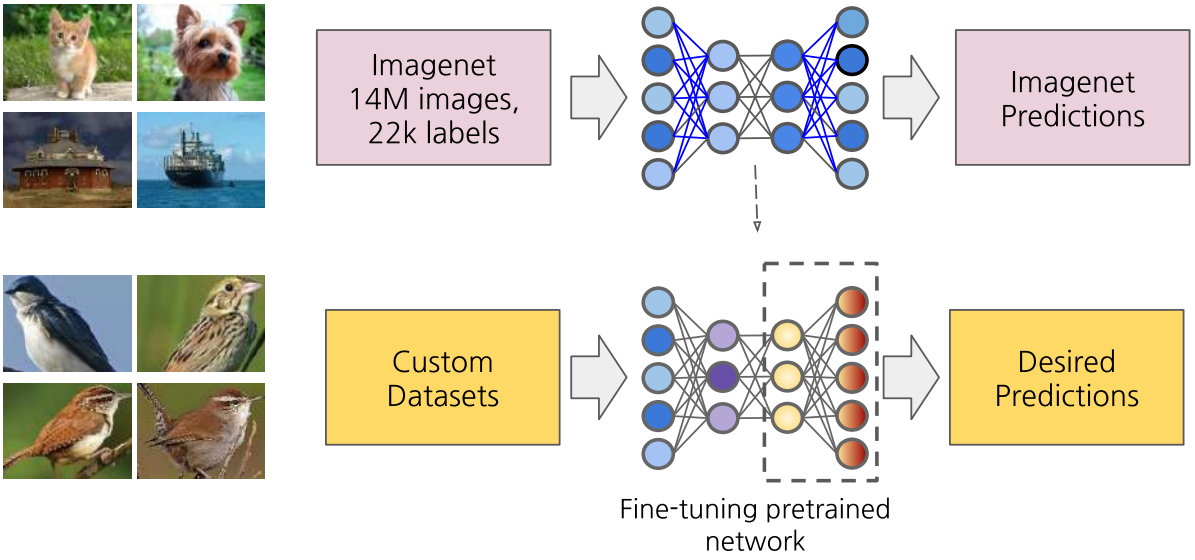


Figure 1-2: Transfer learning schematic

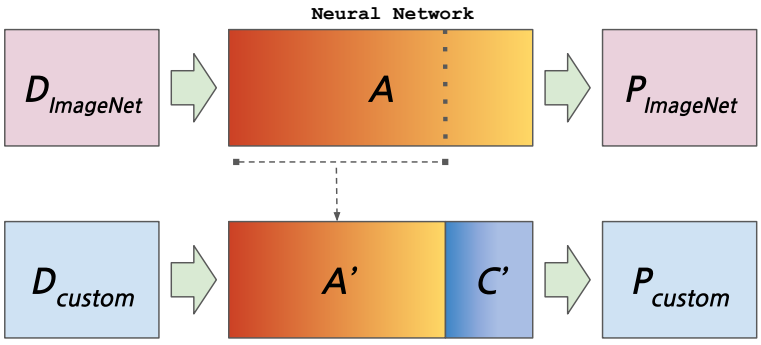


Figure 1-3: Internal representations difference: ImageNet vs. custom task

distribution. An additional problem appears when downstream task is very different from the generic data which has been used to pre-train the network. The intermediate representations learned by ImageNet are not very useful if trying to perform medical image classification. This is shown schematically in Figure 1-3, where the internal representation in hidden layers are shown to be starkly different. To be successfully used, the models have to provide high accuracy. This is difficult, given the aforementioned conditions.

This thesis investigates the problem of fitting pre-trained models efficiently and rapidly for downstream tasks, where the samples are visually homogeneous. The goals of this thesis will be centered on the following questions:

1. Fitting homogeneous image data with high accuracy and repeatability despite of the small number of samples, and high degree of visual complexity.

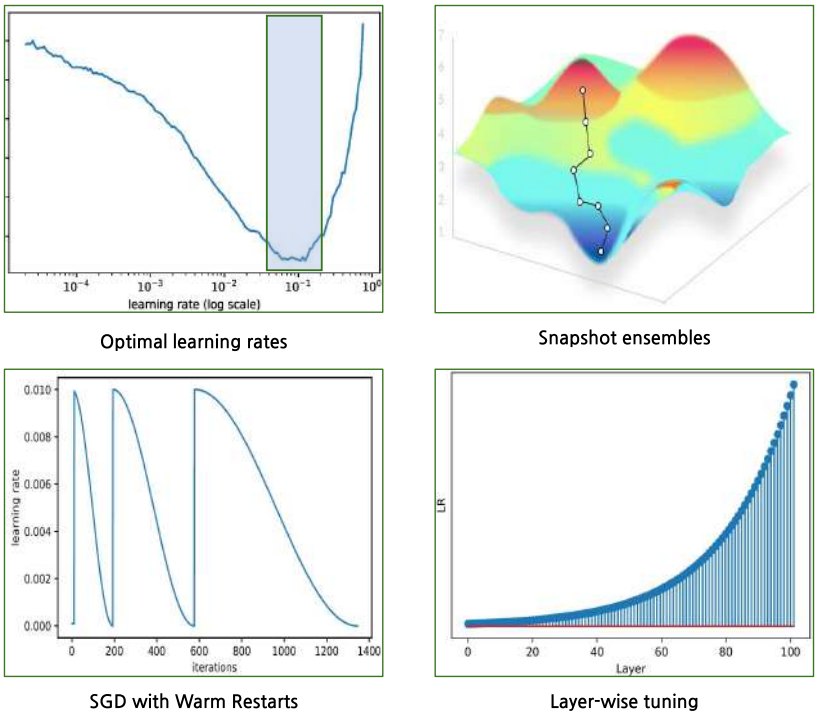


Figure 1-4: Our learning scheme components in a nutshell

2. Understanding patterns of prediction error post-fitting and by simulating real use-case in the application dataset (dermatological images).

To robustly model the homogeneous image information, the thesis has relied on a combination of methods, which are individually tested and verified by academic projects. We use a momentum-variant of SGD with restarts, snapshot ensemble, layer-wise tuning primed with an optimal rate to quickly converge to a good global optimum (See Fig. 1-4). We have further investigated the speedup advantage by this method combined with mixed precision and heuristically explain the reason of robust fit. The details of these experiments are discussed in Chapter 3. We have additionally covered the topic of dataset selection in this chapter, eventually experimenting on three dataset with reliable results. This approximately covers the three sub-types of visual complexity we discussed in Figure 1-1.

With robust fits, we have further investigated the sources of existing error in the target datasets in Chapter 4. We have observed repeating patterns arising from certain label pairs. These have been investigated by class activation mapping techniques such as GradCAM and guided backpropagation [92, 97]. We have discovered a surprising trend in homogeneous images, which is very different from the conventional wisdom gleaned

from ImageNet-styled datasets [32]. The details of these experiments are presented as case studies. This chapter also additionally covers several data curation practices which removes the sources of extrinsic errors. We have demonstrated the effect of these deleterious effects by comparing predictions before and after the adjustments. This section also covers the question of relevance of background and means to remove confounding visual features, which may interfere with model learning.

In Chapter 5, we have discussed the impact of adversarial influence and its significance in model performance. Adversarial inputs need not originate from only malicious actors. They can be the byproduct of environmental factors such as sensor imperfections or digitization. They are capable of significantly changing the predictions without letting the images be visually very different. We have examined the effect of impulse noise on the prediction quality of images as an ablation study. We have further investigated the effect of motion blur and soft-focus by means of smoothing kernels. In this line of study, we have compared the effect of such imperfections with similar perturbations in CIFAR-10. We have discovered that while homogeneous images can be very sensitive to noise, they are quite resilient to blur. These results are explained and summarized in this chapter.

Any model with practical applications will encounter samples which are outside the distribution it was trained on. In the best case scenario, a model will be able to account for the changes in the distribution. But it is usually not the case with fine-grained and homogeneous images. Hence, we have tried to quantify the effects of distribution shift on our target dataset, and made some comparisons with CIFAR-10. Our research shows that small, homogeneous application datasets such as skin images are very highly susceptible to distribution shifts and need some mitigating operation such as domain adaptation to be suitable for deployment.

The final chapter of this thesis (Chapter 6) has discussed the outcomes from our experiments, their implications and future directions this line of research could take. This Introduction and the following Chapter 2 is dedicated to giving a capsule overview of the work related to the thesis topics. It is meant to bring the reader gently up-to-speed with the state of the art advances, issues encountered commonly, and how the specific methodologies are applicable in our current understanding.

Chapter 2

Background Literature

This chapter gives a brief overview of the related topics. Additionally, we summarily discuss the merits, demerits, and applicability to our existing body of work.

2.1 Homogeneous Image Analysis

Computer vision has been at the forefront of several advances in deep learning [60]. Fine-grained classification and localization are longstanding problems that are benefiting from recent improvements. This domain looks into recognizing and analyzing images of multiple subordinate categories of a super-category (a.k.a meta-category). Within the domain of fine-grained image analysis, we make a fine distinction concerning homogeneous images. This category is identifiable by its hallmark absence of strong visual features, such as landmarks, texture & contrast. They often require the aid of a domain expert in evaluating results and creating baselines since the data presentation is not trivial. X-ray plates, skin photographs, satellite imagery, biodiversity tracking, clothes, and apparel product images are commonly encountered examples. Homogeneous image cognition is considered a harder subset within fine-grained image analysis. They are gaining traction in many ML conference venues, with several identification challenges hosted in competition sites. Progresses in fine-grained image analysis have improved workflows in niche domains such as histopathological image segregation [87], breast cancer detection & analysis [67], wildlife monitoring [108], climate change studies [2] and automated retail product identification [113].

Analysis in homogeneous images is difficult from an algorithmic standpoint. This gets

Figure 2-1: Examples in fine-grained image analysis (via Wei et al. (2019))

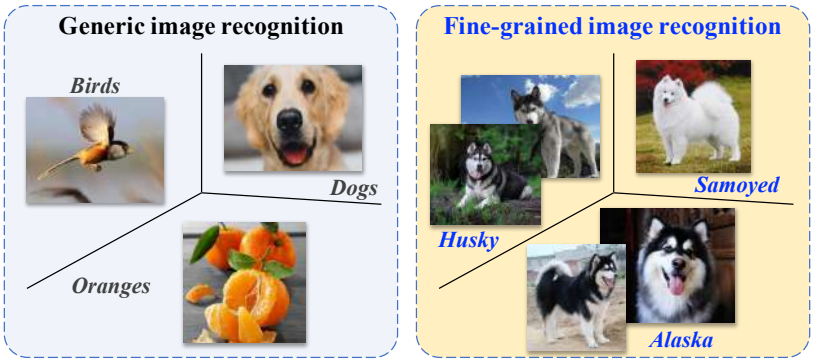


Figure 2-2: Challenges in performing fine-grained classification (via Wei et al. (2019))

compounded when the data quality is not rich. In datasets with coarse meta-categories (such as CIFAR-10, ImageNet classes), the images are visually quite different. They can be easily segregated by feature extraction. Fine-grained classification inputs, however, come from a single meta-category. Figures 2-1 and 2-2 illustrate the basic concept. Homogeneous images go a step further with nearly indistinguishable subordinate categories. The human cognition system relies on a variety of indicators to make assessments and identify objects. Machine learning methods are constrained to the input data and annotations if any. They are further constrained by factors such as image depth, presence or absence of shadow, low resolution, etc., which the human cognition can take care of very easily. Homogeneous multimedia classification grapples not only with a small inter-class variation of visually similar subordinate categories, but also intra-class variation in inputs. Objects from the same class can have differences in color, size, or scale, which can get easily obfuscated by a change of illumination and viewpoint. Deep learning based approaches in this regards can be grouped under the following categories broadly:

1. Approaches using multiple deep networks or sub-networks to localize and categorize images.

2. Approaches using end-to-end feature encoding or feature extractors via convolu-

tional neural networks (CNN)

3. Approaches using attention mechanism to find the discriminative regions for saliency and/or external information.

In this grouping, the first two are constrained by the dataset, using only the information accompanying the images. To better perform image analysis on homogeneous data, scientists have endeavored to extract discriminative semantic parts of homogeneous images. Thereafter the model has relied on intermediate representations to classify the images. Several networks or sub-networks can be used for this approach. These approach has been the cornerstone of **localization-categorization** based approaches. These approaches have benefited from part-level bounding boxes, segmentation masks, and key points localization to locate semantic key parts [126, 62, 115]. These approaches are very labor-intensive to curate and prepare for downstream tasks. Their broad-scale applicability is doubtful, as these are perfected for a single task in hand.

End-to-end feature encodings learn discriminative features directly by developing deep models for fine-grained cognition. CNN based convolutional feature extractors approaches have gained traction in robust classification schemes. They can be extremely powerful in identifying subordinate categories. Deep convolutional feature extractor based approaches offering generic descriptors such as Off-the-shelf [93], ONE [118] and InterActive [119] are extremely popular and feature in several research papers and competitions. These models work well even when data annotations are sparse or not existent. One prominent and successful example is the Bilinear CNN which presents an image as a pooled outer-product of features from two CNNs [64]. A prominent hallmark of end-to-end deep learning methods is the reliance on the depth and optimal fine-tuning of the network. These approaches, although simpler in design, use optimization choices, scheduling, and specialized model learning. Both cost-sensitive algorithmic and data-intensive augmentation & sampling methods are used to solve imbalance conditions in long-tailed distributions whenever required. They are widely applicable to many categories of tasks because of their generic designs. This thesis leverages this approach to produce high quality, consistent results on Oxford-IIIT Pets, and Dermatological lesion data in the absence of accompanying annotations.

Local-attention based mechanism is another prominent way which can aid the aforementioned methods in developing insights into fine-grained classification. Methods

such as Guided backpropagation and Gradient-based class-activation maps use the gradients of a subordinate category flowing into the final and penultimate layers to produce a localization map, which can be useful in understanding or isolating saliency in a fine-grained classification task [92, 97]. The current thesis has also utilized this method to interpret the classification models and understand the hierarchy texture, color, and shape in select fine-grained classification results in skin images.

A smaller subset of approaches, which goes beyond the conventional paradigms, involves external information e.g. web-data, multi-modality data, and human-computer interaction. Free but noisy internet-based multimedia can be used to supplement the quality of training data, as demonstrated by Zhuang et al. and Sun et al. [128, 99]. This approach tries to reduce the negative effect of noisy labels while at the same time overcoming the information gap between the task and the crawled information. A recent trend seen also involves web-data, but using prototypical information, domain adaptation and transfer learning to boost the quality of inductive biases learned in supervised classification tasks [127, 79]. There exist some auxiliary methods also, which can boost the quality of the analysis. These involve creating images in a specific category, such as e.g. CVAE-GAN [4, 36], generating text descriptions from images by visual grounding e.g. StackGAN++ [125], few-shot learning [101] and Contrastive-learning based fine-grained recognition [58]. Although they are gaining traction, we are limiting our scope to end-to-end encoding based methods, which can be made generic and reproducible.

2.2 Transfer Learning and Performance Optimization

2.2.1 Transfer Learning Methodology

Data dependence is a cornerstone in deep learning. For good outcomes, we need a large network in addition to large amounts of data. In many cases, the application in hand may not have a sufficient amount of data to train a network *ab initio*. Deep learning leverages the power of transfer learning to adapt to several tasks. It works on the assumption that deep networks and the brain are similar. They have iterative and continuous abstraction processes. Just like the brain does not require unique networks for each task that we perform, a neural network should be able to re-utilize some layers as feature extractors, and the extracted features should be versatile for downstream tasks. Networks need not

be initialized from scratch and trained on limited data. Instead, they are to be trained on industry-accepted data corpora and thereafter utilized for several downstream tasks by fine-tuning specific layers. This approach has worked well when the pre-training (or pre-task commonly) is performed on a large data corpus, and the downstream task (or target task) is smaller but sufficient to fine-tune the network.

Transfer learning was introduced by Bozinovski et al. for *Perceptrons* in 1976 [10]. They have come a long way since then. Transfer learning today is ubiquitous. In the current body of work, we have utilized serialized network architectures such as ResNet (ResNet-34, ResNet-50, ResNet-101 and ResNet-152) [40], DenseNet (DenseNet-121) [45] and ResNext (ResNext-50 and ResNext-101) [120] to perform fine-grained classification on homogeneous data corpora. Transfer learning on standard, off-the-shelf architectures is one aspect of the plan. To have efficient learning, optimization and scheduling are the other facets to consider.

2.2.2 Learning Speedup Techniques

Training models can be the computational bottleneck in the deep learning pipeline. They often require several hours to days and speedups are valuable. Training deep networks with free parameters is a minimization problem of the type $f : R^n \rightarrow R$. Gradient Descents works to optimize f by iteratively adjusting these n free parameters in the parameter vector $x_t \in R^n$ by using $\nabla f_t(x_t)$ i.e. the gradient from the backward pass. This is denoted by Equation 2.1,

$$x_{t+1} = x_t - \alpha_t \nabla f_t(x_t) \tag{2.1}$$

$$x_{t+1} = x_t - \alpha_t H_t^{-1} \nabla f_t(x_t) \tag{2.2}$$

A more elegant way would be to calculate the inverse Hessian matrix H_t^{-1} as seen in Equation 2.2, but that problem becomes intractable when the number of parameters goes up exponentially e.g. in the computer vision domain. Quasi-Newtonian methods for example L-BFGS [65] and Levenberg-Marquardt algorithm [77], although plausible approaches are not effective for the scale of data that we see in Deep learning & computer vision. Modern approaches have therefore all relied on approximating the Hessian

matrix to achieve speed advantages [54, 123]. Several approaches took another route by simply using momentum information as described by Bengio et al. for Geoffrey Hinton's RMSProp [7, 42] in Equations 2.3 and 2.4

$$\nu_{t+1} = \mu_t \nu_t - \alpha_t \nabla f_t(x_t) \tag{2.3}$$

$$x_{t+1} = x_t + \nu_{t+1} \tag{2.4}$$

where ν_t denotes the velocity vector, α_t is a monotonically decreasing learning rate and μ_t is the momentum term which assimilates weights prior information to the current observation (usually set higher than or equal to 0.9 since the number of epochs tracked by momentum is approximately $\frac{1}{1-\mu}$). One difficulty observed in this optimization scheme was the monotonic learning rate and amount of weight decay regularization on the model parameter vector x_t as the model learning progressed. With gradually smaller learning rates, L2 regularization aiming to shrink the parameters by a higher Froebenius penalty term was seen to be less effective. Loschilov et al. and Smith et al. introduced very similar approaches by which the learning rate was changed periodically which took care of stagnation in saddle points and consequently reducing over-fitting [71, 96, 95]. When the learning rate periodically went up, the effect on parameter regularization increased, despite the choices made on the regularization term λ. Such restarts showed improvements in the convergence performance. Loschilov et al. considered the stochasticity of the mini-batch information by taking averaged gradients and losses periodically.

In their approach, stochastic gradient descent (SGD) [90] was restarted once every T_i epochs, i being the run-index. The periodic learning rate was modulated according to the equation,

$$\alpha_t = \alpha_{min} + \frac{1}{2}\left(\alpha_{max} - \alpha_{min}\right)\left(1 + cos\left(\frac{\pi \cdot T_{cur}}{T_i}\right)\right) \tag{2.5}$$

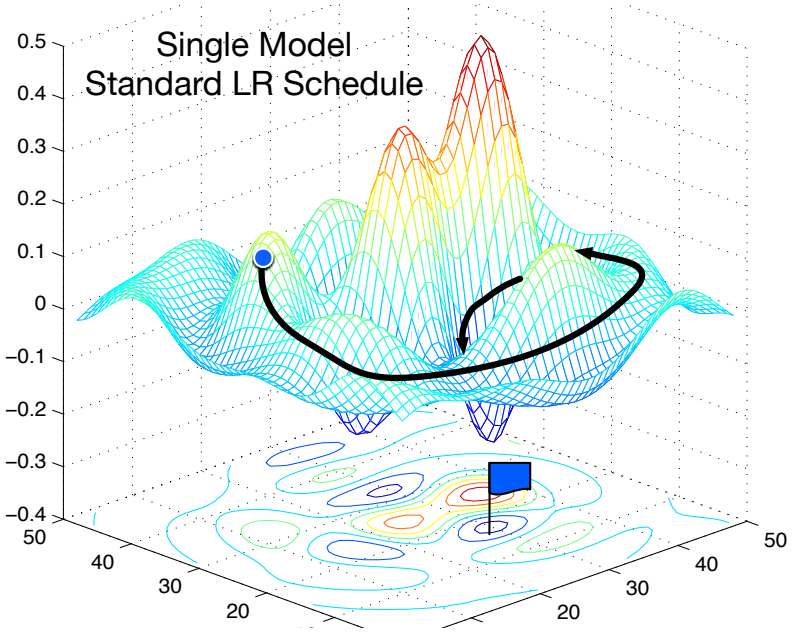
where $(\alpha_{min}, \alpha_{max})$ is the range in which the rate fluctuates, and T_{cur} accounts for how many epochs have passed since last restarting. Because of the cosine envelope in the governing equation, the rate cycle is cosine-annealed i.e. the fluctuation traced a cosine pattern. Loschilov et al. and Smith et al. differed on the nature of cyclical restarts. Whereas Smith et al. [96] provided a more generic approach to cyclically generated learning rates, Loschilov et al. [71] introduced the cosine anneal cycle as the simplest warm restart solution. Further additions also included cycle length multiplication, where the

cosine cycle was elongated to cover an increasingly higher number of epochs for improved convergence. Huang et al. improved the original scheme of SGD-R with warm restarts by adding ensembles of the same network [44]. They trained a single neural network instead of ensemble results from multiple runs, which is still a costly endeavor. The model was expected to converge to several similarly valued local minima along the optimization process [50]. They used this valuable manifold symmetry information in several local minima to get the best results, instead of opting for different experimental runs which could be plagued by differences due to random batches and initial seeds. In their approach, they let SGD converge M times to local minima along the path of the highest gradient and at each minima, the parameters were saved for a test-time ensemble. The loss function periodically jumped out of the local minima due to SGD-R rate cycling until the best values were obtained at convergence. During the test time, one could average the saved snapshots and get much better performance than an individual run. Figure 2-3 shows a schematic of this process.

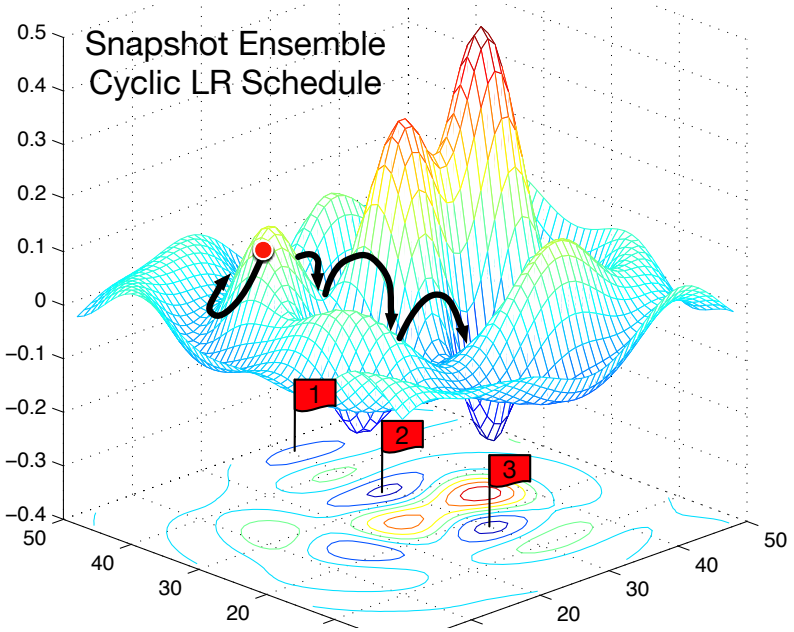
2.2.3 Generalization Performance (ADAM vs. SGD)

Although adaptive methods such as ADAM [54], RMSProp [42], AdaDelta [123] and more recently AMSGrad [89], are becoming the first choice to learn models, they have been observed to not generalize as well as SGD and their non-adaptive variants. There are several hypotheses to this question. Keskar et. al argued that sharp minima are problematic to good generalization [52]. Flat basin in the loss landscape serves better for generalization than sharp drops. They have promoted switching from adaptive methods to SGD (with momentum). Similarly, Wilson et al. pointed out that over-parameterized networks could show very different solutions in adaptive-momentum based methods leading to poor generalization cases [116].

Adaptive methods triumphed over SGD only when the gradients were sparse or small initially– the conditions which SGD-R and snapshot ensembles have demonstrated to avoid well [52]. Reddi et al. claimed that adaptive methods were not reliable when the problem is acutely non-convex. This had very direct implications on the selection of the optimization method in fine-grained analysis. The momentum penalty constrained the loss function to navigate in a small section of the loss landscape, because of which saddle point stagnation become common and true global minima was hard to reach. Another



(a) Conventional SGD with a typical decaying rate schedule converges to one or a handful of minima before finding best fit.



(b) Snapshot ensemble, where the model undergoes several anneal cycles, escaping from multiple local minima due to restarts.

Figure 2-3: SGD vs SGD-R with Snapshot ensemble (Courtesy: Huang et al.)

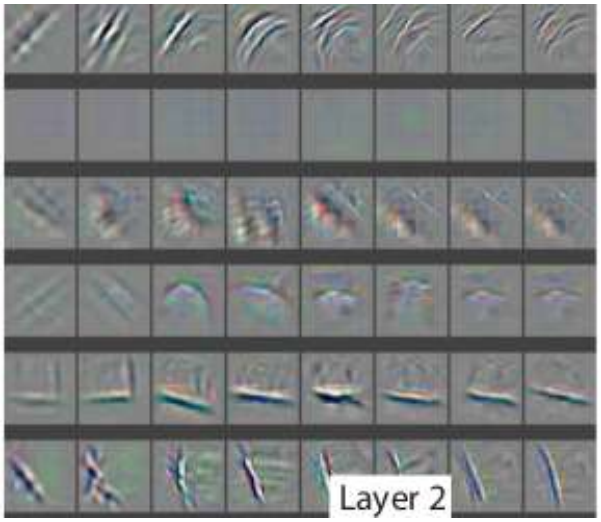
way to look at this in the macroscopic sense was the model's inability to bridge the avoidable bias. A further observation made by Reddi et al. pertained to the fact that some mini-batches provided a large gradient while others did not. The influence of large gradients died out due to the exponential averaging process in the absence of long-term memory [89]. State of the art reproducible results across all vision challenges is often still claimed by SGD and its non-adaptive variants [31, 21].

Hutter et al. have claimed L_2 regularization posed by ADAM is not as effective as that of SGD. They demonstrated that although conventionally weight decay regularization and L_2 regularization are believed to be synonymous, this fact did not hold for adaptive mechanisms. In particular, the L_2 regularization effect was weaker in design than the weight decay regularization. This is possibly the strongest reason why ADAM leads to poorer results than SGD in image classification leaderboards [19]. Weight decay was effective in both ADAM and SGD, and therefore ADAM with explicit weight decay regularization schemes (as seen in AdamW [72] and NovoGrad [35]) worked at par with SGD with momentum (SGD-M).

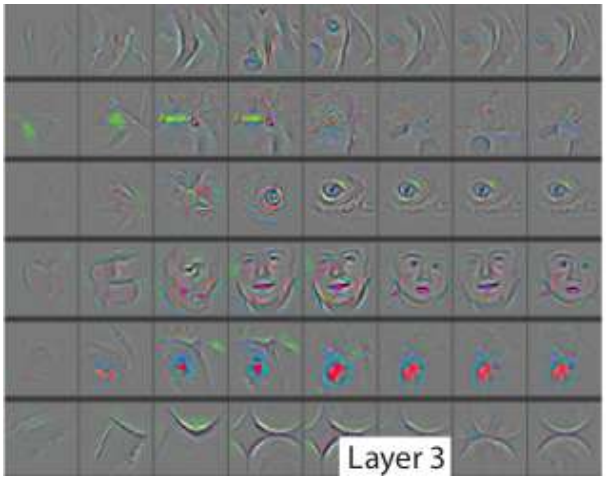
2.2.4 Layer Specific Tuning

Layer-wise enhancement of training was identified to be a major factor in convergence speed by You et. al [122]. In homogeneous image training, this was important since we were trying to enhance the compositional learning in the neural network for difficult image sets. Zeiler et al. were among the first to show a hierarchical growth of features across increasing layer number in CNNs [124]. Figure 2-4 is a example of what a CNN learns when trained with ImageNet data [24].

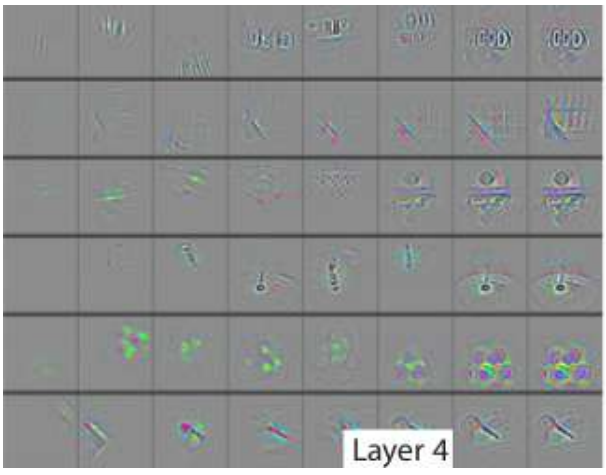
Since transfer learning uses sparse information provided by limited data, it was deemed important to preserve the pre-trained information gleaned in the first few layers [27, 70]. Although we can disregard and use a uniform learning rate scaling for all the layers, it would only increase the time to convergence. Different layers captured specific hierarchies and hence needed tuning separately. Having a differential scaling scheme in learning rate vis-a-vis the layer depth afforded quicker learning with only the final few layers seeing the largest amount of changes, as proposed by Howard et al. [43]. In their method, instead of a single rate for all the layers, every CNN layer was modified by a separate learning rate. If we considered Eq. 2.1 again, x_t (the unified parameter vector



(a)



(b)



(c)

Figure 2-4: Compositional nature of information (Courtesy: Zeiler et al.)

at iteration t) was separated for participating layers i.e. $x_t \rightarrow \{x_t^1 \dots x_t^L\}$, where L denoted the number of layers in the network. Each of these layer vectors x_t^i, were to be conditioned by a learning rate α^i specific to it instead of a universal rate α. Therefore each member of $\{x_t^1 \dots x_t^L\}$ was treated with a unique member of $\{\alpha^1 \dots \alpha^L\}$. With this arrangement, the SGD equation with a cost function $J(\theta)$ could be written as,

$$x_{t+1}^i = x_t^i - \alpha^i \cdot \nabla J(x_t) \tag{2.6}$$

If we factor cyclical learning rate, then the α^i rate becomes variable within a small set bounded by (α_{min}, α^i) instead of $(\alpha_{min}, \alpha_{max})$ seen in Eq. 2.5. In the current body of work, three different approaches were considered. The first involved dividing the network into three sections and using a scaled learning rate for each section. This translated into Eq. 2.7

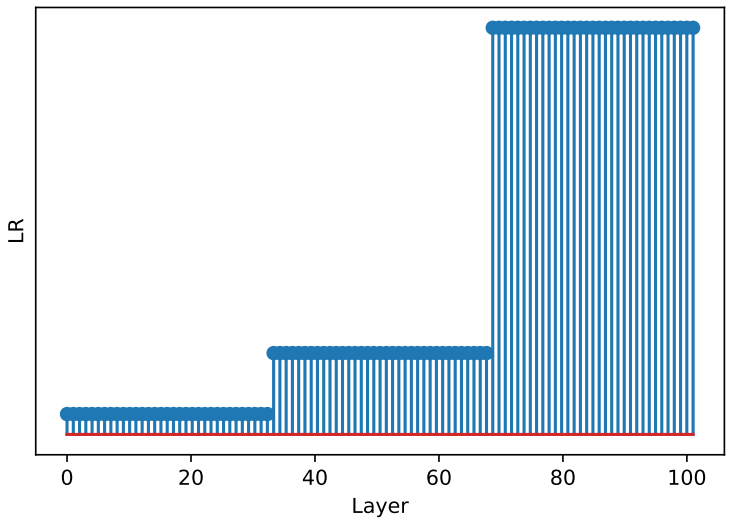
$$\alpha_i = \begin{cases} 0.01\alpha_{opt}, & i \in [1, \frac{L}{3}] \\ 0.1\alpha_{opt}, & i \in [\frac{L}{3}, \frac{2L}{3}] \\ \alpha_{opt}, & i \in [\frac{2L}{3}, L] \end{cases} \tag{2.7}$$

where α_i was the desired rate for i^{th} layer. Since this was crude, a refinement involved linearly scaling the rate between the desired rate (at the final layer) to a hundredth of this rate at the beginning (layer-1) in an L-layer network, as seen in Eq. 2.8

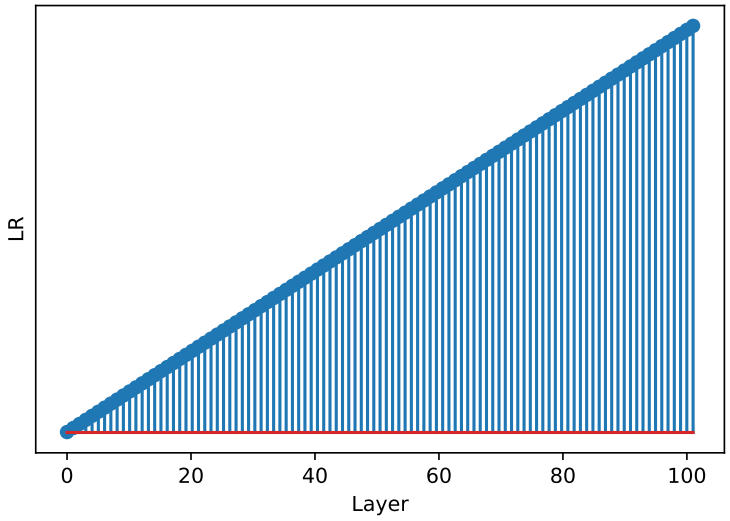
$$\alpha_i = \left[\frac{\alpha_i - 0.01\alpha_i}{L}\right] \cdot i \tag{2.8}$$

The final version, which the current work utilizes, uses a set of learning rates which are log-stepped between a minimum and the desired valued. The layer specific values for L-layer network is given by Eq. 2.9. These three iterations have been schematically demonstrated in Figures 2-5a,2-5b and 2-5c in the order explained.

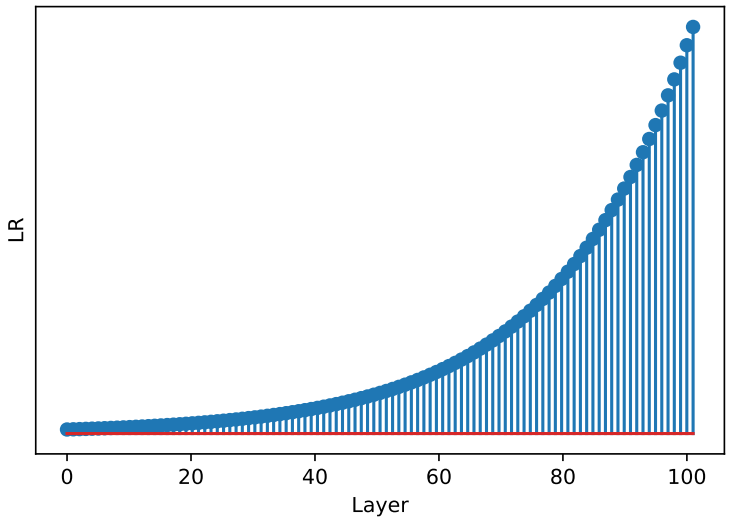
$$\alpha_i = \alpha_{min} \cdot \left(\frac{\alpha_{opt}}{\alpha_{min}}\right)^{\frac{i}{L-1}} \tag{2.9}$$



(a) Block-wise assignment



(b) Linear interpolation-based assignment



(c) Logarithmic interpolation-based assignment

Figure 2-5: Iterations in the development of Discriminative LR

2.3 Adversarial Examples and Effects

Although Deep learning methods have shown great promises in how we analyze and infer information, the sub-domain of adversarial attacks has raised questions on their applicability in real-world problems. This problem was first discussed by Szegedy et al and Goodfellow et al. [103, 37] as inputs that a malicious actor could intentionally design. In their examples, input images were seemingly obvious to the human eye but caused a large prediction change when inferred via a learned model. This highlighted that deep networks learned to make decisions in ways much different than human cognition. Some mitigating work has been published recently, such as the Carlini-Wagner detection [15]. They showed approaches that attempt to check a sample as benign or adversarial before running through the network.

Although most of the adversarial attack sub-domain of ML deals with aggressive strategies to confound models, adversarial examples can present by natural causes as well. Gilmer et al. have shown that randomly corrupted images are capable of fooling classifiers [34]. Additive Gaussian noise [26], translations [3], blur and contrast modulations [41] are capable to creating havoc in prediction outputs. They observed that errors in Gaussian noise patterns suggested the presence of adversarial samples, which could occur even in naturally occurring images. They further investigated methods which prove that any transformation which could improve the distance to *decision boundaries* was capable of providing resilience to adversarial perturbations. This included adversarial training as one of the options. A panel showing commonly encountered adversarial perturbations is shown in Figure 2-6.

Diettrich et al. postulated that most commonly available architectures became better by being successful in learning and predicting using even corrupted data. The difference between predicting clean versus perturbed samples have fairly remained constant. Reduction in mean Corruption Error (mCE) was not due to a difference between noise-free and corrupt samples, but due to better representation learning in modern architectures. The accuracy of models went up and so did mCE, but the gap between them persists. In that sense, Diettrich et al. noted the vulnerability of networks remained unchanged. As of present, conventional pre-processing techniques such as contrast adjustment, histogram equalization, frequency adjustments are worthy candidates in the absence of task-specific dataset transformations.

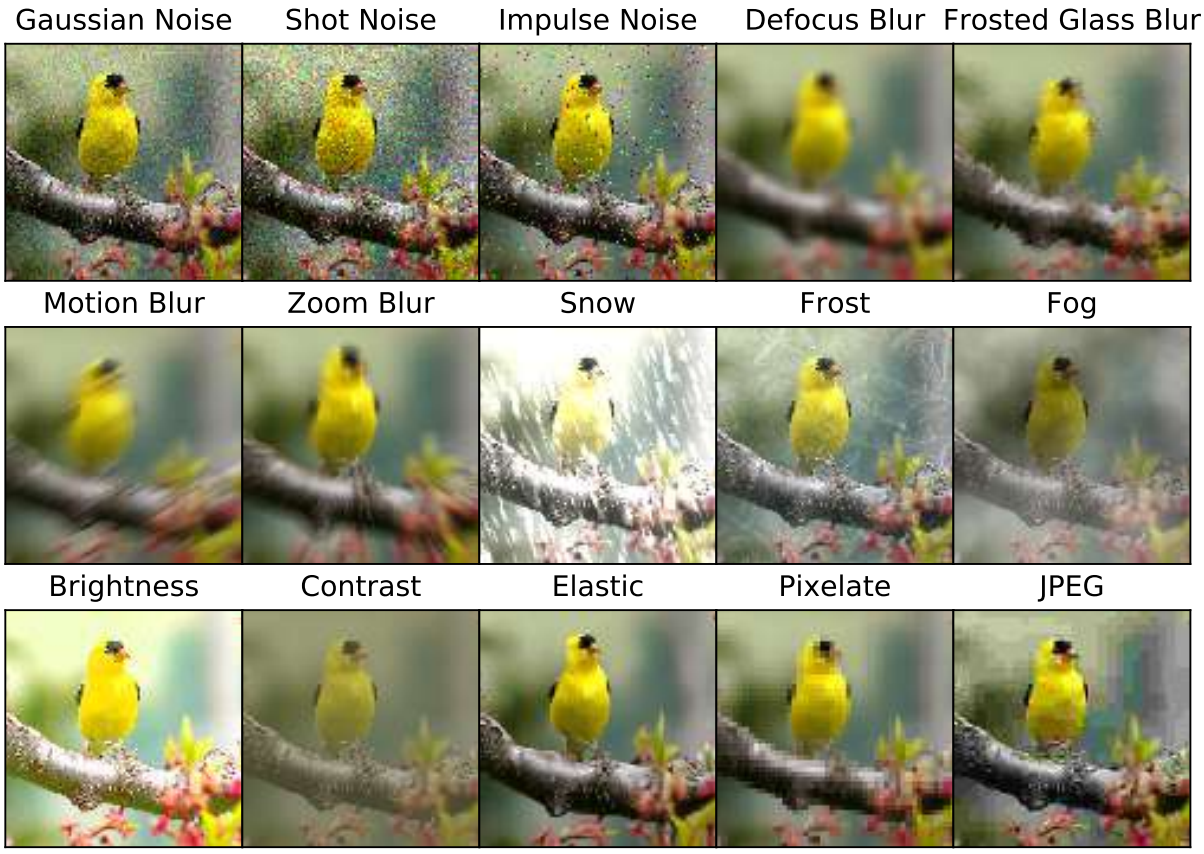


Figure 2-6: Types of common adversarial corruption (Diettrich et al.)

2.4 Deep learning in Dermatological Images

Quality dermatological attention is a growing requirement around the world. Rising populations in emerging economies and a greater incidence of transmissible skin infections have increased the footfall to care providers. Concurrently, an under-supply of dermatologists is being observed, who can treat such conditions. As per the last census in the US, there were only 3.6 dermatologists per 10,000 people [53, 57]. Approximately 2 billion people around the world have some kind of skin ailment, with steady year-over-year increments. They are the fourth most common cause of morbidity [68] with *Melanoma* being the fifth-most invasive form of cancer reported by the National Institutes of Health (NIH) in the US [17].

The spectrum of diseases in dermatology is quite wide. Yet with timely interventions, the survival rate has been seen close to 98% [98]. One of the core concerns regarding treatment is the waiting time to meet a specialist. Faced with long waiting queues, patients are more likely to visit general health practitioners for treatment. Their diagnostic accuracy is between 24–70% for different conditions and concurrency with dermatologist

opinion at around 57% on the average [68, 73]. Even with specialists, opinions of diagnosis can differ at different stages. There is an immense potential to improve the quality of diagnostics and make meaningful impacts.

Detection and automatic grading of dermatological images have seen a huge improvement since the advent of deep learning. They have circumvented the rule-based approaches in traditional computer vision to perform better. Faster processing and cheap cloud-based storage are making inroads into making quality diagnostics available to the public. Skin cancer screening and classification via convolutional neural networks (CNN) [59] has garnered the maximum attention in recent times [28]. Skin cancer previously required close attention and long examinations by qualified medical professionals via a whole-body scan. The early screening rates were traditionally low (16% in men and 13% in women) due to the privacy and modesty of participants. Esteva et al. showed that for *Melanoma*, it was possible to achieve dermatologist-level performance (72.1% Top-1 accuracy) on individual lesion images using an *Inception-v3* network [102], by transfer learning on 129,450 annotated images. Brinker et al. followed up with another study which showed superior sensitivity and specificity when compared with results of a panel of board-certified dermatologists [13, 12]. Haenssle et al. demonstrated that CNNs could in practice out-perform 58 international dermatologists with varying levels of experience (self-reported at 29% beginner, 19% skilled, and 52% experts) [39]. Their CNN, trained only on images and labels, exhibited 0.86 Area under curve (AUROC) values as compared to 0.82 for experts.

Although the ML methodology has gained traction in several domains (e.g. diabetic retinopathy [1]), it has not been validated by the Food and Drug Agency (FDA) in the US for skin applications due to several notable limitations. Beyond Esteva et al. classifying malignant and benign lesions, several other projects have shown exciting progress. Shrivastava et al. have used similar ideas for detecting *Psoriasis* [94]. Park et al. have used a crowd-sourcing aided model for better anomaly prediction [82]. Yang et al. have tried to mimic the dermatologist's criteria by involving representation learning [121]. New research projects are coming up which explore different aspects of dermatological diagnostics such as differential diagnosis [68], histo-pathological analysis [56, 80], dermatological tele-medicine [46, 23] and precision healthcare [17].

The majority of literature using ML for dermatology has explored aggravated condi-

tions such as *Melanoma, Non-melanoma skin cancers (NMSC), Psoriasis, Rosacea* etc. Most generic approaches have relied on using pre-trained models and performing feature extraction for classification or localization [51, 18, 86]. Many studies have tried to understand common skin complaints. Since there is an established shortage of skin doctors, any robust intervention in this domain will alleviate the wait time and channel efforts of the doctors to the cases which require urgent attention. The biggest bottleneck in this respect has been the lack of availability of verified labeled data. SD-198 and its sister dataset, SD-260 (Sec. 3.1.2) are the only large scale dataset which covers most of the common complaints such as *Acne, Alopecia, Erythema, Pigmentation, Ulcers* etc. The current work has also focused on curating a similar dataset (Sec. 3.1.1) and observing the nature of classification thereafter.

In the current work, adversarial effects on dermatological classification have also been investigated. Current datasets capture the lesion (or dermoscopic images if applicable) using high-resolution digital cameras. Deep learning has proved to be quite successful in such pristine images. However, real-world workflows introduce several aberrations. Images can turn out noisy, pixellated, and blurry (soft focus). Pristine images are not truly representative of such conditions and the models are bound to perform poorer given such inputs. However, no study as of yet exists which has documented the extent of performance gains and drops in the presence of adversarial effects.

Chapter 3

Robust Model Learning

Data, be it structured or unstructured, has attributes. Performing machine learning involves creating algorithmic models that can learn representations. These representations are closely linked to attributes. Hence, learning good representations lead to good outcomes on the task at hand.

Human-guided classification methods in vision problems have traditionally involved feature engineering i.e. handpicking attributes in the images which can simplify the classification or detection process. Deep learning, on the other hand, has made a quantum leap in how we capture the relevant features and representations to give superior results. However, they are not without their challenges. Traditional methods such as Support vector machines [20], Logistic regression [75] and Random forests [11] have typically involved a small collection of model variables and *hyperparameters*. These are easy to set and investigate while testing a model's strengths and weaknesses. Model parameters in Deep learning methods run into several million at the least. It is confusing to even understand the role of a group of nodes or layers in analyzing model performance. It is therefore important to have well-designed learning paradigms when working with such *black-box* models.

Robust learning involves making choices in the optimization and scheduling algorithms, in conjunction with choosing appropriate network type and depth. When faced with optimizing millions of parameters, we can only establish the best outcomes by making careful choices in hyperparameter selection. It is very expensive to utilize grid search or beam search. The problem gets more compounded when we are faced with homogeneous data, which lack distinguishing inter-class and intra-class attributes.

3.1 Homogeneous Data

Homogeneity is the lack of clear and distinguishing attributes. Such images are challenging in computer vision problems. In trivial classification, the decision boundaries partitioning the data are simple. In the case of homogeneous data, the classifier has to construct a complicated set of inter-class boundaries in the manifold. The severity of this problem can get compounded when a sufficient amount of data is not available. A primary goal in this dissertation is to understand model design choices that can create operable decision boundaries for small-sized homogeneous data corpus. In the following subsections, we shall briefly discuss a few data sources which are suitable for such learning problems. It is worth noting that medical images, satellite imagery, wildlife, and natural specimen photography are good candidates for this data description.

3.1.1 Exmedio Dermatological Data

The Exmedio Dermatological Dataset [76] is a private dataset composed of ten classes of common skin problems. This data corpus has been made by a systematic collection of user-submitted dermatological images belonging to the East Asian racial type by following out-patient statistics. The samples have been collected with the consent and cooperation of volunteers in Japan. Some additional photographs have been sourced from affiliated medical centers and clinics. All borrowed samples are covered under agreed frameworks of data reuse. The image sizes are variable but usually larger than 200×200 pixels. The image data is adherent to the Joint Photographic Expert Group (JPEG) [111] and Portable Network Graphics (PNG) [91] standards. These photographs have been stripped of their meta-data and labeled by registered clinicians without further modifications. The ten constituting classes are (i) Acne, (ii) Alopecia, (iii) Blister, (iv) Crust, (v) Erythema, (vi) Leukoderma, (vii) Pigmented Maculae, (viii) Tumor, (ix) Ulcer, and (x) Wheal. Table 3.1 presents information about the selected labels and their sizes.

3.1.2 SD-198 Skin Data

The SD-198 dataset (along with SD-128 and SD-260) is a dermatological image corpus that has been used for detecting common skin ailments from photographs [100]. The dataset has 198 labels covering the spectrum of trivial classes such as acne to serious

Label/Class	Samples
Acne	971
Alopecia	681
Blister	690
Crust	639
Erythema	689
Leukoderma	664
P. Macula	717
Tumor	790
Ulcer	782
Wheal	636

Table 3.1: Sample Distribution in Exmedio Dermatological Data.

conditions such as skin cancer. A total of 6584 images are contained in this set, with the sample-frequency between 10 and 60 per class. These photographs have been sourced from *DermQuest* [30], an online database of skin lesions. These images have been further curated for quality and annotated by Sun et al.

3.1.3 Oxford IIIT Pet Images

The Oxford-IIIT Pet Images dataset is a small image corpus of approximately 7400 images [83]. They are distributed across 37 classes of dogs and cat breeds. Almost all of these labels have about two hundred images each. This dataset is a benchmark dataset in various fine-grained classification challenges in conference venues. We have used these images without their annotations or bounding boxes, relying only on the photographic media and class label. The dataset has been split in a ratio of 5:1 for training and validation. This dataset is useful since it allows an opportunity for fine-grained classification on a small photographic data corpus.

3.1.4 CIFAR-10

Since these datasets are small and homogeneous, we have additionally used CIFAR-10 [55] to standardize the learning methodologies before adapting them to our target data. This dataset has ten labels: (i) Airplane, (ii) Automobile, (iii) Bird, (iv) Cat, (v) Deer, (vi) Dog, (vii) Frog, (viii) Horse, (ix) Ship, and (x) Truck. There are 60,000 images in total across the labels uniformly with a recommended split of 5:1.

3.2 Model Learning

Trying to implement learning schemes on homogeneous imbalanced data could be misleading. The number of samples available was fewer compared to standardized computer vision datasets. A smaller data corpus could lead models to exhibit bias or over-fitting easily. The absence of benchmarks makes evaluating training methodology difficult. In application domains such as medical images, we anticipated rapid re-training of the models as a requirement too. Hence, learning paradigms were needed that demonstrate reasonable accuracy, fast convergence, and repeatable performance. Van Horn et al. have demonstrated that practical end-to-end deep learning solutions exist when transferring knowledge in long-tailed datasets [109] i.e. situations where the number of categories could be large but the number of samples per category is small. Therefore, before trying to implement classification for small and homogeneous data, we aimed to design and perfect the learning scheme on standardized data, the benchmarks of which are more widely accessible. We adopted CIFAR-10 for designing our experiments and transfer the best practices subsequently.

Our model was built utilizing PyTorch framework [84], running on a single NVIDIA GPU (Tesla V100 32GB HBM2). We chose ResNet architectures (ResNet-34, ResNet-50, ResNet-101 and ResNet-152), DenseNet-121, ResNext50 and Resnext101 as potential models [40, 45, 120]. The serialized architecture was convenient for learning optimization. It provided a lower memory footprint and less model structure variability as compared to neural architecture search (NAS) focused models such as NASnet [129]. We chose them at the expense of a small deficit in accuracy, over potentially significant improvements in learning speed. Additionally, benchmarks for these popular models were more widely available to compare and validate.

In our learning scheme, the batch size was set at 64 and the Categorical Cross-entropy loss function was used, as described in Eq. 3.1.

$$\mathbb{L}(x, C) = -log\,(x, C) \; = -log\left(\frac{e^{x[C]}}{\sum_{j=1}^{N} e^{x[j]}}\right) \tag{3.1}$$

The x denotes the current sample, and C being a class in the data model. (The cost function is derived from the sum of the individual losses in the mini-batch, i.e. $\sum_x \mathbb{L}(x, C)$). We normalized data with the recommended mean (0.4914, 0.4822, 0.4465)

and standard deviation (0.2023, 0.1994, 0.2010). We chose a more traditional data split of 5:1 between training and validation sets. Due to limited data, the validation set also doubled as the test set. We performed dynamic in-memory augmentation such as center crop, random zoom (at maximum 1.1x scaling), horizontal & vertical flips in the data-loader when selecting mini-batches.

3.2.1 Baseline

To compare training methodology for any improvements, we performed a baseline evaluation of the models under recommended learning settings. Our rationale was to select a data corpus reported commonly in contemporary literature, exhibiting sufficient variety, but smaller than ImageNet [24]. Hence we chose CIFAR-10 as the standard data to build and configure learning schemes. Our classifier was built on PyTorch v1.2 framework with recommended practices such as dynamic augmentation and early stopping. Classic, well-tested architectures such as ResNet-34, ResNet-50, ResNet-101 and ResNet-152 were employed. Additionally, we also tested on newer architectures reporting state of the art in leaderboards, such as DenseNet-121, ResNext-50, and ResNext-101, as mentioned in Section 3.2. A single GPU (NVIDIA V100 16GB HBM2) was used. We normalized the data with the recommended mean and standard deviations for CIFAR-10. The dataset split was done in the ratio of 5:1 into the training and validation set. We performed dynamic augmentation such as crop, zoom, horizontal & vertical flips. Stochastic gradient descent (SGD) and Adaptive Momentum (ADAM) with step decay were used as the optimizer in these tasks. These models were trained to their best validation accuracy, with a single learning rate ($\alpha = 0.01$), until Early stopping halted the process. Except for the default most up-to-date information on learning, no other enhancements were employed. Setting a sufficient number of training epochs was based on prior experience. The results of CIFAR-10 baseline tests are shown in Table 3.2. SGD fares marginally better than ADAM with default configurations in place. However, ADAM is faster than SGD in most cases.

3.2.2 Components of Improved Learning Scheme

In our method, we combined separately available learning improvements into a novel technique intending to improve the speed of convergence and accuracy. We aimed to

Model	Acc. SGD	Time	Acc. ADAM	Time
ResNet-34	90.19%	158 min	90.97%	64 min
ResNet-50	90.48%	325 min	88.56%	109 min
ResNet-101	92.70%	400 min	90.05%	176 min
ResNet-152	90.19%	453 min	89.58%	240 min
DenseNet-121	92.35%	231 min	83.56%	147 min
ResNext-50	92.84%	193 min	90.65%	74 min
Resnext-101	94.96%	441 min	90.81%	191 min

Table 3.2: Baseline test on CIFAR-10

preserve the valuable information from ImageNet pre-training in earlier layers to boost the learning pace and correctness. The components of the modified scheme are as follows:

Learning Rate (LR) Range Test

For optimizing model learning, robust convergence was necessary. One way to achieve this was by tweaking the learning rates throughout the training. But before this, knowledge of good initial rates was required. Conventionally, model learning has involved fixing the learning rate and leaving the optimization of the same to the built-in schedulers. These schedulers monotonically decrease the rate throughout the training. However, if the rate remained too big, the convergence to stable optimum would be impossible. If the rate was too low, that could lead to vanishing gradients. Hence, there were two niches to be optimized: (a) Finding the optimum range of rates and (b) scheduling the rates to give proper weight updates.

To find the best initial rate, we employed a scheme devised by Smith et al. [96]. In this method, the learning rate α was increased rapidly and progressively over several mini-batches, and the batch loss was observed. This batch loss typically decreased until a point of inflection where the losses started increasing again. Using the second moment of the rate estimate, we could find the point where the loss values were the lowest. Although this is not the perfect rate for all mini-batches by definition, the neighborhood of this point (in the logarithmic scale) was a good starting point for the model learning. Figure 3-1 shows several mini-batches being tested on CIFAR-10. The minimum is observed before $\alpha = 10^{-1}$ in Figure 3-2. Although Smith et al. considers this rate good for evaluation, it has been observed that in the case of imbalanced and homogeneous data, it is wiser to

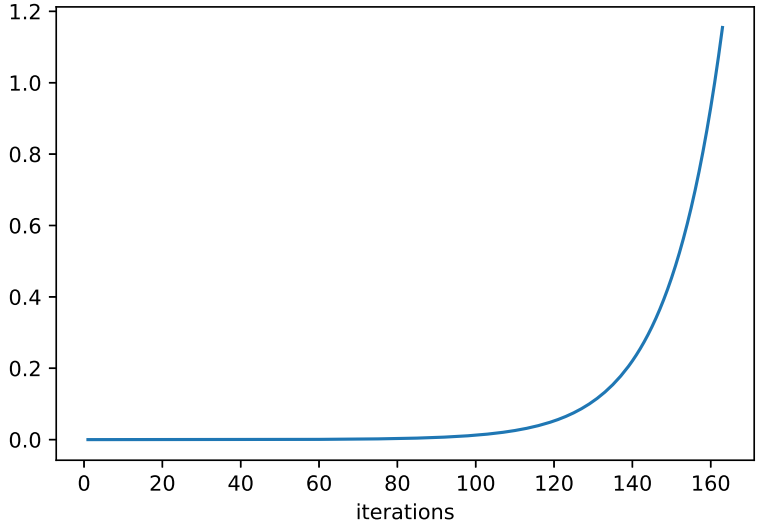


Figure 3-1: Learning rate range test with progressively higher α

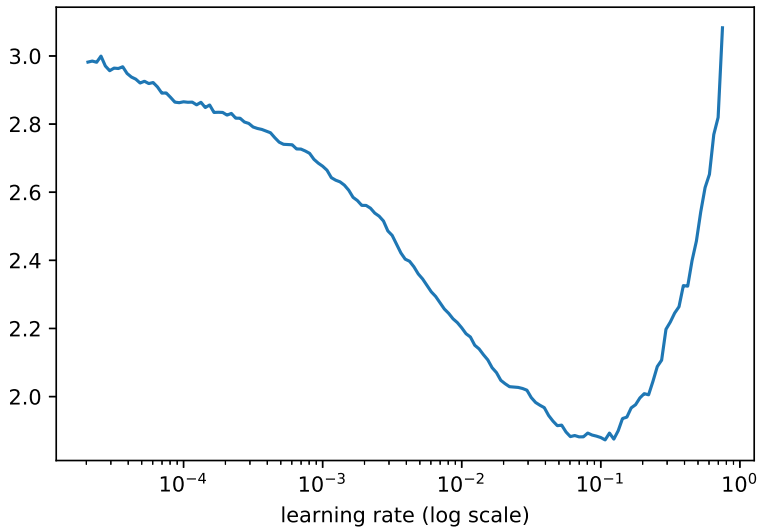


Figure 3-2: Validation loss inflexion point in range test

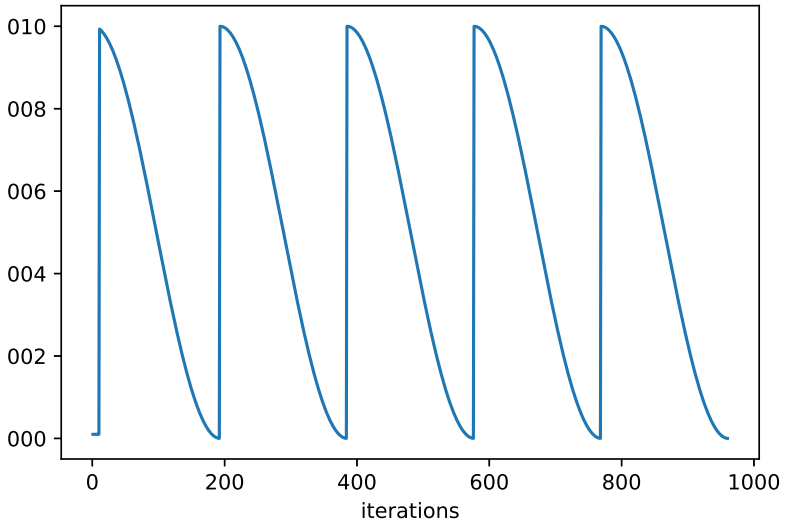


Figure 3-3: Conventional SGD-R with warm restarts per epoch

consider a neighborhood in a bracket in the logarithmic scale.

Stochastic Gradient Descent with Restarts

Our model improvements were based on Stochastic gradients with warm restarts (SGD-R) introduced by Loschilov et al. [71]. In this paper, instead of letting the learning rate to decay over the course of the training, the rate was cycled every epoch from a designated value to a minimum. Cycling of the rate every epoch gave perturbations to the parameters to jump out of local minima, if any. The general equation governing the rate cycling is given in Equation 3.2,

$$\nu_t = \nu_{min} + \frac{1}{2}(\nu_{opt} - \nu_{min})\left(1 + cos\left(\frac{\pi \cdot T_{cur}}{T_i}\right)\right) \tag{3.2}$$

v_t is the rate at any instant t, $(\nu_{opt} - \nu_{min})$ is the range over which the rate gets cycled. T_i is the number of iterations between restarts, and T_{cur} is the number of epochs since the last restart. The learning rate followed a cosine annealed curve between each epoch when $T_i = 1$. The schematic is shown in Figure 3-3. Loschilov et al. have demonstrated a higher convergence over plain Stochastic Gradient Descent with their method.

Discriminative learning rates (DLR) across layers

In a conventional neural network, the learning rate α is applied to the whole network. When working with pre-trained networks (such as pre-trained ImageNet), this may not be

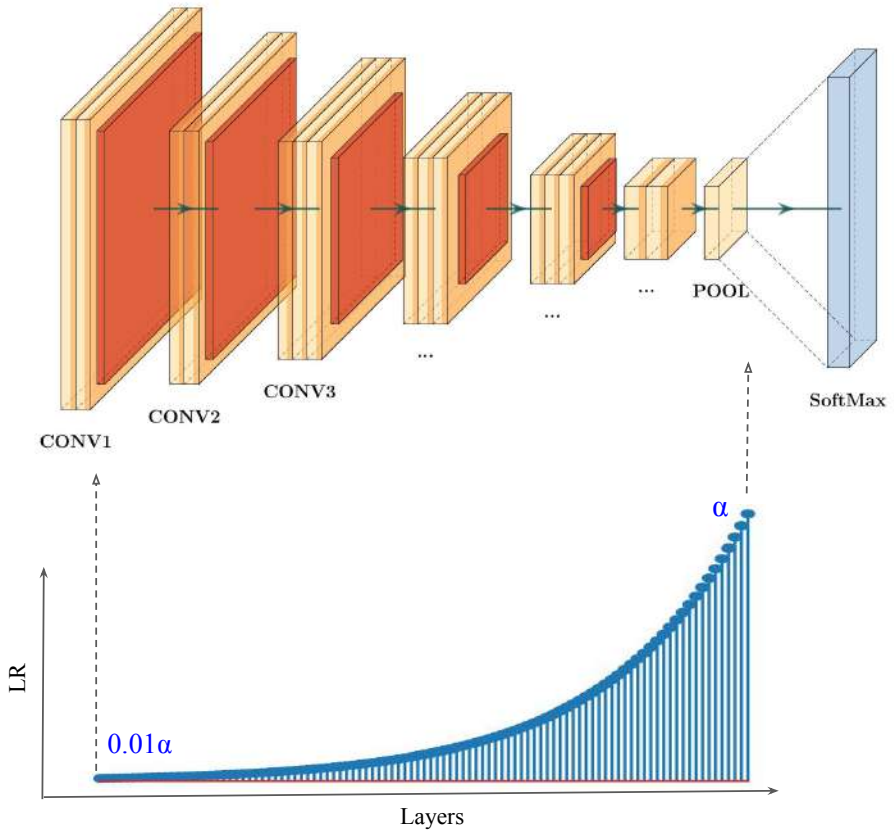


Figure 3-4: Schematic of DLR: Differential layers seeing different LR

very ideal. When a neural network learns representations, the initial few layers capture the most rudimentary geometrical information such as horizontal & vertical lines, corners, and edges. When transfer learning is approached with a unified learning rate scheme, the initial high value (before decay) is enough to destabilize the valuable information contained in these initial layers. As network representations are composed of features of the preceding layers, it could result in poorer learning in the final layers. For this specific reason, vanilla fine-tuning modifying convolutional neural networks (CNN) freeze the network except the final fully connected (FC) layer.

To learn a better representation than just fine-tuning the FC layer of CNN, we used throttling of the network by layer-wise rates [43]. In this discriminative rate scheme, different layers of the network saw different learning rates. The values ranged between the optimum rate α_{opt} applied to the final layer and $0.01\alpha_{opt}$ being used for the first layer. The rates for all other layers in between were logarithmic scaled between $0.01\alpha_{opt} - \alpha_{opt}$ depending on the position. This schema is shown in Figure 3-4.

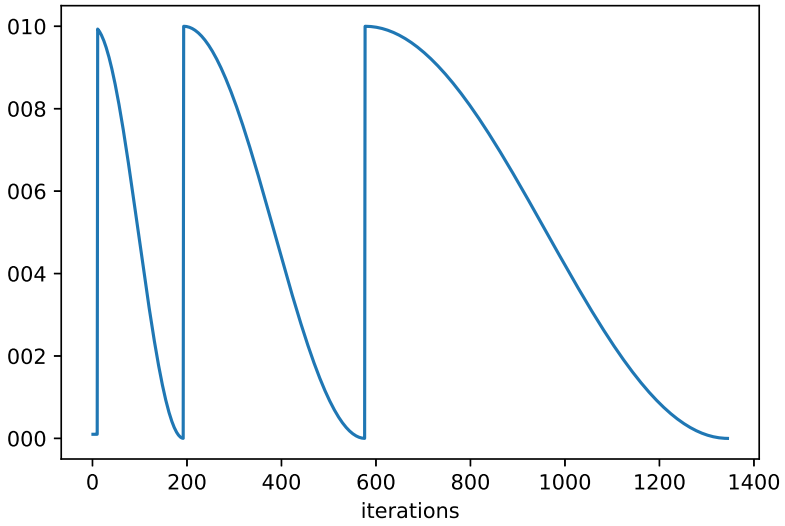


Figure 3-5: Cycle length multiplication in SGD-R

Discriminative learning rates, although computationally more intensive than conventional SGD-R, tended to commence fitting the network at a significantly higher accuracy/lower loss. This was due to the initial layers left relatively undisturbed and prone to very slow changes.

Cycle Length Multiplication (CLM)

In vanilla SGD-R, the cosine annealing of the learning rate cycle reset every epoch to give a warm restart. However, as the model progressively converges to an optimum value we required progressively lesser fluctuation. One idea would be to decrease the amplitude of the cosine to give a damping rate cycle. But this method was tested to detrimental results without much difference to a plain exponential decay (An exponential envelope to a learning function such as sine or cosine, exhibited its property close to a pure exponential pattern). An alternative approach was to reduce the frequency of the cosine cycle to cover increasingly more number of iterations (independent of the epoch) as shown by Loschilov et al. This scheme is illustrated in Figure 3-5, where the cycle length multiplies by a factor of two every restart. With such a rate cycle, the parameters were not disturbed too frequently and but at the same time, they receive a necessary perturbation one in a while to jump out of a local saddle point if any. Heuristically, if the perturbations occurred close to the optimum of the loss landscape, the model could still recover and converge within a few epochs.

Cycle-length multiplication used in conjugation with SGD-R did not exhibit significant differences in the beginning, but Loschilov et al. showed that the effects were more pronounced when learning advanced to a higher number of epochs. SGD-R was oscillating around the convex optimum but SGD-R with cycle length multiplication was able to converge to the optimum loss values by slowing down the learning rate annealing process.

We combined these aforementioned methods into one learning scheme and tested the performance on the CIFAR-10 and our homogeneous set of skin images.

3.2.3 Results on CIFAR-10

Our experiment setup for CIFAR-10 was similar to the baseline. We chose the same batch size as used in Section 3.2.1. The learning rate was deduced from the range test before the model fit. To evaluate the efficacy we concurrently ran vanilla SGD-R, SGD-R optimized with DLR, and SGD-R optimized with DLR & CLM. The model fitting concluded whenever Early stopping forced the gradient updates to end. The results of accuracy and validation loss are shown in Figures 3-6 through 3-9.

For every learning scheme in a model, the stable accuracy attained and elapsed time (rounded to the nearest minute) was recorded. These values are illustrated in Table 3.3. Further, the quality of the classification was also verified by receiver operator characteristics (ROC) and area under the curve (AUROC). Representative curves for four classes have been shown in Figure 3-10. The complete set of plots are available in Appendix A-1. Confusion matrices for model trained on SGD and our method are provided in Figures 3-11 and 3-12, with a comparative matrix between the two illustrated in Figure 3-13.

We also investigated the effect of having mixed-precision computing via third party libraries on the performance of these models. We chose the combined scheme only for comparison. A limited number of models are available via these external libraries. These are shown in Table 3.4. The mixed precision computing automatically scaled the precision of the parameters and variables from 32-bit floating point `(FP32)` to 16-bit floating point `(FP16)` and 16-bit integers `(INT16)` wherever necessary. This reduced the computational complexity, memory overhead, and matrix multiplication times, leading to much faster results at a small risk of over-fit.

For each model, we calculated the performance gain in terms of wall time. We define speedup as the ratio of time recorded in the SGD baseline over model convergence time

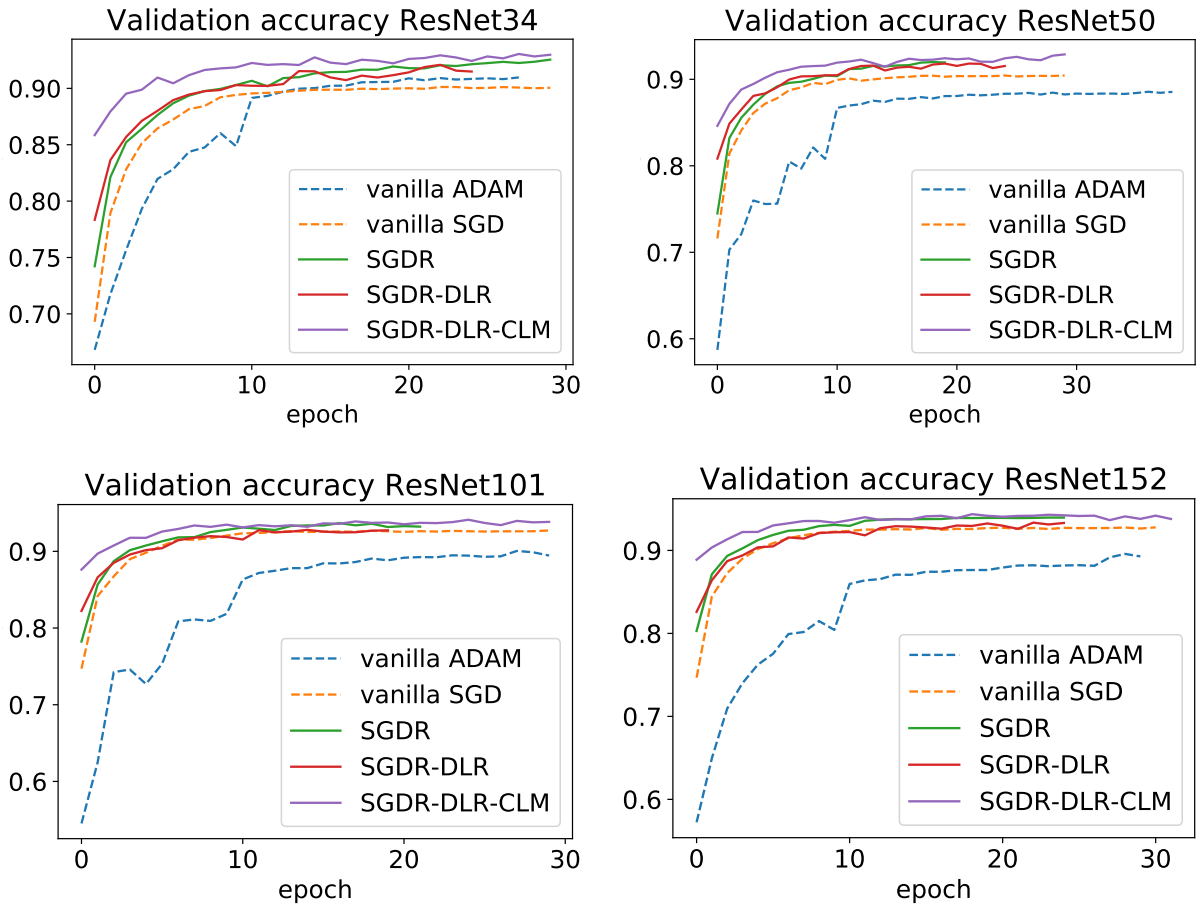


Figure 3-6: Validation accuracy for different ResNet models (CIFAR10)

in our scheme (Equation 3.3). Effective speedup would be seen whenever $\eta \geq 1$. In Table 3.5, η_1 and η_2 refer to the speedup measurements done with conventional precision and automatic mixed-precision respectively.

$$\eta = \frac{\text{Baseline wall time}}{\text{Modified Scheme wall time}} \tag{3.3}$$

Analysis on CIFAR-10 Data

As can be seen in accuracy curves for each model, the SGD optimizer acting as the baseline fared poorer in accuracy in comparison to the SGD-R based modifications. Vanilla ADAM in default configuration fared lower than SGD. This can however be improved by introducing breakpoints and restarting with lower learning rates. In the case of SGD combined with DLR, the fitting commenced at a much higher accuracy than plain SGD-R. They eventually converged to similar accuracy scores in most cases, albeit with minor statistical fluctuations. With SGD-R combined with DLR & CLM, the models not only

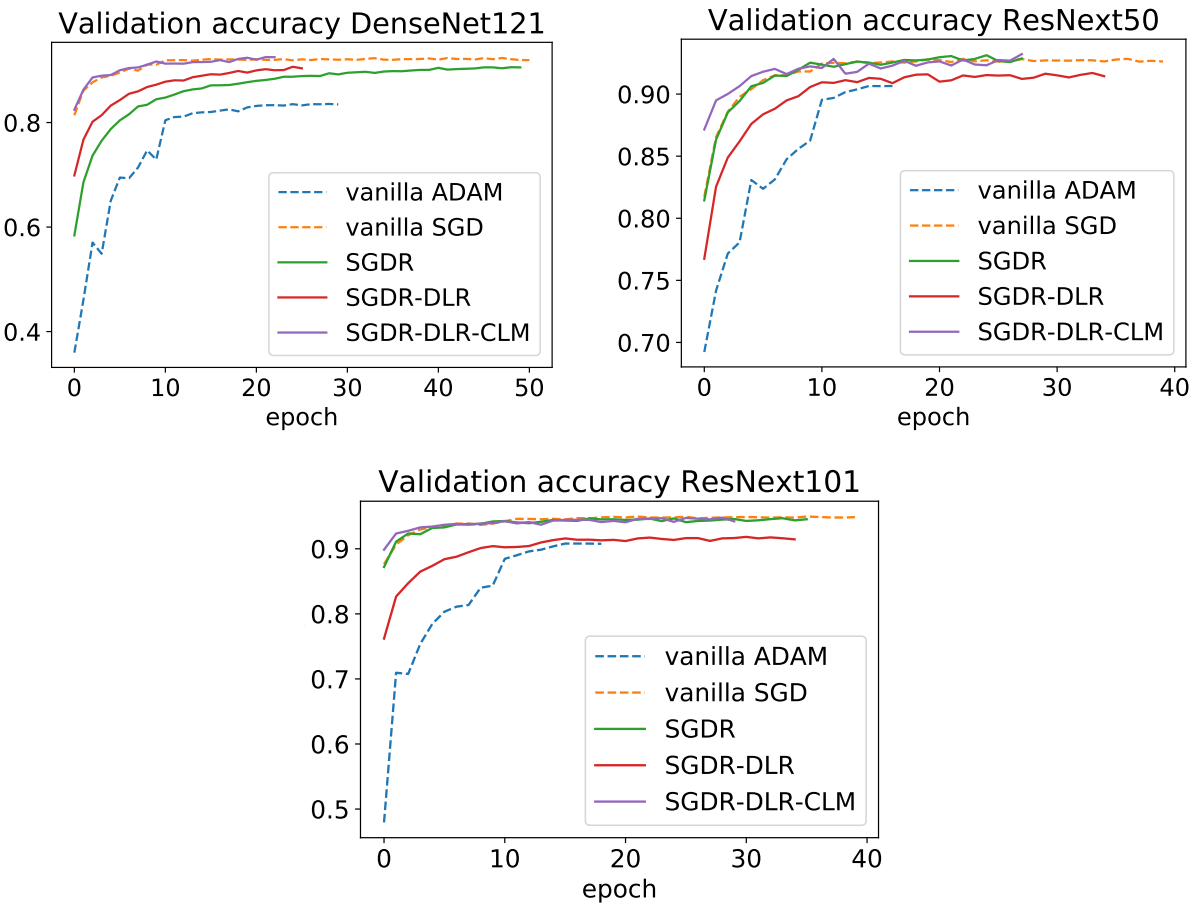


Figure 3-7: Validation accuracy in different learning schemes (CIFAR10)

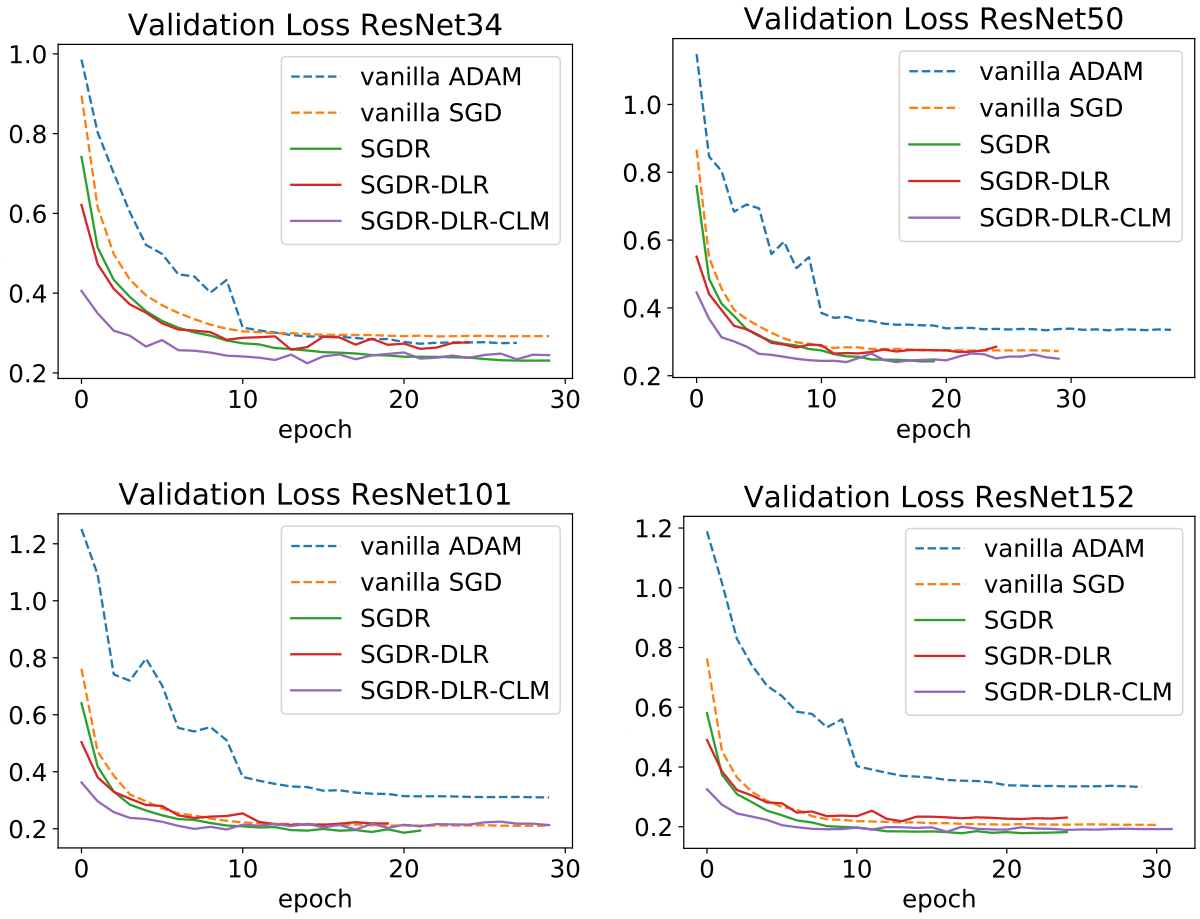


Figure 3-8: Validation loss in different learning schemes (CIFAR10)

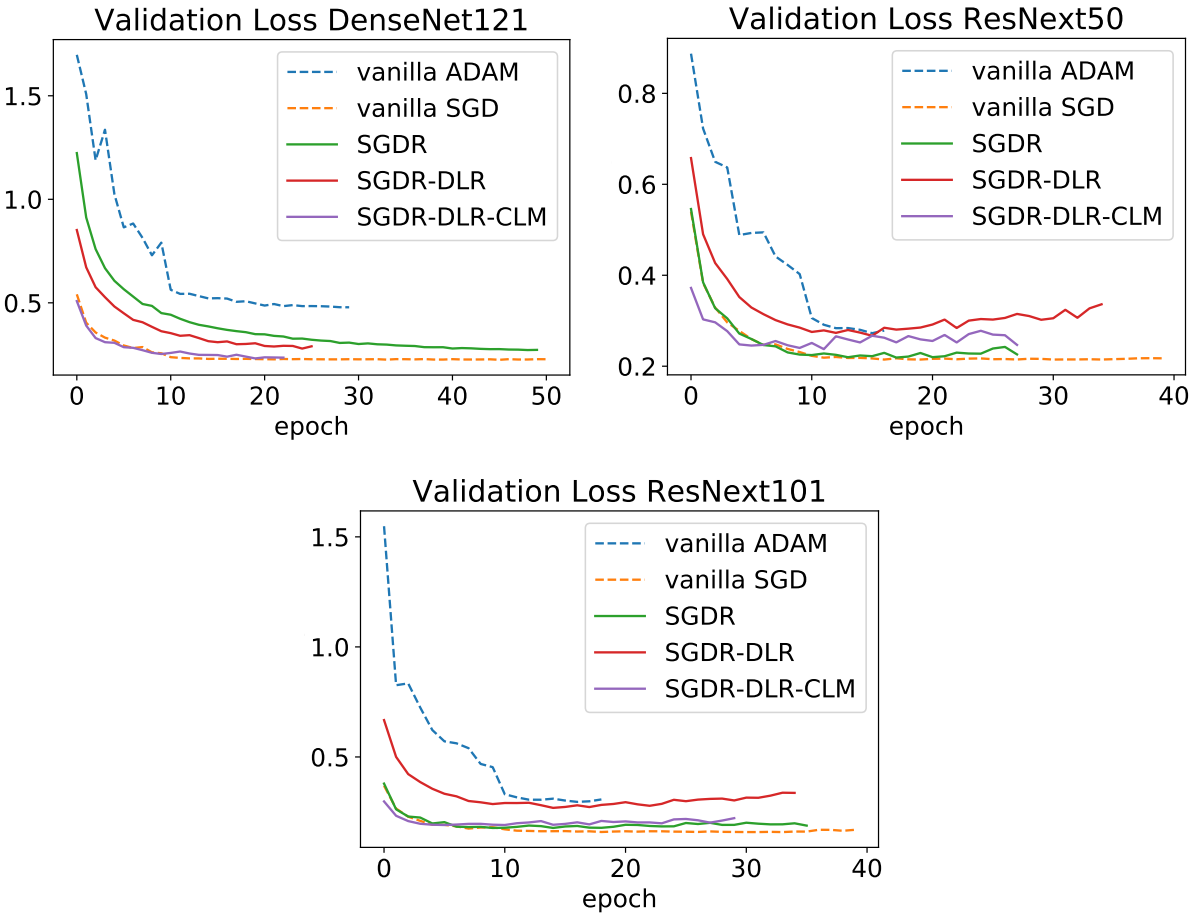


Figure 3-9: Validation loss in different learning schemes (CIFAR10)

Model	**Scheme**	**Top 1 Acc.**	**Time**
ResNet-34	ADAM	90.97%	64 min
	SGD	90.19%	158 min
	SGDR	92.55%	51 min
	SGDR+DLR	92.10%	56 min
	SGDR+DLR+CLM	93.05%	92 min
ResNet-50	ADAM	88.56%	109 min
	SGD	90.49%	325 min
	SGDR	92.00%	56 min
	SGDR+DLR	92.00%	81 min
	SGDR+DLR+CLM	92.88%	183 min
ResNet-101	ADAM	90.05%	176 min
	SGD	92.62%	400 min
	SGDR	93.90%	94 min
	SGDR+DLR	93.10%	122 min
	SGDR+DLR+CLM	94.12%	228 min
ResNet-152	ADAM	89.58%	240 min
	SGD	92.85%	453 min
	SGDR	93.00%	153 min
	SGDR+DLR	93.50%	208 min
	SGDR+DLR+CLM	94.37%	258 min
DenseNet-121	ADAM	83.56%	147 min
	SGD	92.35%	231 min
	SGDR	90.59%	250 min
	SGDR+DLR	90.67%	125 min
	SGDR+DLR+CLM	92.52%	212 min
ResNext-50	ADAM	90.65%	74 min
$(32 \times 4d)$	SGD	92.84%	193 min
	SGDR	93.13%	118 min
	SGDR+DLR	91.70%	155 min
	SGDR+DLR+CLM	93.22%	195 min
ResNext-101	ADAM	90.81%	191 min
$(32 \times 8d)$	SGD	94.96%	441 min
	SGDR	94.70%	385 min
	SGDR+DLR	91.75%	157 min
	SGDR+DLR+CLM	94.80%	381 min

Table 3.3: Model improvement metrics on CIFAR-10

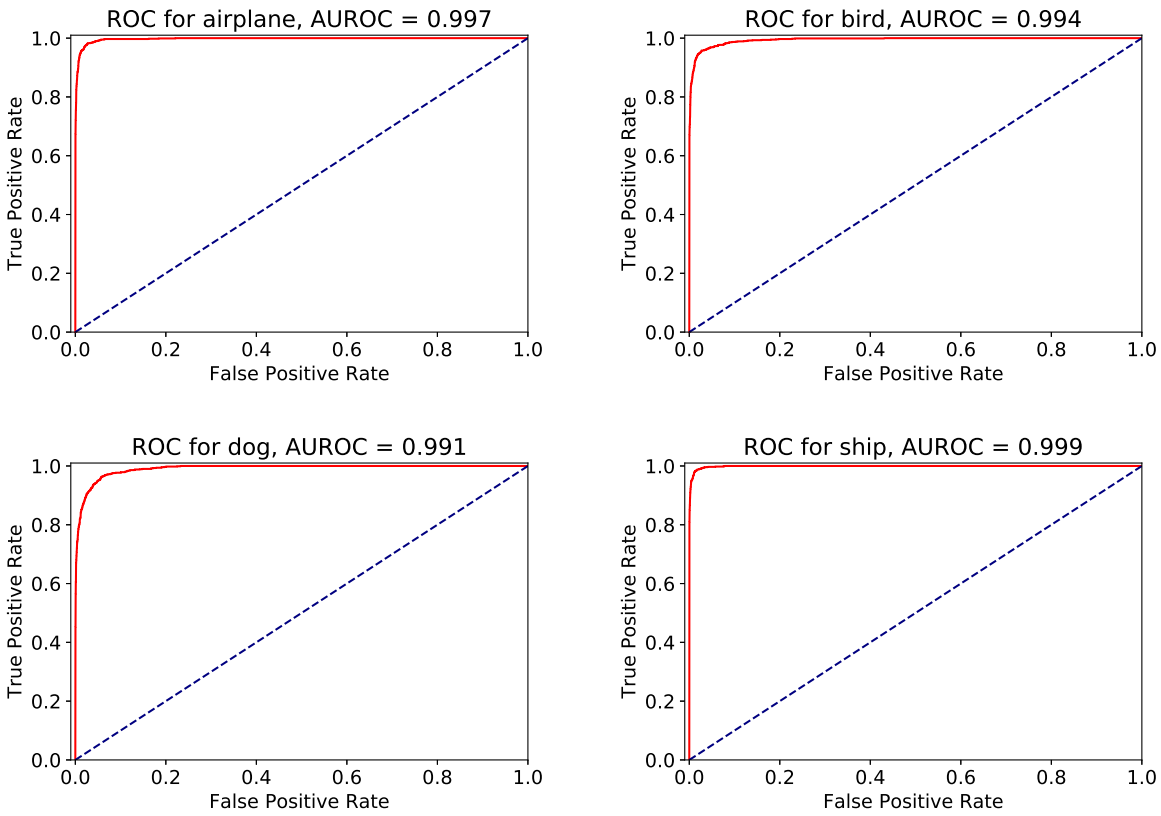


Figure 3-10: ROC curves & AUROC in improved training (CIFAR-10)

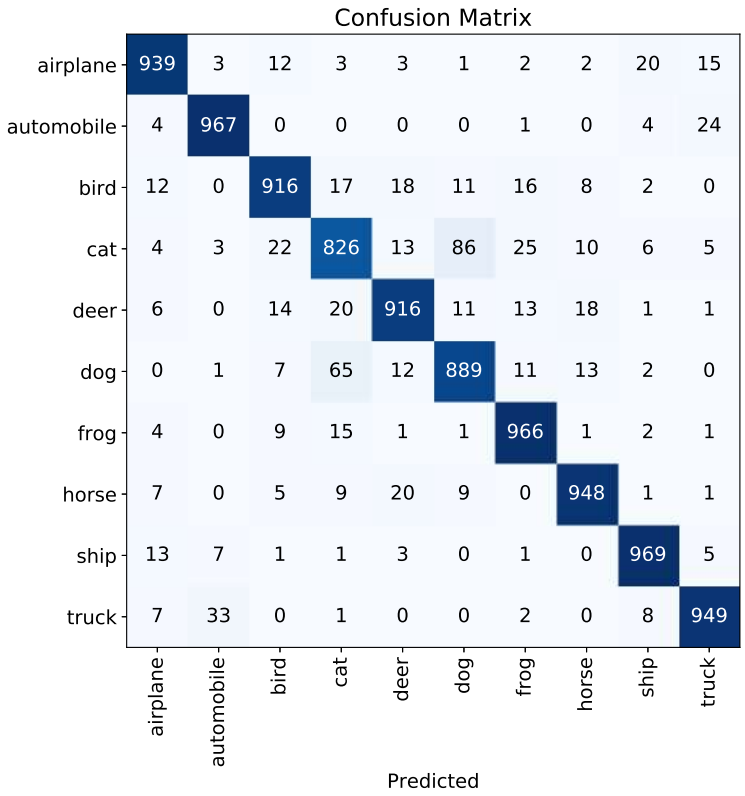


Figure 3-11: Confusion matrix for plain SGD (ResNet152)

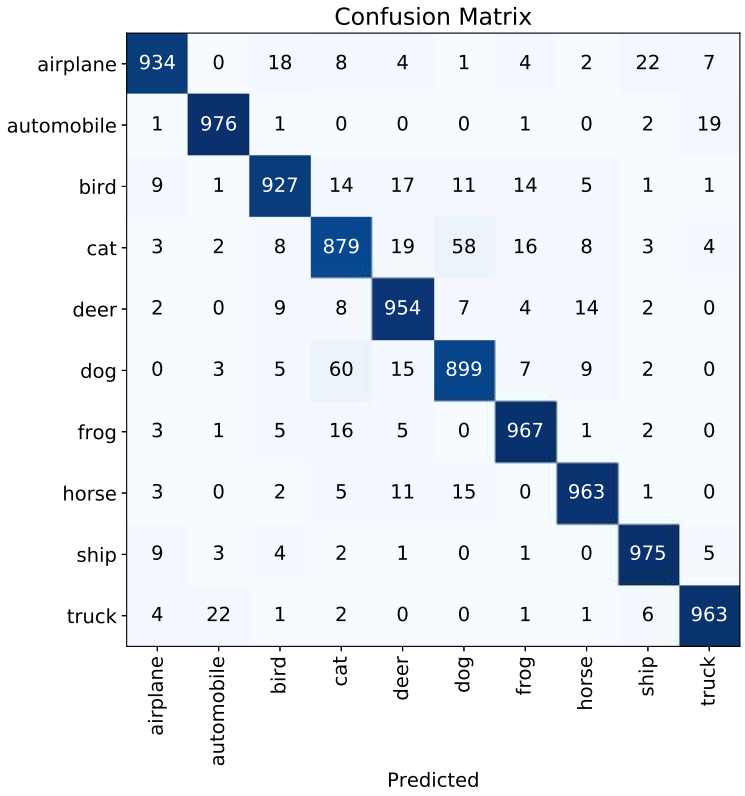


Figure 3-12: Confusion matrix for our training scheme (ResNet152)

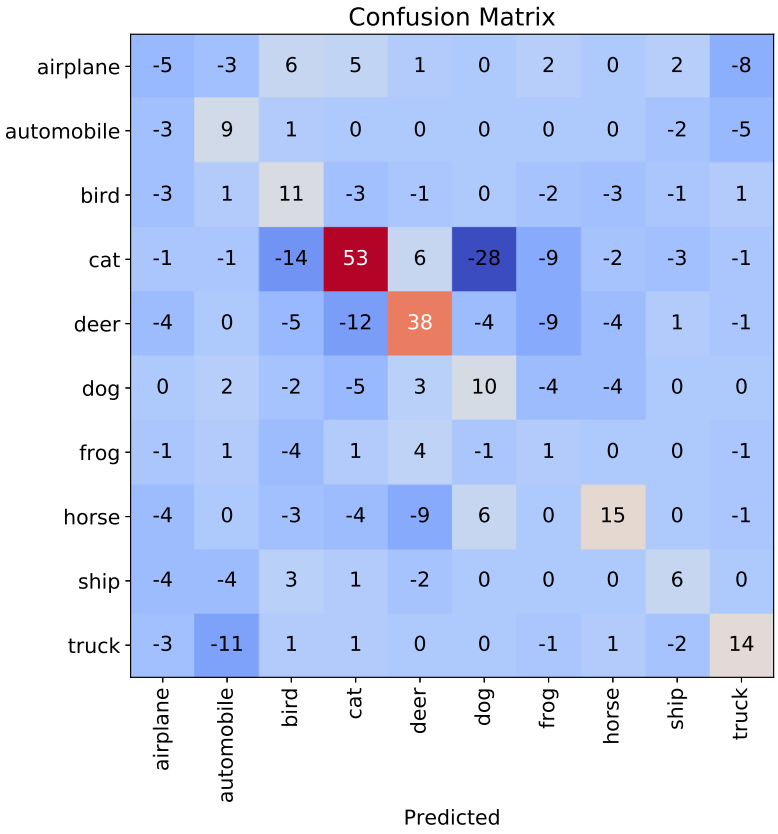


Figure 3-13: Confusion matrix Difference (Our method vs. SGD)

Model	**Acc** (SGD-R + DLR + CLM)	**Time**
ResNet-34	92.2%	26 min
ResNet-50	92.6%	55 min
ResNet-101	92.7%	78 min
ResNet-152	93.9%	112 min
DenseNet-121	93.4%	75 min

Table 3.4: AMP computing for SGDR+DLR+CLM (CIFAR-10)

Model	**Our scheme** (η_2)	**Ours + AMP** (η_2)
ResNet-34	1.71	6.07
ResNet-50	1.77	5.91
ResNet-101	1.75	5.12
ResNet-152	1.75	4.04
DenseNet-121	1.09	3.08

Table 3.5: Model speedup (CIFAR10)

commenced fitting at higher accuracy, but they also finished with a stable accuracy faster than all other methods. Heuristically, the higher initial fit is because of DLR addition and the faster convergence can be attributed to the CLM rate cycling. A similar trend can be observed in the loss curves, where both SGD-R with DLR and SGD-R with DLR & CLM show a much lower initial loss value, and quickly converge to a minimum. The only noteworthy exceptions were the case of newer models, where carefully tuned SGD could come close to the performance of SGD-R combined with DLR & CLM. In the case of ResNext-101, improvements were nearly indistinguishable.

With our modifications, the results for training speedup were striking. In the case of traditional linear ResNet models, we saw a consistent speedup over the baseline SGD, as seen in Table 3.4. These results were further improved when we used automatic mixed-precision via the external libraries, as seen in Table 3.5. The speeding up was higher in simpler models, reaching about being 6 times faster than baseline in best cases. DenseNet being a more complicated model with a large number of skip-back connections saw lesser improvement in both categories.

With a careful selection of hyperparameters, it is possible to train the CIFAR-10 dataset to near state-of-the-art accuracy within 30 epochs. Most models under investigation showed the capacity to learn features to reach convergent scores in the end. This is evident from the Top-1 Accuracy having scores within ± 4 percentage points from the

maximum recorded value. Hence, we conclude that given a certain amount of network depth, it is reasonable to expect high scores no matter how many layers the network has. We do not require very deep networks to get the best accuracy. Careful design of scheduling and optimizer functions can extract the best performance.

When models deal with applications such as wildlife images, medical photographs, or industrial assembly line video-feed, it is an important attribute to be able to train rapidly. Shallower networks prove to be advantageous here. Classification and recognition models built on lightweight, shallow networks but enhanced with good schedulers & optimizers can perform at par with traditionally selected large networks.

3.2.4 Results on Oxford-IIIT Pets Data

With kept a setup similar to the CIFAR-10 experiments (Refer Sec. 3.2.3), we did a model fitting on the Oxford IIIT Pets dataset composed of 37 classes. The baseline test procedure was kept similar to before. Thereafter, we deduced the optimal learning rate from the range test according to Sec. 3.2.2. SGD-R, SGDR & DLR, SGDR & DLR+CLM learning schemes were tested. The results on learning accuracy and loss trends are shown in Figures 3-14 through 3-17. A table summarizing the results from various schemes, including the baseline, is provided in Table 3.6. The ROC and AUROC values for 16 representative labels are shown in Figures 3-18 & 3-19. The rest are provided in Figure A-3 and Figure A-4. A limited number of these models were tested via third-party libraries which enabled mixed precision. The results of these trials are shown in Tables 3.7–3.8. Due to the presence of a large number of classes, confusion matrices cannot be shown effectively and hence excluded from consideration.

Analysis of IIIT Pets Data

In the case of the Oxford-IIIT Pets dataset, the various model accuracies and validation loss curves for the improved scheme were either better or on par with using robustly configured SGD or ADAM optimizers. ADAM proved to have unstable convergence towards stable model performance with heavy perturbations in intermediate epochs, perhaps indicating its unsuitability in fine-grained data. Comparing with highest performing contributors to open fine-grained classification challenges in Kaggle, such as *iMaterialist* and *iNaturalist*, the optimizer of choice has predominantly been SGD.

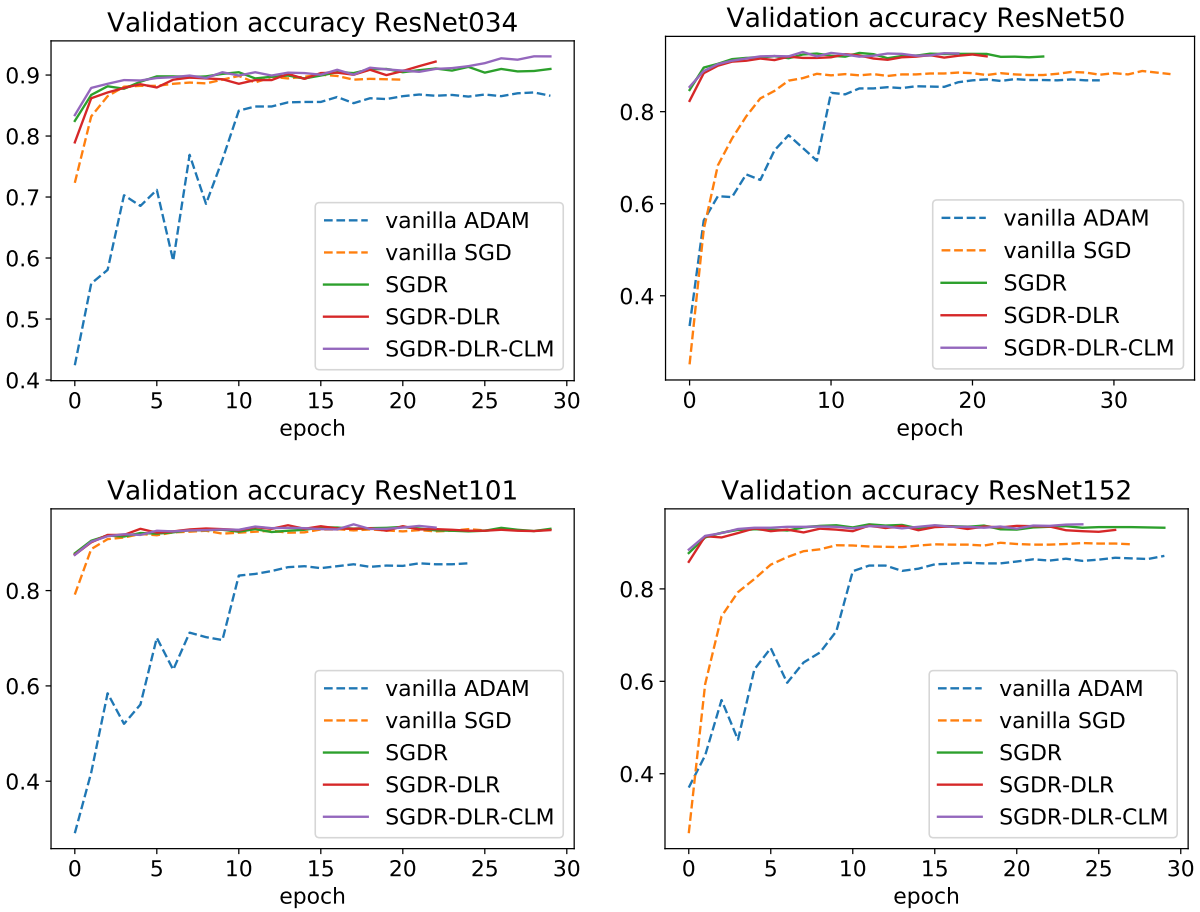


Figure 3-14: Validation accuracy for ResNets (Oxford-IIIT Pets)

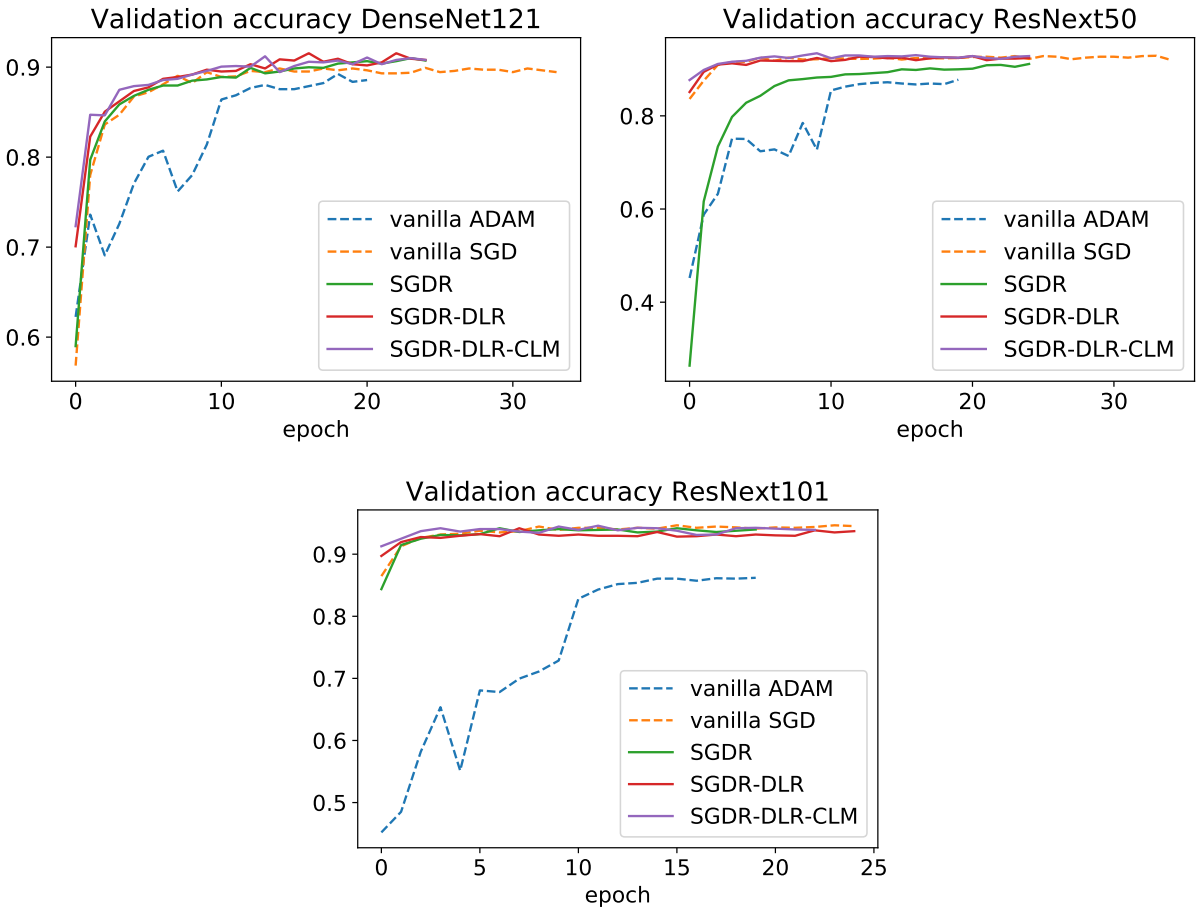


Figure 3-15: Validation accuracy for other networks (Oxford-IIIT Pets)

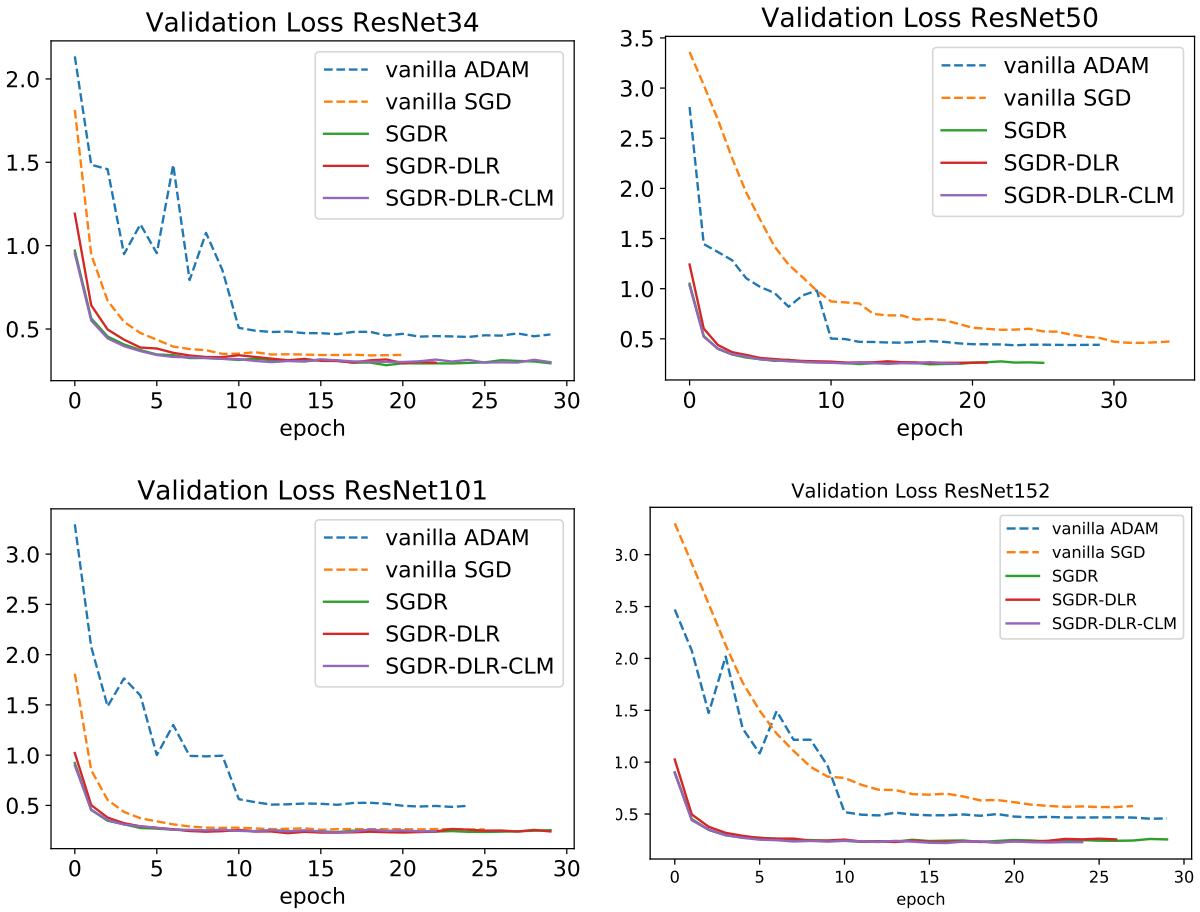


Figure 3-16: Validation loss for ResNets (Oxford-IIIT Pets)

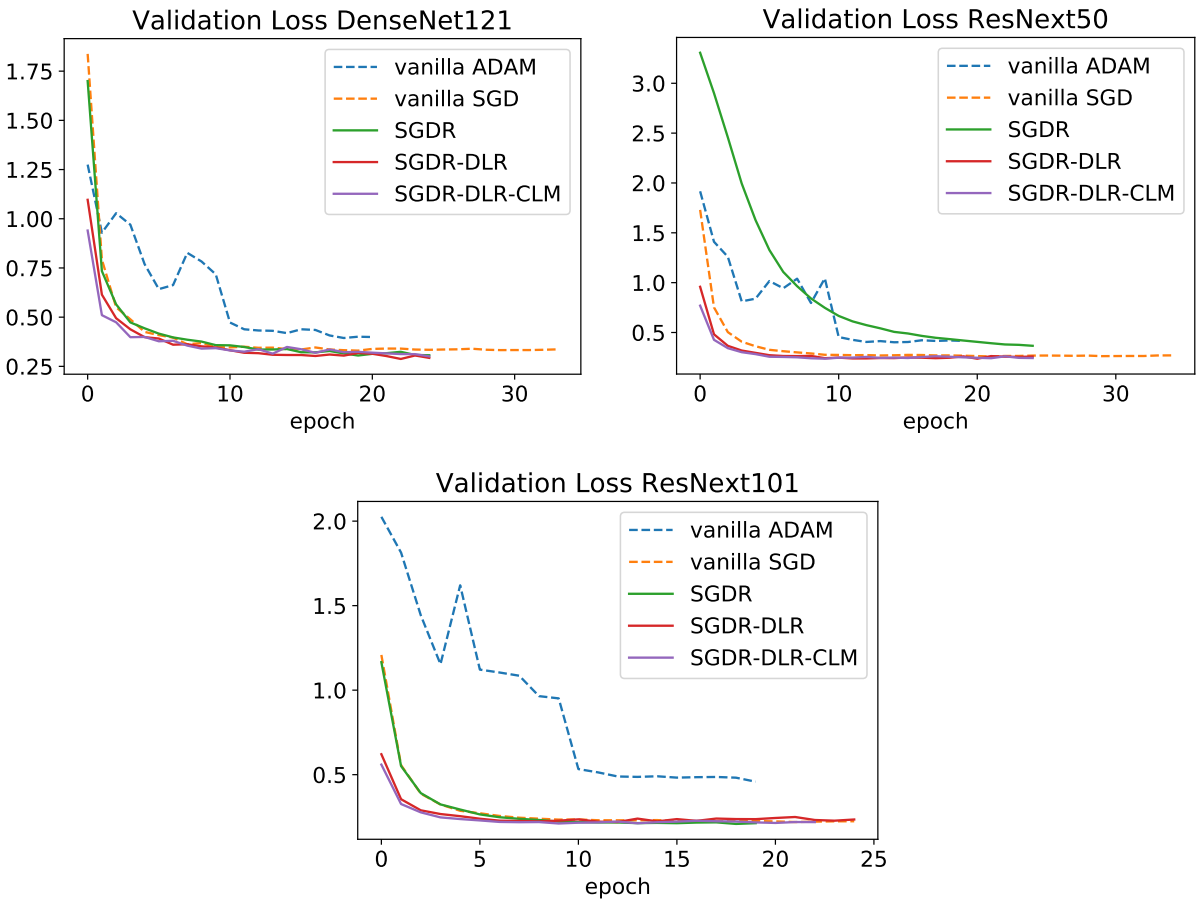


Figure 3-17: Validation loss for other networks (Oxford-IIIT Pets)

Model	**Scheme**	**Top 1 Acc.**	**Time**
ResNet-34	ADAM	87.14%	20 min
	SGD	89.78%	14 min
	SGDR	91.33%	20 min
	SGDR+DLR	92.20%	15 min
	SGDR+DLR+CLM	93.06%	15 min
ResNet-50	ADAM	87.07%	43 min
	SGD	88.83%	24 min
	SGDR	92.55%	18 min
	SGDR+DLR	92.42%	15 min
	SGDR+DLR+CLM	92.63%	11 min
ResNet-101	ADAM	85.72%	23 min
	SGD	92.89%	20 min
	SGDR	92.96%	28 min
	SGDR+DLR	92.70%	22 min
	SGDR+DLR+CLM	93.57%	15 min
ResNet-152	ADAM	87.14%	36 min
	SGD	89.91%	27 min
	SGDR	93.57%	31 min
	SGDR+DLR	92.76%	25 min
	SGDR+DLR+CLM	93.97%	23 min
DenseNet-121	ADAM	89.24%	17 min
	SGD	89.85%	25 min
	SGDR	91.00%	20 min
	SGDR+DLR	91.54%	19 min
	SGDR+DLR+CLM	91.01%	14 min
ResNext-50	ADAM	87.75%	16 min
$(32 \times 4d)$	SGD	92.89%	23 min
	SGDR	91.13%	18 min
	SGDR+DLR	92.42%	16 min
	SGDR+DLR+CLM	92.89%	15 min
ResNext-101	ADAM	86.19%	31 min
$(32 \times 8d)$	SGD	94.65%	37 min
	SGDR	93.10%	30 min
	SGDR+DLR	93.70%	35 min
	SGDR+DLR+CLM	94.24%	31 min

Table 3.6: Model improvement metrics on Oxford-IIIT Pets Data

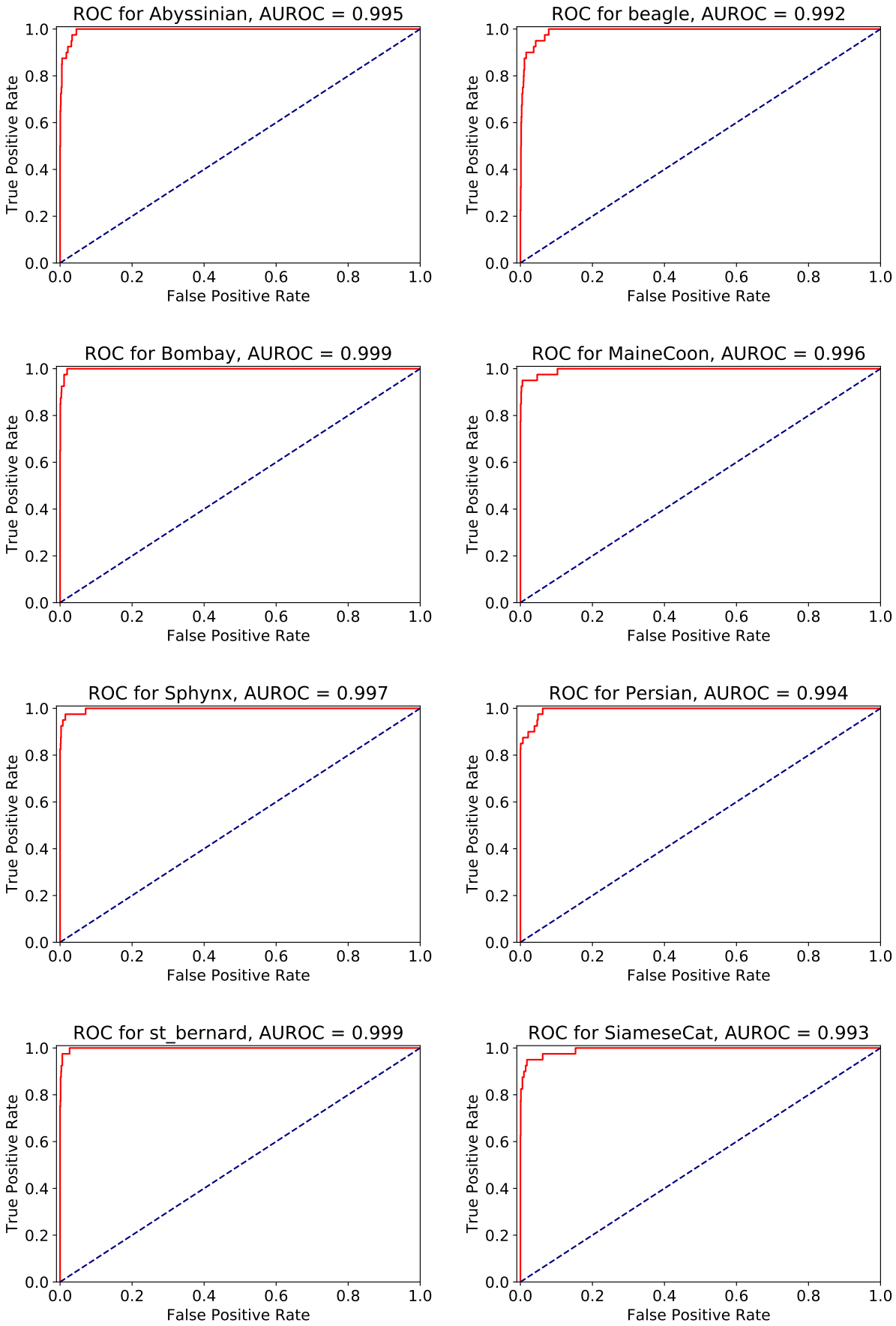


Figure 3-18: ROC curves & AUROC (IIIT Pets Set A)

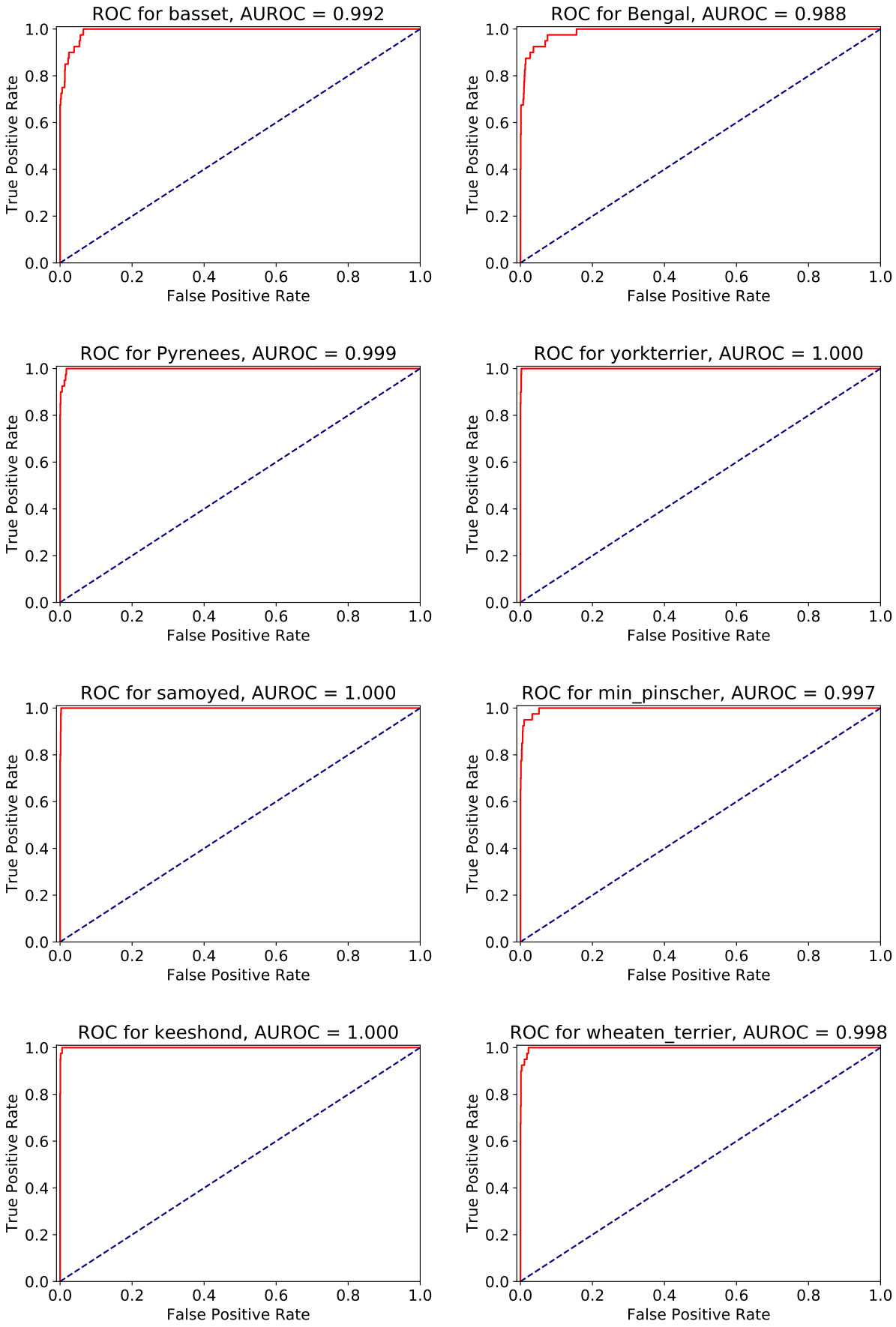


Figure 3-19: ROC curves & AUROC (IIIT Pets Set-B)

Model	**Acc** (SGD-R + DLR + CLM)	**Time**
ResNet-34	93.30%	194 sec
ResNet-50	93.70%	297 sec
ResNet-101	94.58%	410 sec
ResNet-152	94.31%	512 sec
DenseNet-121	93.70%	355 sec

Table 3.7: AMP computing for SGDR+DLR+CLM (IIIT Pets)

Model	**Our scheme** (η_2)	**Ours + AMP** (η_2)
ResNet-34	0.93	4.06
ResNet-50	2.18	4.85
ResNet-101	1.33	2.85
ResNet-152	1.17	3.17
DenseNet-121	1.78	3.84

Table 3.8: Model speedup (IIIT Pets)

With our improvements, different models had Top-1 accuracy in a narrow spread between 91.01%–94.24%. This is approximately only 2.5% less than the current state-of-the-art *Sharpness Aware Minimization* and an order of magnitude faster [29]. The difference between our improvised scheme and plain SGD is between 6–8% on average. With mixed-precision computing, the model fit can be achieved in a few minutes with careful choice of hyperparameters as demonstrated in Tables 3.7 and 3.8. In all the model learning cases, we have verified the fit by receiver operator characteristics (as seen in Figure 3-18 and 3-19.

3.2.5 Results on Exmedio Data

With a setup similar to the CIFAR-10 experiments and adapting from IIIT-Pets dataset (Refer Sec. 3.2.3–3.2.4), we did model fitting on the Exmedio Dermatological data. We ran a baseline test according to the procedure elaborated in Sec. 3.2.1. Thereafter, we deduced the optimal learning rate from the range test according to Sec. 3.2.2. Concurrently, SGD-R, SGDR & DLR, SGDR & DLR+CLM were tested. The results on model accuracy and trends on the loss curves are shown in Figures 3-20 through 3-23. A table summarizing the results from various schemes, including the baseline, is provided in Table 3.9. The ROC and AUROC values for 4 representative labels are shown in Figure 3-24. The complete list of ROC figures are available in Appendix A-2. Representative

confusion matrices for vanilla SGD and our combination method used in ResNet-152 are provided in Figures 3-25 and 3-26 respectively. A comparative confusion matrix showing the difference before and after improvement is shown in Figure 3-27.

As in the previous case, a limited number of these models were tested via third-party libraries which enabled mixed precision. The results of these trials are shown in Tables 3.10–3.11.

Analysis on Exmedio Data

In all the cases of ResNet models, the vanilla optimizers (SGD and ADAM) could only fit up to a certain upper limit. Because of the homogeneous nature of data, the gradient updates slowed quickly and the accuracy curves flattened. SGD-R demonstrated marginally better performance than vanilla SGD. In all the ResNet models, however, SGD-R combined with DLR gave an incrementally better fit and accuracy. The performance of SGD-R combined with DLR and CLM was a notch better even in comparison with SGD-R with DLR. Just like the case of CIFAR-10, the model fit began at a much higher accuracy in SGD-R with DLR & CLM than all other learning schemes. Consequently, it was able to reach the peak accuracy range much quicker.

This was a bit different in the case of newer architectures, which showed a surprising parity in the results obtained even with SGD. The gap encountered by the optimizers were compensated to a good extent by the dense network architectures and presence of parallel nodes. Due to these architectural improvements, the models could reach similar accuracy levels as our modified scheme. In all the models seen, the performance of ADAM optimizer surprisingly marginal to significantly poorer than SGD based optimization process. In the case of ResNext-101, all SGD based optimization schemes performed were near-equivalent. The best results in accuracy were obtained in ResNext-101, although they fit much slower than all other models.

Automatic mixed precision in training could improve the speed of model learning significantly. As seen in Table 3.10, all the models were learned in under 20 minutes. Between 8 to 12 epochs were used in all these cases. Our learning scheme without mixed-precision enabled could achieve between 1.32 to 2.33 times speeding up. The same process with mixed precision enabled the learning to go faster by over 3.5 times in aggregate, as seen in Table 3.11.

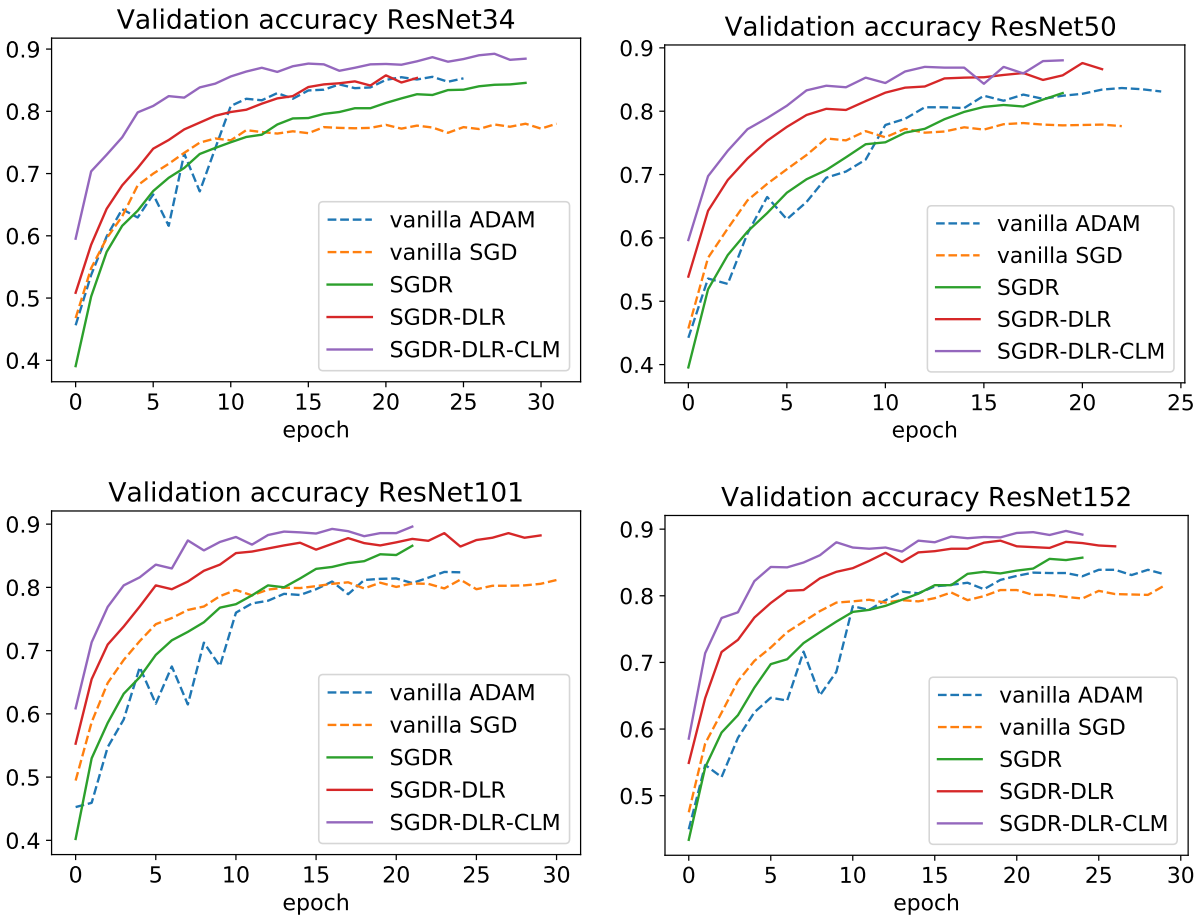


Figure 3-20: Validation accuracy for ResNets (Exmedio)

In the case of fine-grained classification, these results are encouraging. The experiments which have relied on a single learning rate or a narrow bracket thereof can be optimized further if we use SGD-R combined with modifications. We can hope for peak performance if we can introduce breakpoints in the fitting process, where we can scale down the rates incrementally and lead to even better fits than *status quo*.

3.3 Chapter Summary

In deep learning research, the proclivity of practitioners has been to pick models with a large number of layers for better accuracy. Although deeper models perform well, they introduce high computational cost in the learning problem. Higher memory and bandwidth (in terms of GPU frame buffer) are needed to solve the task at hand.

The current work has highlighted that the usability of layers can be improved by robust training. A fewer number of layers can achieve the same performance as very

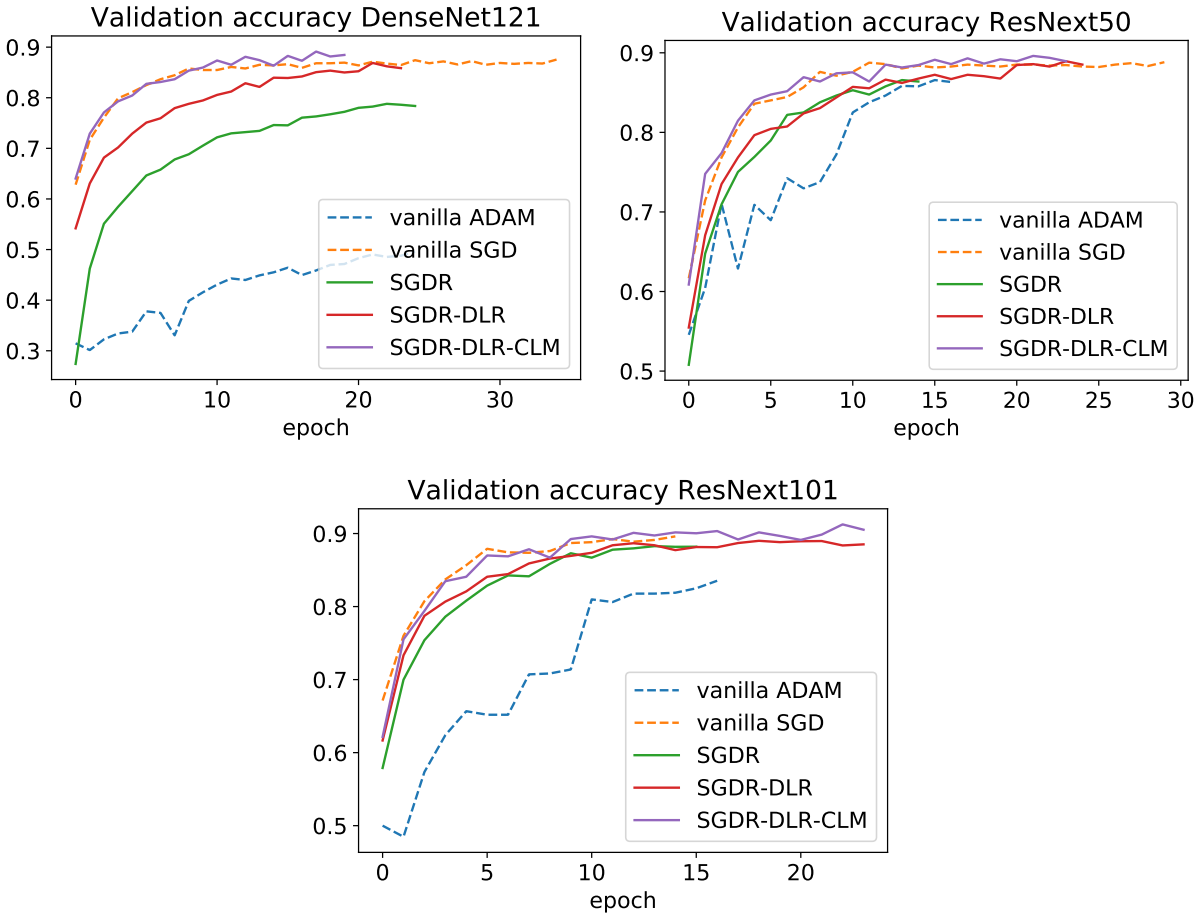


Figure 3-21: Validation accuracy for other networks (Exmedio)

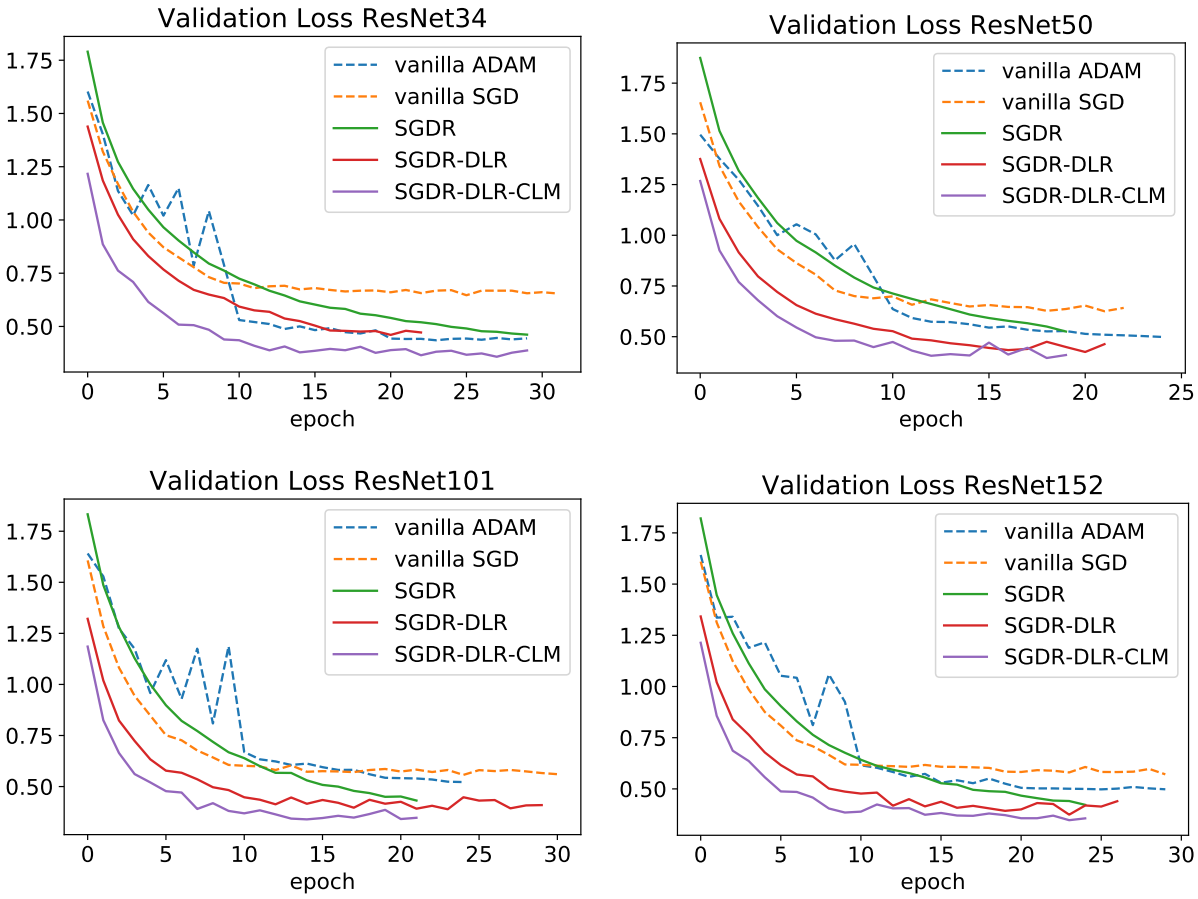


Figure 3-22: Validation loss for ResNets (Exmedio)

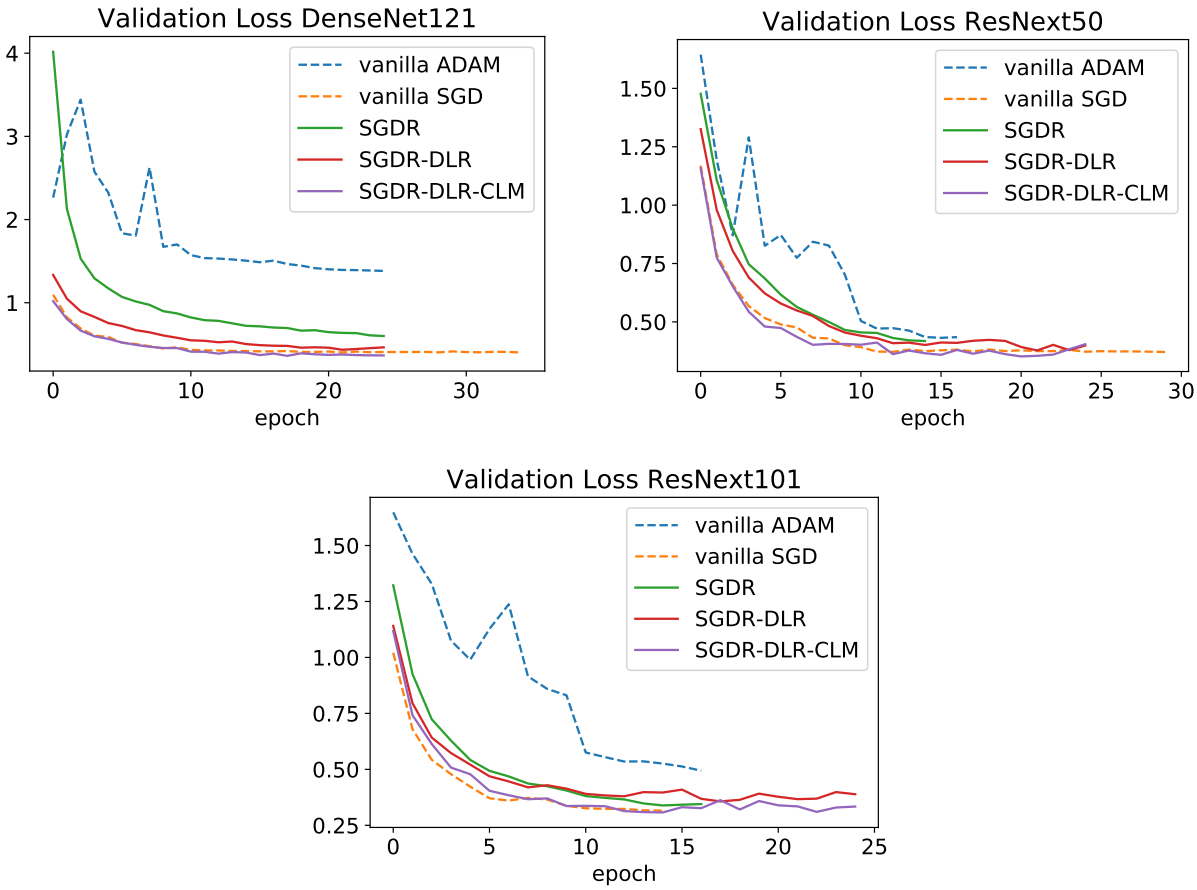


Figure 3-23: Validation loss for other networks (Exmedio)

Model	**Scheme**	**Top 1 Acc.**	**Time**
ResNet-34	ADAM	85.54%	19 min
	SGD	78.07%	46 min
	SGDR	84.02%	20 min
	SGDR+DLR	85.78%	21 min
	SGDR+DLR+CLM	89.24%	31 min
ResNet-50	ADAM	83.65%	22 min
	SGD	78.12%	45 min
	SGDR	82.86%	33 min
	SGDR+DLR	87.60%	23 min
	SGDR+DLR+CLM	88.03%	34 min
ResNet-101	ADAM	82.44%	20 min
	SGD	80.25%	55 min
	SGDR	86.51%	24 min
	SGDR+DLR	87.84%	33 min
	SGDR+DLR+CLM	88.57%	28 min
ResNet-152	ADAM	83.90%	34 min
	SGD	81.50%	65 min
	SGDR	85.70%	41 min
	SGDR+DLR	88.27%	24 min
	SGDR+DLR+CLM	89.55%	47 min
DenseNet-121	ADAM	50.57%	25 min
	SGD	87.54%	28 min
	SGDR	78.79%	15 min
	SGDR+DLR	86.87%	14 min
	SGDR+DLR+CLM	89.12%	12 min
ResNext-50	ADAM	86.58%	12 min
$(32 \times 4d)$	SGD	88.82%	19 min
	SGDR	86.57%	10 min
	SGDR+DLR	88.94%	16 min
	SGDR+DLR+CLM	89.61%	15 min
ResNext-101	ADAM	83.54%	27 min
$(32 \times 8d)$	SGD	89.23%	23 min
	SGDR	88.27%	21 min
	SGDR+DLR	88.94%	32 min
	SGDR+DLR+CLM	91.25%	33 min

Table 3.9: Model improvement metrics on Exmedio Data

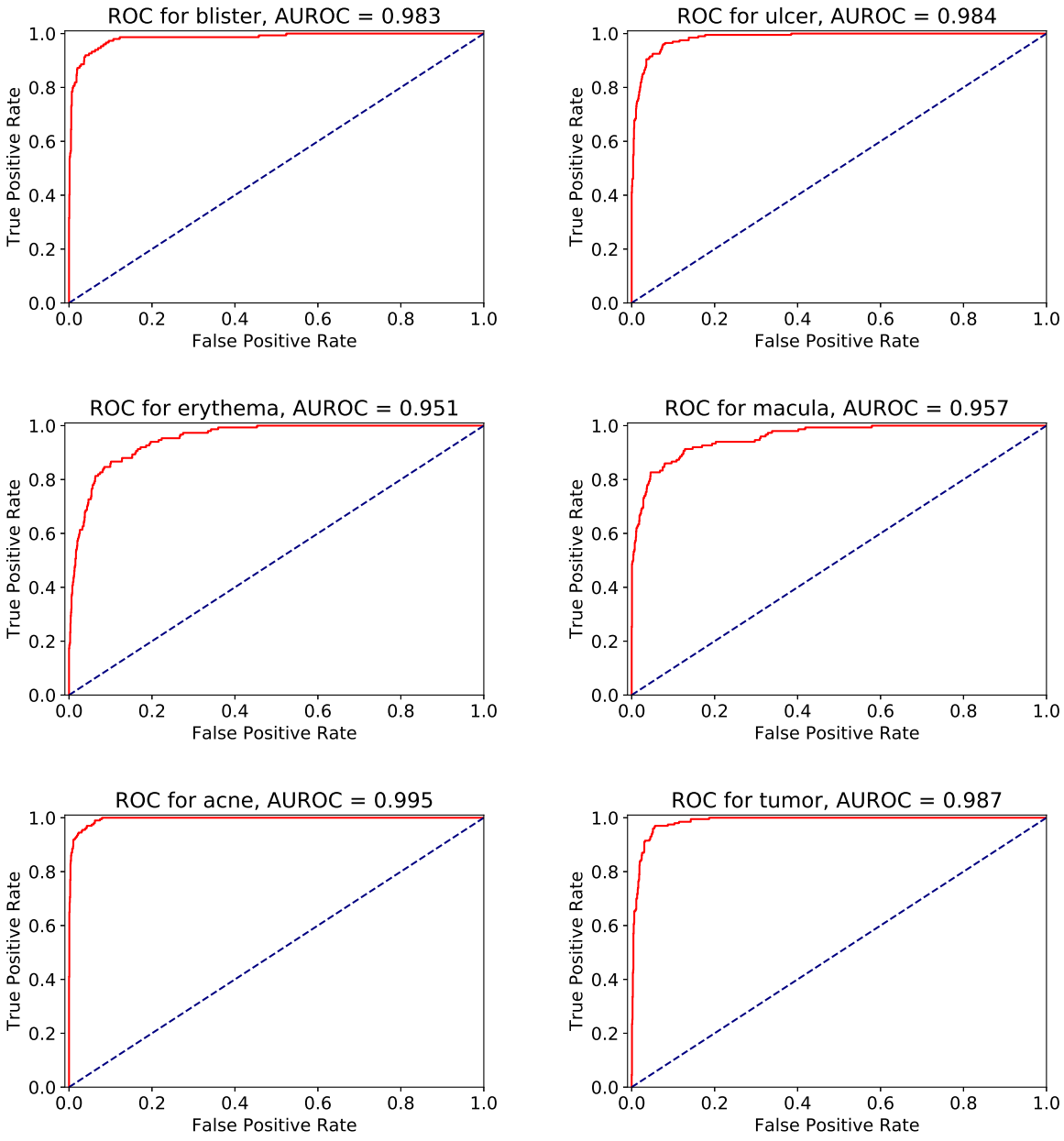


Figure 3-24: ROC curves & AUROC in improved training (Exmedio)

Model	**Acc** (SGD-R + DLR + CLM)	**Time**
ResNet-34	86.1%	12 min
ResNet-50	86.9%	12 min
ResNet-101	89.7%	15 min
ResNet-152	89.4%	19 min
DenseNet-121	88.5%	15 min

Table 3.10: AMP computing for SGDR+DLR+CLM (Exmedio)

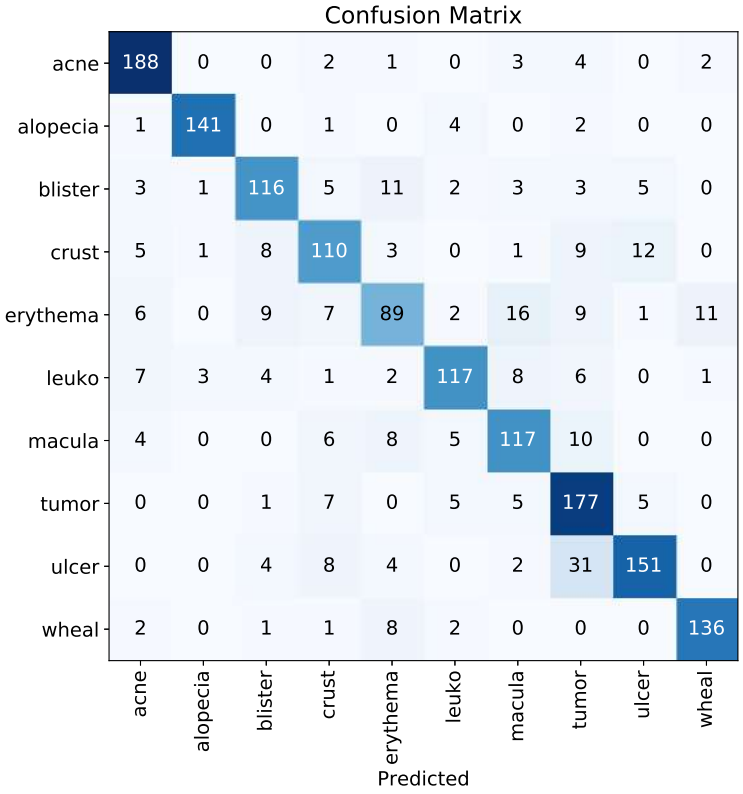


Figure 3-25: Confusion matrix for plain SGD

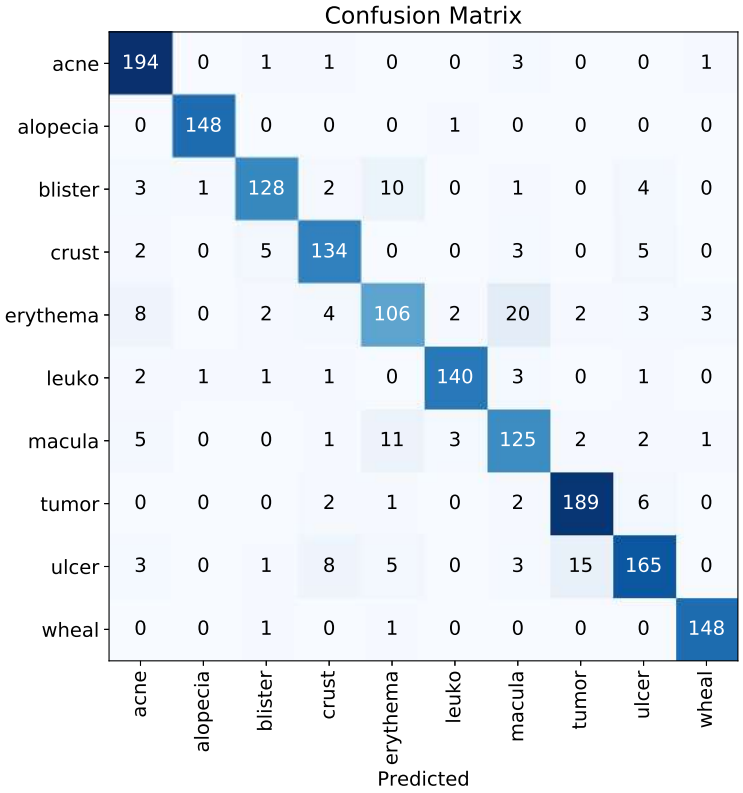


Figure 3-26: Confusion matrix for our training scheme

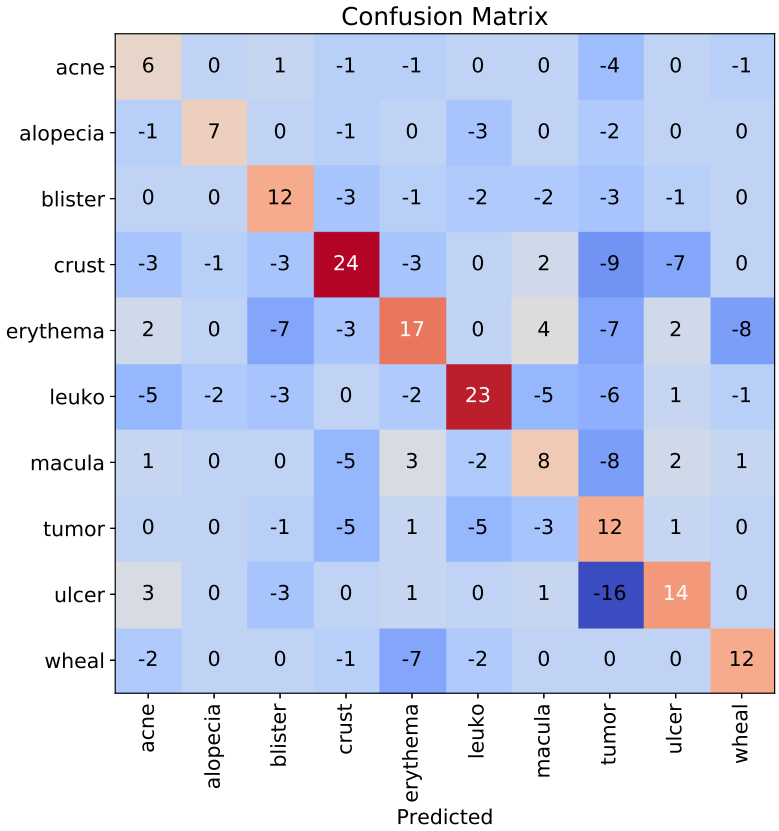


Figure 3-27: Confusion matrix Difference (Our method vs. SGD) (Exmedio)

Model	**Our scheme** (η_2)	**Ours + AMP** (η_2)
ResNet-34	1.48	3.82
ResNet-50	1.32	3.81
ResNet-101	1.95	3.66
ResNet-152	1.38	3.42
DenseNet-121	2.33	1.67

Table 3.11: Model speedup (Exmedio)

deep networks, if properly optimized. We found that the ingredient to a good model fit lay in learning rate cycling and layer-specific tuning. Working across three datasets, we observed that smaller architectures were as competent as large ones in many classification tasks.

The optimization scheme discussed in this chapter also managed to reduce the time to fit the model. Performance close to the state-of-the-art was seen to be realized in a fraction of the time, as compared to the baseline techniques with commonly recommended practices. These metrics saw further improvement with mixed-precision training.

Chapter 4

Preliminary Interpretations

In Chapter 3, we discussed strategies to learn our models to their best fits. We discovered that even with large differences in sizes and architectures, they can be tuned to give near-equal performance by good choices in scheduling and optimizers. We further verified the goodness of fit by checking the loss curves and confusion matrices. However, the peak accuracy obtained by these models were far below the ideal situation. In application cases such as medical diagnostics, the margin of error is very slim. A gap of approximately 10% between peak model performance and perfect detection is hard to ignore.

Human cognition is far superior to machine-learned predictions. That remains the basis of curating image data corpora today. When working with data for machine learning, we work with the assumption that human annotators have done the task with perfection. However, there exists a small margin of error with us, known as *Bayes optimal error* [81]. In lesser-known or custom data, this metric is not readily known. The performance we see with our machine predictions are a combination of Bayes error and the inability to models to fit data perfectly, without *memorizing* the training set.

Although we cannot remove the element of Bayes error in these predictions, we can analyze the failure of models to correctly predict what the human annotators would have otherwise not been mistaken with. With convergent model performance in most cases, we can attribute our errors to the nature of the data itself. In this chapter, we have taken up several case studies on homogeneous image corpus. We try to understand the pattern by quantifying the errors among labels. We try to pick the most egregious cases and rationalize the reason why the model made incorrect predictions. By understanding the rationale behind the errors, we can curate data better for machine learning tasks.

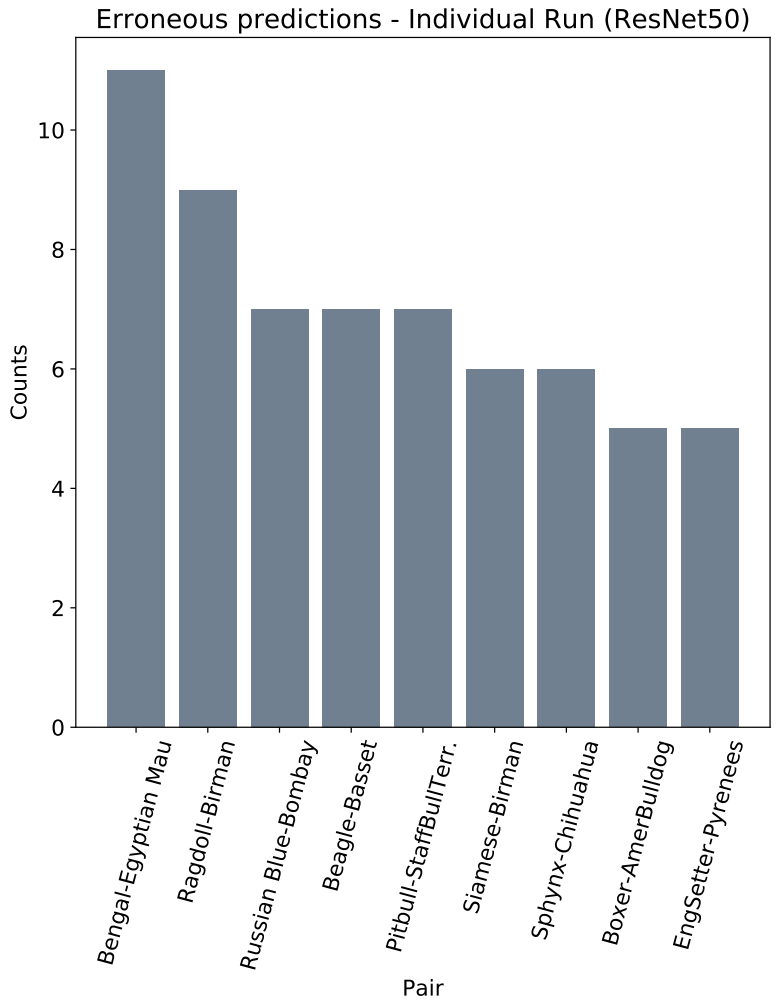


Figure 4-1: Error distribution from a single inference run (Oxford-IIIT Pets)

4.1 Erroneous Label Pairs: Oxford-IIIT Pets

Using repeated fitting on a ResNet-50 model, we conducted individual inference runs on the test set and observed erroneous pairs of labels being generated. An individual run is shown in Figure 4-1. We averaged several such runs and found the prominent pairs which were present in every trial. These are listed in Table 4.1 and shown in Figure 4-2.

In all the cases of misclassification observed in Oxford-IIIT Pets data, the reason was the texture in the sample. The most prominent examples were of *Bengal* cat species categorized as *Egyptian Mau* because of the spots in their coats (Figure 4-3a). Few *Russian Blue* cats were categorized as *Bombay* because of the uniform black color, a distinct characteristic of the latter (Figure 4-3b). The same held for *Ragdoll* cats which have dark snouts (Figure 4-3c). A few images with wrinkled faces of the *Beagle* were easily misjudged as *Basset* hounds, which also have the characteristic wrinkles and drooped

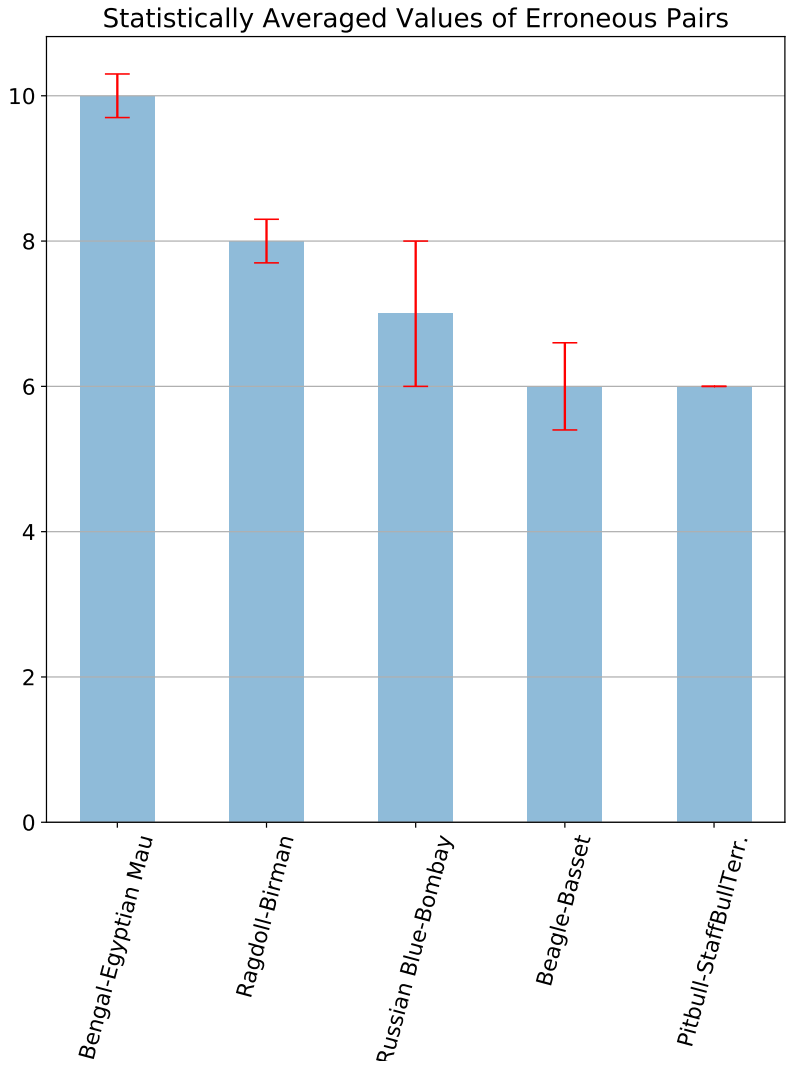


Figure 4-2: Most probable erroneous results

Pair	**Average number of errors**
Bengal and Egyptian Mau	10 ± 0.3
Ragdoll and Birman	8 ± 0.3
Russian Blue and Bombay	7 ± 1
Beagle and Basset	6 ± 0.6
Pitbull and Staffbull Terrier	6

Table 4.1: Statistically averaged list of erroneous pairs

Pair	Average number of errors
Ulcer and *Tumor*	31.75 ± 6.28
Erythema and *P. Macula*	27.25 ± 3.50
Erythema and *Wheal*	19.25 ± 2.75
Crust and *Ulcer*	16.00 ± 3.46

Table 4.2: Statistically averaged list of erroneous pairs

ears (Figure 4-3d). Geirhos et al. hypothesized that texture was the eminent feature in classification for ImageNet data [32]. In this standard corpus, the prediction error patterns were consistent with their observations.

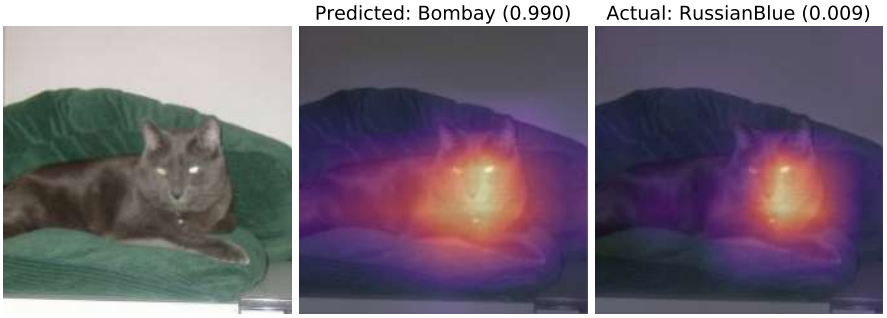
4.2 Erroneous Label Pairs: Skin Images

By training a sufficiently large model (ResNet-152), we tried to observe the pattern of errors in the classification. Some class pairs exhibited more errors than ordinarily to draw our interest in the same. The graph in Figure 4-4 shows the statistics gleaned from a single fitting. Table 4.2 and Figure 4-5 shows the statistics averaged over several such trials to highlight pairs which consistently fared worse than others. These numbers counted the occurrence of the first label predicted as the second, and vice versa. For these select pairs, we have attempted to identify the source of errors by class saliency mapping methods such as GradCAM and Guided Backpropagation (GBP) [92, 97]. Grad-CAM used the gradients of any target label going into the penultimate `CONV` layer in CNN to produce a coarse localization map, which highlighted the relevant regions. By handpicking the labels we needed to investigate as targets, we could investigate what regions the model deemed important for making decisions. By visualizing what the model picked out as differences in skin lesions, we could test if the hypothesis of texture dominance held here as well.

4.2.1 Ulcers and Tumors

Ulcer and *Tumor* had a high degree of error owing to similar planar attributes in dermatological photographs. Most commonly, the surface manifestation of both these classes is about the same. With the depth and curvature perception missing, there was a high

(a)



(b)

(c)

(d)

Figure 4-3: Misprediction in Oxford-IIIT Pet categories

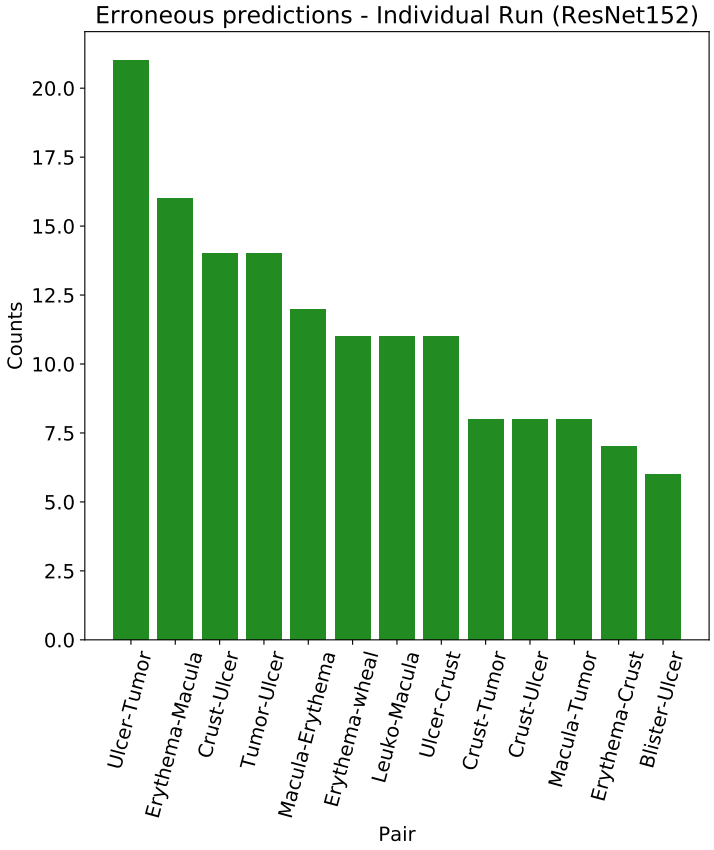


Figure 4-4: Error distribution from a single model fit

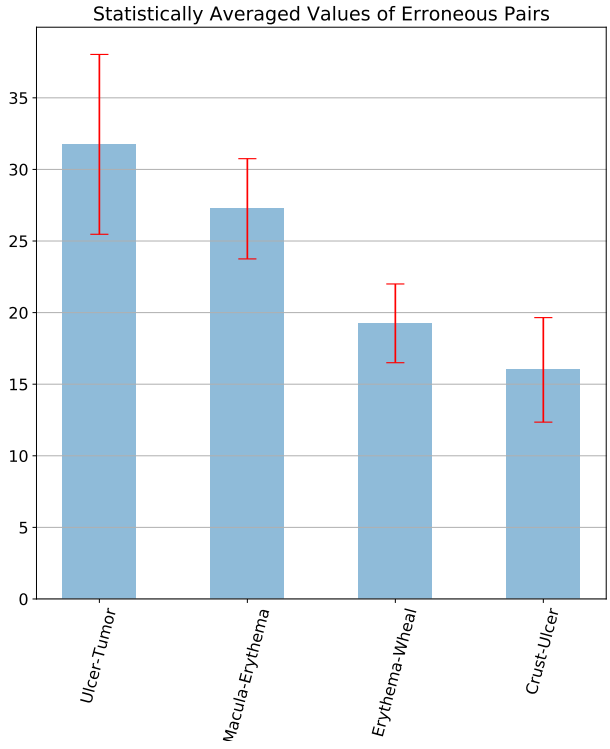


Figure 4-5: Most probable erroneous results (ResNet-152)

scope of error in these labels. In Figures 4-6a and 4-6b, even though the specimens are planar, the presentation gives a sense of depth. In Figures 4-6c and 4-6d, the *Ulcers* have a ring of inflammation which was mistaken for the appearance of a *Tumor*.

In the alternate situation, *Tumors* were predicted as *Ulcers* because of the appearance of a shallow depression or open lesions, which are hallmarks of *Ulcers*. In most cases of wrong predictions, the class activation maps for the correct and incorrect labels were co-located. This made the prediction inherently harder.

In the cases we have investigated so far, the lesion were incorrectly predicted because of greater effect of shape similarity and less due to texture dissimilarity. The co-location and similarity in color also made it confusing for the model to predict the right class.

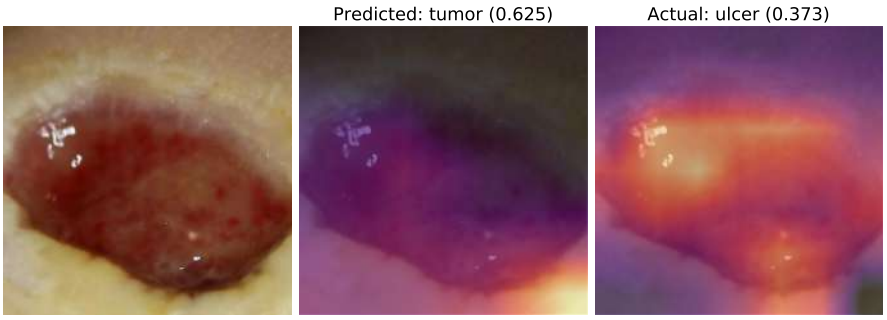
4.2.2 Macula and Erythema

Both *P. Macula* and *Erythema* appeared as pigmentation. *P. Macula* appeared as hyper-pigmented patches in limbs, whereas *Erythema* could appear as red patches anywhere in the body. In both the cases, the lesions were smooth without any remarkable visual attributes. The major difference between them was the color of presentation. In the panels shown in Figures 4-8 and 4-9, the lesions were not any different except for the color. Classic presentation of *Erythema* were as seen in Figures 4-8a and 4-8c.*P. Maculae* are dark brown spots surrounded by some inflammation, as seen in Figures 4-9a, 4-9b and 4-9d.

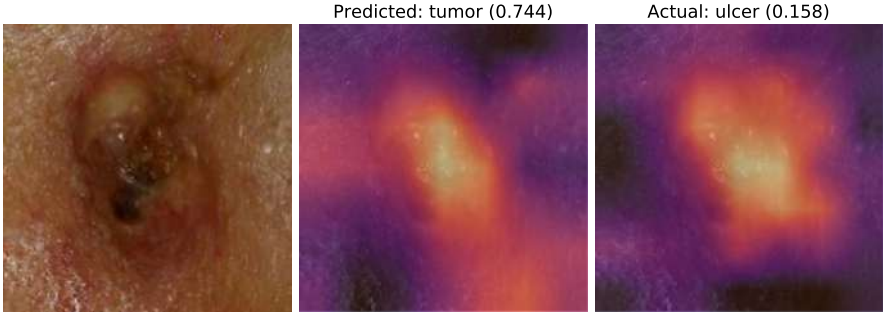
In both the labels, the dominant visual attribute was the high contrast they offered over the pale surrounding skin. Since both appeared geometrically similar without any distinguishing features, they were often confused.

4.2.3 Ulcer and Crust

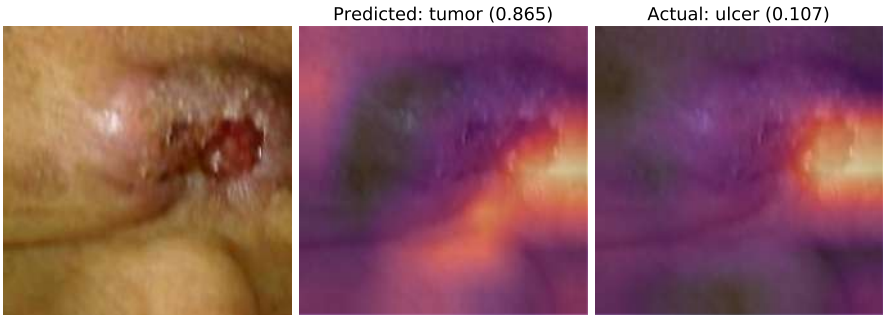
Ulcer and *Crust* proved to be an interesting challenge to the model because these labels are chronologically related. Often *Crust* appears during the healing process off *Ulcers*. There is a very strong visual correlation between these two labels. Without having a diagnostic history, it was difficult to accurately classify the intermediate stages between the two. These were cases where human interpreters could have also have difficulty in categorizing. Figure 4-10 and 4-11 show some wrong categorizations arising from inability to resolve this chronological conundrum.



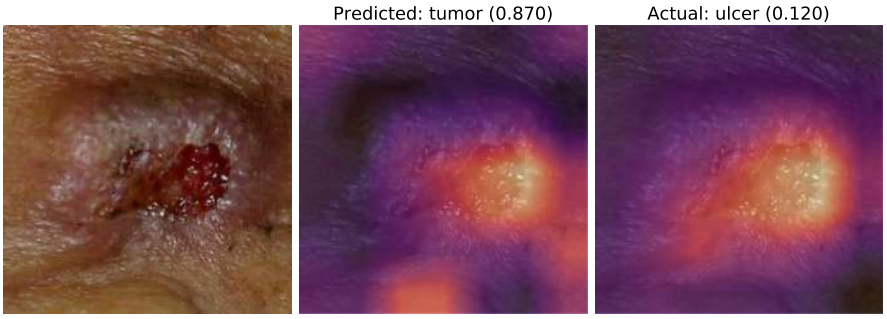
(a)



(b)

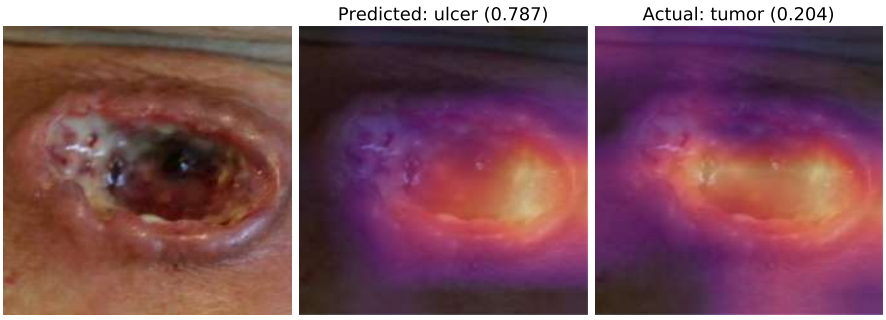


(c)

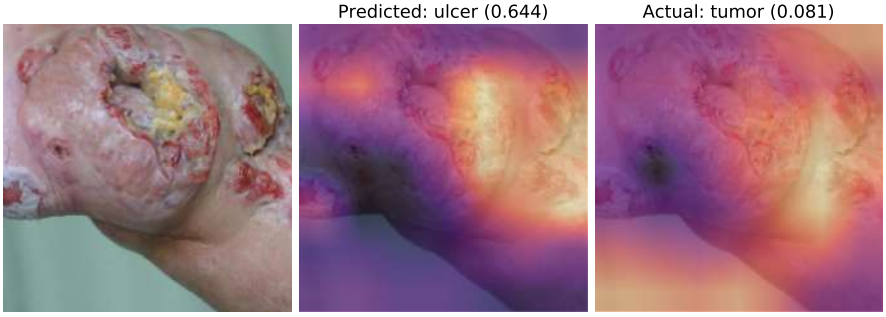


(d)

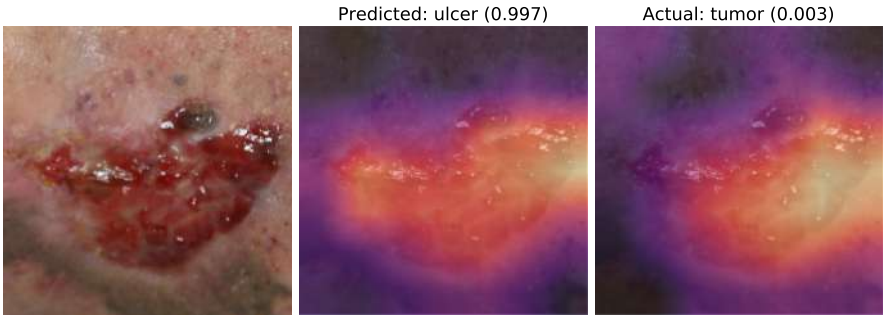
Figure 4-6: Ulcer predicted as Tumor



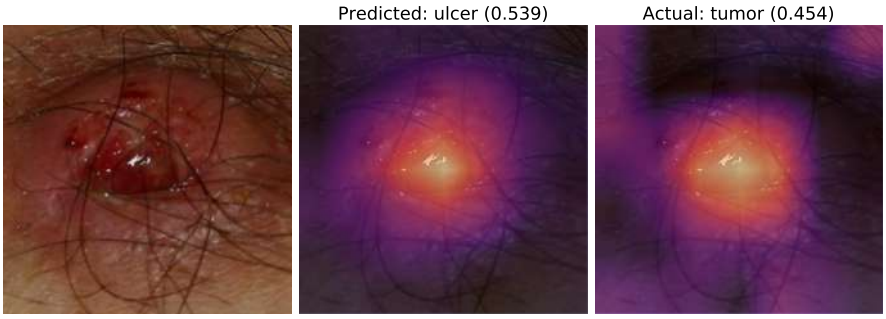
(a)



(b)

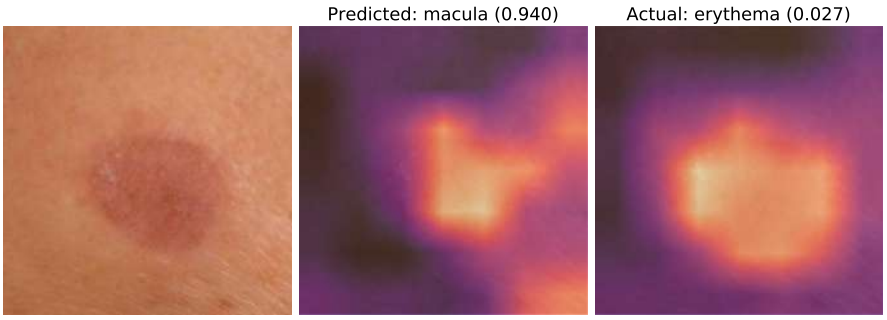


(c)

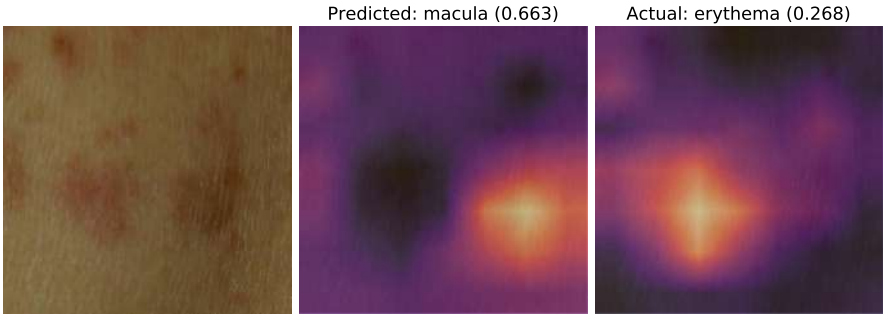


(d)

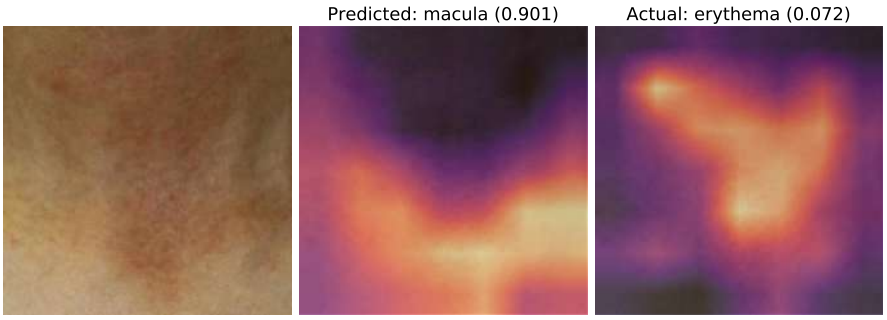
Figure 4-7: Tumors predicted as Ulcers



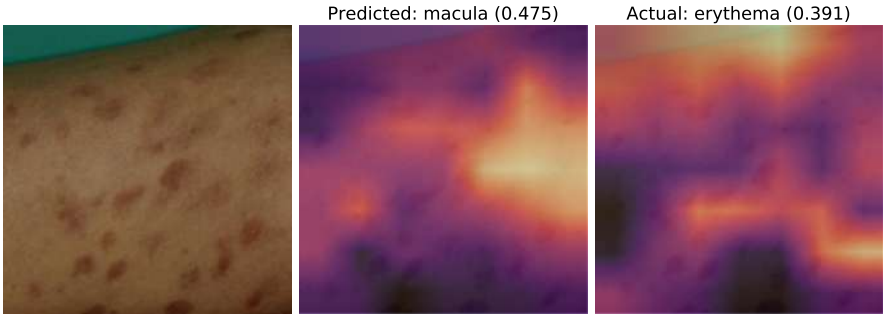
(a)



(b)

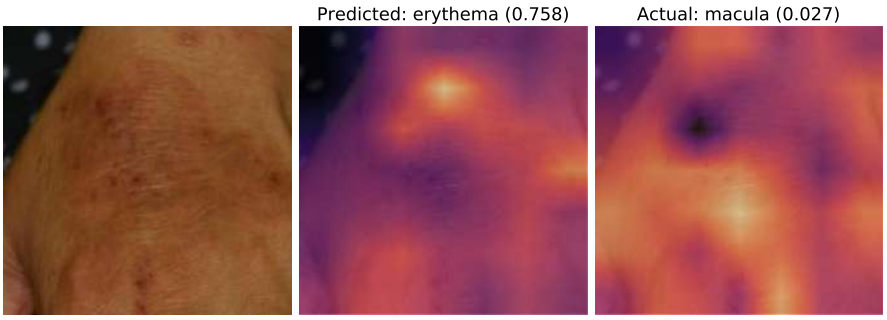


(c)

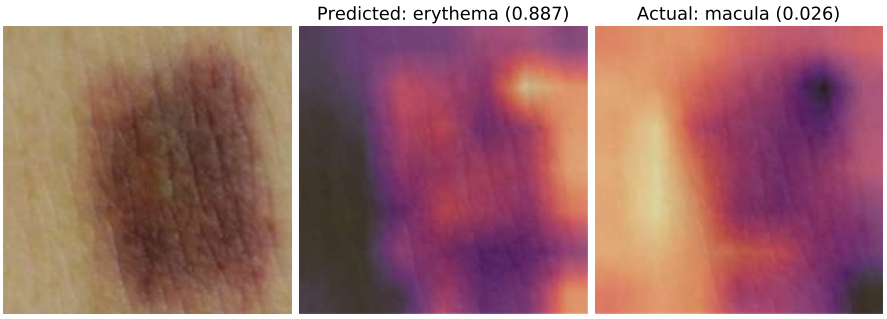


(d)

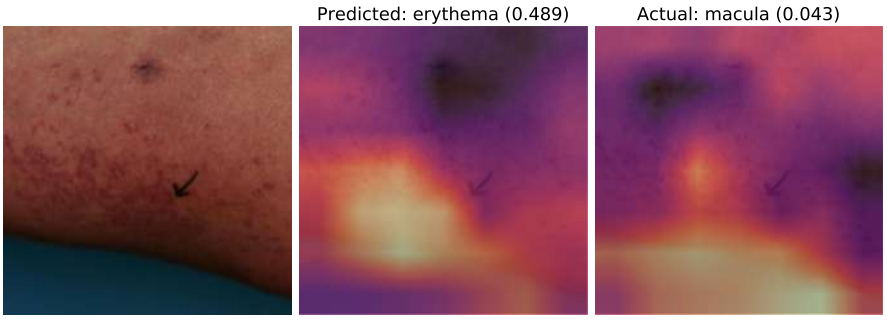
Figure 4-8: Erythema predicted as Maculae



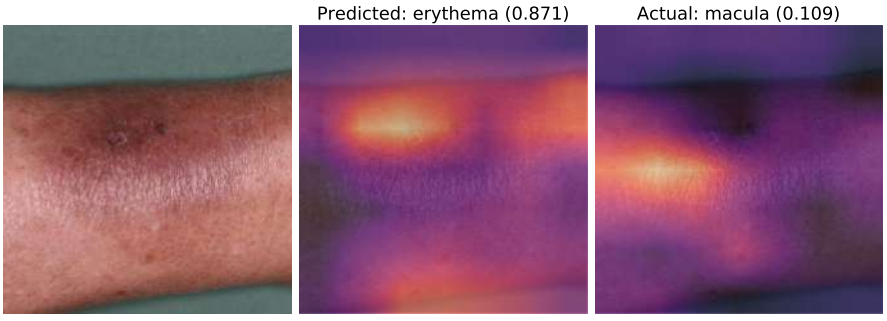
(a)



(b)

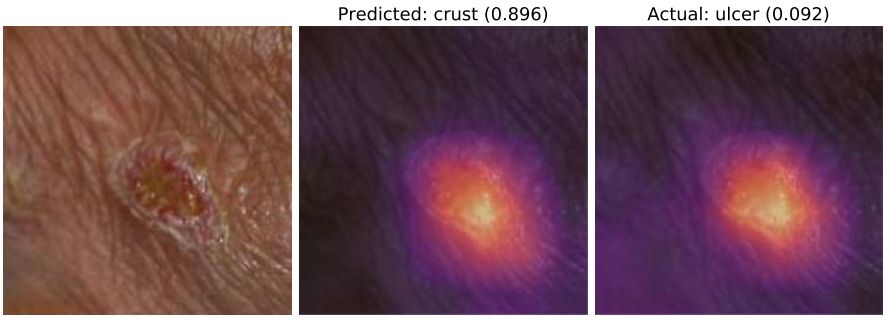


(c)

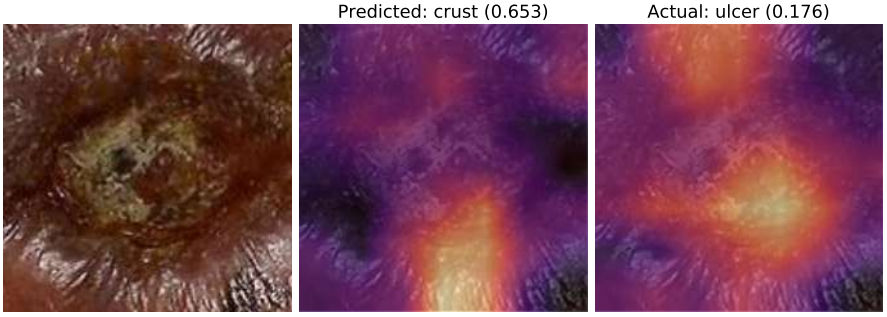


(d)

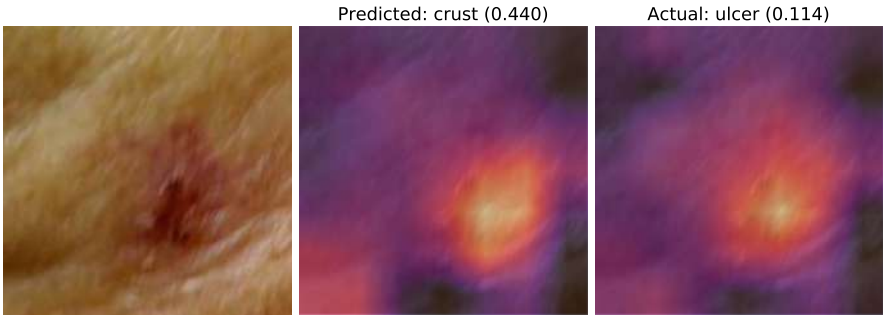
Figure 4-9: Maculae predicted as Erythema



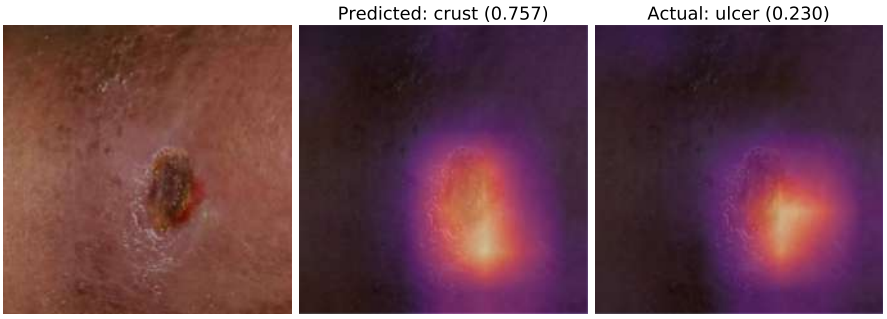
(a)



(b)

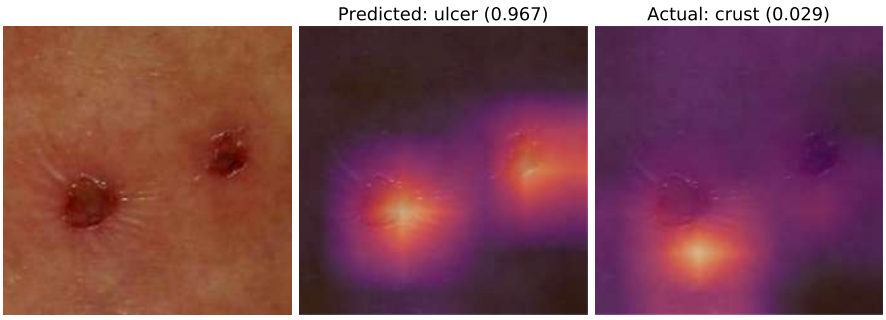


(c)

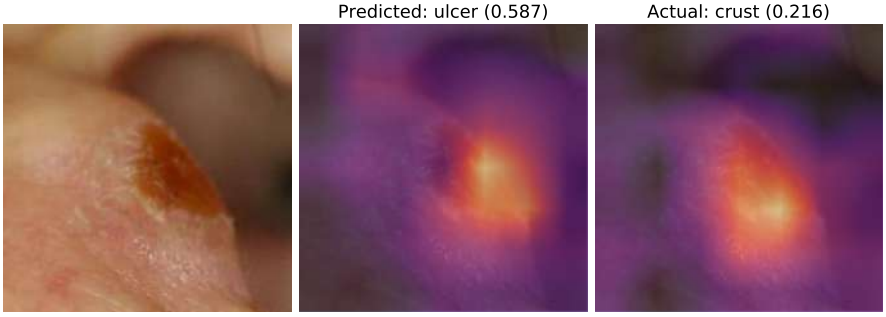


(d)

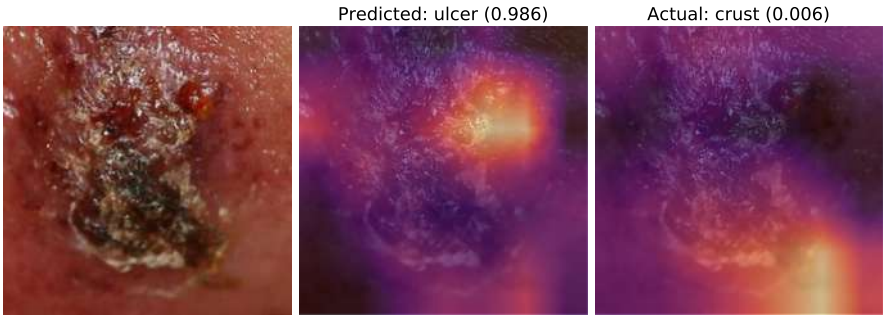
Figure 4-10: Ulcer predicted as Crust



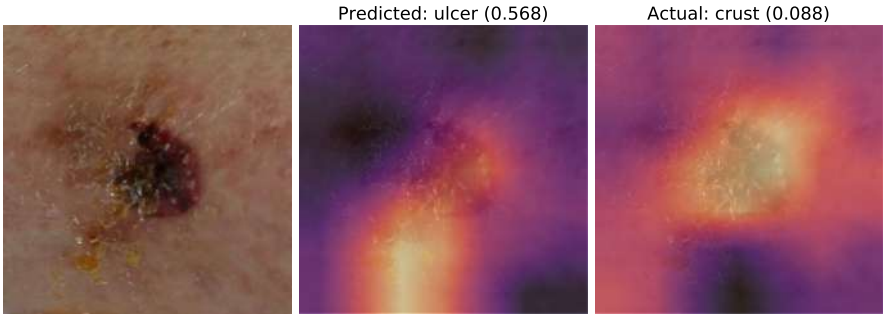
(a)



(b)



(c)



(d)

Figure 4-11: Crust predicted as Ulcer

4.2.4 Erythema and Wheal

The case for *Erythema* and *Wheal* is similar to that of *Ulcer* and *Crust*. *Wheal* is temporary pigmentation of the skin arising from trauma or inflammation. It manifests as slightly discolored patches in the region of impact. Unless subject history is understood, it may look visually similar to mild *Erythema*. This resulted in our model making mistakes in categorizations. Examples of this pair are not illustrated since the distinction is trivial.

4.3 Effect of Image Background

In the class activation maps, it was seen that the highlighted regions often covered some background regions as well. In the case of *Melanoma*, Bisotto et al. had shown that the background information contributed to the prediction in a model [9]. This dataset however had very well-curated images at high magnification, for uniform cancerous lesions. We wanted to investigate whether the background for the skin lesions in common cases also had any effect on making the right predictions. This exercise required we isolated the background from the regions of interest in the images, and see if the model predicted anything similar to the original classes.

To test this hypothesis, the center portion, which usually contained the lesion, was cropped using NumPy from images having confident predictions. This region was replaced with a black rectangular patch, measuring approximately a third of the input's dimensions. The edited images were visually verified to check that lesion was fully or at least 75% occluded. A ResNet-152 model was prepared to its best fit according to procedures demonstrated in Section 3.2.5. Figure 4-12 shows some representative results from these trials.

It was found that if the patch completely covered the lesion in the specimen image, then the prediction was random and unrelated to any attributes present in the original image. For example, Fig. 4-12a and 4-12c are images for which the model correctly predicted their true class. However, when the lesion was covered by the patch, as seen in Fig. 4-12b and 4-12d, the prediction was entirely different and changed in between different learning trials. In most cases, the softmax probability of the true class was not any different from the other classes.

When the lesion was partially covered, the model predicted the true class on many occasions. Such an example is demonstrated in Fig. 4-12e and 4-12f. Hence, the model could learn from minute details in the region of interest. Even if the model cannot establish with certainty, it can predict correctly with a lower softmax probability.

4.4 Presence of Confounders

Marks and highlights made by doctors and clinicians were observed in several samples during the process of data cleaning. These marks were present in several samples across the labels. During the process of model learning, these marks and highlights were a major source of errors in classification as the model learned to pick up these visual markers in making predictions. Since these marks were very close to the lesion themselves, it was difficult to remove them. Representative images of such samples are shown in Figure 4-13.

To alleviate this problem, one of the best approaches would be image in-painting [66]. But since the number of samples was few and the data was densely homogeneous, this approach would not work well. To work around this problem, a more traditional route was undertaken. The image was partitioned into patches and dominant colors present in the image were calculated by a k-means clustering on the quantized RGB space. A binary mask of the image was computed to isolate the confounding marks & highlights from the background. The mask output was used to null the image at the relevant places and the pixels were replaced by the dominant color in the respective patches. In the majority of cases, the process mitigated the problem and the texture appeared smooth. A demonstration of this method is shown by Figures 4-13a and 4-13b. Figures 4-14a–4-14d and 4-14e–4-14h show the process of removing the aberration to provide cleaner outputs.

This method had its caveats. An insufficiency in this approach was seen when the lesion itself was composed of dark colors, such as many cases of *P. Macula*. One such sample is demonstrated in Figure 4-13c.

(a) Actual: Blister (0.743)

(b) Predicted: Acne (0.839)

(c) Actual: Ulcer (0.81)

(d) Predicted: Crust (0.79)

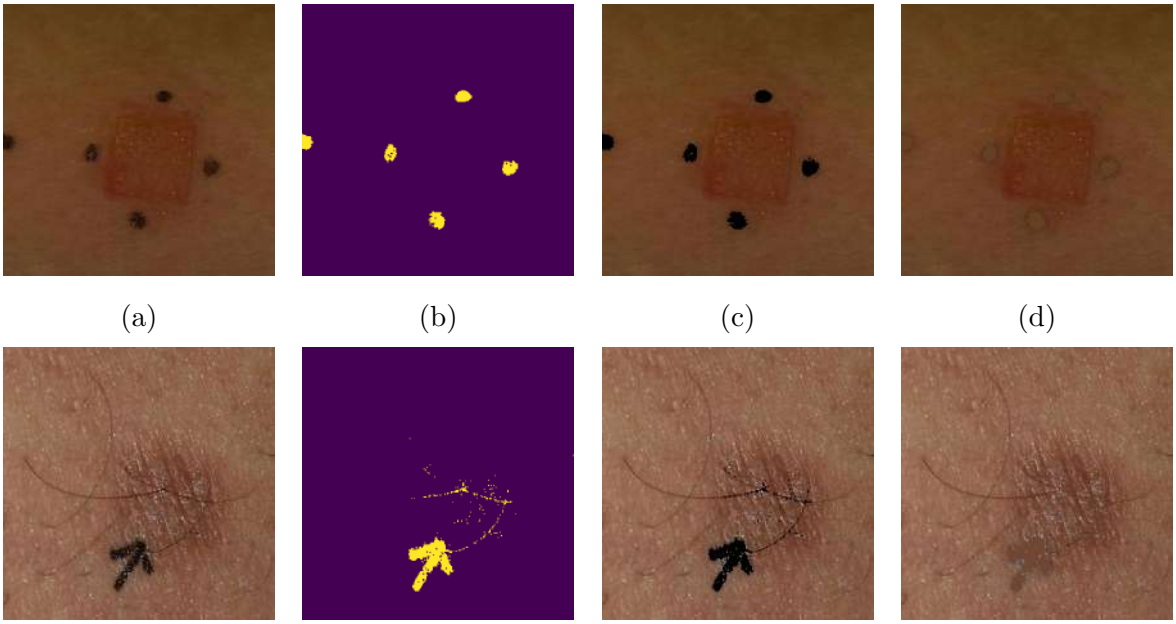
(e) Actual: Tumor (0.99)

(f) Predicted: Tumor (0.38)

Figure 4-12: Testing contribution of background

(a) (b) (c)

Figure 4-13: Presence of confounders in the visual attributes



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4-14: Confounder Removal by k-Means clustering of dominant colors

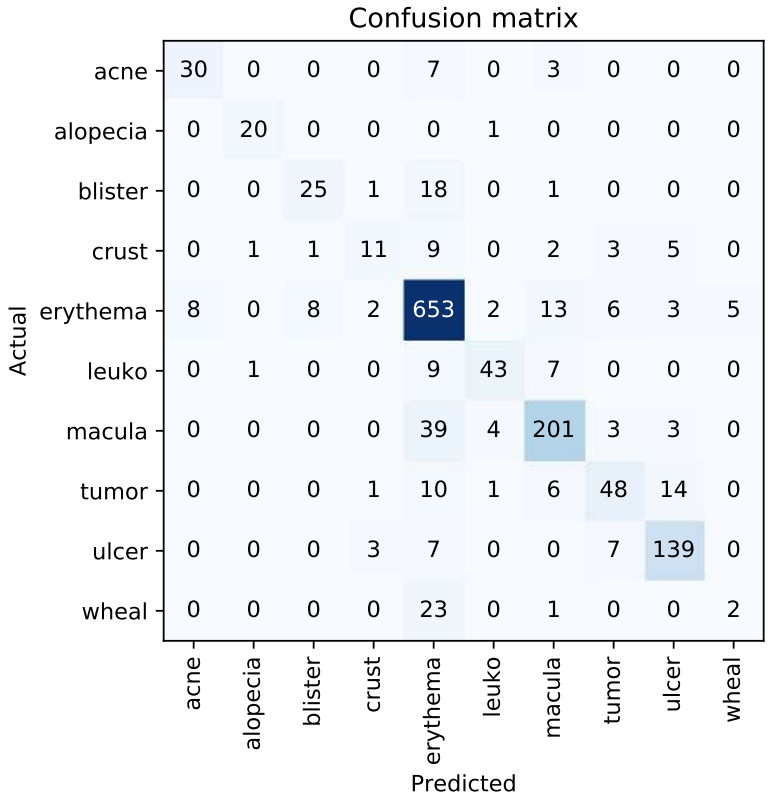


Figure 4-15: Results of poor learning on an unbalanced corpus

4.5 Other Mitigation Strategies

Apart from carefully analyzing the visual complexity in samples for errors, the following strategies have proved beneficial in improving the quality of model predictions.

4.5.1 Balancing Data Distribution

Model learning outcomes are trustworthy when the data classes are being sampled evenly. The learning scores (training accuracy, validation accuracy, etc.) are a weighted representation of the performance across all the classes. They can be believed as relevant indicators only when the class representations are fair and equal. Consider the results of a model learned on data that had not been altered by balancing strategies, in Figure 4-15. This data correlated well with the out-patient case statistics. The dataset however poorly represented the participating classes from the deep learning point of view. The model predicted the dominant class well, but the same could not be said of the other labels. However, if we were to believe a single metric such as validation accuracy, these results could be deceiving.

A balanced dataset exhibits true macro-average. Our dataset from the patient pool was updated frequently during the preliminary phases. Even if these classes were created with uniformity, they needed to be checked at intervals for class imbalance. We combined user-submitted images and clinical samples before performing data augmentation. Care was taken not to place copies of the data in both training and validation sets. If a large number of a particular class was due for addition, we culled some poorer quality data to make room for the new images. Various strategies were explored to perform image augmentation once the raw images were labeled and ready,

1. External libraries such as *imgaug*, *albumentation* were used to create augmented copies of the images which had variable crop location, with slightly different contrast, brightness or color balance [48, 14]

2. PyTorch data loader was customized to create versions of the image in the memory which were flipped, rotated, zoomed, etc. during the batch uptake in model learning.

3. Test Time Augmentation (TTA) was also employed to create several dynamic copies of an image during the testing phase in an epoch. The ensemble average of the predictions was considered as the sample prediction [78].

Skin lesions can be modeled to look very realistic via generative adversarial networks (GAN) [5, 6, 33, 36]. We avoided using them in our studies since expert opinion remains inconclusive about their efficacy in making attribute-rich copies of the data [104].

4.5.2 Improving Field of View (FOV)

Reducing the field of view (FOV) has a profound impact on the quality of detection. Skin lesions rarely occur in isolation. If other visual landmarks are prominent enough, the model may mistake it for the object of interest in the image. Consider Figure 4-16 where the image has not been cropped. The *Erythema* patches have been disregarded for an anatomical landmark which looks like *Pigmented Macula*.

FOV correction where the object of interest covers approximately 50% of the area can significantly improve the prediction quality. According to Sprawls et al., it is even possible to counteract the effect of motion blur [61]. In Figure 4-17a, having a wide FOV

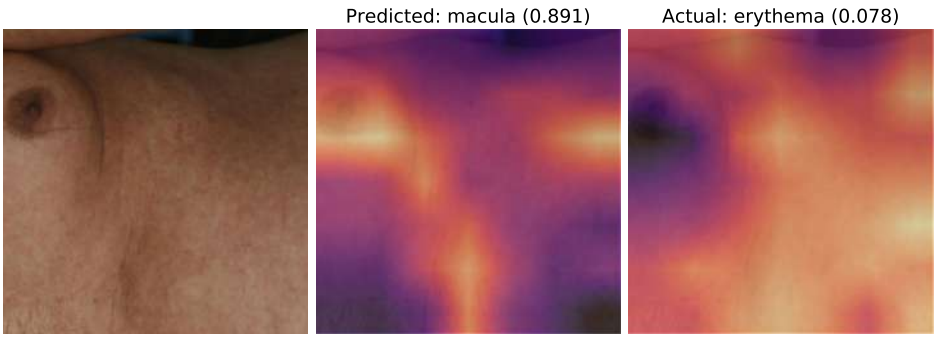


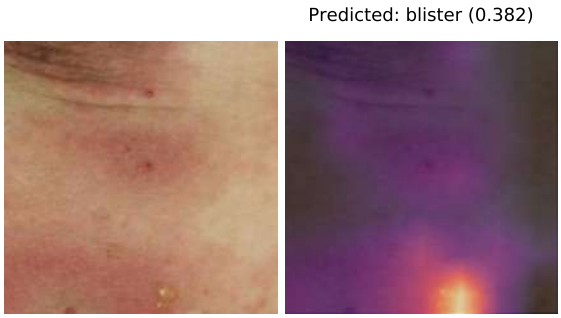
Figure 4-16: Results of poor FOV in dermatological sample

less to 38% accuracy of prediction of the *Blister*. Zooming in to the lesion and capturing the sample with a desirable FOV in Figure 4-17b led to a perfect prediction of the true class.

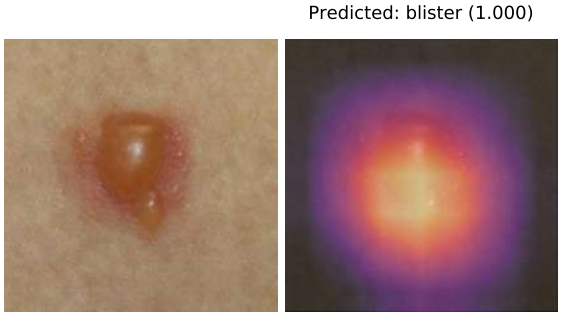
4.5.3 Gamma and Illumination Correction

Illumination and contrast artifacts were present in several images. These occurred due to insufficient or non-uniform illumination when the user took the photographs. The artifacts presented themselves as unnaturally dark or bright regions in the images. Having unnatural shadows in the images led the model to make several miscategorizations. An approach that was followed in alleviating this issue was to do gamma correction on the photographs ($\gamma = 1.2$–1.5). Figure 4-18 shows the before (Fig. 4-18a) and after (Fig. 4-18b) effect of correction on one such sample and the associated prediction change to correct label.

Several images were prone to have *wash-out* effect due to exposure by flash. The flash created high-reflectance patches that obscured the images. Contrast correction was able to reduce some of the effects. However, for the best mitigation capturing the illumination map was necessary [25]. This was done by taking two photographic shots in quick succession - one in ambient conditions and the other with the camera flash. Using this approach, it is hoped that future samples can be corrected more efficiently.

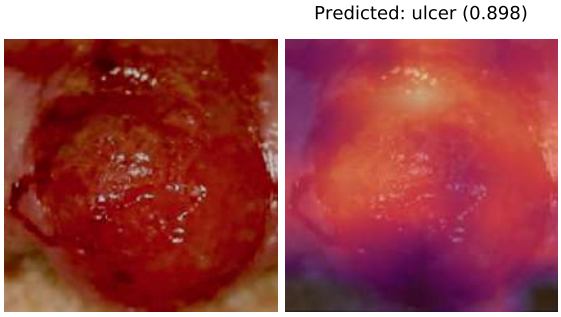


(a) Large FoV capture.

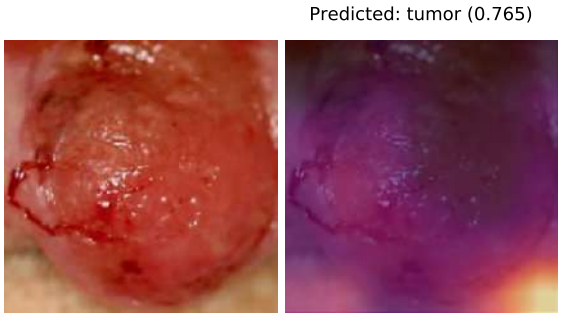


(b) Small FoV capture. Improvement ~3×

Figure 4-17: Improving predictions by FOV reduction



(a) Original Image



(b) After Gamma-balancing

Figure 4-18: Photogrammatic correction by gamma adjustment

4.6 Chapter Summary

In the previous chapter, we demonstrated schemes for producing robust model fit across several architectures. Given some minimum depth, the performance was seen to become architecture agnostic (within small error bounds). Yet, there was a gap between the peak performance we could extract and human-level prowess in label identification.

When experimenting with Oxford-IIIT Pets data, the pattern of classification errors was consistent with the factors described by Geirhos et al. The errors were due to the model confusing some samples due to the texture in the images. Analysis of homogeneous skin lesion images revealed that the gap was due to effects of color, shape, contrast, and visual complexity as well. The chapter investigated some of these effects via case studies. Additionally, skin images were seen to have more imperfections arising out of non-standardized data curation. By applying curation strategies such as FOV reduction, weighted resampling, and illumination correction in skin images, we noticed improvements in the model accuracy.

The presence of confounders and the effect of lesion background were also investigated for any effect on outcomes. The background was observed to have a negligible effect on the accuracy when we tested with samples having confident predictions. The effect of confounders was mitigated by a pre-processing technique which split an image into grid cells and replaced the aberrations with the local dominant color. The model saw an improvement of 1.8 percentage points in Top-1 accuracy with this mitigation.

We surmise that ML-based techniques can deliver only if we improving the algorithm or optimization scheme, as well as curate the data by pre-processing. In a real-world use case, satisfactory outcomes are hard to come by using only end-to-end deep learning processes.

Chapter 5

Adversarial Perturbations and Distribution Shifts

All image samples contain varying degrees of noise. Noise is an inalienable part of any imaging system. Even seemingly pristine images have some amount of sensor noise which remains even after JPEG encoding. Sometimes the sources of noise are more perceptible. During image capture, imperfections or debris on the camera lens can propagate to the digital photograph as specks and spots. If the images are captured by hand-held devices, such as consumer-grade cameras or smartphones, they present a small degree of motion blurring or camera soft-focus as well. These could have a detrimental effect on image cognition.

Computer vision models are sensitive to such imperfections in varying amounts. Goodfellow et al. showed that it was possible to mislead the classifier by imperceptible addition of defects [37]. Robustly trained computer vision models are expected to disregard the presence of small imperfections in producing faithful predictions [16]. When working with new sets of data, we have very little knowledge about the noise content in individual images. We need to be able to assess if such data can be trusted with predictions and the means to assess the quantum of changes by establishing baselines.

In this chapter, the CIFAR-10 dataset has been tested against impulse noise corruption and simulated de-focus. The effect on homogeneous images has been investigated subsequently.

(a) Original (b) Impulse noise corrupted (c) Simulated defocus

Figure 5-1: Examples of imperfections simulated in data

5.1 Simulating Impulse Noise and Motion Blur

To simulate impulse noise which would mimic specks, debris, and poor resolution, we randomly changed up to a known percentage of the image pixels by salt-and-pepper noise with the ratio of black to white noisy pixels at 3:1. The placement of the pixels was random and only depended on the quantity of noise specified. In the ablation studies discussed further in the chapter, we set these noise levels from 1% to 5% and an extremum of 15%.

To simulate motion blur and soft focus, we considered both isotropic (simulating only blur/soft focus) and anisotropic (simulating soft focus with minor movement) modifications. We opted to simulate them by 7×7 uniform Gaussian kernel and 45° directional blurring kernel respectively (For CIFAR-10, the kernel size was 3×3). Figures 5-1b and 5-1c demonstrate the effect of creating noisy replicas and soft-focus images from the original (Figure 5-1a).

5.2 Ablation Study on CIFAR-10

To perform an ablation study on the CIFAR-10 dataset, we created copies of the train and test set with impulse noise between 1%–5% of the total pixels. An extreme case of 15% noise was also considered. We considered the three combinations which exhaustively covered analyzing model performance on noise corrupted data:

- Training with normal images, testing with noise corrupted images
- Training with noise corrupted images, testing with normal images

Test Set Noise	Top-1 Accuracy (%)
0%	94.37
1%	90.35
2%	89.41
3%	88.98
4%	88.53
5%	87.71
15%	89.04

Table 5.1: Aggregate model accuracy with Test set noise

- Training and testing with matched levels of noise corruption.

For testing blur and defocus resilience, we trained with regular images and tested on test sets that were blurred according to the process described in Sec. 5.1. A summary of all these experiments is shown in Table 5.4, in addition to item-wise discussion in the following sections.

5.2.1 Training Set Normal, Testing Set Corrupted

When the train set was normal and the testing set was corrupted with noise amounts (between 1%–5% pixels), an initial drop in accuracy up to 5% percentage points was observed, and thereafter seen steady. The accuracy correlated with the amount of noise present and was stable with minor fluctuations within one percentage point at most. The noise was not steadily degrading prediction quality as long as it was a small percentage of pixels. The results are elaborated in Table 5.1 and Figure 5-2.

5.2.2 Training Set Corrupted, Testing Set Normal

When the train set was corrupted by impulse noise (between 1%–5% pixels) and the testing set left untouched, the drop in accuracy was marginally lower than the previous case. The difference was negligible among all the cases, even with the noise amount increasing steadily. These results are elaborated in Table 5.2 and Figure 5-3.

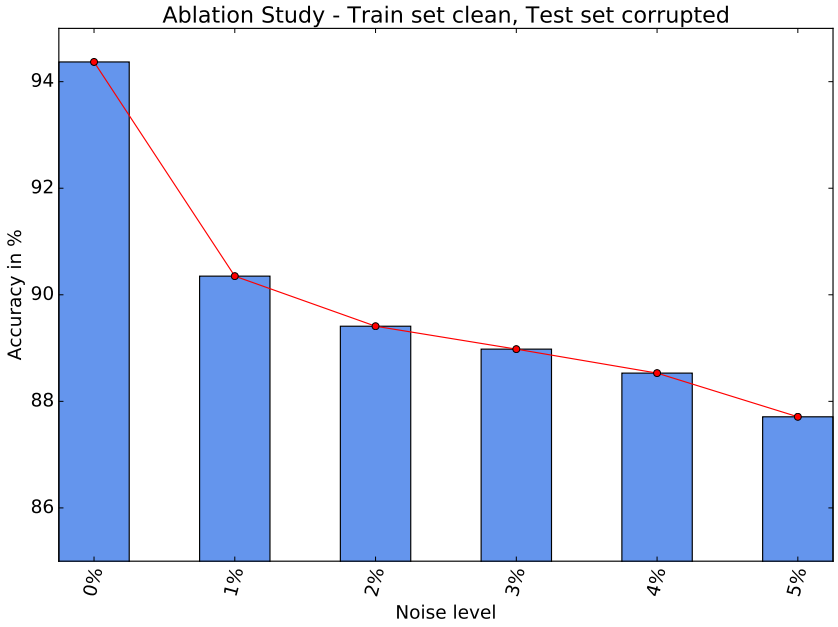


Figure 5-2: Ablation study with varying amount of Test set noise (CIFAR-10)

Train Set Noise	**Top-1 Accuracy (%)**
0%	94.37
1%	90.77
2%	90.64
3%	90.44
4%	90.52
5%	90.36
15%	89.56

Table 5.2: Aggregate model accuracy with Training set noise

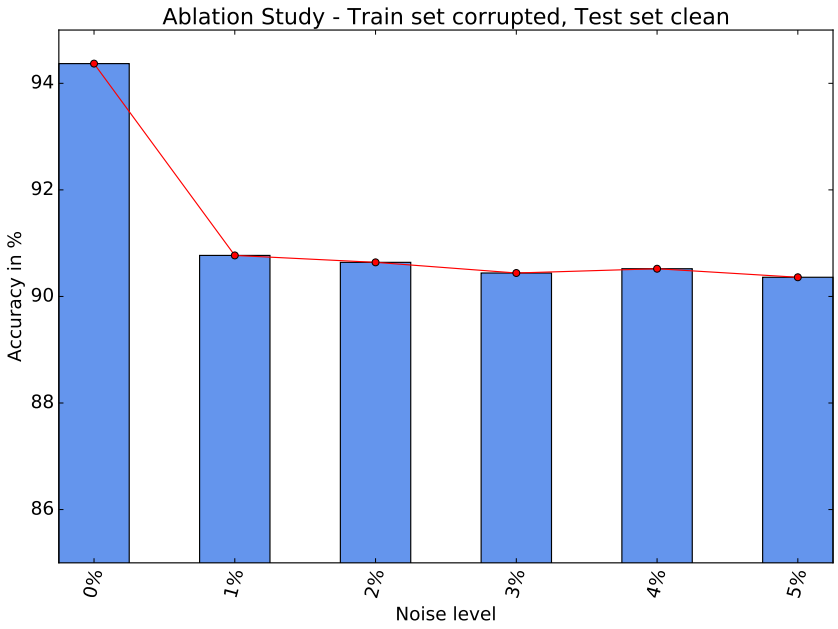


Figure 5-3: Ablation study with varying amount of Training set noise (CIFAR-10)

Matched Noise level	Top-1 Accuracy (%)
0%	94.37
1%	90.62
2%	89.89
3%	89.93
4%	89.19
5%	89.98
15%	89.14

Table 5.3: Aggregate model accuracy with matching noise levels

5.2.3 Both Training and Test Sets Corrupted

When both the sets were corrupted by noise (between 1%–5% pixels) and learning was undertaken with sets having matching levels of noise, the accuracy was seen to be approximately the same within a small bracket of values ($\mu = 89.92\%$, $\sigma = 0.20\%$). The difference was negligible among all the values of noise levels. These results are elaborated in Table 5.3 and Figure 5-4.

5.2.4 Training Set Normal, Test Set Isotropically Blurred

In the case of CIFAR-10, we chose to use a 3×3 Gaussian kernel instead of 7×7 since the image size was 32×32. A larger kernel would have unreasonably blurred the image

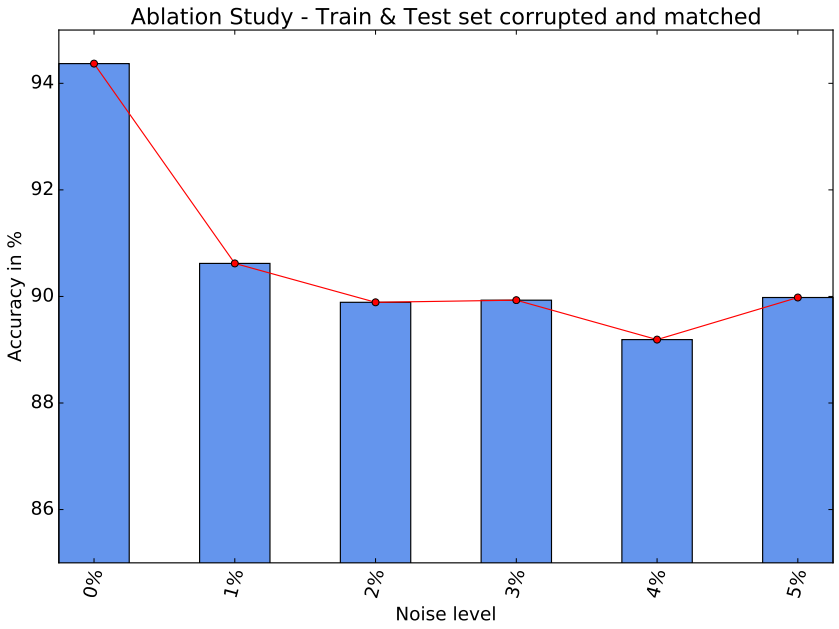


Figure 5-4: Ablation study with matching noise levels in Train & Test (CIFAR-10)

beyond recognition. We used normal images for the training set. The kernel was applied to the test set to simulate the effect of soft-focus capture. The aggregate Top-1 Accuracy of the model was found to be 85.09%. Only *cat*, *dog* and *frog* labels were found to be slightly more confused than the other seven. A confusion matrix of the model is shown in Figure 5-5.

5.2.5 Training Set Normal, Test Set Anisotropically Blurred

Similar to the previous case, we chose a 3×3 kernel. The blur direction was set at 45 degrees by setting the diagonal elements of the kernel appropriately. We used normal images for the training set and the kernel was applied to the test set. The aggregate Top-1 Accuracy of the model was found to be 69.61%. A confusion matrix reflecting the model performance is shown in Figure 5-6. Unlike isotropic blur, directional blur produced more strong class confusions. The differences in the model performance can be seen in a comparative confusion matrix seen in Figure 5-7.

5.3 Ablation Study on Exmedio Skin Data

Similar to the process of introducing imperfections in CIFAR-10 described in Sec. 5.2, data were prepared with varying amounts of impulse noise, isotropic, and anisotropic

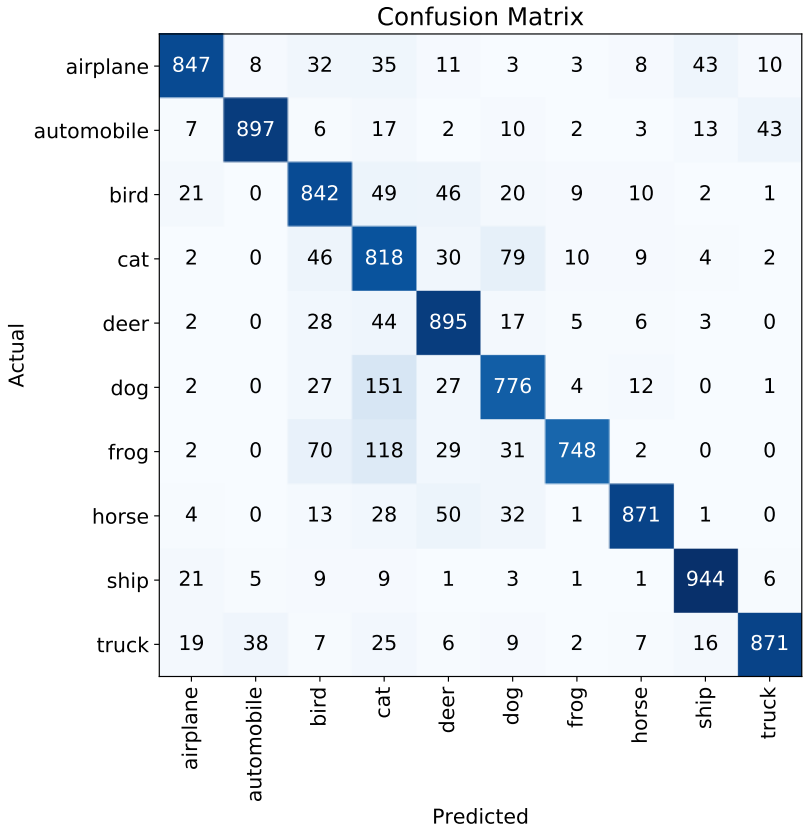


Figure 5-5: Confusion matrix - Test set isotropic blurred (CIFAR-10)

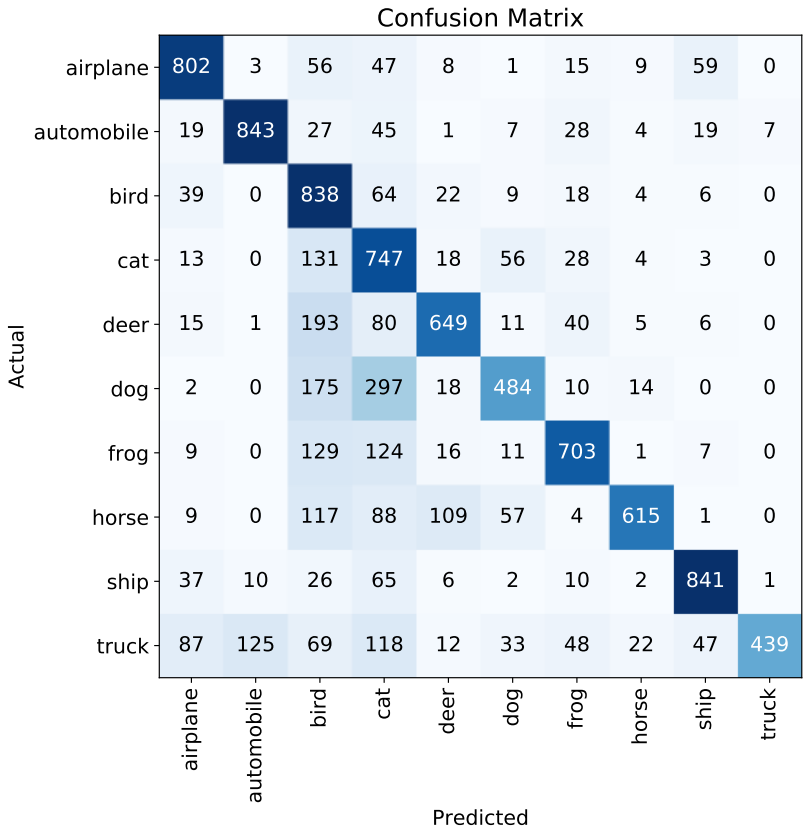


Figure 5-6: Confusion matrix - Test set anisotropic blurred (CIFAR-10)

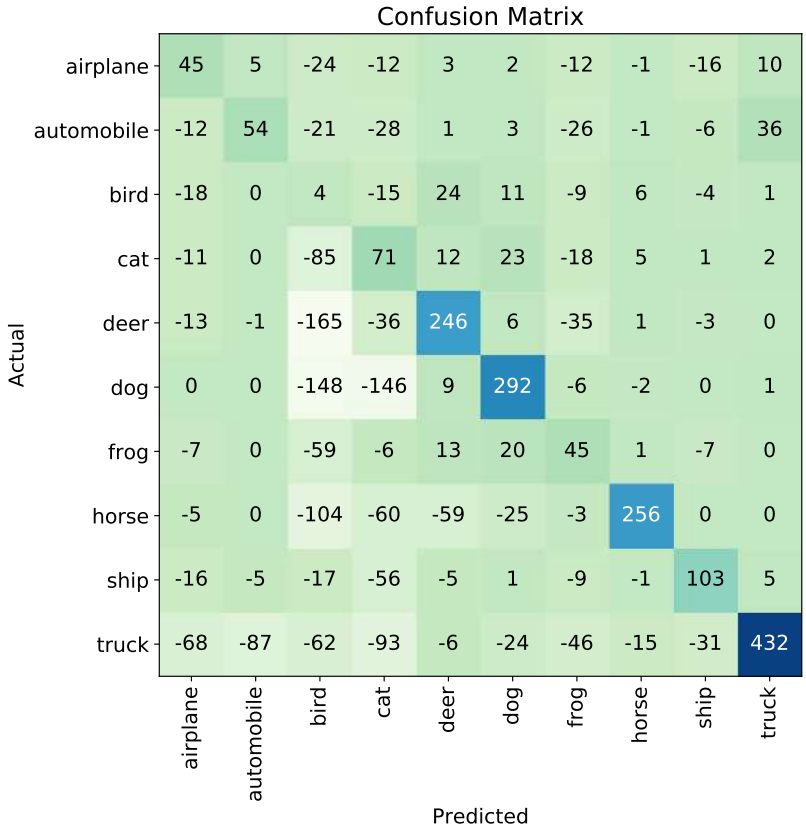


Figure 5-7: Differences in prediction - Isotropic vs. Anisotropic blurring

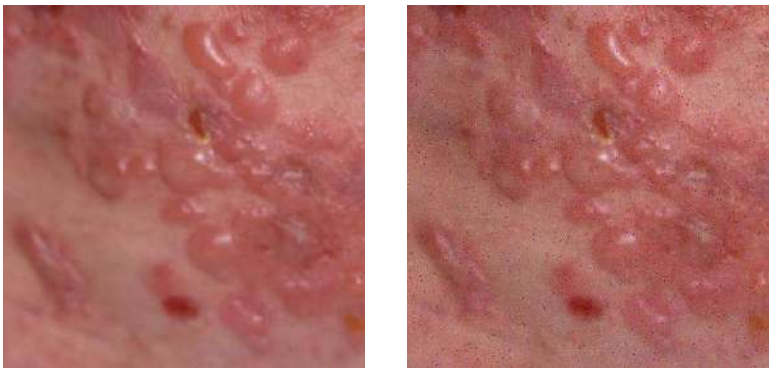
blur. Representative samples from impulse noise addition and soft-focus creation are shown in Figure 5-8 to give readers an estimate of the changes introduced. The ablation study performed on this data is discussed in the following subsections. A summary of all adversarial experiments is shown in Table 5.8, in addition to item-wise discussion in the following sections.

5.3.1 Training Set Normal, Testing Set Corrupted

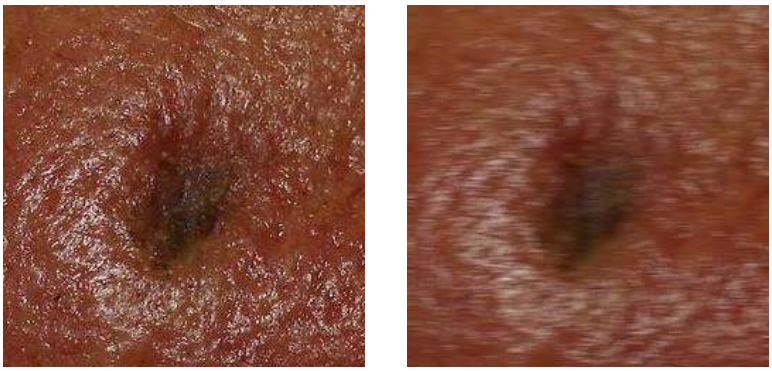
With the train set normal and test set corrupted with varying amounts of noise (1%–5% pixels), an initial drop in accuracy was observed. This drop steadily increased with increasing amounts of noise. These results are shown in Table 5.5 and Figure 5-9.

5.3.2 Training Set Corrupted, Testing Set Normal

When the train set was corrupted by impulse noise (between 1%–5% pixels) and testing set left untouched, the drop in accuracy was marginally lower than the previous case. The phenomenon was similar Sec. 5.2.2, where training with corrupt images showed better



(a) Original (b) Impulse noise corrupted



(c) Original (d) Simulated defocus

Figure 5-8: Examples of imperfections simulated in Exmedio data

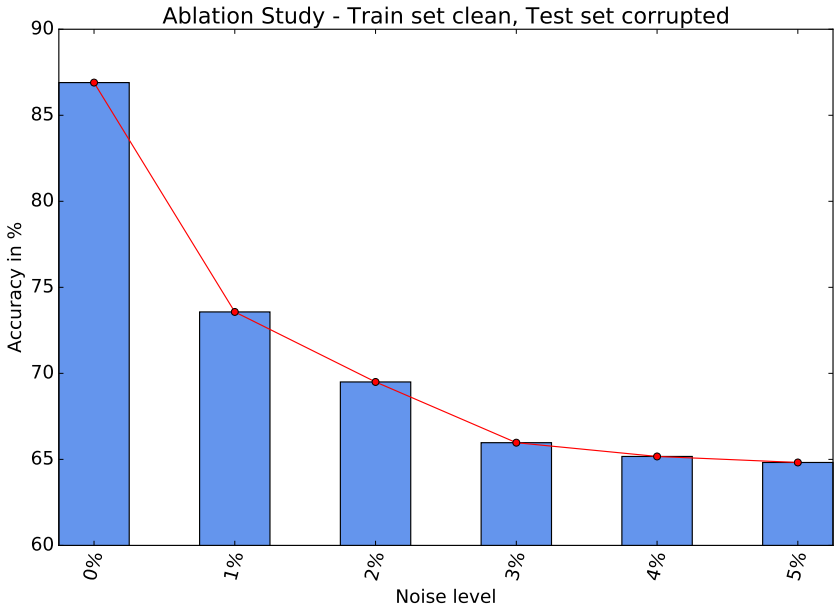


Figure 5-9: Ablation study with varying amount of Test set noise (Exmedio)

Table 5.4: Adversarial tests summary (CIFAR10)

Train Status	**Test Status**	**Top-1 Acc.** (in %)
Clean	Clean	94.37
Clean	1% impulse noise	90.35
	2% impulse noise	89.41
	3% impulse noise	88.98
	4% impulse noise	88.53
	5% impulse noise	87.71
	15% impulse noise	89.04
1% impulse noise	Clean	90.77
2% impulse noise		90.64
3% impulse noise		90.44
4% impulse noise		90.52
5% impulse noise		90.36
15% impulse noise		89.56
1% impulse noise	1% impulse noise	90.62
2% impulse noise	2% impulse noise	89.89
3% impulse noise	3% impulse noise	89.93
4% impulse noise	4% impulse noise	89.19
5% impulse noise	5% impulse noise	89.98
15% impulse noise	15% impulse noise	89.14
Clean	Isotropic Blur	85.09
	Anisotropic Blur	69.61

results than testing. These results are elaborated in Table 5.6 and Figure 5-10.

5.3.3 Both Training and Test Sets Corrupted

When both the sets were corrupted by noise (between 1%–5% pixels) and learning was undertaken with sets having matching levels of noise, the accuracy was stable within a small range ($\mu = 81.98\%$, $\sigma = 0.31\%$). These results are elaborated in Table 5.7 and Figure 5-11.

5.3.4 Training Set Normal, Test Set Isotropically Blurred

We chose 5×5 Gaussian kernel for the isotropic blurring of the images in the test set. We used unaltered images for the model training. The aggregate Top-1 Accuracy of the model was found to be 75%. With blurring, the model was confused to a slightly

Test Set Noise	Top-1 Accuracy (%)
0%	86.90
1%	73.57
2%	69.50
3%	65.97
4%	65.17
5%	64.82
15%	59.05

Table 5.5: Aggregate model accuracy with Test set noise (Exmedio)

Train Set Noise	Top-1 Accuracy (%)
0%	86.90
1%	81.10
2%	80.25
3%	77.64
4%	75.75
5%	74.36
15%	68.40

Table 5.6: Aggregate model accuracy with Training set noise (Exmedio)

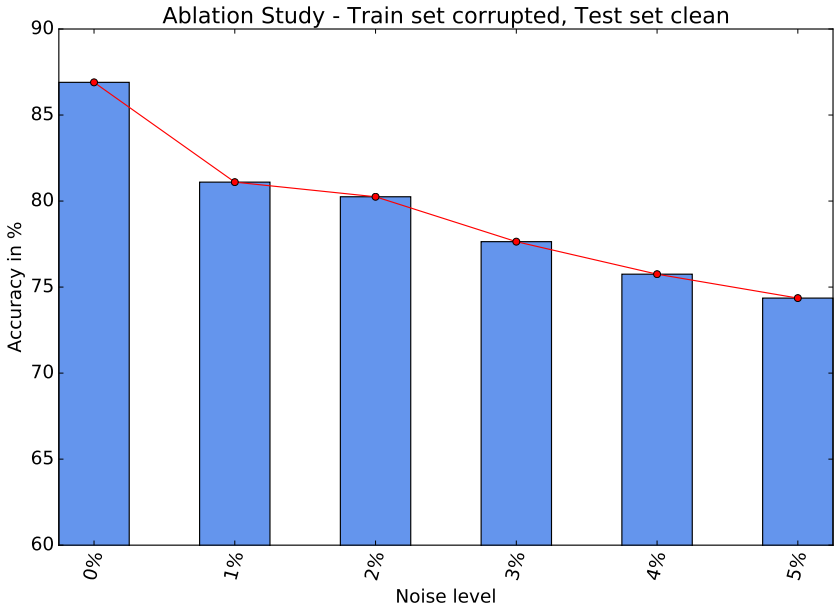


Figure 5-10: Ablation study with varying amount of Training set noise (Exmedio)

Matched Noise level	Top-1 Accuracy (%)
0%	86.90
1%	82.50
2%	81.71
3%	82.13
4%	80.62
5%	81.95
15%	80.07

Table 5.7: Aggregate model accuracy with matching noise levels

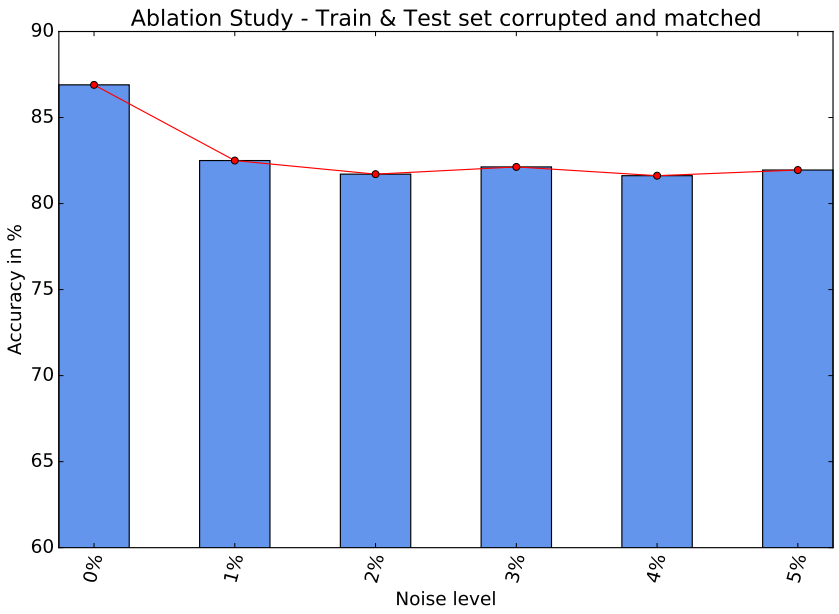


Figure 5-11: Ablation study with matching noise levels in Train & Test (Exmedio)

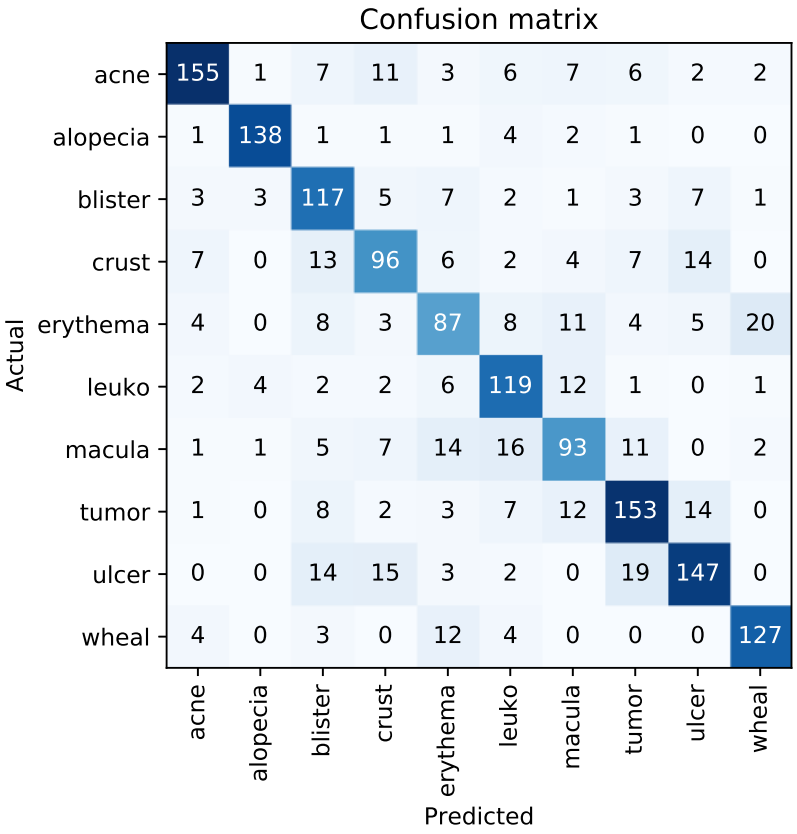


Figure 5-12: Confusion matrix - Test set isotropic blurred (Exmedio)

greater degree in the case of *Ulcer*. This label was confused for *Blister* and *Crust*, which are chronologically very related labels. A confusion matrix of the model is shown in Figure 5-12.

5.3.5 Training Set Normal, Test Set Anisotropically Blurred

The results for the anisotropic blurring of the test set was very remarkable and surprising. Anisotropic blurring saw negligible detriment in comparison with Isotropic blurring, seen previously in Sec. 5.3.4. This was significantly different from the pattern we observed in Sec. 5.2.5 where directional blurring significantly impacted the quality of prediction. The aggregate Top-1 Accuracy of the model was found to be 76%. A confusion matrix reflecting the model performance is shown in Figure 5-13.

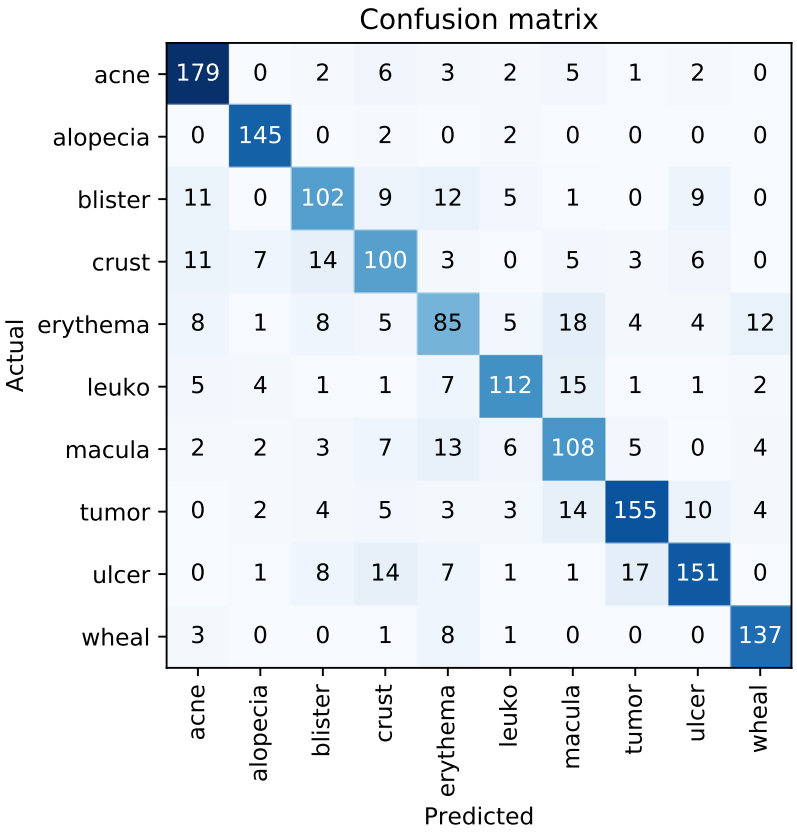


Figure 5-13: Confusion matrix - Test set anisotropic blurred (Exmedio)

5.4 Distribution Shift

Deep learning works best when the training and testing set largely follow the same statistical distribution. A major hurdle for such systems is seen when the conditions of training differ from the actual application in test time. Robustly trained classifiers are meant to perform reasonably over a variety of inputs. The input(s) may not be of the type the model was trained on. Additionally, inputs at test time could belong to a novel class that the model has not learned. Bias gets introduced when making a model learn on any specific dataset. This bias is unavoidable just like noise in images and emerging from the criteria we chose in building the data corpora. When working with critical technologies such as aerospace or medical diagnostics, models have to behave predictably in dealing with unseen inputs or categories. These phenomenon are broadly termed as *dataset shift* or *distribution shifts*. The current section tries to understand the scope of generalization in the models and data seen so far.

Table 5.8: Adversarial tests summary (Exmedio Skin Data)

Train Status	**Test Status**	**Top-1 Acc.** (in %)
Clean	Clean	86.90
Clean	1% impulse noise	73.57
	2% impulse noise	69.50
	3% impulse noise	65.97
	4% impulse noise	65.17
	5% impulse noise	64.82
	15% impulse noise	59.05
1% impulse noise	Clean	81.10
2% impulse noise		80.25
3% impulse noise		77.64
4% impulse noise		75.75
5% impulse noise		74.36
15% impulse noise		68.40
1% impulse noise	1% impulse noise	82.50
2% impulse noise	2% impulse noise	81.71
3% impulse noise	3% impulse noise	82.13
4% impulse noise	4% impulse noise	80.62
5% impulse noise	5% impulse noise	81.95
15% impulse noise	15% impulse noise	80.07
Clean	Isotropic Blur	74.85
	Anisotropic Blur	76.10

5.4.1 Understanding Dataset shift

Dataset shift or Distribution shift occurs when training and test joint distributions are not the same, as shown in Equation 5.1.

$$P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y) \tag{5.1}$$

To better understand dataset shift, let us assume our training data was sampled from some distribution $p_s(x, y)$ and our test data consist of possibly unlabeled examples drawn from some different distribution $p_T(x, y)$. Without much information about p_S and p_T we cannot assume the existence of a robust classifier.

For simplicity, let us assume a binary classification scheme. If the distribution shifts in arbitrary ways, the setup still permits the case where the distribution over inputs remain the same, but the labels are somewhat *flipped*, as seen in the Equations 5.2 and 5.3. It is

still possible to deduce and correctly identify sources of trouble in the model.

$$p_S(x) = p_T(x) \tag{5.2}$$

$$p_S(y|x) = 1 - p_T(y|x) \tag{5.3}$$

In a real situation, covering multiple labels, the problem does not remain this straightforward. The *shift* may spill over to another label as the manifold is not divided between just two labels. Principled approaches to identifying this problem have led to the following categories which researchers group this issue under for homogeneous images,

5.4.2 Covariate Shift

Covariate shift occurs in $X \to Y$ (i.e. X mapping to Y) problems where,

$$P_{\text{train}}(y|x) = P_{\text{test}}(y|x) \tag{5.4}$$

$$P_{\text{train}}(x) \neq P_{\text{test}}(x) \tag{5.5}$$

Covariate shift is perhaps the most often discussed case in the distribution shift problems. It is where the training and test follow different distributions but the functional relation remains the same. In this case, it is understood that the labeling function $P(y|x)$ does not change even though input could change. This shift is due to a shift in the distribution of features. This is the case often dealt with in dealing with fine-grained classification challenges such as pet categories or homogeneous dermatological data.

5.4.3 Labeling Shift

Labeling shift or Prior probability shift happens in $Y \to X$ problems (i.e. Y causing X). It can be mathematically expressed as,

$$P_{\text{train}}(x|y) = P_{\text{test}}(x|y) \tag{5.6}$$

$$P_{\text{train}}(y) \neq P_{\text{test}}(y) \tag{5.7}$$

This case is just the opposite of the covariate distribution shift, where it is assumed

that $P(y)$ could become somewhat variable even if the class conditional $P(y|x)$ values remain fixed. It is a likely scenario when we are confronted with problems where we want to predict the diagnosis, given some description of the manifestation/symptoms, even if the prevalence is changing over time.

There are even cases where covariate and label shift could be co-occurring. It is useful to consider the angle of label shift in managing such distributions.

5.4.4 Concept Shift

Perhaps the hardest of the three, *Concept shift* is defined as,

$$P_{\text{train}}(x|y) \neq P_{\text{test}}(x|y) \tag{5.8}$$

$$P_{\text{train}}(y) = P_{\text{test}}(y) \; \forall \; Y \leftarrow X \tag{5.9}$$

$$P_{\text{train}}(y|x) \neq P_{\text{test}}(y|x) \tag{5.10}$$

$$P_{\text{train}}(x) = P_{\text{test}}(x) \; \forall \; X \leftarrow Y \tag{5.11}$$

This shift is encountered when the definition of the categorization is amenable to change such as visual criteria for a diagnostics problem, where the label and the features are constantly under some kind of change but satisfy an underlying mapping.

5.4.5 Significance to Fine-grained Classification

If we want to design a model for detecting fine-grained classification for diseases, the data would represent a variety of conditions, both severe and non-severe. If we get high accuracy on our model design there is no guarantee it will work well in real-world use-case. The distributions in the training data versus the real-world use case might differ considerably. It would indeed be easy to distinguish between the presence and absence of a condition with high certainty. But due to the test case sampling procedure, there will be a significant intra-class variation which will lead to a high bracket of estimates primarily by covariate shift. A much more subtle situation arises when the distribution is non-stationary and the model is not updated adequately to compensate for its deleterious effects.

5.5 Generalization in CIFAR-10

When considering a standardized data corpus such as CIFAR-10, some generalization accuracy loss is to be expected as well. Recht et al. measured the accuracy of CIFAR-10 classification by using a test set made by separating some sections of the validation set and collating similar images from the Tiny Images Dataset [88, 105]. When performing a test by this new composite set, they found the accuracy to drop between 4–10%. A basic ResNext-29 architecture's accuracy dropped by 6.3%, whereas ResNet-110, ResNet-56 and ResNet-44 dropped by 8.3%, 8.4% and 8.8% respectively.

When discussing the possible reasons for the gap between CIFAR-10 accuracy and the new metrics, the authors have put forward multiple hypotheses.

1. **Statistical Error**: A natural outcome of any randomized trial, no two model tests will yield the same accuracy even with similar inputs.

2. **Duplicates:** The authors note that CIFAR-10 has almost identical images sometimes in train and test images, which could lead to some amount of memorization by the model. Most models have 99% – 100% accuracy in sample prediction on these near-duplicates, increasing the overall accuracy.

3. **Hyperparameter Tuning:** Initial learning rate, weight decay, and dropout were known to significantly impact the accuracy between different models.

4. **Hard Images:** Some images were simply hard for the model to make the predictions on.

5.6 Generalization in Exmedio Skin Data

In real-world scenarios, skin image classifiers are expected to work on diverse sets of input images. A supplied image may be different from the types of images the model learned to discriminate. In medical images, input images may belong to novel classes which would be examples of label shift. Models are expected to behave predictably even when the inputs aren't certain.

As described in Sec. 3.1.2, we chose the SD-198 dataset to test the generalization characteristic of skin lesion classifiers. It was an ideal choice for our study since it was

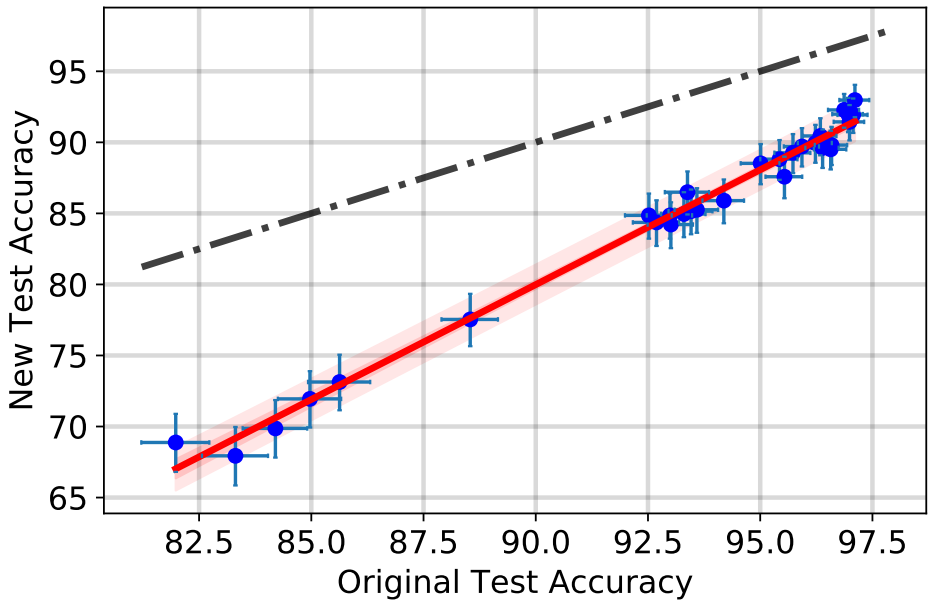


Figure 5-14: Model Performance in CIFAR-10 generalization, Recht at al. (2018)

also composed of user-submitted images. Since a one-to-one correspondence did not exist between classes in this dataset and our collected images, a dermatologist helped group relevant labels to our experimental design. This grouping is illustrated in Table 5.9.

Approximately a hundred samples were selected at random from these composite new classes and tested with our model (The only exception being *Wheal* which had only 71 samples). We trained a ResNet-50 model to optimal accuracy on our dataset and performed inference on a hybrid set. Poor generalization was observed, with only 32% Top-1 aggregate accuracy. Only *Acne* and *Alopecia* fared properly with 70% and 73% recall values respectively. Classifier bias was seen to favor simpler labels such as *Acne*, and *Blister* dominantly over others. The differences between different models lay between two percentage points. Figure 5-15 and 5-16 show the confusion matrices for this test on the ResNet-50 model.

5.7 Post-hoc Generative Model Evaluation for Augmentations

Data memorization is a significant issue with small data corpora which can potentially lead to over-fitting and prediction biases. Traditionally, it has been dealt with via using data augmentation. However, in several cases including medical images, care must be

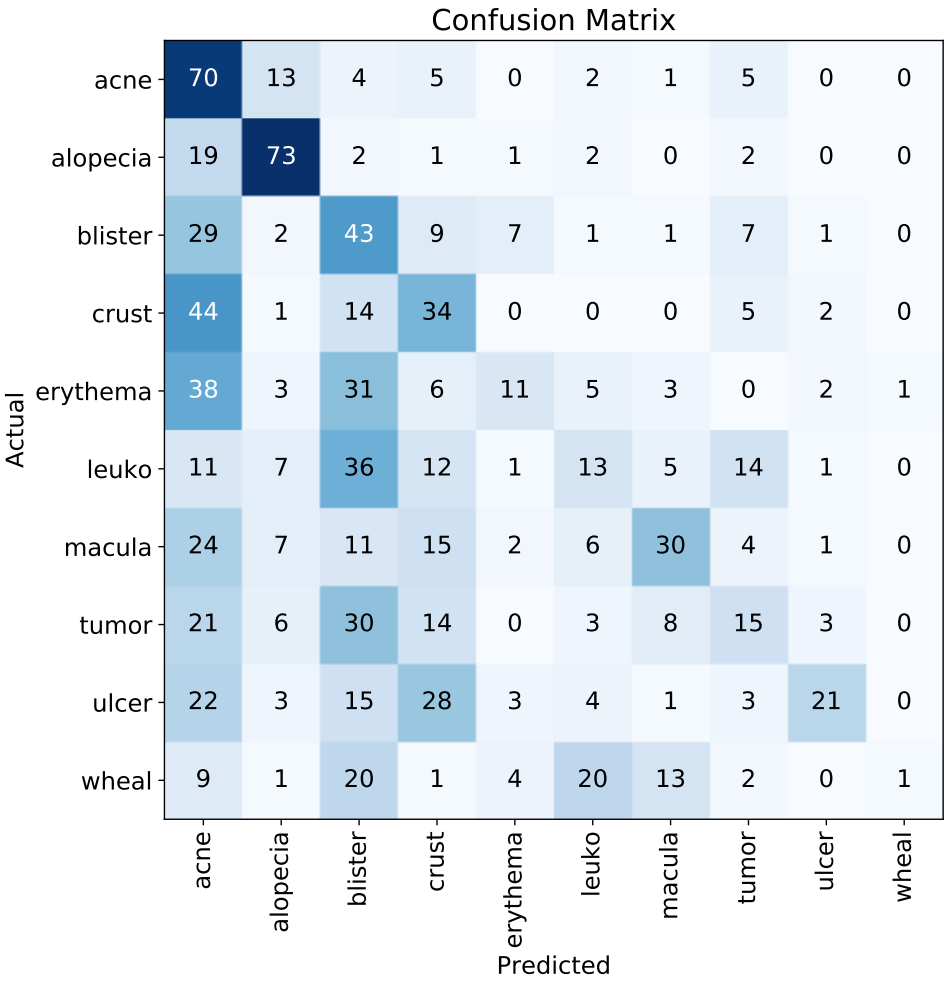


Figure 5-15: Model Performance in SD-198 generalization

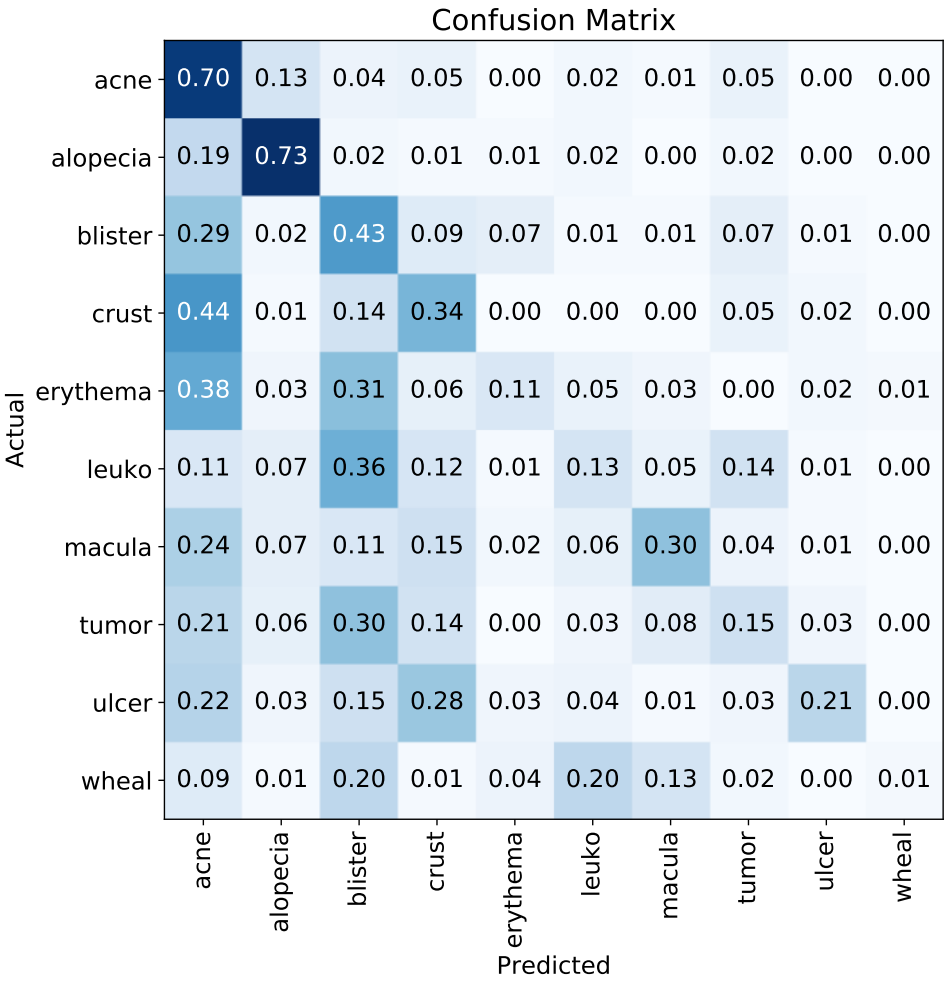


Figure 5-16: Model Performance in SD-198 generalization

Table 5.9: SD-198 grouped for distribution shift study.

Label	SD-198 Classes
Acne	Acne Keloidalis Nuchae, Acne Vulgaris, Steroid Acne, Favre Racouchot, Nevus Comedonicus, Pomade Acne
Alopecia	Alopecia Areata, Androgenetic Alopecia, Follicular Mucinosis, Kerion, Scar Alopecia.
Blister	Dyshidrosiform Eczema, Hailey Disease, Herpes Simplex, Herpes Zoster, Varicella, Mucha Habermann disease
Crust	Angular Cheilitis, Bowen's Disease, Impetigo
Erythema	Acute Eczema, Candidiasis, Erythema Ab Igne, Ery. Annulare Centrifigum, Ery. Craqule, Ery. Multiforme, Rosacea, Exfoliative Erythroderma
Leukoderma	Balanitis Xerotica Obl., Beau's Lines, Halo Nevus, Leukonychia, Pityriasis Alba, Vitiligo
P. Macula	Actinic Solar Damage, Becker's Nevus, Blue Nevus, Cafe Au Lait Macula, Compound Nevus, Congenital Nevus Dermatosis Nigra, Epidermal Nevus, Green Nail
Tumor	Angioma, Apocrine Hydrocystoma, Lipoma, Dermatofibroma, Digital Fibroma, Fibroma, Leiomyoma
Ulcer	Aphthous Ulcer, Behcet's Disease, Ulcer, Stasis Ulcer, Mal Perforans, Pyoderma Gangrenosum, Syringoma
Wheal	Urticaria, Stasis Edema

taken to perform augmentation. Although we can use affine transformation efficiently, several other means such as color, hue, deformation, and geometrical transformations cannot be used for dataset augmentation. Recently, there has been a lot of momentum in the computer vision domain to perform augmentation by generative adversarial networks (GAN) [36, 33].

However, this method is not without its problems. Some recent research points to the fact that generative output may not truly represent the fine visual attributes a real image has. The quality of synthetic images is very strongly related to the way the parent labels are sampled [38]. To test whether such images can truly be representative of real images, we needed to perform a post-hoc analysis of the generative model creating synthetic samples. Our first step in this process was to implement a GAN model based on label pairs that exhibited high concurrency in a real clinical setting. The investigative rationale was straightforward: If synthetic samples are created from two labels in a predetermined

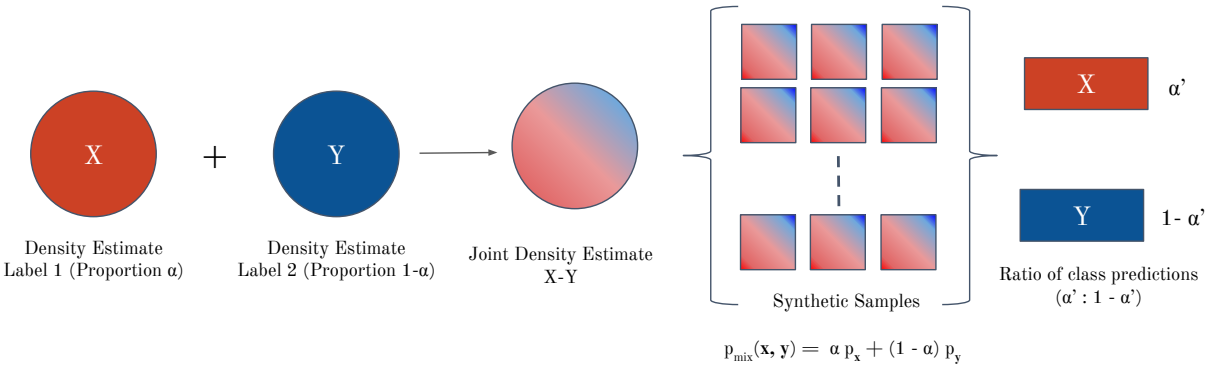


Figure 5-17: Schematic for post-hoc model evaluation for data augmentations

proportion, the representation in latent space would be a quasi-linear combination of the density estimates of parent samples. When predicting a large number of such synthetic images, the ensemble accuracy should closely reflect the mixing ratio of these labels in their average posterior probability of getting a label. Figure 5-17 is a schematic of our hypothesis.

Using approximately equal training data from *Acne* and *Blister*, which are known to exist together frequently, we designed a binary classifier. We also created several novel samples by a Progressively growing GAN (ProGAN) framework from these aforementioned labels [33, 49]. We trained the model repeatedly and calculated the posterior probabilities for the synthetic samples each time. This ensemble performance closely followed the distribution but was not accurate enough. These results are summarized in Table 5.10. It was observed that the representation learning indicated by posterior probabilities are close but different between each run. The model was unable to guess the label ratio in the binary representation within acceptable bounds. Augmenting the dataset with such synthetic samples could bias the model. This error could amplify in the presence of more classes. The margin of error in healthcare is slim and hence it is unlikely serious practitioners should adopt this approach.

Table 5.10: Result of synthetic data testing from generative models

Model accuracy	Label-1 Estimate	Label-2 Estimate
88.9%	58.75%	41.25%
88.4%	22.50%	77.50%
90.4%	45.00%	55.00%
88.5%	57.50%	42.50%

5.8 Chapter Summary

Adversarial perturbations have been well-studied in the case of popular datasets. Medical datasets have seen limited experimentation to investigate their adversarial resilience and benchmarks are rarely published. This chapter investigated the effects of two of the most common imperfections commonly seen (impulse noise and soft-focus blur) on the model performance in Exmedio skin images.

We performed a set of ablation studies with increasing content of impulse noise. Both training and test sets were subject to corruption. The model performance was seen to drop by at least ten percentage points when the test set was corrupted. Adversarial training improved the stability of models across several noise levels. The best performance was seen when the noise content statistically matched between the training and test sets. The pattern of noise-induced accuracy degradation was similar in skin images and CIFAR-10. However, skin images were more susceptible as compared to CIFAR-10 images.

In the case of simulated soft-focus induced by Gaussian kernels, CIFAR-10 was seen to be less resilient than skin images. Skin images saw a lower drop in accuracy versus CIFAR-10. The differences between isotropic and anisotropic blurring were found to be negligible. We investigated the generalization performance on SD-198 images by testing a model robustly trained on Exmedio Skin images. The results were far poorer than those conducted for CIFAR-10 by Recht et al. (2018). The model seemed to favor visually simpler labels in detection.

Current research discussions involve using generative models such as GANs to synthesize photo-realistic samples which could be used for augmentations. Although this approach has worked well for trivial images such as ImageNet or CIFAR-10 categories, it has not been thoroughly investigated for medical data augmentation. We created synthetic samples using Progressive GAN (ProGAN) for two commonly co-morbid skin conditions. We evaluated the statistical properties of the generated samples by a classifier

trained on real images. Upon investigation, we found the ratio of mixing to vary quite differently between trials. We deduce that using such samples as augmentation could inadvertently bias the classifiers and hence should be avoided.

Chapter 6

Conclusion

With deep learning methods, our goals to understand and automate cognitive tasks have become higher. Deep learning methodologies have discarded the rule-based approaches which were prevalent until a decade ago to provide remarkable results. The consensus to achieve better performance has been to deploy larger models on a higher quantity of data. It has been facilitated by the falling cost of computing, cheap storage, and ubiquitous sensors capturing information from the digital world around us. However, these methods have missed the aspect of scalability. Results are on par with human performance only when annotated data is abundant. Real-world collection may not always be at the scale of ImageNet or WikiQA. Additionally, computing at a scale of several dozen to thousands of GPU is not within the reach of every individual researcher. Transfer learning has been able to mitigate some of the issues of building ML models from scratch. Fine-tuning these models for custom application still requires careful consideration.

Model learning improvements

When looking for learning improvements in object classification and detection tasks using pre-trained models, adopting complex models with deeper layers is the current trend. Larger and complicated models typically capture a lot of information through their hidden layers, but this improvement comes at a cost of increased fitting time, memory requirements, and computational complexity. Instead of making the model larger, the current approach aimed at increasing the usability of layers by better model fitting. Chapter 3 demonstrated a paradigm involving a combination of layer-wise rate throttling and cosine rate annealing with cycle length multiplication. Although both the methods have existed

separately, it was seen that combining the two could have a better effect on model learning by being able to commence fitting at a much higher rate as well as reaching optimum accuracy (or low loss values) very quickly. The training could consequently be reduced to a few epochs or requiring multiple restarts with manual tuning in between. Although primarily designed keeping ResNets with its skip connections in mind, this method was effective even in networks such as DenseNets (having numerous skips pointing back to base layer) and ResNexts (which have split-transform-merge parallel architectures).

With this combination scheme, we could get a nominal speedup of about 1.7 times in classifying standardized datasets such as CIFAR-10 and between 1.3–1.9 times in custom, smaller datasets such as Exmedio Skin Image dataset. These scores were further improved by using automatic mixed-precision learning, where the learning time was about 5 times faster while classifying CIFAR-10 and between 3.42–3.82 times in Skin image classification. We can summarize the advantages of this paradigm with the following points.

- Reduced fitting time in comparison to conventional methods by using differential rates and cycle length multiplication.
- Higher acceleration per epoch towards convergence i.e. reduced the number of epochs needed to achieve the same level of accuracy.
- Lesser memory footprint by increasing the usability of layers by tighter fits.

With this learning paradigm, we observed a small model like ResNet-34 having accuracy in the neighborhood of larger models (e.g. ResNet-152). This lessens the requirement to adopt very deep models for specific tasks. This approach is beneficial from three standpoints as well:

1. Older GPU devices with lesser memory bandwidth can be used for getting good results with this paradigm. A major bottleneck with older generations of GPU is the memory bandwidth. Newer and larger models cannot be used in them without running out of space. Deep learning has seen the requirement to use newer GPUs for larger computational matrices and better results. If smaller models fit by this paradigm could provide performance in the neighborhood of state-of-the-art in benchmarks, they could lead to a longer retention of GPUs. With newer GPU devices costing upwards of USD 4000, this amounts to significant savings. The

Although the results in the thesis are derived from using a NVIDIA V100™(32GB SXM2) this learning paradigm has been tested to work well with models on NVIDIA 1080Ti™(Released Q1 2017, except ResNet-152, DenseNet, and ResNext models), NVIDIA Titan XP ™(Released Q1 2017 11GB DDR5 RAM, except ResNet-152 & ResNext models), and NVIDIA Titan-V ™(Released Q1 2018 12GB HBM2, except ResNext models).

2. Model learning on cloud platforms could be run at a lower cost. This is beneficial for users and research groups operating on shoestring budgets. For example, this method could train CIFAR-10 to more than 94% accuracy for $ 3.6 USD as compared to $9.9 USD at the minimum (`p2.xlarge` AWS instance equipped with K80 GPU, `us-east-1` Q4 2020 pricing).

3. Smaller models trained to high accuracy on server GPUs can be utilized without further modifications on edge computing devices such as NVIDIA Jetson/Xavier™which have limited frame buffer to store model parameters.

Model interpretations and errors

Chapter 4 explored the nature of classification errors in homogeneous images. After robust model training, it was observed that there was still a gap between best model performance and error-free detection in all the models that trained. Although an accepted part of any statistical system, this fact poses serious questions in application domains such as healthcare, where the margin of error is slim. Robustly learned systems are expected to provide high consistent quality of service, with low error margins.

In the case of Oxford-IIIT Pet dataset discussed in Section 4.1, GradCAM analysis showed the nature of classification to follow the similar trends published by previous studies. The same was not true in the case of homogeneous skin data. Some pairs exhibited a much higher error-rate than usual ($\geq$ 10 per run or approximately 5% of the label meant for inference). Many of these classification errors followed a pattern apart from the presence or absence of texture. Upon investigation of these labels, it was discovered that these miscategorizations emerged due to strong shape symmetry, color symmetry & insufficient chronological information in addition to previously discovered properties regarding texture. Fine-grained images looked further into initial layers for

separation in the latent manifold, as compared to CIFAR-10 or ImageNet published by earlier studies.

During the process of creating such high-performance classifiers, the deleterious effects of poor quality inputs could be alleviated by using few curating strategies. The most important of them was found to be individual class representation. The supervised learning paradigm relies on a cornerstone of learning representations well by equal representation of the labels. In cases where the labels were unevenly distributed, two strategies could be employed:

1. Data augmentation, which created fixed copies from the set of images that were present. Commonly available tools such as ImgAug [48] could handle this well. Precautions were to be kept in mind to choose the type of augmentations to perform.

2. Weighted sampling which selected data inversely proportional to its representation. Hence, if a label was over-represented it was loaded less often by the software routines, as compared to another label which could be sparsely present. This method was adopted and further refined to involve augmentations in the data-loading module itself. Care was taken not to involve schemes such as color jitter, deformation, etc in data such as skin lesion images since these could affect the quality of learning.

Further, other strategies such as illumination correction and FOV reduction were shown to sufficiently improve the quality of learning. Confounding markers were seen in $\geq$ 400 images in the train set and approximately 150 images in the test set of the skin image dataset. These were sufficiently resolved by transforming images into grids of patches and replacing the aberrations in a patch with the respective dominant color calculated by k-Means clustering (of the color-space). There is further scope of improvement in this preprocessing step by identifying better masking thresholds. In the current research, these were set by manual analysis with some trial and error in finding the right RGB values. Since the skin tone in the current dataset was mostly uniform, outcomes were consistent. However, if more racial cohorts were to be used, it could have posed challenges. Some degree of automation could be immensely beneficial in such situations. A promising direction would be to investigate the R-channel instead of RGB values, in designing masking thresholds across different skin tones.

Since skin images are extremely homogeneous, it was not well understood if the background played any part in the correct detection. Previous studies involving dermoscopic images indicated the background was necessary to detect *Melanoma*. However, it was not understood for commonly encountered skin problems. This hypothesis was tested by occluding the main features in high-confidence samples and running inference. It was seen that background seemed to have a tenuous connection to presenting a good estimation.

Adversarial Robustness and Generalization

In Chapter 5, we tried to deduce the extent to which model performance can be degraded in skin images due to imperfections and perturbations in homogeneous data. It is well understood in the machine learning domain that noise induces learning instabilities. This phenomenon has been well demonstrated by scores of papers with conventional data. It has not been investigated reliably in the medical images domain. Our goal was to understand if discriminative models can be reliable, within certain bounds to carry investigations.

We added impulse noise to our data (1%–5% and 15%) in different ablation schemes for the benchmark CIFAR-10 and Exmedio Skin data. When we modeled CIFAR-10 with adversarial noise, the only disadvantageous case observed was in the case of introducing noise in the testing data. Training with corrupted images increased resilience in both clean and noise-corrupted images. The model accuracy was observed to be in a narrow band ($\mu, \sigma = 81.98 \pm 0.31\%$). The same was not seen in skin lesion images. Accuracy rapidly fell by about a maximum of 20% when testing with corrupted images. Some improvement was observed with adversarial training, which improved the gap by 5% (i.e. a 15% drop). Stable performances were once again seen in a narrow band when noise matched statistically between training and testing sets ($\mu, \sigma = 81.49 \pm 0.86$). Blur corruption was more significant in the case of CIFAR-10 but the effect was negligible in Skin data, showing some resilience due to homogeneity. This was true when testing for isotropic as well as anisotropic blur.

We do not yet fully understand how adversarial inputs affect model performance. This is a topic of ongoing discussion in the broader community [117]. It is not certain if medical images can be suitably processed to remove such artifacts without inducing unavoidable changes. However, we can assume that model stability can be streamlined if

we can estimate the noise content in test images and match the training corpus accordingly. Statistically matched levels of noise have been shown to provide stable prediction outcomes, if not the best outcomes. Wan et al. demonstrated a noise estimator based on histogram [112]. This method is limited to grayscale images. Such methods could be extended to RGB images with suitable alterations as demonstrated by Malinski et al [74].

Although CIFAR-10 has shown a drop in accuracy by 4–5% in previously published generalization tests, our trial on labels from the SD-198 dataset to Exmedio data yielded far poorer results. It was observed that in homogeneous skin images, the propensity of the model was to choose visually simpler labels, the reasons for which are not well understood. To alleviate such issues, we also considered using synthetic samples generated by GANs. However, likelihood inference and posterior probability estimates yielded that the sampling from the true label is not stable enough to present accurate statistical information to the model for data balancing. With this discovery, we can safely assume that synthetic samples are not sufficient for data augmentation in sparse labels. It cannot alleviate long-tailed distribution since any sample that we synthesize could end up with a model bias towards visually simpler categories.

Outlook for medical image analysis

The earliest questions on computation solving the issue of medical image analysis were posed by Lodwick et al [69]. Computer-aided diagnosis has come a long way since then. Scientists are hopeful that deep learning will one day solve long-standing automation issues. Unlike other computer science sub-disciplines, there is no Turing test for medical image comprehension. ML performance lags other domains because the available data is lesser and visually complex. The label complexity is another impeding factor in overall success.

ImageNet-like leaderboards have enjoyed a lot of success because the data is well-curated and adopted by the community. Training data is crucial to successful applications. In niche domains, it is harder to obtain and requires expert knowledge to generate labels. In many cases, the labels are not clean and annotated. Large publicly available, well-curated datasets are still scarce. Despite of this, several semi-public datasets have emerged in recent times such as Stanford CheXpert [47], NIH Pulmonary X-ray

dataset [114] and HAM10000 skin cancer dermoscopic set [106]. The availability of such datasets is slowly expanding the scope of deep learning in medical applications.

The amount of data is not the only pertinent variable. Data complexity is an important issue to consider for good precision and recall performance in these models. Image quality is an important aspect to consider. Although signal-to-noise (SNR) is referred to in most cases as the comparison metric, most studies have so far not included the effect of artifacts, soft-focus blur, and diversity in the performance of imaging devices. In the absence of standard protocols regarding the size of the image, resolution, and placement of the camera, datasets see a lot of variability in inputs. Acquisition parameters such as radiance and post-processing techniques can lead to very different outcomes. With data involving more attributes, every feature dimension that we add pushes the model closer to the *curse of dimensionality*. Medical images such as X-rays or skin images have a very uniform background. Relevant features can be small and easy to miss unless care is taken to design special architectures [110, 85]. Skin lesion diagnosis application was investigated with self-supervised methods as well, such as MixMatch and SCAN [8, 107]. The outcomes were sub-par as compared to the model learning with supervision. There are many issues in achieving human-level performance:

1. Absence of label granularity doesn't help to evaluate the severity of the condition (or staging). Several conditions can be misjudged due to lack of such information with confusions arising from chronologicity.

2. Bounding boxes are not very relevant often since contours, textures, and colors are important. Pixel-level segmentation is impractical at the moment since the images could be even multi-dimensional or a time-series.

3. The number of classes to consider is very large. Dermatology has over 2000 conditions across 60 different disease families. Instead of being discrete, several of them form a part of the spectrum with shared symptoms and attributes.

4. Diseases such as skin images rarely occur in isolation. A human evaluator can disregard unnecessary information easily. The same is not true in the case of ML models. We have discussed such issues in Section 4.5.2.

In addition to the statistical issues discussed, the quality of labeling has a profound impact on the outcomes. The process of annotation sees a significant amount of inter-

operator variability (i.e. difference in opinions between different clinicians) and intra-operator variability. (variability within an individual clinician decision). Insufficient clinical reports also add to the difficulty in diagnosis. In the current study, a single racial type (East Asian) has been investigated. With the introduction of additional racial cohorts, the variability in skin tone will negatively impact the precision of such models. Section 5.6 demonstrates one such example, where we tested our model performance on samples obtained from a different geographical location [100].

Deep learning has shown tremendous success in retinal images [22], chest X-rays, and multi-modal data having ample clinical reports. However, their performance falls short when the data is scarce or the inputs complex and non-standardized. Long-tailed distributions are not easily modeled. Instead of unsupervised learning, supervised and few-shot learning perhaps could be the only viable means in the future. In light of these deficiencies, deep learning models can be valuable in assisting physicians in verifying diagnosis and picking up attributes that the human eyes could have missed. There is a long road yet to be traveled to reach an unassisted, reliable diagnosis with deep learning.

Publications

Archival Venues

1. **Improving image classifiers for small datasets by learning rate adaptations.**
   Sourav Mishra, Toshihiko Yamasaki, and Hideaki Imaizumi.
   16^{th} International Conference on Machine Vision Applications (MVA)
   May 27–30, 2019. DOI: 10.23919/MVA.2019.8757890
   (Selected for Oral Presentation)

2. **Interpreting fine-grained dermatological classification by deep learning**
   Sourav Mishra, Hideaki Imaizumi, and Toshihiko Yamasaki.
   Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 16–June 20, 2019.
   (Selected for Oral Presentation)

3. **Robustness of deep learning models in dermatological evaluation: A critical assessment**
   Sourav Mishra, Subhajit Chaudhury, Hideaki Imaizumi, and Toshihiko Yamasaki.
   IEICE Transactions on Informations and Systems,Vol. E104-D (03), March 2021.

Non-archived and national venues

1. **Assessing Robustness of Deep learning Methods in Dermatological Workflow**
   ACM Conference on Health Inference and Learning (CHIL), 2020
   Sourav Mishra, Subhajit Chaudhury, Hideaki Imaizumi, and Toshihiko Yamasaki
   (Selected for Workshop Spotlight)

2. **Supervised classification of dermatological diseases by DNN**
   Meeting on Image Recognition and Understanding (MIRU), 2018
   Sourav Mishra, Hiromi Hirano, Hideaki Imaizumi, and Toshihiko Yamasaki

3. **Implementing & interpreting fine-grained classification of Dermoscopic images by Deep Learning**
   Meeting on Image Recognition and Understanding (MIRU), 2019
   Sourav Mishra, Hideaki Imaizumi, and Toshihiko Yamasaki

4. **Man + Machine: The future of precision evaluation in dermoscopy**
   Winter Session 2019, Institute of Telecommunication Engineers (ITE), Japan
   Sourav Mishra and Toshihiko Yamasaki
   (Winner: Best Student Paper (2019))

Appendix A

Additional Figures

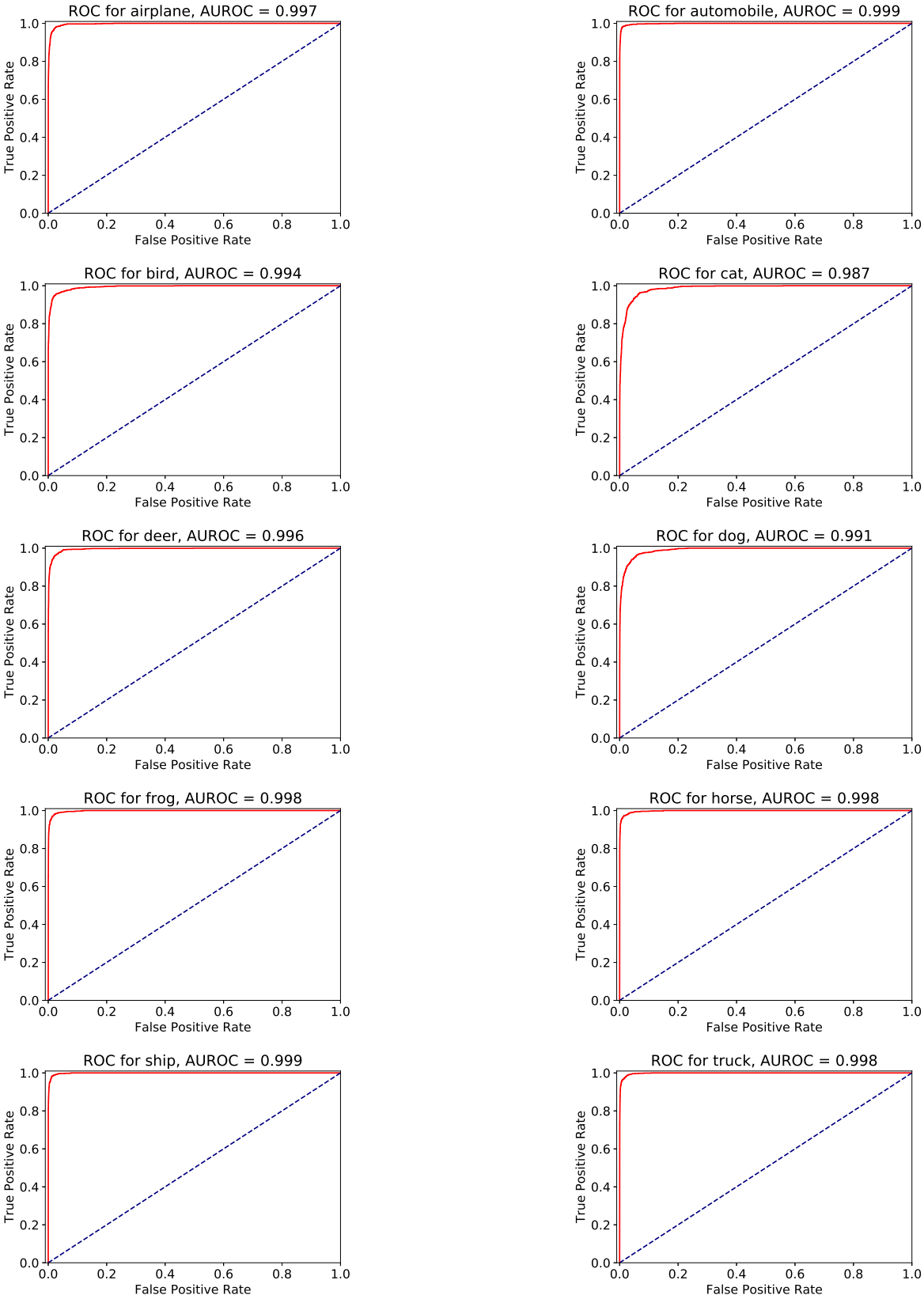


Figure A-1: ROC curves and AUROC (CIFAR10)

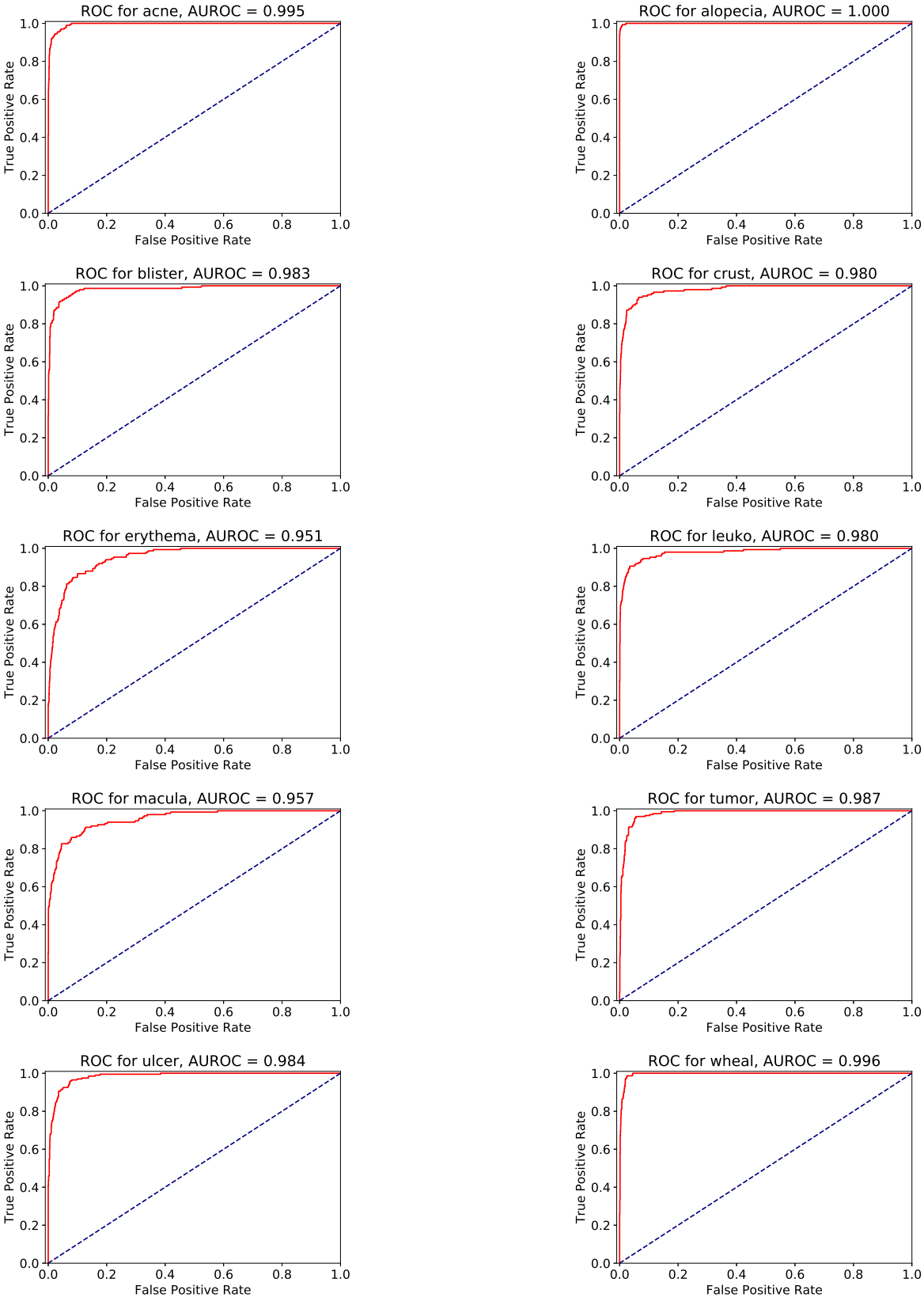


Figure A-2: ROC curves and AUROC (Exmedio)

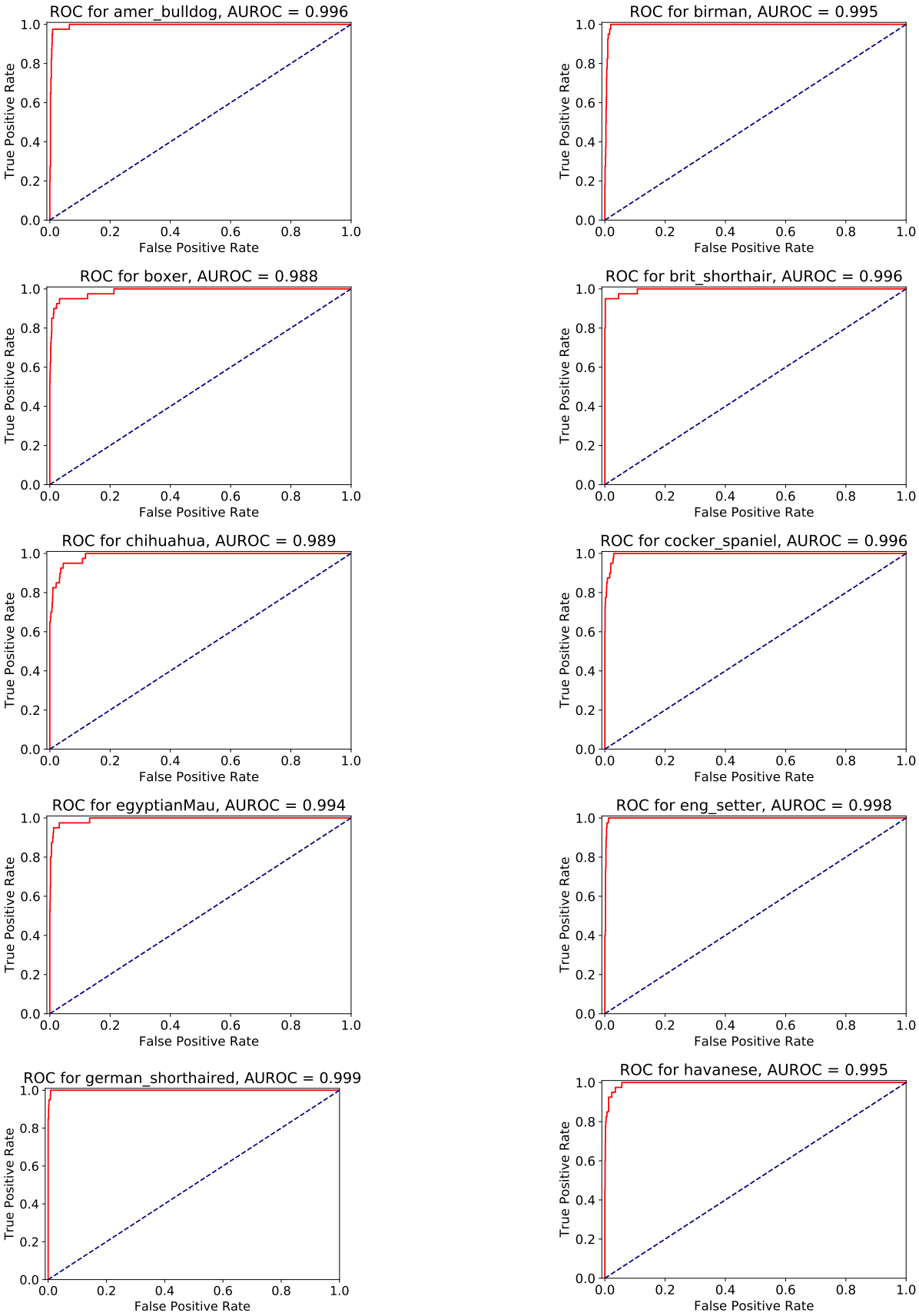


Figure A-3: ROC curves and AUROC (IIIT Pets) Set A

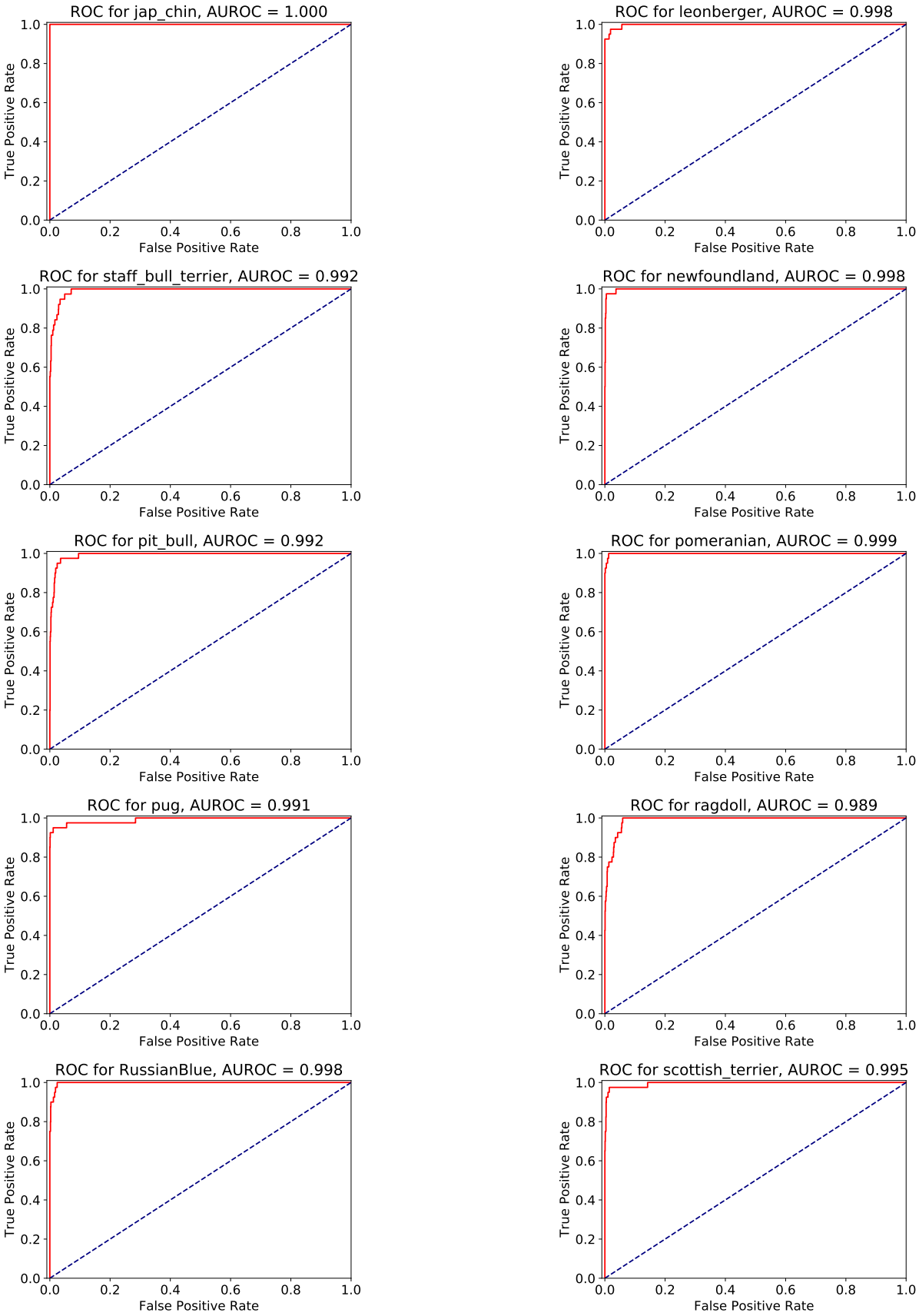


Figure A-4: ROC curves and AUROC (IIIT Pets) Set B

Bibliography

[1] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Investigative ophthalmology & visual science*, 57(13):5200–5206, 2016.

[2] Sina Ardabili, Amir Mosavi, Majid Dehghani, and Annamária R Várkonyi-Kóczy. Deep learning and machine learning in hydrological processes climate change and earth systems a systematic review. In *International Conference on Global Research and Education*, pages 52–62. Springer, 2019.

[3] Aharon Azulay and Yair Weiss. Why do deep convolutional networks generalize so poorly to small image transformations? *J. Mach. Learn. Res.*, 20:184:1–184:25, 2019.

[4] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017.

[5] C. Baur, Shadi Albarqouni, and Nassir Navab. Generating highly realistic images of skin lesions with GANs. In *International Skin Imaging Collaboration Workshop, MICCAI*, 2018.

[6] C. Baur, Shadi Albarqouni, and Nassir Navab. Melanogans: High resolution skin lesion synthesis with gans. *ArXiv*, abs/1804.04338, 2018.

[7] Yoshua Bengio. RMSprop and equilibrated adaptive learning rates for nonconvex optimization. *ArXiv preprint arXiv:1502.04390*, 2015.

[8] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 5049–5059, 2019.

[9] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (de) constructing bias on skin lesion datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 0–0, 2019.

[10] Stevo Bozinovski. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica*, 44(3), 2020.

[11] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[12] Titus J Brinker, Achim Hekler, Alexander H Enk, Carola Berking, Sebastian Haferkamp, Axel Hauschild, Michael Weichenthal, Joachim Klode, Dirk Schadendorf, Tim Holland-Letz, et al. Deep neural networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer*, 119:11–17, 2019.

[13] Titus J Brinker, Achim Hekler, Alexander H Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Tim Holland-Letz, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113:47–54, 2019.

[14] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[15] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[16] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069*, 2018.

[17] Stephanie Chan, Vidhatha Reddy, Bridget Myers, Quinn Thibodeaux, Nicholas Brownstone, and Wilson Liao. Machine learning in dermatology: Current applications, opportunities, and limitations. *Dermatology and therapy*, pages 1–22, 2020.

[18] Noel Codella, Junjie Cai, Mani Abedini, Rahil Garnavi, Alan Halpern, and John R Smith. Deep learning, sparse coding, and svm for melanoma recognition in dermoscopy images. In *International workshop on machine learning in medical imaging*, pages 118–126. Springer, 2015.

[19] Cody Coleman, Deepak Narayanan, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. Dawnbench: An end-to-end deep learning benchmark and competition. *Training*, 100(101):102, 2017.

[20] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[21] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[22] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O' Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

[23] Itaru Dekio, Eisuke Hanada, Yuko Chinuki, Tatsuya Akaki, Mitsuhiro Kitani, Yuko Shiraishi, Sakae Kaneko, Minao Furumura, and Eishin Morita. Usefulness and economic evaluation of adsl-based live interactive teledermatology in areas with

shortage of dermatologists. *International journal of dermatology*, 49(11):1272–1275, 2010.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer vision and Pattern recognition (CVPR)*, pages 248–255. IEEE, 2009.

[25] Jeffrey M. DiCarlo, F. Xiao, and B. Wandell. Illuminating illumination. In *Color Imaging Conference*, 2001.

[26] Samuel Dodge and Lina Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *2017 26th international conference on computer communication and networks (ICCCN)*, pages 1–7. IEEE, 2017.

[27] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning (ICML)*, pages 647–655, 2014.

[28] 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):115–118, 2017.

[29] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization, 2020.

[30] SA Galderma. Dermquest image library, 2014.

[31] Xavier Gastaldi. Shake-shake regularization. *arXiv preprint arXiv:1705.07485*, 2017.

[32] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[33] A. Ghorbani, V. Natarajan, David Coz, and Y. Liu. Dermgan: Synthetic generation of clinical skin images with pathology. In *ML4H, Neural Information Processing Systems*, 2019.

[34] Justin Gilmer, Nicolas Ford, Nicholas Carlini, and Ekin Cubuk. Adversarial examples are a natural consequence of test error in noise. In *International Conference on Machine Learning (ICML)*, pages 2280–2289, 2019.

[35] Boris Ginsburg, Patrice Castonguay, Oleksii Hrinchuk, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, Huyen Nguyen, Yang Zhang, and Jonathan M. Cohen. Stochastic gradient methods with layer-wise adaptive moments for training of deep networks, 2020.

[36] Ian J. Goodfellow, Jean Pouget-Abadie, M. Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[37] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[38] Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. In *Advances in Neural Information Processing Systems*, pages 11058–11070, 2019.

[39] Holger A Haenssle, Christine Fink, R Schneiderbauer, Ferdinand Toberer, Timo Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, Luc Thomas, A Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer vision and Pattern recognition*, pages 770–778, 2016.

[41] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

[42] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning Lecture 6a: Overview of Mini-batch gradient descent.

[43] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, (ACL)*, pages 328–339. Association for Computational Linguistics, 2018.

[44] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

[45] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4700–4708, 2017.

[46] Hideaki Imaizumi, Akio Watanabe, Hiromi Hirano, Masatoshi Takemura, Hideyuki Kashiwagi, and Shinichiro Monobe. Hippocra: Doctor-to-doctor teledermatology consultation service towards future AI-based diagnosis system in japan. In *2017 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pages 51–52. IEEE, 2017.

[47] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[48] Alexander Jung. Imgaug documentation. *Readthedocs. io, Jun*, 25, 2019.

[49] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[50] Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural information processing systems*, pages 586–594, 2016.

[51] Jeremy Kawahara, Aicha Ben Taieb, and Ghassan Hamarneh. Deep features to classify skin lesions. In *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*, pages 1397–1400. IEEE, 2016.

[52] Nitish Shirish Keskar and Richard Socher. Improving generalization performance by switching from adam to sgd. *arXiv preprint arXiv:1712.07628*, 2017.

[53] Alexa Boer Kimball and Jack S Resneck Jr. The us dermatology workforce: a specialty remains in shortage. *Journal of the American Academy of Dermatology*, 59(5):741–745, 2008.

[54] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[55] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The CIFAR-10 dataset. *online: http://www.cs.toronto.edu/kriz/cifar.html*, 55, 2014.

[56] Maxime W Lafarge, Erik J Bekkers, Josien PW Pluim, Remco Duits, and Mitko Veta. Roto-translation equivariant convolutional networks: Application to histopathology image analysis. *arXiv preprint arXiv:2002.08725*, 2020.

[57] Rosilene Canzi Lanzini, Robyn S Fallen, Judy Wismer, and Hermenio C Lima. Impact of the number of dermatologists on dermatology biomedical research: a canadian study. *Journal of cutaneous medicine and surgery*, 16(3):174–179, 2012.

[58] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907-193934, 2020.

[59] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[60] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[61] Dongwoo Lee, Haesol Park, I. Park, and Kyoung Mu Lee. Joint blind motion deblurring and depth estimation of light field. *ArXiv*, abs/1711.10918, 2018.

[62] Di Lin, Xiaoyong Shen, Cewu Lu, and Jiaya Jia. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1666–1674, 2015.

[63] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[64] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 1449–1457, 2015.

[65] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[66] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018.

[67] Weihuang Liu, Mario Juhas, and Yang Zhang. Fine-grained breast cancer classification with bilinear convolutional neural networks (bcnns). *Frontiers in Genetics*, 11, 2020.

[68] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, pages 1–9, 2020.

[69] Gwilym S Lodwick. Computer-aided diagnosis in radiology: A research plan. *Investigative Radiology*, 1(1):72–80, 1966.

[70] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3431–3440, 2015.

[71] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017.

[72] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[73] Brooke A Lowell, Catherine W Froelich, Daniel G Federman, and Robert S Kirsner. Dermatology in primary care: prevalence and patient disposition. *Journal of the American Academy of Dermatology*, 45(2):250–255, 2001.

[74] Lukasz Malinski and Bogdan Smolka. Self-tuning fast adaptive algorithm for impulsive noise suppression in color images. *Journal of Real-Time Image Processing*, 17(4):1067–1087, 2020.

[75] Scott Menard. *Applied logistic regression analysis*, volume 106. Sage Publishers, 2002.

[76] Sourav Mishra, Hideaki Imaizumi, and Toshihiko Yamasaki. Interpreting fine-grained dermatological classification by deep learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[77] Jorge J Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978.

[78] Nikita Moshkov, Botond Mathe, Attila Kertesz-Farkas, Reka Hollandi, and Peter Horvath. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific reports*, 10(1):1–7, 2020.

[79] Li Niu, Ashok Veeraraghavan, and Ashutosh Sabharwal. Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7171–7180, 2018.

[80] Thomas George Olsen, B Hunter Jackson, Theresa Ann Feeser, Michael N Kent, John C Moad, Smita Krishnamurthy, Denise D Lunsford, and Rajath E Soans. Diagnostic performance of deep learning algorithms applied to three common diagnoses in dermatopathology. *Journal of pathology informatics*, 9, 2018.

[81] Manfred Opper and David Haussler. Generalization performance of bayes optimal classification algorithm for learning a perceptron. *Phys. Rev. Lett.*, 66:2677–2680, May 1991.

[82] Andrew J Park, Justin M Ko, and Robert A Swerlick. Crowdsourcing dermatology: Dataderm, big data analytics, and machine learning technology. *Journal of the American Academy of Dermatology*, 78(3):643–644, 2018.

[83] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3498–3505, 2012.

[84] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural information processing systems*, pages 8026–8037, 2019.

[85] Hans Pinckaers, Bram van Ginneken, and Geert Litjens. Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. *arXiv preprint arXiv:1911.04432*, 2019.

[86] Victor Pomponiu, Hossein Nejati, and N-M Cheung. Deepmole: Deep neural networks for skin mole lesion classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 2623–2627. IEEE, 2016.

[87] Łukasz Rączkowski, Marcin Możejko, Joanna Zambonelli, and Ewa Szczurek. Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning. *Scientific reports*, 9(1):1–12, 2019.

[88] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451*, 2018.

[89] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond, 2019.

[90] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of Mathematical Statistics*, pages 400–407, 1951.

[91] Greg Roelofs and Richard Koman. *PNG: the definitive guide*. O'Reilly & Associates, Inc., 1999.

[92] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International conference on computer vision (ICCV)*, pages 618–626, 2017.

[93] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on Computer vision and Pattern recognition (CVPR) workshops*, pages 806–813, 2014.

[94] Vimal K Shrivastava, Narendra D Londhe, Rajendra S Sonawane, and Jasjit S Suri. Reliable and accurate psoriasis disease classification in dermatology images using comprehensive feature space in machine learning paradigm. *Expert Systems with Applications*, 42(15-16):6184–6195, 2015.

[95] Leslie N Smith. No more pesky learning rate guessing games. *arXiv preprint arXiv:1506.01186*, 2015.

[96] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472, 2017.

[97] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[98] Robert S Stern. Prevalence of a history of skin cancer in 2007: results of an incidence-based model. *Archives of dermatology*, 146(3):279–282, 2010.

[99] Xiaoxiao Sun, Liyi Chen, and Jufeng Yang. Learning from web data using adversarial discriminative neural networks for fine-grained classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 273–280, 2019.

[100] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 206–222. Springer, 2016.

[101] Xin Sun, Hongwei Xv, Junyu Dong, Huiyu Zhou, Changrui Chen, and Qiong Li. Few-shot learning for domain-specific fine-grained image classification. *IEEE Transactions on Industrial Electronics*, 2020.

[102] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[103] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *ArXiv preprint arXiv:1312.6199*, 2013.

[104] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

[105] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[106] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5:180161, 2018.

[107] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–285. Springer, 2020.

[108] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 8769–8778, 2018.

[109] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

[110] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pages 210–218. Springer, 2018.

[111] Gregory K Wallace. The JPEG still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992.

[112] Yi Wan, Qiqiang Chen, and Yan Yang. Robust impulse noise variance estimation based on image histogram. *IEEE Signal Processing Letters*, 17(5):485–488, 2010.

[113] Wenyong Wang, Yongcheng Cui, Guangshun Li, Chuntao Jiang, and Song Deng. A self-attention-based destruction and construction learning fine-grained image classification method for retail product recognition. *Neural Computing and Applications*, 32(18):14613–14622, 2020.

[114] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2097–2106, 2017.

[115] Xiu-Shen Wei, Chen-Wei Xie, Jianxin Wu, and Chunhua Shen. Mask-cnn: Localizing parts and selecting descriptors for fine-grained bird species categorization. *Pattern Recognition*, 76:704–714, 2018.

[116] Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in neural information processing systems*, pages 4148–4158, 2017.

[117] Rey Reza Wiyatno, Anqi Xu, Ousmane Dia, and Archy de Berker. Adversarial examples in modern machine learning: A review, 2019.

[118] Lingxi Xie, Richang Hong, Bo Zhang, and Qi Tian. Image classification and retrieval are one. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 3–10, 2015.

[119] Lingxi Xie, Liang Zheng, Jingdong Wang, Alan L Yuille, and Qi Tian. Interactive: Inter-layer activeness propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 270–279, 2016.

[120] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1492–1500, 2017.

[121] Jufeng Yang, Xiaoxiao Sun, Jie Liang, and Paul L. Rosin. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[122] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks, 2017.

[123] Matthew D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[124] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013.

[125] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

[126] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based R-CNNs for fine-grained category detection. In *European conference on computer vision*, pages 834–849. Springer, 2014.

[127] Yabin Zhang, Hui Tang, and Kui Jia. Fine-grained visual categorization using meta-learning optimization with sample selection of auxiliary data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248, 2018.

[128] Bohan Zhuang, Lingqiao Liu, Yao Li, Chunhua Shen, and Ian Reid. Attend in groups: a weakly-supervised deep learning framework for learning from web data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1878–1887, 2017.

[129] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on Computer vision and Pattern recognition*, pages 8697–8710, 2018.