Curriculum Learning in Neural Networks: A Neurobiological Perspective on Model Performance

Freya Zhang
Cheriton School of Computer Science
University of Waterloo
freya.zhang@uwaterloo.ca

Toshani Nath
Cheriton School of Computer Science
University of Waterloo
t2nath@uwaterloo.ca

Abstract

Understanding how the order of data presentation affects the learning efficiency and performance of neural networks is an open area of research. Taking inspiration from the way humans and animals acquire knowledge, we hypothesize that models trained using curriculum learning - ordering data from easiest to hardest - will perform better, and biologically inspired networks will exhibit better convergence. To investigate this, we trained a perceptron on image data presented in forward curriculum (from clean to noisy) and random ordering. We varied learning rates and output activation functions (ReLU, softmax, sigmoid) to assess their impact. Curriculum learning consistently improved test accuracy over random ordering. The sigmoid network - which was more biologically plausible - achieved the highest accuracy and was the only one to show clear convergence of test loss. In contrast, ReLU and softmax models failed to converge, even with lower learning rates. However, curriculum learning did not show improvement in the perceptron - larger networks may be required to fully realize its advantages. These results support curriculum learning as a biologically inspired strategy that mirrors developmental learning processes in the brain and offers a promising direction for future work with more biologically realistic models.

Introduction

Imagine handing a university-level calculus textbook to a baby and expecting them to understand it. Just as humans need to build foundational knowledge before tackling more complex concepts, machine learning models - often created taking inspiration from biological processes - may also benefit from a similar learning progression. This concept is known as curriculum learning (CL), which proposes that training models by gradually increasing the complexity of the data can lead to more efficient learning and improved performance [6]. Similar to how humans are presented with information ordered simplest to hardest through the education system, we can organize data in a more meaningful order when training models [6].

We hypothesize that training with a forward curriculum (easy-to-hard) will result in better performance and faster convergence than random data ordering. We also hypothesize that more biologically plausible models will benefit more from curriculum learning and converge faster. Originally formalized by Bengio et al., curriculum learning has demonstrated its effectiveness across domains including computer vision, natural language processing, healthcare prediction, and reinforcement learning [8] [2]. For example, in noisy or complex settings, focusing on easier examples early in training can improve accuracy and robustness by avoiding overfitting [8]. CL has also been applied to spiking neural networks (SNNs), achieving greater results than "current state-of-the-art experimental results" on all the data sets involved [7]. However, despite the biological inspiration behind neural networks, most models continue to be trained on randomly shuffled data [6].

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

In this study, we focused on shallow models, beginning with a perceptron, then transitioning to a multilayer perceptron (MLP) with varied architectures. We created the training curricula by adding increasing amounts of noise to a base dataset which allows for varying levels of difficulty. Each model was tested on two training data orderings - random and forward - with additional variation in learning rate and activation functions (ReLU, softmax, sigmoid). Through this, we aim to see not only how CL affects performance, but also whether biologically realistic models respond differently than artificial neuron models.

An interesting aspect of our experiment is that although CL was introduced over a decade ago, it is still not widely adopted. To our knowledge, though CL has been studied in deep networks like CNNs and ResNets, relatively fewer studies have looked into its impact on simpler and more explainable models. Our results aim to shed light on whether training strategies like CL that are inspired by biology are beneficial for simple models, especially when paired with models that also mimic biological processes.

Methods

We investigate the effects of curriculum learning on multilayer perceptrons (MLPs) by introducing two primary interventions: (1) ordering training data by difficulty, and (2) varying the activation function used in the model's output layer[1]. Our goal is to examine how these factors influence convergence speed and final model performance, particularly under biologically inspired settings.

Our baseline model is a fully connected MLP with three hidden layers. Each hidden layer uses 1000, 500, and 50 neurons respectively, followed by non-linear activation. The input dimension corresponds to flattened 28×28 MNIST images (784 features), and the output layer produces class scores for ten digit classes. We vary the final activation function across three model variants: ReLU, Sigmoid, and Softmax. While ReLU and Softmax are common in machine learning, Sigmoid is more biologically plausible, modeling the saturating non-linear behaviour observed in actual neurons.

Each model variant is trained under two data ordering regimes. In the curriculum learning condition, training data is presented in a progression from easy to hard. We simulate difficulty (5 stages) by adding increasing levels of noise (0, 0.1, 0.2, 0.3, 0.4) to the MNIST dataset: clean images are considered easy, while high-noise images are harder to classify. In the randomized baseline, the same dataset is shuffled, removing any meaningful order.

To evaluate performance, we measure four primary metrics: training and testing accuracy, training and testing loss, the number of epochs until test accuracy plateaus, and the number of epochs until test loss converges or diverges. Convergence is defined as the point at which a metric (e.g., accuracy or loss) shows minimal or no improvement for a fixed number of epochs. Each experiment is repeated across five independent trials to ensure statistical robustness, and we report both mean and standard deviation across runs.

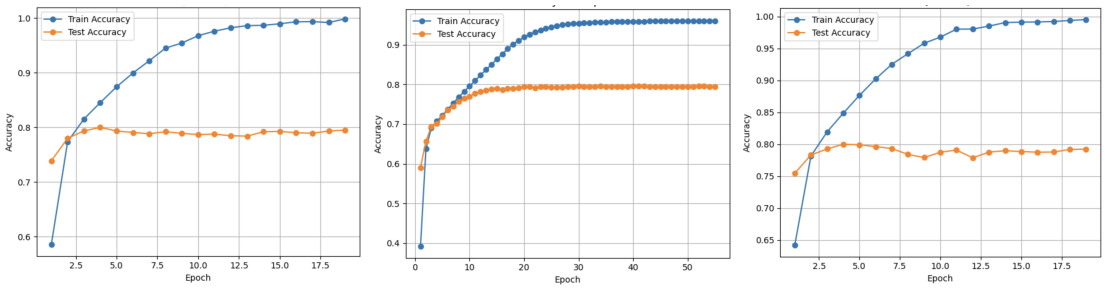
While the model architecture is not inherently more complex than standard MLPs, our study introduces variation through both biologically motivated architectural changes (i.e., the use of the Sigmoid function) and a biologically inspired training strategy (curriculum learning). These interventions allow us to explore how aligning learning mechanisms with biological principles might affect the efficiency and robustness of artificial neural networks.

Results

Train and Test Accuracy

We first examined model performance under randomized training order (no curriculum learning) across three output activation functions: ReLU, Sigmoid, and Softmax. Each model was trained on noisy MNIST data for a maximum of 100 epochs. Note that early stopping was implemented to save training resources, and training is stopped if the test accuracy does not improve from the best test accuracy for more than 15 epochs. We examined training and test accuracy for the ReLU, Sigmoid, and Softmax output models under random data ordering.

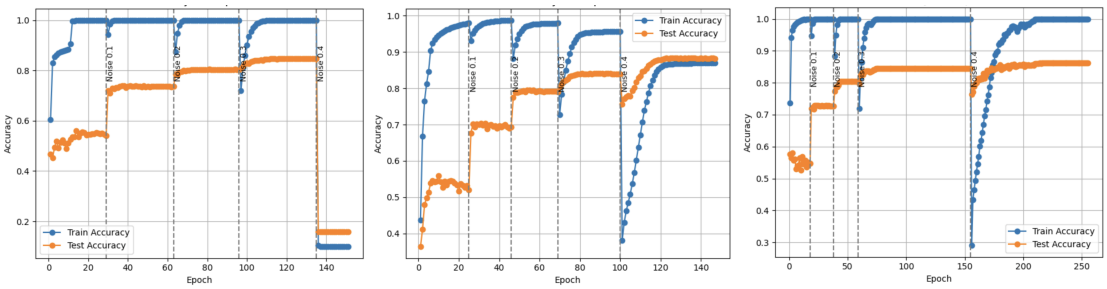
[1]Code available at: `https://github.com/ToshaniN/CurriculumLearning_ToshaniNath_FreyaZhang`



(a) ReLU (b) Sigmoid (c) Softmax

Figure 1: Training and test accuracy under no curriculum learning for ReLU, Sigmoid, and Softmax output activations (learning rate = 0.03).

The ReLU and Softmax models showed rapid increases in training accuracy, reaching nearly 100% within 20 epochs, while test accuracy plateaued early around an average of 79% and remained flat, indicating overfitting (Figures 1a and 1c). The Sigmoid model exhibited slower but steadier learning (Figure 1b). Test accuracy improved gradually, converging around 79% by epoch 30 with a smaller gap between train and test performance. This suggests better generalization, likely due to its smoother, biologically inspired activation function.



(a) ReLU (b) Sigmoid (c) Softmax

Figure 2: Training and test accuracy under curriculum learning for ReLU, Sigmoid, and Softmax output activations (learning rate = 0.03).

When trained using curriculum learning (progressing from low to high noise), all models showed clear staged improvements in both training and test accuracy aligned with the difficulty transitions (Figure 2). The ReLU model achieved near-perfect training accuracy early on but failed to generalize to higher-noise levels, with test accuracy dropping sharply after the final transition. This suggests sensitivity to abrupt changes in data complexity.

The Sigmoid model, while slower to converge initially, exhibited more stable and consistent improvements across all difficulty stages. Its final test accuracy remained high even after transitioning to noisier data, indicating better generalization under curriculum.

The Softmax model performed similarly to Sigmoid, with high training accuracy throughout and a steady increase in test accuracy across stages. Unlike ReLU, neither Sigmoid nor Softmax experienced a collapse in test performance after transitioning to harder data, suggesting that curriculum learning helps mitigate overfitting for these models.

Figure 3 compares final test accuracy across all models and training regimes. Both Sigmoid and Softmax models showed clear improvement when trained with curriculum learning, increasing from 79.5% to 88.3% and 79.2% to 86.2%, respectively. In contrast, the ReLU model performed worse under curriculum learning, with accuracy dropping from 79.5% to just 15.8%. This drop occurred during the final training stage with the highest noise level (0.4), where the model's performance collapsed. Notably, the best test accuracy ReLU achieved during training was around 84%, suggesting it was unable to adapt to increasingly difficult data despite strong early performance.

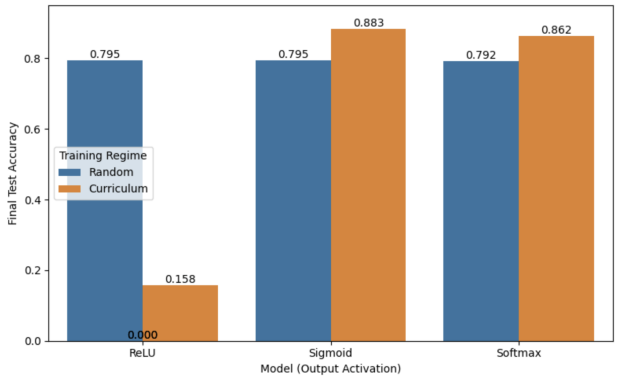
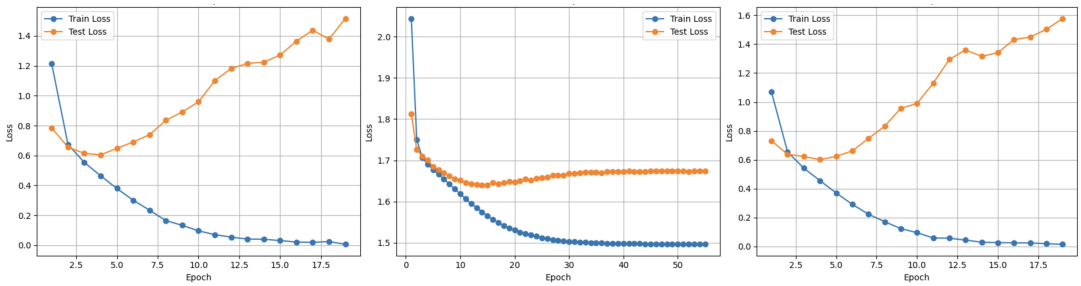


Figure 3: Final test accuracies for models ReLU, Sigmoid, and Softmax as their final layer activation function.

Train and Test Loss

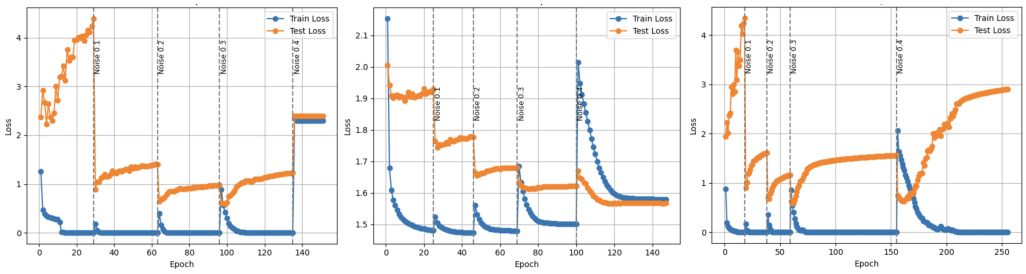
Next, we will focus on the training and test loss. When trained with a learning rate of 0.03 on



(a) ReLU (b) Sigmoid (c) Softmax

Figure 4: Training and test loss under no curriculum learning for ReLU, Sigmoid, and Softmax output activations (learning rate = 0.03).

randomly ordered data, only the Sigmoid model demonstrated clear convergence in both training and test loss (Figure 4). In contrast, both the ReLU and Softmax models showed diverging test loss curves despite steadily decreasing training loss (Figures 4a and 4c), indicating severe overfitting. These results reinforce earlier findings: ReLU and Softmax models are more prone to memorizing training data, while Sigmoid—though slower to optimize—achieves more stable and generalizable learning under high learning rates without curriculum support.



(a) ReLU (b) Sigmoid (c) Softmax

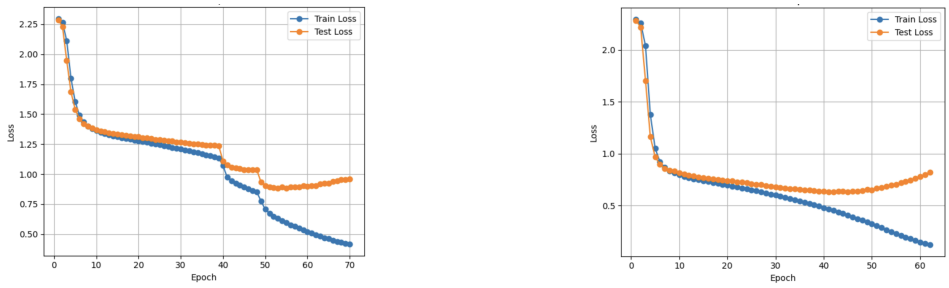
Figure 5: Training and test loss under curriculum learning for ReLU, Sigmoid, and Softmax output activations (learning rate = 0.03).

Under curriculum learning with a learning rate of 0.03, all models exhibited stage-wise transitions in loss corresponding to increasing noise levels (Figure 5). For the ReLU and Softmax models,

test loss initially decreased but increased again in the later half of each stage, suggesting instability, especially when high-difficulty data was introduced. In contrast, the Sigmoid model showed consistent convergence across all stages, with both training and test loss gradually decreasing. Notably, more epochs were required during higher-difficulty stages across all models, indicating that the model relied more on challenging samples to refine its weights. This supports the idea that later, harder data contributes more meaningfully to learning under curriculum regimes.

Lower Learning Rate

To assess whether previous divergence was due to an overly high learning rate, we retrained the ReLU and Softmax models without curriculum learning using a reduced learning rate of 0.001. As shown in Figure 6, both models exhibited convergent behaviour: test loss decreased steadily before flattening.



(a) ReLU (b) Softmax

Figure 6: Training and test loss under no curriculum learning for ReLU and Softmax output activations (learning rate = 0.001).

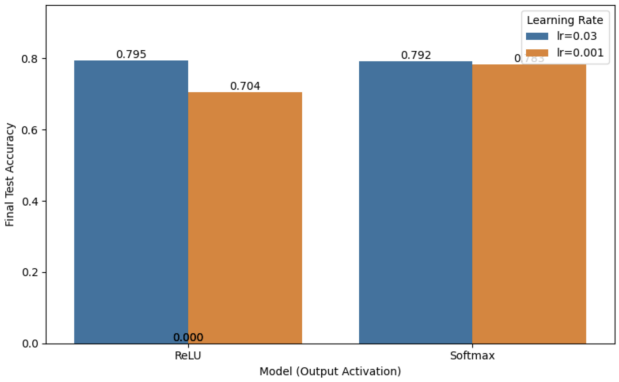
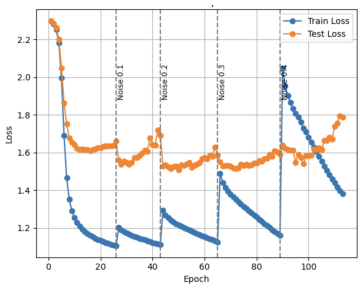


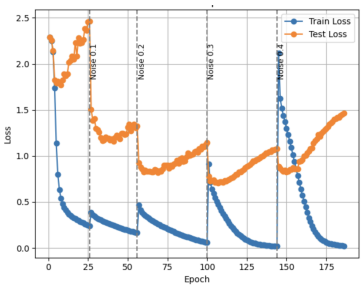
Figure 7: Final accuracies for models ReLU and Softmax as their final layer activation function, trained under no curriculum learning.

Although lowering the learning rate improved loss convergence for ReLU and Softmax models without curriculum learning, it did not improve final test accuracy—in fact, performance slightly dropped (Figure 7). This suggests that while a lower learning rate stabilizes training, it can slow optimization or limit generalization capacity, especially in the absence of structured data presentation. Convergence alone, therefore, does not guarantee improved performance.

Even with a reduced learning rate of 0.001, both ReLU and Softmax models showed divergence in test loss when trained using curriculum learning (Figure 8). While training loss decreased reliably across all stages, test loss increased, especially after the final transition to the noisiest data (0.4 noise). This suggests that curriculum learning, despite its staged structure, did not prevent overfitting in these models and may have amplified sensitivity to late-stage high-difficulty samples. These results contrast with the stable convergence seen in the Sigmoid model, highlighting that curriculum learning can interact negatively with certain activation functions, particularly under high noise conditions.



(a) ReLU

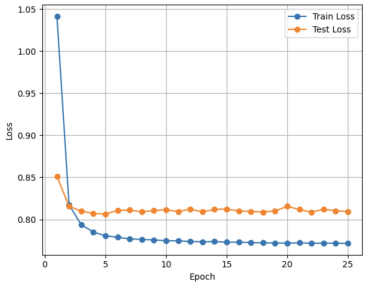


(b) Softmax

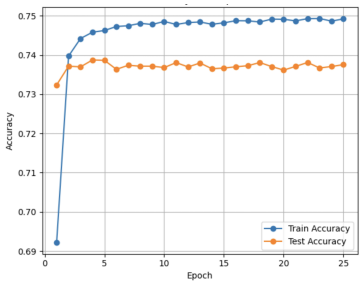
Figure 8: Training and test loss under curriculum learning for ReLU and Softmax output activations (learning rate = 0.001).

Perceptron

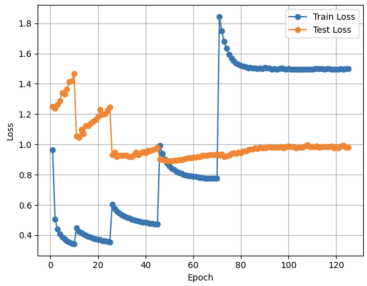
We also evaluated curriculum learning on a single-layer Perceptron (784 neurons). Without curriculum, the model achieved a final test accuracy of 73.75%, with both training and test loss converging smoothly (Figure 9, top). In contrast, when trained using curriculum learning, final test accuracy decreased to 67.95% despite steady improvements during early stages (Figure 9, bottom).



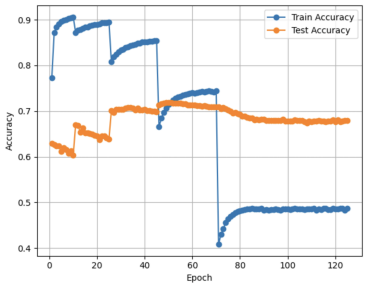
(a) No Curriculum - Loss



(b) No Curriculum - Accuracy



(c) Curriculum - Loss



(d) Curriculum - Accuracy

Figure 9: Training and test performance for Perceptron with and without curriculum learning (learning rate = 0.03).

These results suggest that curriculum learning may be less effective or even detrimental for simpler models like the Perceptron, which lack the representational capacity to benefit from staged difficulty progression.

Discussion

Consistent with the claims of Soviany et al. (2022) and Wang et al. (2021), we found that curriculum learning significantly improved model performance compared to randomly ordered data. Across all the models tested, the forward curriculum resulted in higher test accuracies - especially in the sigmoid model which reached 88% test accuracy compared to 86% on the softmax model, and 84% on the ReLU model (however the ReLU model accuracy dropped to 15% after reaching the highest difficulty data). These test accuracies overall showed that the models improved over training with randomly shuffled data, which produced around 79% accuracy for all models. This aligns with the idea that training on simpler examples early on helps avoid overfitting and leads to more robust generalization in the future, especially in noisy and complex environments [8].

We also saw that the sigmoid model, which can be considered the most biologically realistic, showed the best performance and convergence. Unlike the ReLU and softmax models, which required a lower learning rate (0.001), the sigmoid model was able to achieve convergence with a learning rate of 0.03. These results suggest that biologically inspired models may benefit more from curriculum learning, supporting our hypothesis, and consistent with the findings of Tang et. al. (2023).

We noticed that extremely simple models like the perceptron do not appear to benefit much from curriculum learning - the test accuracy produced by the forward curriculum was lower. This was somewhat contradictory to our expectations, especially since another study had listed the perceptron among the machine learning models that benefited from CL [1]. This result may however align with another literature cited earlier which suggested that CL is more helpful in noisy and complex settings [8]. It is possible that the perceptron's limited capacity may have allowed it to learn the easier data early on, but made it unable to handle the harder data later in the training. Differences in the image datasets that were used to conduct the experiment may also have had an impact, other studies may have used data with different dimensionalities. This result can support the argument that forward curriculum is not always beneficial, as the easier samples can slow down the training process by offering small gradients when they're learnt [3]. The opposite strategy, ordering data from hardest-to-easiest, has proven beneficial in certain situations [3].

Overall summary of observations

Random Training (lr=0.03):

- ReLU and Softmax overfit quickly with training accuracy $\approx$100% and testing accuracy $\approx$79%.
- Sigmoid learned steadily with similar test accuracy but slightly better generalization.

Curriculum Learning (lr=0.03):

- Sigmoid and Softmax improved test accuracy — 88.3% and 86.2% respectively.
- ReLU achieved 84% but collapsed under high noise with 15.8% test accuracy.

Loss Trends (lr=0.03):

- Sigmoid showed consistent convergence.
- ReLU and Softmax diverged in test loss under both training methods, especially with curriculum.

Lower Learning Rate (0.001):

- Stabilized ReLU and Softmax loss curves under no CL but not under curriculum learning.
- No accuracy improvement observed.

Perceptron:

- Performed better without curriculum — 73.75% accuracy with random training vs 67.95% with curriculum — showing limited benefit of CL for simpler models.

Further developments

While our current work uses manually added noise to define data difficulty, future work could explore more dynamic approaches like automatic curriculum learning strategies. One promising direction is

implementing confidence-aware loss, where the model dynamically adjusts sample weights during training based on its prediction confidence, placing more emphasis on "easy" samples early on [7]. Though we simulated a similar effect by varying both noise levels and training epochs per difficulty level, a dynamic, feedback-driven approach may be more effective. Alternatively, a teacher-student framework could also allow for real-time scheduling based on the student's learning progress.

We did our study using random and easy-to-hard ordered data. Another curriculum we can explore in the future to compare results against is the anti-curriculum, where training examples are sorted in hard-to-easy ordering, as in, in decreasing order of difficulty [4]. This may shed further insights on the impact of data orderings.

We are interested in implementing curriculum learning in spiking neural networks (SNNs), which allow for greater biological plausibility and have shown promising results in recent studies [7]. Similar to the study conducted by Hunsberger et. al., which showed that improved performance can be achieved by integrating LIF neurons, we would also like to try using LIF neurons in an SNN [5]. Combining CL with SNNs could further enhance learning efficiency while providing insights into how biologically inspired models acquire knowledge. We are also interested in exploring CL in deep convolutional neural networks (CNNs), which have also shown improved performance when trained with curriculum derivatives Self Paced Learning (SPL) and Self Paced Learning with Diversity (SPLD) [1]. As these are less biologically plausible than spiking models, it would be quite insightful to compare their results with those of an SNN trained with CL.

References

[1] Vanya Avramova. Curriculum learning with deep convolutional neural networks. Master's thesis, KTH, School of Computer Science and Communication (CSC), 2015.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery.

[3] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. Active bias: Training more accurate neural networks by emphasizing high variance samples, 2018.

[4] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. *CoRR*, abs/1904.03626, 2019.

[5] Eric Hunsberger and Chris Eliasmith. Spiking deep networks with LIF neurons. *CoRR*, abs/1510.08829, 2015.

[6] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *CoRR*, abs/2101.10382, 2021.

[7] Lingling Tang, Jiangtao Hu, Hua Yu, Surui Liu, and Jielei Chu. Learning spiking neural network from easy to hard task, 2023.

[8] Xin Wang, Yudong Chen, and Wenwu Zhu. A comprehensive survey on curriculum learning. *CoRR*, abs/2010.13166, 2020.