

Approximate Fisher Information Matrix to Characterize the Training of Deep Neural Networks

Zhibin Liao¹, Tom Drummond, Ian Reid, and Gustavo Carneiro¹

Abstract—In this paper, we introduce a novel methodology for characterizing the performance of deep learning networks (ResNets and DenseNet) with respect to training convergence and generalization as a function of mini-batch size and learning rate for image classification. This methodology is based on novel measurements derived from the eigenvalues of the approximate Fisher information matrix, which can be efficiently computed even for high capacity deep models. Our proposed measurements can help practitioners to monitor and control the training process (by actively tuning the mini-batch size and learning rate) to allow for good training convergence and generalization. Furthermore, the proposed measurements also allow us to show that it is possible to optimize the training process with a new dynamic sampling training approach that continuously and automatically change the mini-batch size and learning rate during the training process. Finally, we show that the proposed dynamic sampling training approach has a faster training time and a competitive classification accuracy compared to the current state of the art.

Index Terms—Machine learning, deep learning, neural networks, stochastic gradient descent, Fisher information matrix, neural network training characterisation

1 INTRODUCTION

DEEP learning networks (a.k.a. DeepNets), especially the recently proposed deep residual networks (ResNets) [1], [2] and densely connected networks (DenseNets) [3], are achieving extremely accurate classification performance over a broad range of tasks. Large capacity deep learning models are generally trained with stochastic gradient descent (SGD) methods [4], or any of its variants, given that they produce good convergence and generalization at a relatively low computational cost, in terms of training time and memory usage. However, a successful SGD training of DeepNets depends on a careful selection of mini-batch size and learning rate, but there are currently no reliable guidelines on how to select these hyper-parameters.

Recently, Keskar et al. [5] proposed numerical experiments to show that large mini-batch size methods converge to sharp minimisers of the objective function, leading to poor generalisation, and small mini-batch size approaches converge to flat minimisers. In particular, Keskar et al. [5] proposed a new sensitivity measurement based on an exploration approach that calculates the largest value of the objective function within a small neighbourhood. Even though very relevant to our work, that paper [5] focuses only on mini-batch size and does not elaborate on the dynamic sampling training method, i.e., only shows the

rough idea of a training algorithm that starts with a small mini-batch and then suddenly switches to a large mini-batch. Other recent works characterise the loss function in terms of their local minima [6], [7], [8], which is interesting but does not provide a helpful guideline for characterizing the training procedure.

In this paper, we introduce a novel methodology for characterising the SGD training of DeepNets [1], [3] with respect to mini-batch sizes and learning rate for image classification. These experiments are based on the efficient computation of the eigenvalues of the approximate Fisher information matrix (hereafter, referred to as Fisher matrix) [10], [11]. In general, the eigenvalues of the Fisher matrix can be efficiently computed (in terms of memory and run-time complexities), and they are usually assumed to approximate of the Hessian spectrum [10], [11], [12], [13], which in turn can be used to estimate the objective function shape. In particular, Jastrzkebski et al. [13] show that the Fisher matrix (referred to as the sample covariance matrix in [13]) approximates well the Hessian matrix when the model is realisable—that is, when the model's and the training data's conditional probability distributions coincide. In theory, this happens when the parameter is close to the optimum. In a deep learning context, this means that the Fisher matrix can be a reasonable approximation of the Hessian matrix at the end of the training (assuming sufficient training has been done), but there is no clear functional approximation to guarantee such approximation through the entire training. Nevertheless, in this work we show empirical evidence that the properties of the Fisher matrix can be useful to characterizing the SGD training of DeepNets.

The proposed characterisation of SGD training is based on spectral information derived from the Fisher matrix: 1)

• Z. Liao, I. Reid, and G. Carneiro are with the Australian Centre for Robotic Vision, University of Adelaide, Adelaide, SA 5005, Australia.
E-mail: {zhibin.liao, ian.reid, gustavo.carneiro}@adelaide.edu.au.

• T. Drummond is with the Australian Centre for Robotic Vision, Monash University, Clayton, VIC 3800, Australia. E-mail: tom.drummond@monash.edu.

Manuscript received 12 Oct. 2017; revised 21 Sept. 2018; accepted 12 Oct. 2018. Date of publication 16 Oct. 2018; date of current version 3 Dec. 2019.
(Corresponding author: Zhibin Liao.)

Digital Object Identifier no. 10.1109/TPAMI.2018.2876413

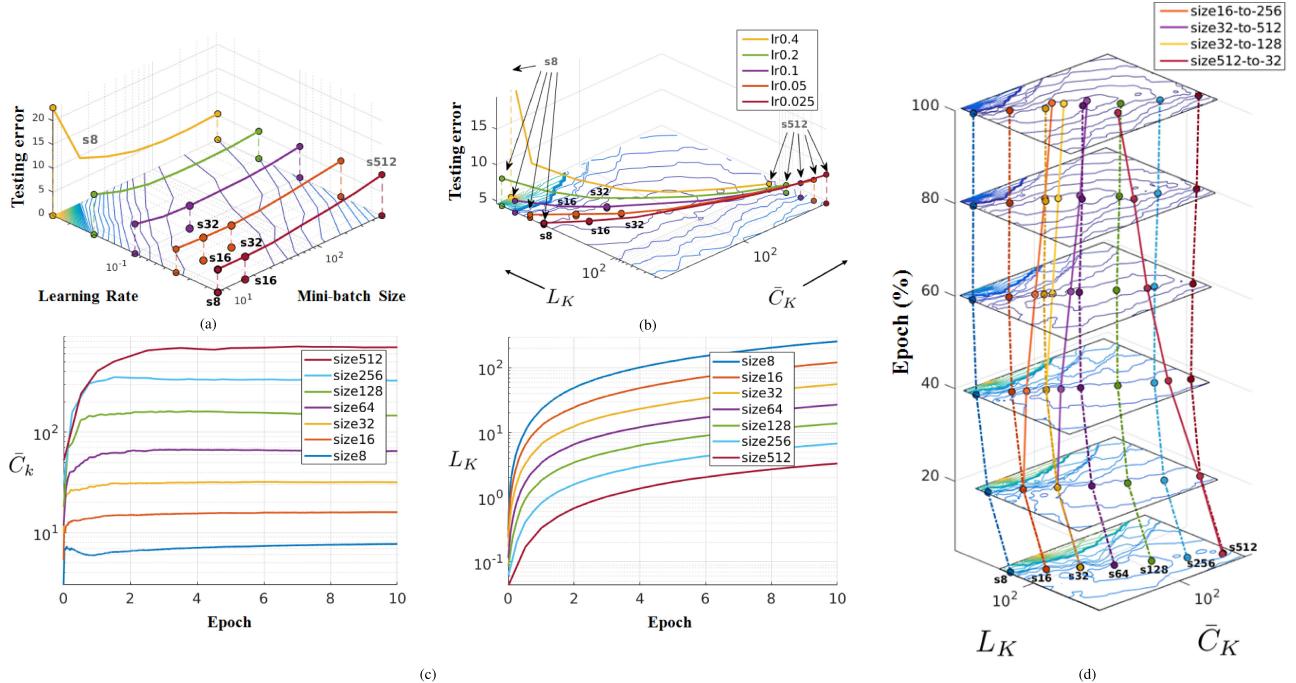


Fig. 1. The evaluation of using different learning rates and mini-batch sizes (a) and corresponding \bar{C}_K and L_K values (b) at the last epoch for the testing set of CIFAR-10 [9]. For (a) and (b), the intermediate points (i.e., mini-batch sizes) on each lr line are ignored for clarity, except for the top five configurations that produce lowest testing errors. The stability of the proposed measures over the first 10 epochs is shown in (c), illustrated by a subset of the models with $lr=0.1$. Finally, in (d), the subset of the models in (c) has been used as “beacons” to guide the dynamic sampling training (they share the same colors), i.e., we tune the runtime mini-batch size during the training in order to push the \bar{C}_K and L_K values close to the optimum region in order to achieve accurate classification and fast training – in this example, $s32\text{-to-}128$ is located closer to the optimum region center than the other dynamic sampling models: $s16\text{-to-}256$, $s32\text{-to-}512$, and $s512\text{-to-}32$ - also $s32\text{-to-}128$ shows the lowest testing error amongst these four models (hereafter, we use the following notation to represent the model hyper-parameter values: $s\{\text{mini - batchsize}\}-lr\{\text{learningratevalue}\}$).

the running average of the condition number of the Fisher matrix \bar{C}_K (see (7) for the definition); and 2) the weighted cumulative sum of the energy of the Fisher matrix L_K (see (8) for the definition). We observe that \bar{C}_K and L_K enable an empirically consistent characterisation of various models trained with different mini-batch sizes and learning rate. The motivation of our work is that current hyper-parameter selection procedures rely on validation performance, where the reason why some values are optimal are not well studied, making hyper-parameter selection (particularly on training DeepNets with large-scale datasets) a subjective task that heavily depends on the developer’s experience and the “factory setting” of the training scripts posted on public repositories. In Fig. 1a, we show an example of hyper-parameter selection with respect to different mini-batch sizes (from 8 to 512) and different learning rates (from 0.025 to 0.4, as shown by the lines marked by different colours) for the testing set¹ of CIFAR-10 [9], recorded from a trained ResNet model, where the five configurations with the lowest testing errors are highlighted. In comparison, in Fig. 1b we show the \bar{C}_K and L_K values of the configurations above computed at the final training epoch, showing that the optimal configurations are clustered together in the measurement space and form an optimum region at the center of the \bar{C}_K versus L_K graph (we define this optimum region to be formed by the level set with minimum error value in the contour plot).

¹ This paper does not pursue the best result in the field, so all models are trained with identical training setup, and we do not try to do model selection using the validation set.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 18, 2025 at 18:19:52 UTC from IEEE Xplore. Restrictions apply.

In Fig. 1c, we show that the proposed measures are stable in terms of the relative positions of \bar{C}_K and L_K values even during early training epochs, which means that they can be used to predict the performance of new configurations within a few epochs. From Fig. 1c, we can see that regions of performance are formed based on the mini-batch size used in the training process, where relatively small mini-batches tend to produce more effective training process, at the expense of longer training times. A natural question that can be made in this context is the following: is it possible to reduce the training time with the use of mini-batches of several sizes, and at the same time achieve the classification accuracy of training processes that rely exclusively on small mini-batches? Fig. 1d shows that the answer to this question is positive, where the proposed \bar{C}_K and L_K values can be used to guide dynamic sampling—a method that dynamically increases the mini-batch size during the training, by navigating the training procedure in the landscape of \bar{C}_K and L_K . The dynamic sampling approach has been suggested before [5], [14], but we are unaware of previous implementations. Our approach has a faster training time and a competitive accuracy result compared to the current state of the art on CIFAR-10, CIFAR-100 [9], SVHN [15], and MNIST [16] using recently proposed ResNet and DenseNet models.

2 LITERATURE REVIEW

In this section, we first discuss stochastic gradient descent (SGD) [4], inexact Newton and quasi-Newton methods [14], [17], [18], as well as (generalized) Gauss-Newton

methods [19], [20] the natural gradient method [21], and scaled gradient iterations such as RMSprop [22] and AdaGrad [23]. Then we discuss other approaches that rely on numerical experiments to measure key aspects of SGD training [5], [6], [7], [8], [24].

SGD training [4] is a common iterative optimization method that is widely used in deep neural networks training. One of the main goals of SGD is to find a good balance between stochastic and batch approaches to provide a favourable trade-off with respect to per-iteration costs and expected per-iteration improvement in optimising an objective function. The popularity of SGD in deep learning lies in the tolerable computation cost with acceptable convergence speed. Second-order methods aim to improve the convergence speed of SGD by re-scaling the gradient vector in order to compensate for the high non-linearity and ill-conditioning of the objective function. In particular, Newton's method uses the inverse of the Hessian matrix for re-scaling the gradient vector. This operation has complexity $O(N^3)$ (where N is the number of model parameters, which is usually between $O(10^6)$ and $O(10^7)$ for modern deep learning models), which makes it infeasible. Furthermore, the Hessian must be positive definite for Newton's method to work, which is not a reasonable assumption for the training of deep learning models.

In order to avoid the computational cost above, several approximate second-order methods have been developed. For example, the Hessian-free conjugate gradient (CG) [25] is based on the fact it only needs to compute Hessian-vector products, which can be efficiently calculated with the \mathcal{R} -operator [26] at a comparable cost to a gradient evaluation. This Hessian-free method has been successfully applied to train neural networks [27], [28]. Quasi-Newton methods (e.g., the BFGS [14], [17]) take an alternative route and approximate the inversion of Hessian with only the parameter and gradient displacements in the past gradient iterations. However, the explicit use of the approximation matrix is also infeasible in large optimization problems, where the L-BFGS [18] method is proposed to reduce the memory usage. The (Generalized) Gauss-Newton method [19], [20] approximates Hessian with the Gauss-Newton matrix. Another approximate second-order method is the natural gradient method [21] that uses the inverse of the Fisher matrix to make the search quicker in the parameters that have less effect on the decision function [14]. Without estimating the second-order curvature, some methods can avoid saddle points and perhaps have some degree of resistance to near-singular curvature [14]. For instance, AdaGrad [23] keeps an accumulation of the square of the gradients of past iterations to re-scale each element of the gradient, so that parameters that have been infrequently updated are allowed to have large updates, and frequently updated parameters can only have small updates. Similarly, RMSProp [22] normalises the gradient by the magnitude of recent gradients. Furthermore, Adadelta [29] and Adam [30] improve over AdaGrad [23] by taking more careful gradient re-scaling schemes.

Given the issues involved in the development of (approximate) second-order methods, there has been some interest in the implementation of approaches that could characterize the functionality of SGD optimization. Lee et al. [8] show that SGD converges to a local minimiser rather than a

saddle point (with models that are randomly initialised). Soudry and Carmon [7] provide theoretical guarantees that local minima in multilayer neural networks loss functions have zero training error. In addition, the exact Hessian of the neural network has been found to be singular, suggesting that methods that assume non-singular Hessian are not to be used without proper modification [24]. Goodfellow et al. [31] found that state-of-the-art neural networks do not encounter significant obstacles (local minima, saddle points, etc.) during the training. In [5], a new sensitivity measurement of energy landscape is used to provide empirical evidence to support the argument that training with large mini-batch size converges to sharp minima, which in turn leads to poor generalization. In contrast, small mini-batch size converges to flat minima, where the two minima are separated by a barrier, but performance degenerates due to noise in the gradient estimation. Sagun et al. [12] trained a network using large mini batches first, followed by the use of smaller mini batches, and their results show that such barrier between these two minima does not exist, so the sharp and flat minima reached by the large and small mini batches may actually be connected by a flat region to form a larger basin. Jastrzkebski et al. [13] found out that the ratio of learning rate to batch size plays an important role in SGD dynamics, and large values of this ratio lead to flat minima and (often) better generalization. In [32], Smith and Le interpret SGD as the discretisation of a stochastic differential equation and predict that an optimum mini-batch size exists for maximizing test accuracy, which scales linearly with both the learning rate and the training set size. Smith et al. [33] demonstrate that decaying learning rate schedules can be directly converted into increasing batch size schedules, and vice versa, enabling training towards large mini-batch size. Finally in [34], Goyal et al. manage to train in one hour a ResNet [1] on ImageNet [35] using mini-batches of size 8K – this model achieved a competitive result compared to another ResNet trained using mini-batches of size 256.

The dynamic sampling of mini-batch size has been explored in machine learning, where the main focus lies in tuning the mini-batch size in order to improve convergence. Friedlander and Schmidt [36] show that an increasing sampling size can maintain the steady convergence rates of batch gradient descent (or steepest decent) methods, and the authors also prove linear convergence of such method w.r.t. the number of iterations. In [37], Byrd et al. present a gradient-based dynamic sampling strategy, which heuristically increases the mini-batch size to ensure sufficiently progress towards the objective value descending. The selection of the mini-batch size depends on the satisfaction of a condition known as the *norm test*, which monitors the norm of the sample variance within the mini-batch. Similarly, Bollapragada et al. [38] propose an approximate inner product test, which ensures that search directions are descent directions with high probability and improves over the norm test. Furthermore, Metel [39] presents dynamic sampling rules to ensure that the gradient follows a descent direction with higher probability—this depends on a dynamic sampling of mini-batch size that reduces the estimated sample covariance. De et al. [40] empirically evaluate the dynamic sampling method and observe that it can outperform classic SGD when

the learning rate is monotonic, but it is comparable when SGD has fine-tuned learning rate decay.

Our paper can be regarded as a new approach to characterise SGD optimization, where our *main contributions* are: 1) *new efficiently computed measures derived from the Fisher matrix* that can be used to *explain the training convergence and generalization of DeepNets with respect to mini-batch sizes and learning rates*, and 2) *a new dynamic sampling algorithm that has a faster training process and competitive classification accuracy* compared to recently proposed deep learning models.

3 METHODOLOGY

In this section, we assume the availability of a dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{|\mathcal{D}|}$, where the i th image $\mathbf{x}_i : \Omega \rightarrow \mathbb{R}$ (Ω denotes image lattice) is annotated with the label $y_i \in \{1, \dots, C\}$, with C denoting the number of classes. This dataset is divided into the following mutually exclusive sets: training $\mathcal{T} \in \mathcal{D}$ and testing $\mathcal{S} \in \mathcal{D}$.

The ResNet model [1] is defined by a concatenation of residual blocks, with each block defined by

$$r_l(\mathbf{v}_l) = f_r(\mathbf{v}_l, \mathbf{W}_l) + \mathbf{v}_l, \quad (1)$$

where $l \in \{1, \dots, L\}$ indexes the residual blocks, \mathbf{W}_l denotes the parameters for the l th block, \mathbf{v}_l is the input, with the image input of the model being represented by $\mathbf{v}_1 = \mathbf{x}$, $f_r(\mathbf{v}_l, \mathbf{W}_l)$ represents a residual unit containing a sequence of linear and non-linear transforms [41], and batch normalisation [42]. Similarly, the DenseNet model [3] is defined by a concatenation of dense layers, with each layer defined by

$$d_l(\mathbf{v}_l) = f_d([\mathbf{v}_1, \dots, \mathbf{v}_l], \mathbf{W}_l), \quad (2)$$

where $[.]$ represents the concatenation operator, $f_d([\dots], \mathbf{W}_l)$ contains a sequence of transformations and normalisations similar to f_r of (1).

The full model is defined by

$$f(\mathbf{x}, \theta) = f_{out} \circ b_L \circ \dots \circ b_1(\mathbf{x}), \quad (3)$$

where \circ represents the composition operator, $b \in \{r, d\}$ represents the choice of computation block, $\theta \in \mathbb{R}^P$ denotes all model parameters $\{\mathbf{W}_1, \dots, \mathbf{W}_L\} \cup \mathbf{W}_{out}$, and $f_{out}(\cdot)$ is a linear transform parameterised by weights \mathbf{W}_{out} with a softmax activation function that outputs a value in $[0, 1]^C$ indicating the confidence of selecting each of the C classes. The training of the model in (3) minimizes the multi-class cross entropy loss $\ell(\cdot)$ on the training set \mathcal{T} , as follows:

$$\theta^* = \arg \min_{\theta} \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \ell(y_i, f(\mathbf{x}_i, \theta)). \quad (4)$$

The SGD training minimizes the loss in (4) by iteratively taking the following step:

$$\theta_{k+1} = \theta_k - \frac{\alpha_k}{|\mathcal{B}_k|} \sum_{i \in \mathcal{B}_k} \nabla \ell(y_i, f(\mathbf{x}_i, \theta_k)), \quad (5)$$

where \mathcal{B}_k is the mini-batch for the k th iteration of the minimisation process. As noted by Keskar et al. [5], the shape of the loss function can be characterized by the spectrum of the $\nabla^2 \ell(y_i, f(\mathbf{x}_i, \theta_k))$, where sharpness is defined by the magnitude of the eigenvalues. However, the loss function sharpness alone is not enough to characterize SGD training

because it is possible, for instance, to adjust the learning rate in order to compensate for possible generalization issues of the training process [13], [33], [34]. In this paper, we combine information derived not only from the spectrum of $\nabla^2 \ell(y_i, f(\mathbf{x}_i, \theta_k))$, but also from the learning rate to characterize SGD training. Given that the computation of the spectrum of $\nabla^2 \ell(y_i, f(\mathbf{x}_i, \theta_k))$ is infeasible, we approximate the Hessian by the Fisher matrix (assuming the condition explained in Section 1) [10], [11], [13]—the Fisher matrix is defined by

$$\mathbf{F}_k = \left(\nabla \ell(y_{i \in \mathcal{B}_k}, f(\mathbf{x}_{i \in \mathcal{B}_k}, \theta_k)) \nabla \ell(y_{i \in \mathcal{B}_k}, f(\mathbf{x}_{i \in \mathcal{B}_k}, \theta_k))^T \right), \quad (6)$$

where $\mathbf{F}_k \in \mathbb{R}^{P \times P}$.

The calculation of \mathbf{F}_k in (6) depends on the Jacobian $\mathbf{J}_k = \nabla \ell(y_{i \in \mathcal{B}_k}, f(\mathbf{x}_{i \in \mathcal{B}_k}, \theta_k))$, with $\mathbf{J}_k \in \mathbb{R}^{P \times |\mathcal{B}_k|}$. Given that $\mathbf{F}_k = \mathbf{J}_k \mathbf{J}_k^T \in \mathbb{R}^{P \times P}$ scales with $P \in [O(10^6), O(10^7)]$ and that we are only interested in the spectrum of \mathbf{F}_k , we can compute instead $\tilde{\mathbf{F}}_k = \mathbf{J}_k^T \mathbf{J}_k \in \mathbb{R}^{|\mathcal{B}_k| \times |\mathcal{B}_k|}$ that scales with the mini-batch size $|\mathcal{B}_k| \in [O(10^1), O(10^2)]$. Note that the rank of $\tilde{\mathbf{F}}_k$ and \mathbf{F}_k is at most $|\mathcal{B}_k|$, which means that the spectra of $\tilde{\mathbf{F}}_k$ and \mathbf{F}_k are the same given that both will have at most $|\mathcal{B}_k|$ non-zero eigenvalues.

The *first measure* proposed in this paper is the *running average of the truncated condition number* of $\tilde{\mathbf{F}}_k$, defined by

$$\bar{C}_K = \frac{1}{K} \sum_{k=1}^K c_k, \quad (7)$$

where K denotes the epoch number, and $c_k = \frac{\sigma_{\max}(\mathcal{E}_k)}{\sigma_{\min}(\mathcal{E}_k)}$ represents the ratio between the largest to the smallest non-zero singular value of \mathbf{J}_k (i.e., we refer to this ratio as the truncated condition number), with \mathcal{E}_k denoting the set of non-zero eigenvalues computed from $\tilde{\mathbf{F}}_k$ [43], $\sigma_{\max}(\mathcal{E}_k) = \max(\mathcal{E}_k)^{\frac{1}{2}}$, and $\sigma_{\min}(\mathcal{E}_k) = \min(\mathcal{E}_k)^{\frac{1}{2}}$. This measure is used to describe the empirical truncated conditioning of the gradient updates observed during the training process. In Figs. 2a and 2c, we show that c_k is a noisy measure and unfit for characterizing the training, but \bar{C}_K is more stable, which means that it is able to rank the training procedures more reliably.

The *second measure* is the *weighted cumulative sum of the energy of the Fisher matrix* $\tilde{\mathbf{F}}_k$, computed by

$$L_K = \sum_{k=1}^K l_k, \quad (8)$$

where $l_k = \frac{\alpha_k}{|\mathcal{B}_k|} \left(\text{Tr}(\tilde{\mathbf{F}}_k) \right)^{\frac{1}{2}}$, $\text{Tr}(\cdot)$ represents the trace operator, $\text{Tr}(\tilde{\mathbf{F}}_k)$ approximates the Laplacian, defined by $\text{Tr}(\nabla^2 \ell(y_i, f(\mathbf{x}_i, \theta_k)))$, which measures the energy of the approximate Fisher matrix by summing its eigenvalues, and the factor $\frac{\alpha_k}{|\mathcal{B}_k|}$ (with $|\mathcal{B}_k|$ denoting mini-batch size and α_k representing learning rate) is derived from the SGD learning in (5) – this factor in (8) represents a design choice that provides the actual contribution of the energy of the approximate Fisher matrix at the k th epoch. Note that in (8), we apply the square root operator in order to have the magnitude of the values of L_K similar to \bar{C}_K in (7).

3.1 Model Selection

We observe that deep models trained with different learning rates and mini-batch sizes have values for \bar{C}_K and L_K that are

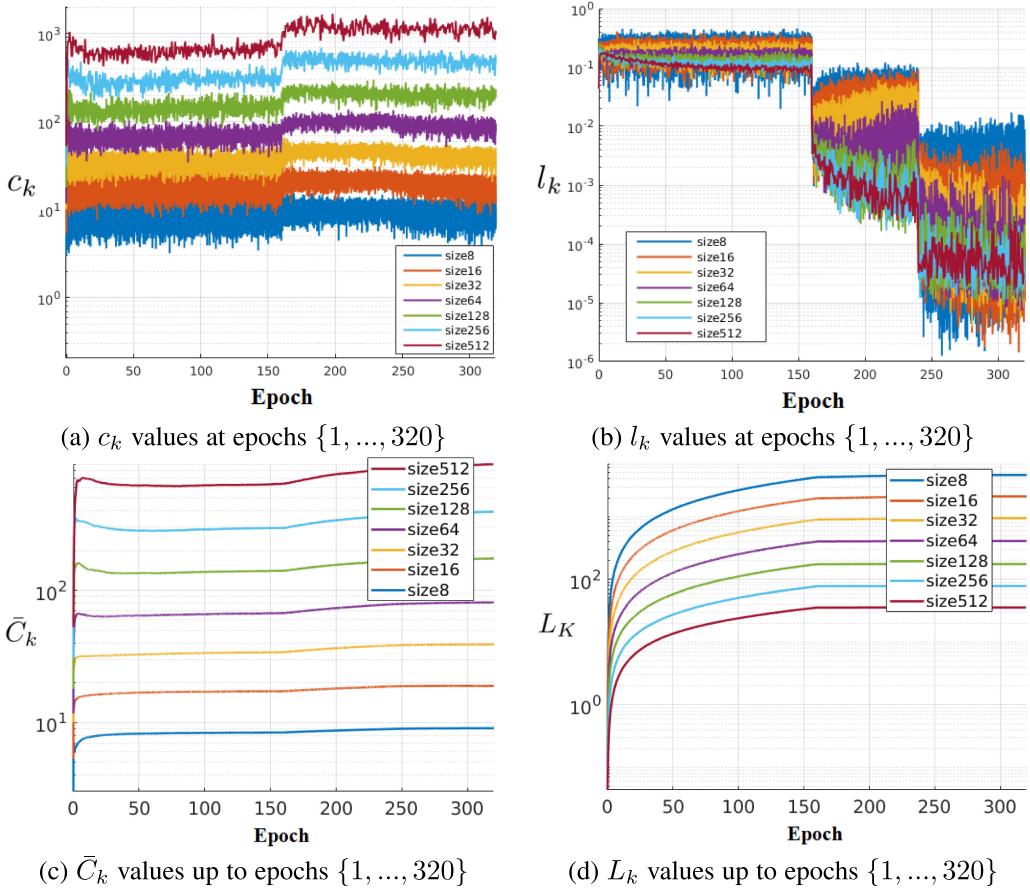


Fig. 2. An illustration of $\{c, l\}_k$ (the first row) and $\{C, L\}_k$ (the second row) of tested ResNet models. These models are trained on CIFAR-10 with the configurations of mini-batch sizes $|\mathcal{B}_k| = \{8, \dots, 512\}$ and initial learning rate $\alpha_1 = 0.1$, α_k is reduced 10 fold at $k \in 161$ st epoch and $k \in 241$ st epochs. In general, the sampled measures $\{c, l\}_k$ are noisy, leading to our proposal of the cumulative measures $\{C, L\}_k$ to assess the entire training procedure. The x -axis is shown by epoch instead of iteration to align the readings of different training configurations because the use of small mini-batch sizes increases the number of iterations to complete an epoch.

stable, as displayed in Fig. 1c (showing first training epochs) and in Figs. 2c and 2d (all training epochs). This means that models and training procedures can be reliably characterized early in the training process, which can significantly speed up the assessment of new models with respect to their mini-batch size and learning rate. For instance, if a reference model produces a good result, and we know its \bar{C}_K and L_K values for various epochs, then new models must navigate close to this reference model – see for example in Fig. 1b that s32-lr0.1 produces good convergence and generalization with test error $4.78\% \pm 0.05\%$, so new models must try to navigate close enough to it by adjusting the mini-batch size and learning rate. Indeed, we can see that the two nearest configurations are s16-0.025 and s32-lr0.05, with test error $4.76\% \pm 0.11\%$ and $4.67\% \pm 0.22\%$ respectively.

3.2 Dynamic Sampling

Dynamic sampling [36], [37] is a method that is believed to improve the convergence rate of SGD by reducing the noise of the gradient estimation with a gradual increase of the mini-batch size over the training process (this method has been suggested before [36], [37], but we are not aware of previous implementations). It extends SGD by replacing the fixed size mini-batches \mathcal{B}_k in (5) with a variable size mini-batch. The general idea of this method [36], [37] is that the initial noisy gradient estimations from small mini-batches explore

a relatively flat energy landscape without falling into sharp local minima. The increase of mini-batch sizes over the training procedure provides a more robust gradient estimation and, at the same time, drives the training into sharper local minima that appear to have better generalization properties.

With respect to our proposed measures \bar{C}_K, L_K in (7) and (8), we notice that dynamic sampling breaks the relative stability between curves of fixed mini-batch sizes, as displayed in Fig. 1d. In general, we note that the application of dynamic sampling allows the curves to move from the region of the original batch size to the region of the final batch size, which means that the training process can be adapted to provide a good trade-off between training speed and accuracy, taking into account that larger mini-batches tend to train faster. Therefore, we believe that the idea of starting with small and continuously increasing the mini-batches [36], [37] is just partially true because our results provides evidence not only for such idea, but it also shows that it is possible to start with large mini-batches and continuously decrease them during the training in order to achieve good convergence and generalization.

4 EXPERIMENTS

The experiments are carried out on four commonly evaluated benchmark datasets: CIFAR-10 [9], CIFAR-100 [9], SVHN [15],

and MNIST [16]. CIFAR-10 and CIFAR-100 datasets contain 60,000 32×32 -pixel coloured images, where 50,000 are used for training and 10,000 for testing. SVHN and MNIST are digits recognition datasets where SVHN is a large-scale dataset with over 600,000 RGB street view house number plate images and MNIST has 70,000 grayscale hand-written digits.

We test our methodology using ResNet [1] and DenseNet [3]. More specifically, we rely on a 110-layer ResNet [1] for CIFAR-10/100 and SVHN datasets, including 54 residual units, formed by the following operators in order: 3×3 convolution, batch normalisation [42], ReLU [41], 3×3 convolution, and batch normalisation. This residual unit empirically shows better performance than previously proposed residual units (also observed in [44] in parallel to our own work). We use the simplest skip connection with no trainable parameters. We also test a DenseNet [3] with 110 layers, involving three stack of 3 dense blocks, where each block contains 18 dense layers. By tuning the DenseNet growth rate (i.e., 8, 10, 14) of each dense block, we manage to composite this DenseNet to have 1.77 million parameters, versus to the 1.73 million in the ResNet. Due to the simplicity of the MNIST dataset, we use an 8-layer ResNet and an 8-layer DenseNet, including 3 residual units and 6 dense layers (also grouped into 3 dense blocks). For SGD, we use 0.9 for momentum, and the learning rate decay is performed in multiple steps: the initial learning rate is subject to the individual experiment setup, but followed by the same decay policy that decays by 1/10 at the 50 percent training epochs, and by another 1/10 at 75 percent epochs. That is 161st and 241st epoch on CIFAR-10/100, and 21st and 31st epoch on SVHN and MNIST, where the respective training duration are 320 and 40 epochs. All training uses data augmentation, as described by He et al. [1]. The scripts for replicating the experiment results are publicly available.²

For each experiment, we measure the training and testing classification error, and the proposed measures \bar{C}_K (7) and L_K (8) – the reported results are actually the mean result obtained from five independently trained models (each model is randomly initialised). All experiments are conducted on an NVidia Titan-X and K40 gpus without the multi-gpu computation. In order to obtain $\tilde{\mathbf{F}}_k$ in a efficient manner, the explicit calculation of \mathbf{J}_k is obtained with a modification of the Torch [45] NN and cuDNN libraries (convolution, batch normalisation and fully-connected modules) to acquire the Jacobian $\mathbf{J}_k = \nabla \ell(y_t, f(\mathbf{x}_t, \theta_k))$ during back-propagation. By default, the torch library calls NVidia cuDNN library in backward training phase to compute the gradient w.r.t the model parameters, where the cuDNN library does not explicitly retain \mathbf{J}_k . For each type of the aforementioned torch modules, our modification breaks the one batch-wise gradient computation call to $|\mathcal{B}_k|$ individual calls, one for each sample, and then collects the per-sample gradients to form \mathbf{J}_k . Note that the memory complexity to store \mathbf{J}_k scales by $|\mathcal{B}_k|$ times the number of model parameters, which is acceptable for the 1.7 million parameters in the deep models. Note that \mathbf{J}_k is formed by iterating over the training samples (in the torch programming layer), which is a slow process given that the underlying NVidia cuDNN library is not open-sourced to be modified to compute the Jacobian within the cuda

programming layer directly. We handle this inefficiency by computing $\tilde{\mathbf{F}}_k$ at intervals of 50 mini-batches, resulting in a sampling rate of $\approx 2\%$ of training set, so the additional time cost to form \mathbf{J}_k is negligible during the training. The memory required to store the full \mathbf{J}_k can be reduced by computing $\tilde{\mathbf{F}}_k = \sum_l \mathbf{J}_{(k,l)}^T \mathbf{J}_{(k,l)}$ for any layer l with trainable model parameters, where $\mathbf{J}_{(k,l)}$ presents the rows of \mathbf{J}_k with respect to the parameters of layer l . This leaves the memory footprint to be only $O(|\mathcal{B}_k|^2)$ for $\tilde{\mathbf{F}}_k$.

The training and testing values of the trained models used to plot the figures in this section are listed in the supplementary material, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TPAMI.2018.2876413>. At last, due to the higher memory usage of the DenseNet model (compared to ResNet with same configuration), mini-batch size 512 cannot be trained with the 110-layer DenseNet on a single GPU, therefore is excluded from the experiment.

4.1 Mini-Batch Size and Learning Rate

In Fig. 3, we show our first experiment comparing different mini-batch sizes and learning rates with respect to the training and testing errors, and the proposed measures \bar{C}_K (7) and L_K (8). The grid values of each 2-D contour map is generated by averaging five nearest error values. In this section, we refer the error of a model as the final error value obtained when the training procedure is complete. In general, the main observations for all datasets and models are: 1) each configuration has a unique \bar{C}_K and L_K signature, where no configuration overlays over each other in the space; 2) $|\mathcal{B}_k|$ is directly proportional to \bar{C}_K and inversely proportional to L_K ; 3) α_k is directly proportional to \bar{C}_K and L_K ; and 4) small \bar{C}_K and large L_K indicate poor training convergence, and large \bar{C}_K and small L_K show poor generalization, so the best convergence and generalization requires a small value for both measures. Recently, Jastrzkebski et al. [13] claimed that large $\alpha_k/|\mathcal{B}_k|$ ratio exhibits better generalization property in general. Our results show that this is true up to a certain value for this ratio. In particular, we do observe that the models that produce the top five test accuracy have similar $\alpha_k/|\mathcal{B}_k|$ ratio values (this is clearly shown in the supplementary material, available online), but for very large ratios, when $|\mathcal{B}_k| \in \{8, 16\}$ and $\alpha_k = 0.4$, then we noticed that convergence issues start to appear. This is true because beyond a certain increase in the value of L_K , $\tilde{\mathbf{F}}_k$ becomes rank deficient, so that in some epochs (mostly in the initial training period), the smallest eigenvalues get too close to zero, causing some of the c_k values to be large, increasing the value of \bar{C}_K and making the model ill-conditioned.

For both models on CIFAR-10, we mark the top five configurations in Figs. 3a and 3d with lowest testing errors, where the best ResNet model is configured by *s32-lr0.05* with $4.67\% \pm 0.22\%$, and the best DenseNet is denoted by *s16-lr0.025* with $4.82\% \pm 0.07\%$. This shows that on CIFAR-10, the optimal configurations are with small $|\mathcal{B}_k| \in \{8, \dots, 32\}$ and small $\alpha_k \in \{0.025, \dots, 0.1\}$. On CIFAR-100, both models show similar results, where the best ResNet model has configuration *s16-lr0.05* with error $23.39\% \pm 0.13\%$, and the best DenseNet is configured as *s8-lr0.025* with error $22.90\% \pm 0.47\%$. Note that on CIFAR-100, the optimal configurations are with the same small range of $|\mathcal{B}_k|$ and small α_k .

2. <https://github.com/zhibiniao89/fisher.info.mat.torch>

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on January 18, 2025 at 18:19:52 UTC from IEEE Xplore. Restrictions apply.

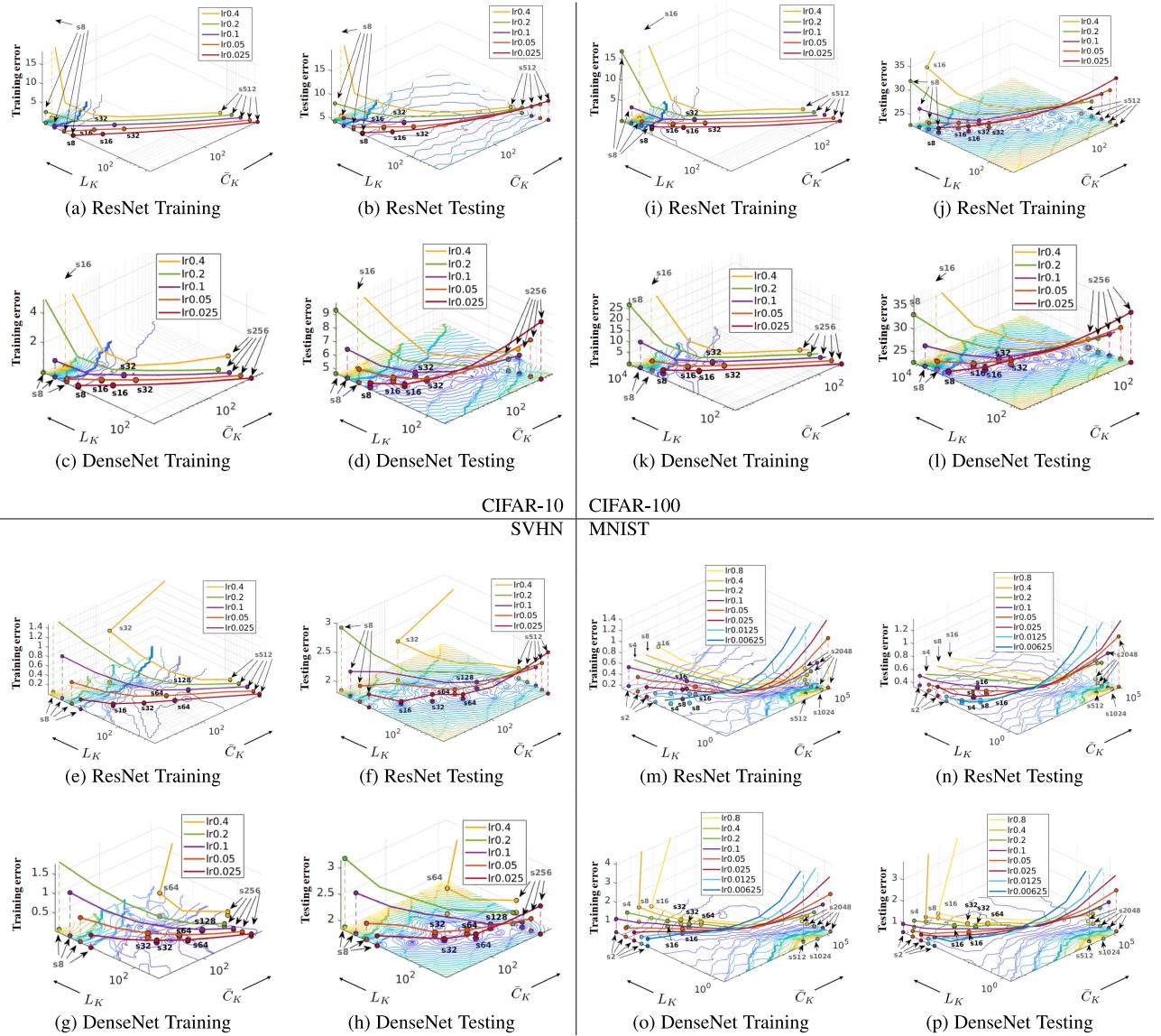


Fig. 3. This graph shows how mini-batch sizes $|B_k|$ and initial learning rate α_k affect the training performance, which is also related to the proposed measures \bar{C}_K and L_K , on four common benchmarks and two model architectures. We connect the models that use the same α_k value. Each line omits the intermediate points for clarity except the top five points over all configurations with lowest testing errors. The gray texts, i.e., $s[8, 2048]$, indicate the $|B_k|$ value at each end of the α_k -connected lines.

This similarity of the best configurations between CIFAR models is expected because of the similarity in the image data. We may also conclude that the range of optimal \bar{C}_K and L_K value is not related to the number of classes in the dataset.

Both models show similar results on SVHN, where the top ResNet result is reached with $s16-lr0.025$ that produced an error of $1.86\% \pm 0.03\%$, while the best DenseNet accuracy is achieved by $s32-lr0.025$ with error $1.89\% \pm 0.01\%$. Compared to CIFAR experiments, it is clear that the optimum region on SVHN is “thinner”, where the optimal configurations are with $|B_k| \in \{16, \dots, 128\}$ and $\alpha_k \in \{0.025, \dots, 0.1\}$, which appears to shift noticeably towards larger $|B_k|$ values. However, compared to the size of the dataset (i.e., SVHN is $10\times$ larger than CIFARS) such $|B_k|$ values are still relative small, so that the optimal ratio of $|B_k|$ with respect to the size of dataset is actually smaller than the ratio observed for the CIFAR experiments. Note that the errors on SVHN are final testing errors, where we found that the lowest

testing error of each individual model usually occurs between 22 and 25 epochs, and the remaining training gradually overfits the training set, making the final testing error worse by 0.2 percent, on average. However, we do not truncate the training in order to keep the consistency of training procedures. Finally, on MNIST we test a wider $|B_k| \in \{2, \dots, 2048\}$ and wider $\alpha_k \in \{0.00625, \dots, 0.8\}$. The best ResNet model is $s16-lr0.1$ with error $0.36\% \pm 0.02\%$, and DenseNet is $s32-lr0.4$ with error $0.54\% \pm 0.02\%$. The optimum region of MNIST on ResNet is with small $|B_k| \in \{4, \dots, 16\}$ and small $\alpha_k \in \{0.0125, \dots, 0.1\}$. On the other hand, the optimum region of MNIST on DenseNet is with slightly large $|B_k| \in \{16, \dots, 64\}$ and large $\alpha_k \in \{0.1, \dots, 0.4\}$, showing a divergence on the optimal region w.r.t the architecture. We emphasize this is due to a large difference in the model parameters, where the 8-layer MNIST ResNet model has 70,000 parameters while the 8-layer DenseNet model has 10,000 parameters. The main conclusion of

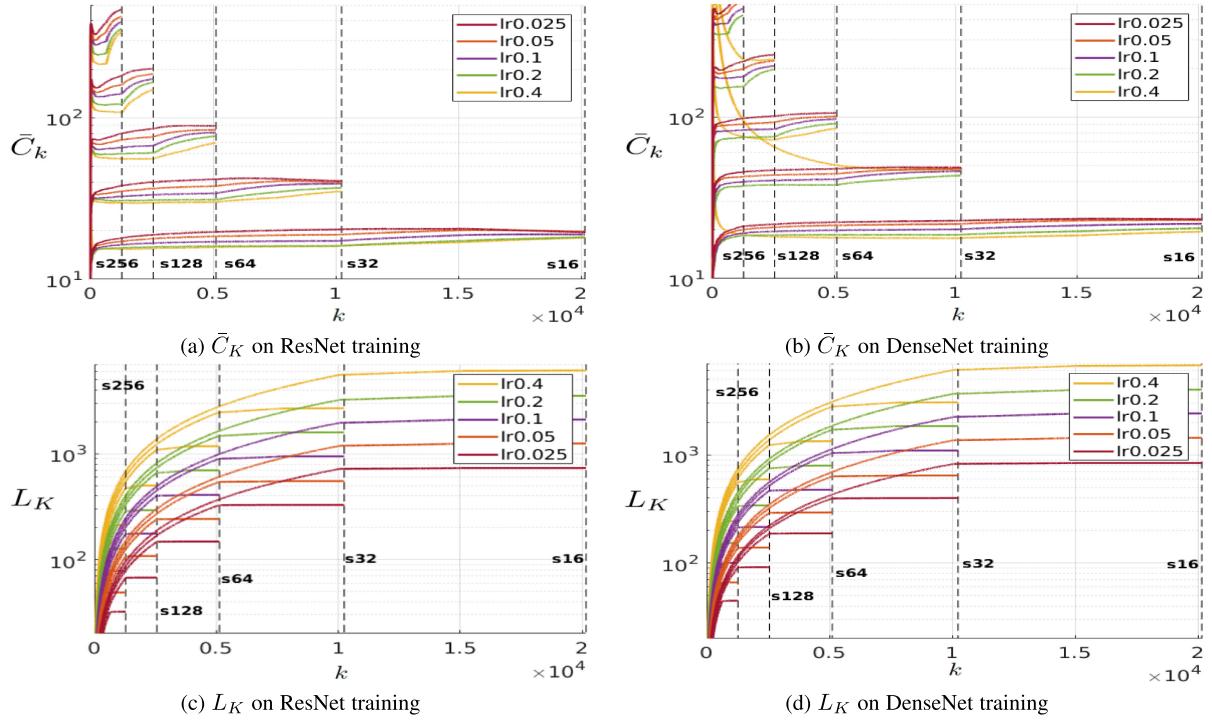


Fig. 4. The proposed measurements (a) and (b) \bar{C}_K and (c) and (d) L_K for the training of the (a) and (c) ResNet and (b) and (d) DenseNet on CIFAR-10, as a function of the number of training iterations (instead of epochs). The black dotted vertical lines indicate the last iterations for the respective experiments with the same batch size (the results of $s\{8, 512\}$ are excluded to avoid a cluttered presentation).

this experiment is that the effect of mini-batch size and initial learning rate are tightly related, and can be used to compensate one another to move the \bar{C}_K and L_K values to different places in the measurement space.

4.2 Functional Relations between Batch Size, Learning Rate, and the Proposed Measures

It is worth noticing the functional relations between hyperparameters and the proposed measures (due to space restrictions, we only show results for ResNet and DenseNet on CIFAR-10). In Figs. 4a and 4b, we observe that \bar{C}_K tends to cluster at similar values for training processes performed with the same mini-batch sizes, independently of the learning rate. On the other hand, in Figs. 4c and 4d, we notice that L_K is more likely to cluster at similar values for training processes performed with the same learning rate, independently of the mini-batch size, particularly at the first half of the training. The major difference between ResNet and DenseNet regarding these functional relations is shown in Fig. 4c, where the learning rate = 0.4 results in poor convergence during the first half of the training process.

4.3 Dynamic Sampling

In Fig. 5, we show the runtime analysis of different dynamic sampling alternatives, and how they affect the values of \bar{C}_K , L_K , as well as the classification errors. The dynamic sampling divides the training process into five stages, each with equal number of training epochs and using a particular mini-batch size, i.e., $s32\text{-to-}512$ uses the mini-batch size sequence $\{32, 64, 128, 256, 512\}$, $s512\text{-to-}32$ uses $\{512, 256, 128, 64, 32\}$, and $s16\text{-to-}64$ uses $\{16, 16, 32, 32, 64\}$. In each dynamic sampling experiment, the first number indicates the initial mini-batch size, and the second indicates the final mini-batch size, and $\text{-}\emptyset$ or -MS indicates

whether it uses a *multi-step dynamic sampling approach*. More specifically, dynamic sampling can be performed over the whole training procedure (indicated by the symbol $\text{-}\emptyset$), or within each particular value of learning rate, where the sampling of mini-batch sizes is done over each different learning rate value (denoted by the symbol -MS). All experiments below use an initial learning rate of 0.1.

Beacons. The “beacon” models are the $s\{16, \dots, 512\}\text{-}lr0.1$ models from Fig. 3. In general, the beacon models accumulate L_K faster during the first half of the training procedure than they do during the second half. On the other hand, the \bar{C}_K measure appears to be more stable during the first half of the training. However, during the second half of the training, we observe that \bar{C}_K grows on CIFARs (see Fig. 2) but decreases on SVHN and MNIST.

Dynamic Sampling. Note in Fig. 5 that the dynamic sampling training procedures tend to push \bar{C}_K and L_K away from the initial mini-batch size region towards the final mini-batch size region (with respect to the respective mini-batch size beacons). Such travel on L_K is faster in the first half of the training procedure than it is in the second half since the growth of L_K is subject to the learning rate (which decays at 50 and 75 percent of the training process). However, from (7), we know that \bar{C}_K is not affected by the learning rate, so it can travel farther towards the final mini-batch size beacon during the second half of training procedure. For instance, on CIFAR-10 experiment for ResNet, the $s32\text{-to-}512$ and $s512\text{-to-}32$ models use the same amount of mini-batch sizes during the training and have the same final \bar{C}_K value but different L_K values. In Fig. 5a ResNet panel, notice that $s32\text{-to-}512$ is close to the optimum region, showing a testing error of $5.07\% \pm 0.21\%$, but $s512\text{-to-}32$ is not as close, showing a testing error of $5.56\% \pm 0.09\%$.

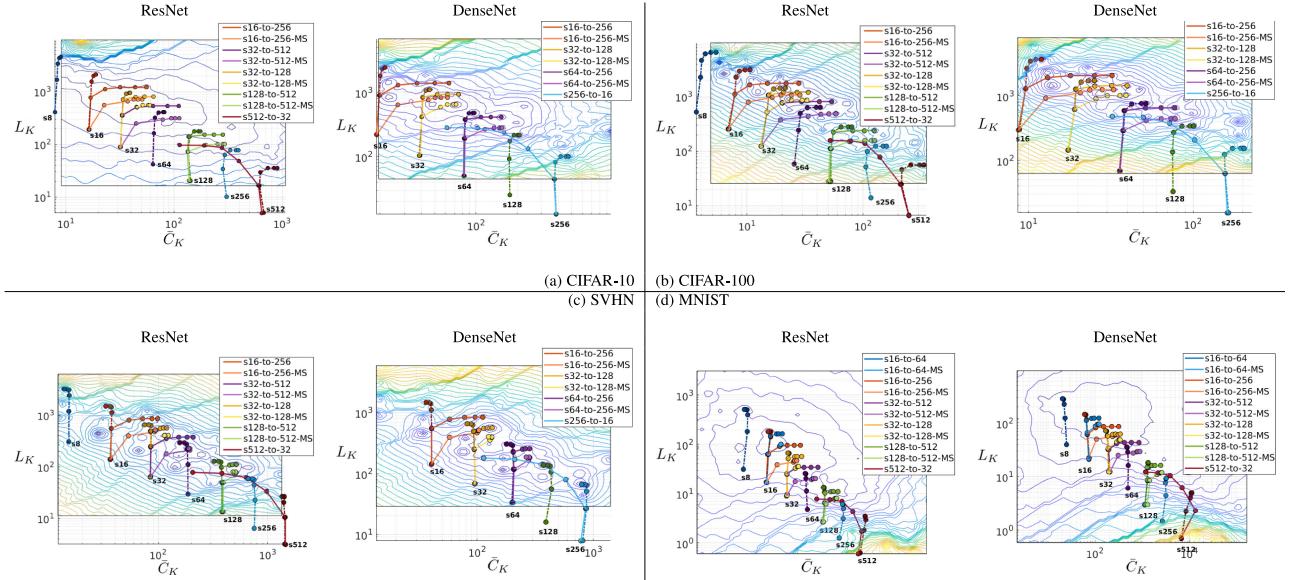


Fig. 5. This graph illustrates the “travelling history” of \bar{C}_K and L_K of several dynamic sampling models. Each model is represented by a curve, where the plotted \bar{C}_K and L_K values are extracted from $\{5, 20, 40, 60, 80, 100\}\%$ of the total training epochs, forming six points on each line. The “beacon” models are the corresponding $s\{16, \dots, 512\}$ -lr.0.1 models from Fig. 3 (for each dataset), which are denoted by dotted curves (not represented in the legend). The mini-batch size of each dotted line is marked at the place where the training is at 5 percent epochs. Furthermore, each dynamic sampling method is designed to share the initial mini-batch size with one of the beacons, so we can observe how they move from the “roots” in the graph.

for ResNet (the two-sample t -test results in a p-value of 0.0037, which means that these two configurations produce statistically different test results, assuming a p-value of 0.05 threshold). In Fig. 5, we see similar trends for all models and datasets. Another important observation from Fig. 5 is that all dynamic sampling training curves do not lie on any of the curves formed by the beacon models—in fact these dynamic sampling training curves move almost perpendicularly to the beacon models’ curves.

We compare the best beacon and dynamic sampling models in Table 1. In general, the results show that dynamic sampling allows a faster training and a similar classification accuracy, compared with the fixed sampling training of the beacons. In Fig. 6, we show the training and testing error curves (as a function of number of epochs) for the $s32$ -to- 512 , $s512$ -to- 32 , and the beacon ResNet models. Note that the characterisation of such models using such error measurements is much less stable, when compared with the proposed \bar{C}_K and L_K .

Dynamic Sampling for Multi-step Learning Rate Decay. Following the intuition that learning rate decay causes the training process to focus on specific regions of the energy landscape, we test if the dynamic sampling should be performed within each particular value of learning rate or over all training epochs and decreasing learning rates, as explained above. This new approach is marked with -MS in Fig. 5, where for each learning rate value, we re-iterate through the sequence of mini-batch sizes of the corresponding dynamic sampling policy. In addition, the training and testing measures for the pair of models $s32$ -to- 512 and $s32$ -to- 512 -MS on ResNet have been shown in leftmost and rightmost columns of Fig. 6 to clarify the difference between the training processes of these two approaches. Since the new multi-step learning and the original dynamic sampling share the same amount of units of mini-batch sizes during training (which also means that they consume similar amount of training

time), then the \bar{C}_K values for both approaches are expected to be similar. However, the L_K values for the two approaches may have differences, as displayed in Fig. 5. In general, the -MS policy is more effective at limiting the growth of L_K compared to the original dynamic sampling counterpart, pushing the values closer to the optimum region of the graph. In Table 1, we notice that the -MS policy produces either significantly better or comparable classification accuracy results, compared to \emptyset and consumes a similar amount of training time. Furthermore, both -MS and \emptyset achieve a similar classification accuracy compared to the best beacon models, but with faster training time.

4.4 Toy Problem

In Fig. 7, we show that the proposed \bar{C}_K and L_K measurements can also be used to quantify the training of a simple multiple layer perceptron (MLP) network. The MLP network has two fully-connected hidden layers, each with 500 ReLU activated nodes. The output layer has 10 nodes that are activated with softmax function, allowing it to work as a classifier for the MNIST dataset. The entire amount of model parameters is 0.77M and the training procedure shares the same training hyper-parameter settings (i.e., momentum and learning rate schedule) of the ResNet and DenseNet models. It can be observed from Fig. 7 that the relative \bar{C}_K and L_K readings for each pair of learning rate and mini-batch size combination is similar to the their counterparts in the ResNet and DenseNet experiments.

5 DISCUSSION AND CONCLUSION

The take-home message of this paper is the following: training deep networks, and in particular ResNets and DenseNets, is still an art, but the use a few efficiently computed measures from SGD can provide substantial help in the selection of model parameters, such as learning rate

TABLE 1
The Comparison between the Best Beacon Model **s#** (at $lr=0.1$), and the Best Dynamic Sampling Models **- \emptyset** , and **-MS**

Model (best of each)		CIFAR-10			CIFAR-100		
		s#	- \emptyset	-MS	s#	- \emptyset	-MS
ResNet	Name	s32	s32-to-128	s32-to-128-MS	s32	s32-to-128	s32-to-128-MS
	Test Error	$4.78\% \pm 0.05\%$	$4.90\% \pm 0.05\%$	$4.76\% \pm 0.13\%$	$23.46\% \pm 0.21\%$	$23.90\% \pm 0.31\%$	$23.69\% \pm 0.34\%$
	p-value versus s#	—	0.0048	0.72	—	0.090	0.35
	p-value - \emptyset versus -MS	—	—	0.029	—	—	0.34
	Training Time (h)	7.7	7.0	7.1	7.9	7.1	7.1
DenseNet	Name	s32	s64-to-256	s32-to-128-MS	s32	s32-to-s128	s32-to-128-MS
	Test Error	$4.96\% \pm 0.12\%$	$5.03\% \pm 0.03\%$	$4.63\% \pm 0.10\%$	$23.26\% \pm 0.12\%$	$23.68\% \pm 0.02\%$	$23.92\% \pm 0.11\%$
	p-value versus s#	—	0.38	0.022	—	0.048	0.0022
	p-value - \emptyset versus -MS	—	—	0.0029	—	—	0.1689
	Training Time (h)	16.1	14.1	15.4	16.1	14.7	14.8
ResNet	Model (best of each)	SVHN			MNIST		
	Name	s128	s128-to-512	s32-to-512-MS	s16	s16-to-64	s16-to-64-MS
	Test Error	$1.93\% \pm 0.04\%$	$1.91\% \pm 0.01\%$	$1.90\% \pm 0.04\%$	$0.36\% \pm 0.02\%$	$0.39\% \pm 0.02\%$	$0.34\% \pm 0.02\%$
	p-value versus s#	—	0.58	0.51	—	0.11	0.18
	p-value - \emptyset versus -MS	—	—	0.69	—	—	0.033
DenseNet	Training Time (h)	9.5	8.9	9.4	0.14	0.11	0.11
	Name	s128	s64-to-256	s64-to-256-MS	s8	s32-to-128	s16-to-64-MS
	Test Error	$1.93\% \pm 0.09\%$	$2.00\% \pm 0.06\%$	$2.03\% \pm 0.02\%$	$0.61\% \pm 0.06\%$	$0.57\% \pm 0.03\%$	$0.60\% \pm 0.01\%$
	p-value versus s#	—	0.31	0.16	—	0.43	0.78
	p-value - \emptyset versus -MS	—	—	0.54	—	—	0.28
	Training Time (h)	21.9	20.4	20.5	0.26	0.06	0.10

and mini-batch sizes, leading to good training convergence and generalization. One possible way to further utilize the proposed \bar{C}_K and L_K in order to achieve a good balance between training convergence and generalization is to dynamically tune batch size and learning rate so that the

\bar{C}_K and L_K measurements do not increase too quickly because this generally means that the training process left the optimal convergence/generalization region.

In conclusion, we proposed a novel methodology to characterise the performance of two commonly used DeepNet

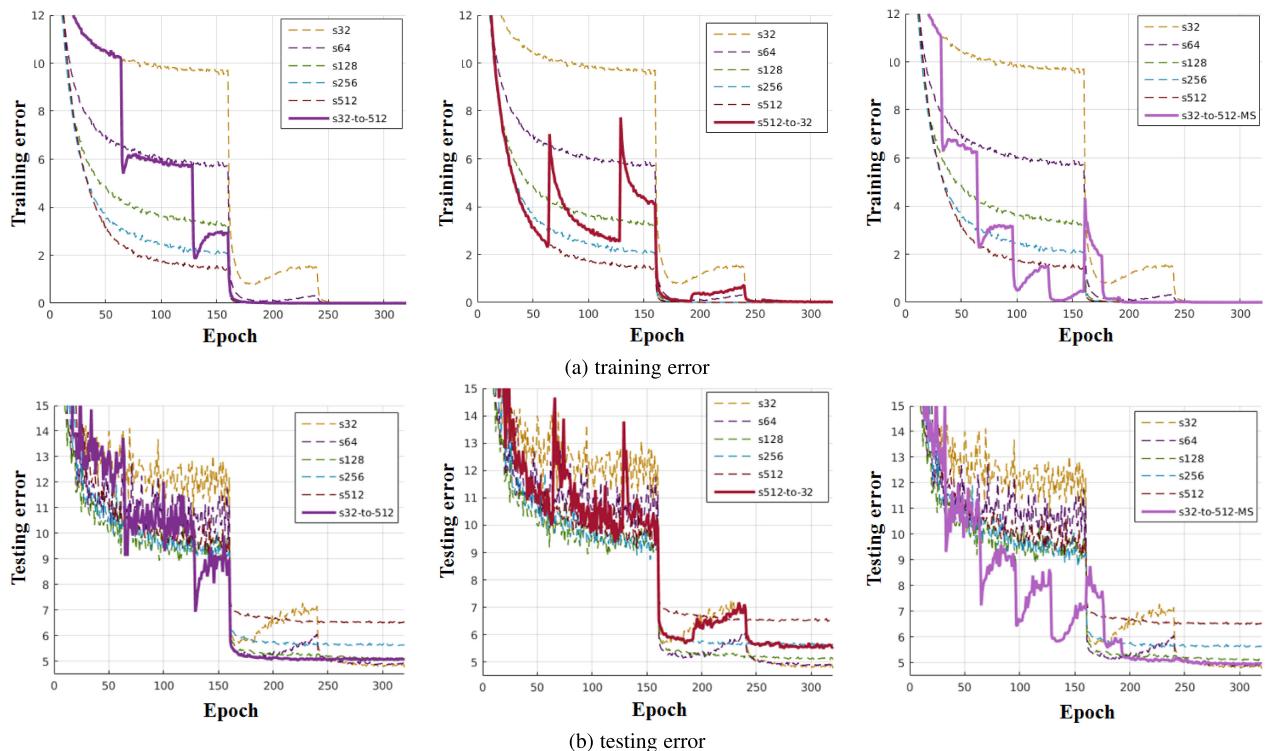


Fig. 6. The (a) training and (b) testing error on CIFAR-10 of selected **s32-to-512** (left), **s512-to-32** (center) dynamic sampling models, and the **s32-to-512-MS** (right) multi-step learning rate decay variant.

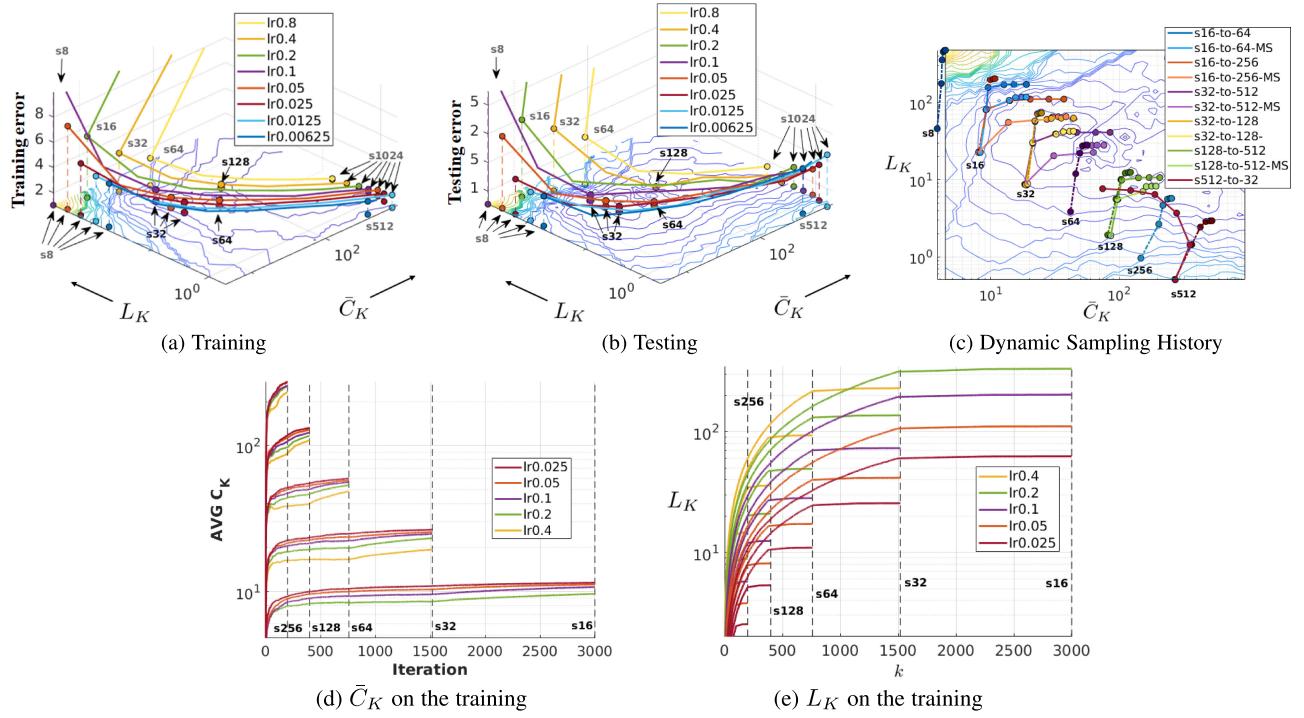


Fig. 7. The proposed measurements quantifying the training of a toy input-500-500-output MLP network.

architectures regarding training convergence and generalisation as a function of mini-batch size and learning rate. This proposed methodology defines a space that can be used for guiding the training of DeepNets, which led us to propose a new dynamic sampling training approach. We believe that the newly proposed measures will help researchers make important decisions about the DeepNets structure and training procedure. We also expect that this paper has the potential to open new research directions on how to assess and predict top performing DeepNets models with the use of the proposed measures (\bar{C}_K and L_K) and perhaps on new measures that can be proposed in the future.

ACKNOWLEDGMENTS

This work was partially edited while Zhibin Liao was a post-doctoral research fellow with the Robotics and Control Laboratory at the University of British Columbia. This work was support by the Australian Research Council through grants DP180103232, CE140100016, and FL130100102.

REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [2] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 646–661.
- [3] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *CVPR*, vol. 1, no. 2, p. 3, 2017.
- [4] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, pp. 400–407, 1951.
- [5] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–16.
- [6] E. Littwin and L. Wolf, "The loss surface of residual networks: Ensembles and the role of batch normalization," *arXiv:1611.02525*, 2016.
- [7] D. Soudry and Y. Carmon, "No bad local minima: Data independent training error guarantees for multilayer neural networks," *arXiv:1605.08361*, 2016.
- [8] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Proc. Conf. Learn. Theory*, 2016, pp. 1246–1257.
- [9] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, vol. 1, no. 4, p. 7, 2009.
- [10] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. Chayes, L. Sagun, and R. Zecchina, "Entropy-SGD: Biasing gradient descent into wide valleys," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–19.
- [11] J. Martens, "New insights and perspectives on the natural gradient method," *arXiv:1412.1193*, 2014.
- [12] L. Sagun, U. Evci, V. U. Guney, Y. Dauphin, and L. Bottou, "Empirical analysis of the hessian of over-parametrized neural networks," in *Proc. Int. Conf. Learn. Representations Workshop*, 2018, pp. 1–14.
- [13] S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey, "Width of minima reached by stochastic gradient descent is influenced by learning rate to batch size ratio," *Int. Conf. Artif. Neural Netw.*, pp. 392–402, 2018.
- [14] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.
- [15] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, 2011, Art. no. 5.
- [16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
- [17] R. Fletcher, *Practical Methods of Optimization*. Hoboken, NJ, USA: Wiley, 2013.
- [18] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM J. Sci. Comput.*, vol. 16, no. 5, pp. 1190–1208, 1995.
- [19] D. P. Bertsekas, "Incremental least squares methods and the extended Kalman filter," *SIAM J. Optimization*, vol. 6, no. 3, pp. 807–822, 1996.

- [20] N. N. Schraudolph, "Fast curvature matrix-vector products," in *Proc. Int. Conf. Artif. Neural Netw.*, 2001, pp. 19–26.
- [21] S.-I. Amari, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, no. 2, pp. 251–276, 1998.
- [22] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," COURSERA: *Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.
- [23] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. Jul, pp. 2121–2159, 2011.
- [24] L. Sagun, L. Bottou, and Y. LeCun, "Singularity of the hessian in deep learning," arXiv:1611.07476, 2016.
- [25] S. Wright and J. Nocedal, *Numerical Optimization*. Berlin, Germany: Springer, 1999, pp. 67–68.
- [26] B. A. Pearlmutter, "Fast exact multiplication by the Hessian," *Neural Comput.*, vol. 6, no. 1, pp. 147–160, 1994.
- [27] J. Martens, "Deep learning via Hessian-free optimization," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 735–742.
- [28] R. Kiros, "Training neural networks with stochastic hessian-free optimization," in *Proc. Int. Conf. Learn. Representations*, 2013, pp. 1–11.
- [29] M. D. Zeiler, "ADADELTA: An adaptive learning rate method," arXiv:1212.5701, 2012.
- [30] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.
- [31] I. J. Goodfellow, O. Vinyals, and A. M. Saxe, "Qualitatively characterizing neural network optimization problems," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–20.
- [32] S. L. Smith and V. Quoc, "A Bayesian perspective on generalization and stochastic gradient descent," in *Proc. 2nd Workshop Bayesian Deep Learn.*, 2017, pp. 1–4.
- [33] S. L. Smith, P.-J. Kindermans, and Q. V. Le, "Don't decay the learning rate, increase the batch size," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–11.
- [34] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch SGD: Training ImageNet in 1 hour," arXiv:1706.02677, 2017.
- [35] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.
- [36] M. P. Friedlander and M. Schmidt, "Hybrid deterministic-stochastic methods for data fitting," *SIAM J. Sci. Comput.*, vol. 34, no. 3, pp. A1380–A1405, 2012.
- [37] R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu, "Sample size selection in optimization methods for machine learning," *Math. Program.*, vol. 134, no. 1, pp. 127–155, 2012.
- [38] R. Bollapragada, R. Byrd, and J. Nocedal, "Adaptive sampling strategies for stochastic optimization," arXiv:1710.11258, 2017.
- [39] M. R. Metel, "Mini-batch stochastic gradient descent with dynamic sample sizes," arXiv:1708.00555, 2017.
- [40] S. De, A. Yadav, D. Jacobs, and T. Goldstein, "Big batch SGD: Automated inference using adaptive batch sizes," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1504–1513.
- [41] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.
- [42] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
- [43] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*. Oxford, U.K.: Clarendon, 1965.
- [44] Training and investigating residual nets. [Online]. Available: <http://torch.ch/blog/2016/02/04/resnets.html> Training and investigating Residual Nets, Accessed on: Mar. 13, 2017.
- [45] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A Matlab-like environment for machine learning," in *Proc. BigLearn NIPS Workshop*, 2011, pp. 1–6.



Zhibin Liao received the bachelor of computer science (1st class Hons) degree from the University of Adelaide and the PhD degree from the Australian Centre for Robotic Vision, University of Adelaide. He is currently a post-doctoral research fellow with the University of British Columbia. His research interests include machine learning, computer vision, and medical imaging analysis.



Tom Drummond received the BA (Hons) degree in mathematics, the MA degree from the University of Cambridge, and the PhD degree in computer science from Curtin University. He is a professor of electrical and computer systems engineering with Monash University. His research interests include computer vision, augmented reality, robotics, and machine learning.



Ian Reid is a professor of computer science with the University of Adelaide. He was formerly a professor of engineering science with the University of Oxford. His research interests span a wide range of topics in computer vision and machine learning, with a particular interest in using cameras as sensors for robotic applications such as for autonomous vehicles.



Gustavo Carneiro is an associate professor with the School of Computer Science, University of Adelaide. He has a strong track record in the fields of computer vision, medical image analysis, and machine learning, and has held distinguished positions with the academia and industry.

▷ For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.