On The Power of Curriculum Learning in Training Deep Networks

Guy Hacohen [1 2] **Daphna Weinshall** [1]

Abstract

Training neural networks is traditionally done by providing a sequence of random mini-batches sampled uniformly from the entire training data. In this work, we analyze the effect of curriculum learning, which involves the non-uniform sampling of mini-batches, on the training of deep networks, and specifically CNNs trained for image recognition. To employ curriculum learning, the training algorithm must resolve 2 problems: (i) sort the training examples by difficulty; (ii) compute a series of mini-batches that exhibit an increasing level of difficulty. We address challenge (i) using two methods: transfer learning from some competitive "teacher" network, and bootstrapping. In our empirical evaluation, both methods show similar benefits in terms of increased learning speed and improved final performance on test data. We address challenge (ii) by investigating different pacing functions to guide the sampling. The empirical investigation includes a variety of network architectures, using images from CIFAR-10, CIFAR-100 and subsets of ImageNet. We conclude with a novel theoretical analysis of curriculum learning, where we show how it effectively modifies the optimization landscape. We then define the concept of an ideal curriculum, and show that under mild conditions it does not change the corresponding global minimum of the optimization function.

1. Introduction

In order to teach complex tasks, teachers are often required to create a curriculum. A curriculum imposes some order on the concepts that constitute the final task, an order which typically reflects their complexity. The student is then gradually introduced to these concepts by increasing complexity, in order to allow her to exploit previously learned concepts and thus ease the abstraction of new ones. But the use of a curriculum is not limited to complex tasks. When teaching a binary classification task, for example, teachers tend to present typical examples first, followed by the more ambiguous examples (Avrahami et al., 1997).

In many traditional machine learning paradigms, a target function is estimated by a learner (the "student") using a set of training labeled examples (provided by the "teacher"). The field of curriculum learning (CL), which is motivated by the idea of a curriculum in human learning, attempts at imposing some structure on the training set. Such structure essentially relies on a notion of "easy" and "hard" examples, and utilizes this distinction in order to teach the learner how to generalize easier examples before harder ones. Empirically, the use of CL has been shown to accelerate and improve the learning process (e.g. Selfridge et al., 1985; Bengio et al., 2009) in many machine learning paradigms.

When establishing a curriculum for human students, teachers need to address two challenges: (i) Arrange the material in a way that reflects difficulty or complexity, a knowledge which goes beyond what is available in the training set in most machine learning paradigms. (ii) Attend to the pace by which the material is presented – going over the simple ideas too fast may lead to more confusion than benefit, while moving along too slowly may lead to boredom and unproductive learning (Hunkins & Ornstein, 2016). In this paper, we study how these principles can be beneficial when the learner is a neural network.

In order to address the first challenge, Weinshall et al. (2018) introduced the idea of curriculum learning by transfer. The idea is to sort the training examples based on the performance of a pre-trained network on a larger dataset, fine-tuned to the dataset at hand. This approach was shown to improve both the speed of convergence and final accuracy for convolutional neural networks, while not requiring the manual labeling of training data by difficulty.

In our work, we address both challenges. We begin by decomposing CL into two separate - but closely related - subtasks and their corresponding functions. The first, termed *scoring function*, determines the "difficulty" or "complexity"

[1]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 91904, Israel [2]Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem 91904, Israel. Correspondence to: Guy Hacohen <guy.hacohen@mail.huji.ac.il>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

of each example in the data. The *scoring function* makes it possible to sort the training examples by difficulty, and present to the network the easier (and presumably simpler) examples first. Scoring is done based on transfer learning as in Weinshall et al. (2018), or bootstrapping as explained below. The second function, termed *pacing function*, determines the pace by which data is presented to the network. The pace may depend on both the data itself and the learner.

We show that the use of different pacing functions can indirectly affect hyper-parameters of the neural network, reminiscent of increased learning rate. As Weinshall et al. (2018) did not employ parameter tuning in their empirical study, the improvement they report might be explained by the use of inappropriate learning rates. As part of our work, we repeat their experimental paradigm while optimizing the network's hyper-parameters (with and without cross-validation), showing improvement in both the speed of convergence and final accuracy in a more reliable way. We then extend these results and report experiments on larger datasets and architectures, which are more commonly used as benchmarks.

We continue by analyzing several *scoring* and *pacing* functions, investigating their inter-dependency and presenting ways to combine them in order to achieve faster learning and better generalization. The main challenge is, arguably, how to obtain an effective *scoring function* without additional labeling of the data. To this end we investigate two approaches, each providing a different estimator for the target *scoring function*: (i) Knowledge transfer as in (Weinshall et al., 2018), based on transfer learning from networks trained on the large and versatile Imagenet dataset. (ii) Bootstrapping based on self-tutoring - we train the network without curriculum, then use the resulting classifier to rank the training data in order to train the same network again from scratch. Unlike curriculum by transfer, bootstrapping does not require access to any additional resources. Both *scoring functions* maintain the required property that prefers points with a lower loss with respect to the target hypothesis. The aforementioned functions are shown in Section 3 to speed up learning and improve the generalization of CNNs.

We investigate three *pacing functions*. (i) *Fixed exponential pacing* presents the learner initially with a small percentage of the data, increasing the amount exponentially every fixed number of learning iterations. (ii) *Varied exponential pacing* allows the number of iterations in each step to vary as well. (iii) *Single-step pacing* is a simplified version of the first protocol, where mini-batches are initially sampled from the easiest examples (a fixed fraction), and then from the whole data as usual. In our empirical setup, the three functions have comparable performance.

In Section 4 we conclude with a theoretical analysis of the effects of curriculum learning on the objective function of neural networks. We show that curriculum learning modifies the optimization landscape, making it steeper while maintaining the same global minimum of the original problem. This analysis provides a framework by which apparently conflicting heuristics for the dynamic sampling of training points can coexist and be beneficial, including SPL, boosting and hard data mining, as discussed under *previous work*.

Previous work. Imposing a curriculum in order to speed up learning is widely used in the context of human learning and animal training (Skinner, 1958; Pavlov, 2010; Krueger & Dayan, 2009). In many application areas, it is a common practice to introduce concepts in ascending order of difficulty, as judged by either the human teacher or in a problem dependent manner (e.g. Murphy et al., 2008; Zaremba & Sutskever, 2014; Amodei et al., 2016). With the rebirth of deep learning and its emerging role as a powerful learning paradigm in many applications, the use of CL to control the order by which examples are presented to neural networks during training is receiving increased attention (Graves et al., 2016; 2017; Florensa et al., 2017).

In a closely related line of work, a pair of teacher and student networks are trained simultaneously, where mini-batches for the student network are sampled dynamically by the teacher, based on the student's output at each time point. As opposed to our method, here the curriculum is based on the current hypothesis of the student, while achieving improved performance for corrupted (Jiang et al., 2018) or smaller (Fan et al., 2018) datasets. Improvement in generalization over the original dataset has not been shown.

In some machine learning paradigms, which are related to CL but differ from it in an essential manner, mini-batches are likewise sampled dynamically. Specifically, in Self-Paced Learning (SPL- Kumar et al., 2010), boosting (Freund et al., 1996), hard example mining (Shrivastava et al., 2016) and even active learning (Schein & Ungar, 2007), mini-batches are sampled at each time point based on the ranking of the training examples by their difficulty with respect to the **current hypothesis** of the model. Thus they differ from CL, which relies on the ranking of training points by their difficulty with respect to some **target hypothesis**.

Confusingly, based on the same ephemeral ranking, SPL advocates the use of easier training examples first, while the other approaches prefer to use the harder examples. Still, all approaches show benefit under different empirical settings (Chang et al., 2017; Zhang et al., 2017). This discrepancy is analyzed in Weinshall et al. (2018), where it is shown that while it is beneficial to prefer easier points with respect to the target hypothesis as advocated by CL, it is at the same time beneficial to prefer the more difficult points with respect to the current hypothesis in agreement with hard data mining and boosting, but contrary to SPL. In contrast, our theoretical analysis (Section 4) is consistent with both heuristics being beneficial under different circumstances.

2. Curriculum Learning

Curriculum learning as investigated here deals with the question of *how to use prior knowledge about the difficulty of the training examples*, in order to sample each mini-batch non-uniformly and thus boost the rate of learning and the accuracy of the final classifier. The paradigm of CL is based on the intuition that it helps the learning process when the learner is presented with simple concepts first.

2.1. Notations and Definitions

Let $\mathbb{X} = \{X_i\}_{i=1}^N = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ denote the data, where $\boldsymbol{x}_i \in \mathbb{R}^d$ denotes a single data point and $y_i \in [K]$ its corresponding label. Let $h_\vartheta : \mathbb{R}^d \to [K]$ denote the target classifier (or learner), and mini-batch $\mathbb{B} \subseteq \mathbb{X}$ denote a subset of $\mathbb{X}$. In the most common training procedure, which is a robust variant of Stochastic Gradient Descent (SGD), h_ϑ is trained sequentially when given as input a sequence of mini-batches $[\mathbb{B}_1, ..., \mathbb{B}_M]$ (Shalev-Shwartz & Ben-David, 2014). The common approach – denoted henceforth *vanilla* – samples each mini-batch $\mathbb{B}_i$ uniformly from $\mathbb{X}$. Both in the common approach and in our work, the size of each mini-batch remains constant, to be considered as a hyper-parameter defining the learner.

We measure the difficulty of example $(\boldsymbol{x}_i, y_i)$ by its minimal loss with respect to the set of optimal hypotheses under consideration. We define a *scoring function* to be any function $f : \mathbb{X} \to \mathbb{R}$, and say that example $(\boldsymbol{x}_i, y_i)$ is more difficult than example $(\boldsymbol{x}_j, y_j)$ if $f(\boldsymbol{x}_i, y_i) > f(\boldsymbol{x}_j, y_j)$. Choosing f is the main challenge of CL, as it encodes the prior knowledge of the teacher.

We define a *pacing function* $g_\vartheta : [M] \to [N]$, which may depend on the learner h_ϑ. The *pacing function* is used to determine a sequence of subsets $\mathbb{X}_1^{'}, ..., \mathbb{X}_M^{'} \subseteq \mathbb{X}$, of size $|\mathbb{X}_i^{'}| = g_\vartheta(i)$, from which $\{\mathbb{B}_i\}_{i=1}^M$ are sampled uniformly. In CL the i-th subset $\mathbb{X}_i^{'}$ includes the first $g_\vartheta(i)$ elements of the training data when sorted by the *scoring function* f in ascending order. Although the choice of the subset can be encoded in the distribution each $\mathbb{B}_i$ is sampled from, adding a *pacing function* simplifies the exposition and analysis.

2.2. Curriculum Learning Method

Together, each *scoring function* f and *pacing function* g_ϑ define a curriculum. Any learning algorithm which uses the ensuing sequence $[\mathbb{B}_i]_{i=1}^M$ is a **curriculum learning algorithm**. We note that in order to avoid bias when picking a subset of the N examples for some N, it is important to keep the sample balanced with the same number of examples from each class as in the training set. Pseudo-code for the CL algorithm is given in Alg. 1.

In order to narrow down the specific effects of using a *scoring function* based on ascending difficulty level, we examine two control conditions. Specifically, we define 2 additional *scoring functions* and corresponding algorithms: (i) The **anti-curriculum algorithm** uses the *scoring function* $f' = -f$, where the training examples are sorted in *descending order* of difficulty; thus harder examples are sampled before easier ones. (ii) The **random-curriculum algorithm** (henceforth denoted **random**) uses a *scoring function* where the training examples are randomly scored.

Algorithm 1 Curriculum learning method

Input: *pacing function* g_ϑ, *scoring function* f, data $\mathbb{X}$.
Output: sequence of mini-batches $\left[\mathbb{B}_1^{'}, ..., \mathbb{B}_M^{'}\right]$.
sort $\mathbb{X}$ according to f, in ascending order
$result \leftarrow []$
for all $i = 1, ..., M$ **do**
  $size \leftarrow g_\vartheta(i)$
  $\mathbb{X}_i^{'} \leftarrow \mathbb{X}[1, ..., size]$
  uniformly sample $\mathbb{B}_i^{'}$ from $\mathbb{X}^{'}$
  append $\mathbb{B}_i^{'}$ to $result$
end for
return $result$

2.3. Scoring and Pacing Functions

We evaluate two *scoring functions*: (i) *Transfer scoring function*, computed as follows: First, we take the Inception network (Szegedy et al., 2016) pre-trained on the ImageNet dataset (Deng et al., 2009) and run each training image through it, using the activation levels of its penultimate layer as a feature vector (Caruana, 1995). Second, we use these features to train a classifier and use its confidence score as the *scoring function* for each image[1]. (ii) *Self-taught scoring function*, computed as follows: First, we train the network using uniformly sampled mini-batches (the *vanilla* method). Second, we compute this network's confidence score for each image to define a *scoring function*[2].

The *pacing function* can be any function $g_\vartheta : [M] \to [N]$. However, we limit ourselves to monotonically increasing functions so that the likelihood of the easier examples can only decrease. For simplicity, g_ϑ is further limited to staircase functions. Thus each *pacing function* is defined by the following hyper-parameters, where *step* denotes all the learning iterations during which g_ϑ remains constant: *step_length* - the number of iterations in each *step*; *increase* - an exponential factor used to increase the size of the data used for sampling mini-batches in each *step*; *start-*

[1] Similar results are obtained when using different confidence scores (e.g, the classifier's margin), different classifiers (e.g, linear SVM), and different teacher networks (e.g, VGG-16 (Simonyan & Zisserman, 2014) or ResNet (He et al., 2016)), see Suppl A.

[2] *Self-taught* can be used repeatedly, see Suppl B.

ing_percent - the fraction of the data in the initial *step*. An illustration of these parameters can be seen in Fig. 1.

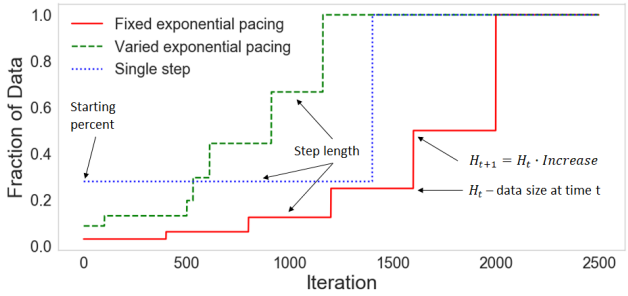


Figure 1. Illustration of the 3 *pacing functions* used below, showing the different hyper-parameters that define each of them (see text). The values of the hyper-parameters used in this illustration were chosen arbitrarily, for illustration only.

We evaluate three *pacing functions*: (i) *Fixed exponential pacing* has a fixed *step_length*, and exponentially increasing data size in each *step*. Formally, it is given by:

$$g_\vartheta(i) = \min\left(starting_percent \cdot inc^{\lfloor \frac{i}{step_length} \rfloor}, 1\right) \cdot N$$

(ii) *Varied exponential pacing* is the same as (i), while allowing *step_length* to vary. This method adds additional hyper-parameters, but removes the need to re-tune the learning rate parameters (see Suppl A). (iii) *Single step pacing* entails the simplification of the staircase function into a step function, resulting in fewer hyper-parameters. Formally:

$$g_\vartheta(i) = starting_percent^{\mathbf{1}[i < step_length]} \cdot N$$

3. Empirical Evaluation

Methodology.[3] We define 6 test cases: **Case 1** replicates the experimental design described in (Weinshall et al., 2018), by using the same dataset and network architecture. The dataset is the "small mammals" super-class[4] of CIFAR-100 (Krizhevsky & Hinton, 2009) - a subset of 3000 images from CIFAR-100, divided into 5 classes. The neural network is a moderate size hand-crafted convolutional network, whose architecture details can be found in Suppl C. **Cases 2** and **3** adopt the same architecture while extending the datasets to the entire CIFAR-10 and CIFAR-100 datasets. **Cases 4** and **5** use a well known public-domain VGG-based architecture[5] (Simonyan & Zisserman, 2014; Liu & Deng, 2015), to classify the CIFAR-10 and CIFAR-100 datasets. **Case 6** adopts the same architecture as cases 1-3, trained with a subset of 7 classes of cats (see Suppl C) from the ImageNet dataset.

[3]All the code used in the paper is available at https://github.com/GuyHacohen/curriculum_learning

[4]Other super-classes achieve similar results, see Suppl A.

[5]Code available at https://github.com/geifmany/cifar-vgg.

Hyper-parameter tuning. As in all empirical studies involving deep learning, the results are quite sensitive to the values of the hyper-parameters, hence parameter tuning is required. In practice, in order to reduce the computation time of parameter tuning, in the curriculum conditions, we performed first grid-search on the curriculum hyper-parameters, followed by a second grid-search on the learning rate parameters, thus avoiding the need to tune a large number of parameters at once. In the non-curriculum conditions, a full 1-stage grid-search was performed. In addition, we varied only the first 2 *step_length* instances in the *varied exponential pacing* condition. Accordingly, *fixed exponential pacing*, *varied exponential pacing* and *single step pacing* define 3, 5 and 2 new hyper-parameters respectively, henceforth called the *pacing* hyper-parameters.

With CL, the use of a *pacing function* affects the optimal values of other hyper-parameters, the learning rate in particular. Specifically, the *pacing function* significantly reduces the size of the dataset at the beginning of the learning, which has the concomitant effect of increasing the effective learning rate at that time. As a result, for a fair comparison, when using the *fixed exponential* or the *single step pacing functions*, the learning rate must be tuned separately for every test condition. This tuning is missing in Weinshall et al. (2018), whose results may therefore be tainted by their arbitrary choice of learning rate. Using the *varied exponential pacing function* can overcome the need for learning rate re-tuning, while adding 2 hyper-parameters (see Suppl A).

As traditionally done (e.g Simonyan & Zisserman, 2014; Szegedy et al., 2016; He et al., 2016), we set an initial learning rate and decrease it exponentially every fixed number of iterations. This method gives rise to 3 learning rate hyper-parameters which require tuning: (i) the initial learning rate; (ii) the factor by which the learning rate is decreased; (iii) the length of each step with constant learning rate[6].

Cross-validation. In grid search, parameters are chosen based on performance on the test set. To avoid contamination of the conclusions, all results were cross-validated, wherein the hyper-parameters are chosen based on performance on a validation set before being used on the test set. For more details on the steps we took to ensure a fair comparison when employing a grid search, see Suppl B.

3.1. Curriculum by Transfer

Case 1: A moderate size network is trained to distinguish 5 classes from CIFAR-100, which are members of the same super-class as defined in the original dataset. Results are shown in Fig. 2. Curriculum learning is clearly and significantly beneficial - learning starts faster, and converges to a better solution. We observe that the performance of CL with

[6]Other tuning methods achieve similar results, see Suppl B.

a random *scoring function* is similar to *vanilla*, indicating that the main reason for the improvement achieved by CL is due to its beneficial *transfer scoring function*. In fact, although tuned separately, the learning rate hyper-parameters for both the *random* and the *curriculum* test conditions are very similar, confirming that the improved performance is due to the use of an effective *transfer scoring function*.

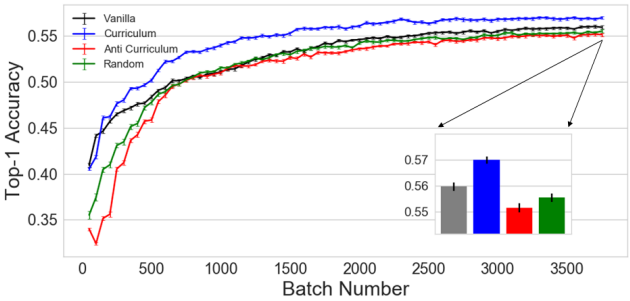


Figure 2. Results in **case 1**, with Inception-based *transfer scoring function* and *fixed exponential pacing function*. Inset: bars indicating the average final accuracy in each condition over the last few iterations. Error bars indicate the STE (STandard Error of the mean) after 50 repetitions. The curriculum method (in blue) reaches higher accuracy faster, and converges to a better solution.

To check the robustness of these results, we repeated the same empirical evaluation using different super-classes of CIFAR-100, with similar results (see Suppl A). Interestingly, we note that the observed advantage of CL is more significant when the task is more difficult (i.e. lower *vanilla* test accuracy). The reason may be that in easier problems there is a sufficient number of easy examples in each mini-batch even without CL. Although the results reported here are based on transfer from the Inception network, we are able to obtain the same results using scoring functions based on transfer learning from other large networks, including VGG-16 and ResNet, as shown in Suppl A.

Cases 2 and 3: Similar empirical evaluation as in **case 1**, using the same moderate size network to classify two benchmark datasets. The results for CIFAR-10 are shown in Fig. 4b and for CIFAR-100 in Fig. 3. Like before, test accuracy with curriculum learning increases faster and reaches better final performance in both cases, as compared to the *vanilla* test condition. The beneficial effect of CL is larger when classifying the more challenging CIFAR-100 dataset.

Cases 4 and 5: Similar empirical evaluation as in **case 1**, using a competitive public-domain architecture. Specifically, we use the Inception-based *transfer scoring function* to train a VGG-based network (Liu & Deng, 2015) to classify the CIFAR-10 and CIFAR-100 datasets. Differently from the previous cases, here we use the *varied exponential pacing function* with a slightly reduced learning rate, as it

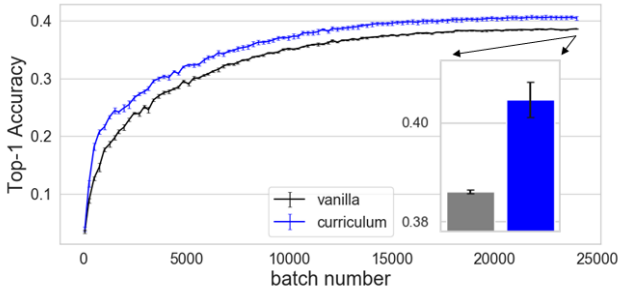
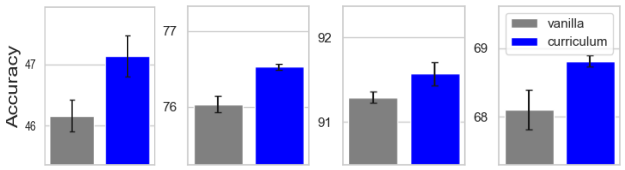


Figure 3. Results in **case 3**, CIFAR-100 dataset, with Inception-based *transfer scoring function* and *fixed exponential pacing function*. Inset: zoom-in on the final iterations, for better visualization. Error bars show STE after 5 repetitions.



(a) ImageNet Cats **(b)** Cifar10 **(c)** Cifar10 VGG **(d)** Cifar100 VGG

Figure 4. Curriculum by transfer learning. Bars indicate the average final accuracy, and error bars indicate the STE. We performed 25 repetitions in (a), 5 in (b) and 3 in (c,d). (a) Cats subset of ImageNet. (b) CIFAR-10, trained on a small network. (c, d) CIFAR-10 and CIFAR-100 respectively, trained on the VGG network.

has the fewest hyper-parameters to tune since learning rate parameters do not need to be re-tuned, an important factor when training large networks. Results for CIFAR-10 are shown in Fig. 4c and for CIFAR-100 in Fig. 4d; in both cases, no data augmentation has been used. The results show the same qualitative results as in the previous cases; CL gives a smaller benefit, but the benefit is still significant.

Case 6: Similar to **case 1**, using the same moderate size network to distinguish 7 classes of cats from the ImageNet dataset (see Suppl C for details). The results are shown in Fig. 4a. Again, the test accuracy in the curriculum test condition increases faster and achieves better final performance with curriculum, as compared to the *vanilla* test condition.

3.2. Curriculum by Bootstrapping

The *self-taught scoring function* is based on the loss of training points with respect to the final hypothesis of a trained network - the same network architecture pre-trained without a curriculum. Using this scoring function, training is re-started from scratch. Thus defined, *curriculum by bootstrapping* may seem closely related to the idea of *Self-Paced Learning* (SPL), an iterative procedure where higher weights are given to training examples that have lower cost with respect to the current hypothesis. There is, however, a very

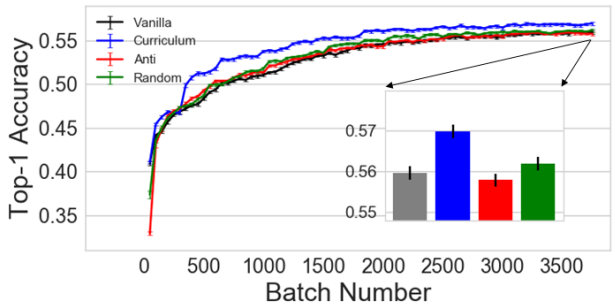


Figure 5. Results in **case 1**, with Inception-based *transfer scoring function* and *Single step pacing function*. Inset: bars indicating the average final accuracy in each condition over the last few iterations. Error bars indicate the STE after 50 repetitions.

important difference between the methods: SPL determines the *scoring function* based on the loss with respect to the current hypothesis (or network), while *bootstrapping CL* scores each point by its loss with respect to the target hypothesis. SPL, it appears, has not yet been introduced to deep neural networks in a way that benefits accuracy.

To compare the *self-taught scoring function* and the *self-paced scoring function*, we investigate their effect on CL in the context of test **case 1**. Final accuracy results are shown in Fig. 7 (the entire learning curves are depicted in Suppl C, Fig. 12). As expected, we see that *bootstrapping CL* improves test accuracy throughout the entire learning session. On the other hand, CL training using the *self-paced scoring function* decreases the test accuracy throughout. This decrease is more prominent at the beginning of the learning, where most of the beneficial effects of the curriculum are observed, suggesting that the *self-paced scoring function* can significantly delay learning.

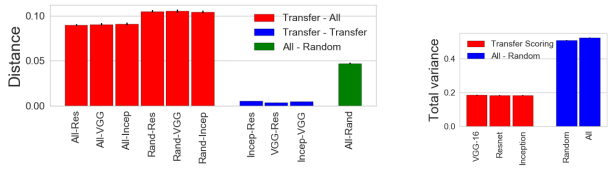
3.3. Alternative Pacing Functions

Single step pacing. Curriculum learning can be costly, and it affects the entire learning protocol via the *pacing function*. At the same time, we observed empirically that the main effect of the procedure seems to have taken place at the beginning of training. This may be due, in part, to the fact that the proposed *scoring function* f is based on transfer from another network trained on a different dataset, which only approximates the unknown *ideal scoring function*. Possibly, since the *scoring function* is based on one local minimum in a complex optimization landscape which contains many local minima, the score given by f is more reliable for low scoring (easy) examples than high scoring (difficult) examples, which may be in the vicinity of a different local minimum.

Once again we evaluate **case 1**, using the *transfer scoring function* and the *single step pacing function*. We see improvement in test accuracy in the curriculum test condition, which resembles the improvement achieved using the *exponential pacing*. Results are shown in Fig. 5. It is important to note that this *pacing function* ignores most of the prior knowledge provided by the *scoring function*, as it only uses a small percent of the easiest examples, and yet it achieves competitive results. Seemingly, in our empirical setup, most of the power of CL lies at the beginning of training.

3.4. Analysis of Scoring Function

In order to analyze the effects of transfer based *scoring functions*, we turn to analyze the gradients of the network's weights w.r.t the empirical loss. We evaluate the gradients using a pre-trained *vanilla* network in the context of **case 1**. First, for each method and each *scoring function*, we collect the subset of points used to sample the first mini-batch according to the *pacing function* $g_\vartheta(1)$[7]. For comparison, we also consider the set of all training points, which are used to compute the exact gradient of the empirical loss in batch learning using GD. We then compute the corresponding set of gradients for the training points in each of these subsets of training points, treating each layer's parameters as a single vector, and subsequently estimate the gradients' mean and total variance[8]. We use these measurements to evaluate the coherence of the gradients in the first mini-batch of each *scoring function*. The Euclidean distance between the mean gradient in the different conditions is used to estimate the similarity between the different *scoring functions*, based on the average preferred gradient. We compare the set of gradients defined by using three *transfer scoring functions*, which differ in the teacher network used for scoring the points: 'VGG-16', 'ResNet', and 'Inception'. We include in the comparison the gradients of the *random scoring function* denoted 'Random', and the gradients of the whole batch of training data denoted 'All'. Results are shown in Fig. 6.



(a) Distances between mean gradients. (b) Total variance.

Figure 6. (a) Distance between the mean gradient direction of preferred examples under different *scoring functions*. Each bar corresponds to a pair of mean gradients in two different conditions, see text. (b) The total variance of each set of gradients.

We see in Fig. 6a - blue bars - that the average gradient vectors, computed based on the 3 *transfer scoring functions*,

[7] In this experiment $g_\vartheta(1)$ is set such that it corresponds to 10% of the data or 250 examples. This number was set arbitrarily, with similar qualitative results obtained for other choices.

[8] Total variance denotes the trace of the covariance matrix.

are quite similar to each other. This suggests that they are pointing towards nearby local minima in parameters space. We also see - green bar - that the average gradient vector computed using a random subset of examples resembles the exact empirical gradient computed using all the training data. This suggests that a random subset provides a reasonable estimate of the true empirical gradient. The picture changes completely when we compute - red bars - the distance between the average gradient corresponding to one of the 3 transfer *scoring functions*, and the average random gradient or the empirical gradient. The large distances suggest that CL by transfer stirs the weights towards different local minima in parameter space as compared to *vanilla* training.

We see in Fig. 6b that the total variance for the 3 transfer *scoring functions* is much smaller than the total variance of some random subset of the whole training set. This intuitive result demonstrates the difference between training with easier examples and training with random examples, and may – at least partially – explain the need for a different learning rate when training with easier examples.

3.5. Summary of Results

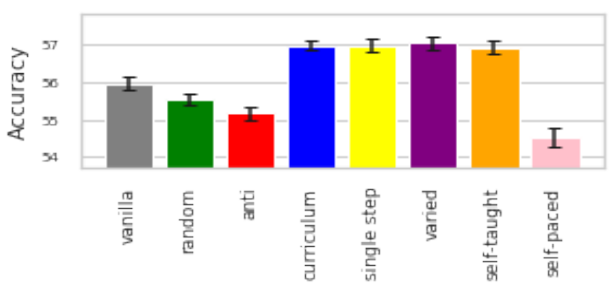


Figure 7. Results in **case 1**, bars showing final accuracy in percent for all test conditions. Error bars indicate STE after 50 repetitions.

Fig. 7 summarizes the main results presented in this section, including: curriculum with an Inception-based *scoring function* for (i) *fixed* exponential pacing (denoted *curriculum*), (ii) *single step* pacing, and (iii) *varied* exponential pacing. It also shows curriculum with fixed exponential pacing for (iv) *self-taught* scoring, and (v) *self-paced* scoring. In addition, we plot the control conditions of *vanilla*, *anti*-curriculum, and *random*. The bars depict the final accuracy in each condition. All the curriculum conditions seem to improve the learning accuracy throughout the entire learning session while converging to similar performance, excluding the self-paced *scoring function* which impairs learning. While different conditions seem to improve the final accuracy in a similar way, the results of *curriculum by transfer* are easier to obtain, and are more robust (see Suppl C for details).

4. Theoretical Analysis

Let $\mathcal{H}$ denote a set of hypotheses h_ϑ defined by the vector of hyper-parameters ϑ. Let $L_\vartheta(X_i)$ denote the loss of hypothesis h_ϑ when given example X_i. In order to compute the best hypothesis $h_{\tilde{\vartheta}}$ from the data, one commonly uses the Empirical Risk Minimization (ERM) framework where[9]

$$\begin{aligned} \mathcal{L}(\vartheta) &= \hat{\mathbb{E}}[L_\vartheta] = \frac{1}{N}\sum_{i=1}^{N} L_\vartheta(X_i) \\ \tilde{\vartheta} &= \arg\min_{\vartheta} \mathcal{L}(\vartheta) \end{aligned} \tag{1}$$

$\mathcal{L}(\vartheta)$ denotes the empirical loss given the observed data, thus defining the Risk of choosing hypothesis h_ϑ. (1) can be rewritten as follows:

$$\begin{aligned} \tilde{\vartheta} &= \arg\min_{\vartheta} \sum_{i=1}^{N} L_\vartheta(X_i) = \arg\max_{\vartheta} \exp(-\sum_{i=1}^{N} L_\vartheta(X_i)) \\ &= \arg\max_{\vartheta} \prod_{i=1}^{N} e^{-L_\vartheta(X_i)} \triangleq \arg\max_{\vartheta} \prod_{i=1}^{N} \alpha P(\vartheta|_{X_i}) \end{aligned}$$

Thus ERM can be interpreted as Maximum Likelihood (ML) estimation with probability defined by the loss function as $P(\vartheta|_X) \propto e^{-L_\vartheta(X)}$.

The choice of loss $L_\vartheta(X)$, and the choice of the estimation framework used to select some optimal hypothesis $h_{\tilde{\vartheta}}$, are somewhat arbitrary. In a similar manner we may choose to maximize the average *Utility* $U_\vartheta(X) = e^{-L_\vartheta(X)}$ of the observed data, which is defined as follows

$$\begin{aligned} \mathcal{U}(\vartheta) &= \hat{\mathbb{E}}[U_\vartheta] = \frac{1}{N}\sum_{i=1}^{N} U_\vartheta(X_i) \triangleq \frac{1}{N}\sum_{i=1}^{N} e^{-L_\vartheta(X_i)} \\ \tilde{\vartheta} &= \arg\max_{\vartheta} \mathcal{U}(\vartheta) \end{aligned} \tag{2}$$

The ERM formulation defined in (1) is different from the empirical utility maximization formulation defined in (2). Both formulations can be similarly justified from first principles.

The *scoring function* in curriculum learning effectively provides a Bayesian prior for data sampling. This can be formalized as follows:

$$\begin{aligned} \mathcal{U}_p(\vartheta) &= \hat{\mathbb{E}}_{\boldsymbol{p}}[U_\vartheta] = \sum_{i=1}^{N} U_\vartheta(X_i)p(X_i) = \sum_{i=1}^{N} e^{-L_\vartheta(X_i)} p_i \\ \tilde{\vartheta} &= \arg\max_{\vartheta} \mathcal{U}_p(\vartheta) \end{aligned} \tag{3}$$

Above $p_i = p(X_i)$ denotes the induced prior probability, which is determined by the *scoring function* and *pacing function* of the curriculum algorithm. Thus $p(X_i)$ will always

[9] $\hat{A}$, for any operator A, denotes the empirical estimate of A.

be a non-increasing function of the difficulty level of X_i. In our algorithm, $p(X_i) = \frac{1}{M}$ for M training points whose difficulty score is below a certain threshold, and $p(X_i) = 0$ otherwise. The threshold is determined by the *pacing function* which drives a monotonic increase in the number of points M, thus changing the optimization function in a corresponding manner.

From (3), $\mathcal{U}_p(\vartheta)$ is a function of ϑ which is determined by the correlation between two random variables, $U_\vartheta(X)$ and $p(X)$. We rewrite (3) as follows

$$\begin{aligned}\mathcal{U}_p(\vartheta) &= \sum_{i=1}^{N}(U_\vartheta(X_i) - \hat{\mathbb{E}}[U_\vartheta])(p_i - \hat{\mathbb{E}}[p]) + N\,\hat{\mathbb{E}}[U_\vartheta]\hat{\mathbb{E}}[p] \\ &= \hat{\mathrm{Cov}}[U_\vartheta, p] + N\,\hat{\mathbb{E}}[U_\vartheta]\hat{\mathbb{E}}[p] = \mathcal{U}(\vartheta) + \hat{\mathrm{Cov}}[U_\vartheta, p]\end{aligned} \tag{4}$$

This proves the following result:

Proposition 1 *The difference between the expected utility when computed with and without prior* $\boldsymbol{p}$ *is the covariance between the two random variables* $U_\vartheta(X)$ *and* $p(X)$.

Curriculum learning changes the landscape of the optimization function over the hyper-parameters ϑ from $\mathcal{U}(\vartheta)$ to $\mathcal{U}_p(\vartheta)$. Intuitively, (4) suggests that if the induced prior probability $\boldsymbol{p}$, which defines a random variable over the input space $p(X)$, is positively correlated with the optimal utility $U_{\tilde{\vartheta}}(X)$, and more so than with any other $U_\vartheta(X)$, then the gradients in the direction of the optimal parameter $\tilde{\vartheta}$ in the new optimization landscape may be overall steeper.

More precisely, assume that $\tilde{\vartheta}$ maximizes the covariance between $p_\vartheta(X)$ and the utility $U_\vartheta(X)$, namely

$$\arg\max_\vartheta \mathcal{U}(\vartheta) = \arg\max_\vartheta \hat{\mathrm{Cov}}[U_\vartheta, p] = \tilde{\vartheta} \tag{5}$$

Proposition 2 *For any curriculum satisfying (5):*

1. $\tilde{\vartheta} = \arg\max_\vartheta \mathcal{U}(\vartheta) = \arg\max_\vartheta \mathcal{U}_p(\vartheta)$
2. $\mathcal{U}_p(\tilde{\vartheta}) - \mathcal{U}_p(\vartheta) \geq \mathcal{U}(\tilde{\vartheta}) - \mathcal{U}(\vartheta) \quad \forall\vartheta$

Proof can be found in Suppl C. We conclude that when assumption (5) holds, the modified optimization landscape induced by curriculum learning has the same global optimum $\tilde{\vartheta}$ as the original problem. In addition, the modified optimization function in the parameter space ϑ has the property that the global maximum at $\tilde{\vartheta}$ is more pronounced.

We define an *ideal curriculum* to be the prior corresponding to the optimal hypothesis (or one of them, if not unique):

$$p_i = \frac{e^{-L_{\tilde{\vartheta}}(X_i)}}{C}, \quad C = \sum_{i=1}^{N} e^{-L_{\tilde{\vartheta}}(X_i)}$$

From[10] (4):

$$\mathcal{U}_p(\vartheta) = \mathcal{U}(\vartheta) + \frac{1}{C}\mathrm{Cov}[U_\vartheta, U_{\tilde{\vartheta}}]$$

The utility at the optimal point $\tilde{\vartheta}$ in parameter space is:

$$\mathcal{U}_p(\tilde{\vartheta}) = \mathcal{U}(\tilde{\vartheta}) + \frac{1}{C}\mathrm{Cov}[U_{\tilde{\vartheta}}, U_{\tilde{\vartheta}}] = \mathcal{U}(\tilde{\vartheta}) + \frac{1}{C}\mathrm{Var}[U_{\tilde{\vartheta}}] \tag{6}$$

In any other point

$$\begin{aligned}\mathcal{U}_p(\vartheta) &= \mathcal{U}(\vartheta) + \frac{1}{C}\mathrm{Cov}[U_\vartheta, U_{\tilde{\vartheta}}] \\ &\leq \mathcal{U}(\tilde{\vartheta}) + \frac{1}{C}\sqrt{\mathrm{Var}[U_\vartheta]\mathrm{Var}[U_{\tilde{\vartheta}}]}\end{aligned} \tag{7}$$

Note that if $\mathrm{Var}[U_\vartheta] = b\ \forall\vartheta$ for some constant b, then assumption (5) immediately follows from (7):

$$\mathcal{U}_p(\vartheta) \leq \mathcal{U}(\tilde{\vartheta}) + \frac{1}{C}\sqrt{b^2} = \mathcal{U}_p(\tilde{\vartheta}) \implies \tilde{\vartheta} = \arg\max_\vartheta \mathcal{U}_p(\vartheta)$$

Therefore

Corollary 1 *When using the ideal curriculum, Proposition 2 holds if the variance of the utility function is roughly constant in the relevant range of plausible parameter values.*

From (6) and (7) we can also conclude the following

Proposition 3 *When using the ideal curriculum*

$$\mathcal{U}_p(\tilde{\vartheta}) - \mathcal{U}_p(\vartheta) \geq \mathcal{U}(\tilde{\vartheta}) - \mathcal{U}(\vartheta) \quad \forall\vartheta : \mathrm{Cov}[U_\vartheta, U_{\tilde{\vartheta}}] \leq \mathrm{Var}[U_{\tilde{\vartheta}}]$$

This implies that the optimization landscape is modified to amplify the difference between the optimal parameters vector and all other parameter values whose covariance with the optimal solution (the covariance is measured between the induced prior vectors) is small, and specifically smaller than the variance of the optimum. In particular, this includes all parameters vectors which are uncorrelated (or negatively correlated) with the selected optimal parameter vector.

Discussion: Training a network based on its current hypothesis ϑ_t can be done in one of 2 ways: (i) using a prior which is monotonically increasing with the current utility, as suggested by self-paced learning; (ii) using a prior monotonically decreasing with the current utility, as suggested by hard data mining or boosting. Our analysis suggests that as long as the curriculum is positively correlated with the optimal utility, it can improve the learning; hence both strategies can be effective in different settings. It may even be possible to find a curriculum which is directly correlated with the optimal utility, and that outperforms both methods.

[10] Henceforth we will assume that $N \to \infty$, so that the estimation symbol $\hat{a}$ can be omitted.

Acknowledgements

This work was supported in part by a grant from the Israel Science Foundation (ISF), MAFAT Center for Deep Learning, and the Gatsby Charitable Foundations.

References

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pp. 173–182, 2016.

Avrahami, J., Kareev, Y., Bogot, Y., Caspi, R., Dunaevsky, S., and Lerner, S. Teaching by examples: Implications for the process of category acquisition. *The Quarterly Journal of Experimental Psychology: Section A*, 50(3): 586–606, 1997.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48. ACM, 2009.

Caruana, R. Learning many related tasks at the same time with backpropagation. In *Advances in neural information processing systems*, pp. 657–664, 1995.

Chang, H.-S., Learned-Miller, E., and McCallum, A. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Advances in Neural Information Processing Systems*, pp. 1002–1012, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. Ieee, 2009.

Fan, Y., Tian, F., Qin, T., Li, X.-Y., and Liu, T.-Y. Learning to teach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HJewuJWCZ.

Florensa, C., Held, D., Wulfmeier, M., Zhang, M., and Abbeel, P. Reverse curriculum generation for reinforcement learning. In *Conference on Robot Learning*, pp. 482–495, 2017.

Freund, Y., Schapire, R. E., et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pp. 148–156. Citeseer, 1996.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471, 2016.

Graves, A., Bellemare, M. G., Menick, J., Munos, R., and Kavukcuoglu, K. Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1311–1320. JMLR. org, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hunkins, F. P. and Ornstein, A. C. *Curriculum: Foundations, principles, and issues*. Pearson Education, 2016.

Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pp. 2309–2318, 2018.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Krueger, K. A. and Dayan, P. Flexible shaping: How learning in small steps helps. *Cognition*, 110(3):380–394, 2009.

Kumar, M. P., Packer, B., and Koller, D. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pp. 1189–1197, 2010.

Liu, S. and Deng, W. Very deep convolutional neural network based image classification using small training sample size. In *Pattern Recognition (ACPR), 2015 3rd IAPR Asian Conference on*, pp. 730–734. IEEE, 2015.

Murphy, R. R., Tadokoro, S., Nardi, D., Jacoff, A., Fiorini, P., Choset, H., and Erkmen, A. M. Search and rescue robotics. In *Springer handbook of robotics*, pp. 1151–1173. Springer, 2008.

Pavlov, P. I. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. *Annals of neurosciences*, 17(3):136, 2010.

Schein, A. I. and Ungar, L. H. Active learning for logistic regression: an evaluation. *Machine Learning*, 68(3):235–265, 2007.

Scholkopf, B., Sung, K.-K., Burges, C. J., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE transactions on Signal Processing*, 45 (11):2758–2765, 1997.

Selfridge, O. G., Sutton, R. S., and Barto, A. G. Training and tracking in robotics. In *IJCAI*, pp. 670–672, 1985.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shrivastava, A., Gupta, A., and Girshick, R. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 761–769, 2016.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Skinner, B. F. Reinforcement today. *American Psychologist*, 13(3):94, 1958.

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

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Weinshall, D., Cohen, G., and Amir, D. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International Conference on Machine Learning (ICML)*, volume 36, 2018.

Zaremba, W. and Sutskever, I. Learning to execute. *arXiv preprint arXiv:1410.4615*, 2014.

Zhang, D., Yang, L., Meng, D., Xu, D., and Han, J. Spftn: A self-paced fine-tuning network for segmenting objects in weakly labelled videos. In *IEEE CVPR*, pp. 4429–4437, 2017.