Eurographics Workshop on Visual Computing for Biology and Medicine (2023)
T. Höllt and D. Jönsson (Editors)

CDF-Based Importance Sampling and Visualization for Neural Network Training

Supplementary material

Alex Knutsson†, Jakob Unnebäck†, Daniel Jönsson, and Gabriel Eilertsen

Linköping University, Sweden

1. Introduction

This supplementary material provides results on two additional common datasets, to provide information on the behavior of the proposed CDF-based importance sampling on a wider range of classification tasks.

2. Supplementary results

Here, we report additional results on the MNIST and CIFAR-10 datasets, as well as the optimal thresholds selected for the starting criteria for each dataset.

2.1. Datasets

The method is evaluated on four datasets with varying complexity. In addition to the Two Circles and Camelyon datasets, we present results for the MNIST and CIFAR-10 datasets here. Together, the four datasets present a gradual increase in complexity, ranging from a synthetic toy dataset to medical imaging.

MNIST [LCB98] consists of 60K images in 28×28 pixels resolution, with 10 different classes of hand-written digits between 0-9. For testing of the final accuracy, there is an additional 10K images.

CIFAR-10 [KNH] contains 10 different classes of objects. There are 50K training images in 32×32 pixels resolution. The classification problem is significantly more challenging compared to MNIST. For testing of the final accuracy, there is an additional 10K images.

2.2. Training setup

- The MNIST network uses a learning rate of 0.01. It consists of two convolutional layers, two max-pooling layers, and two linear layers. The total number of weights is 21,840. The loss function used is cross entropy.
- The CIFAR-10 network uses a learning rate of 0.001. It consists of three convolutional layers, three max-pooling layers, one batch norm layer, and two linear layers. The total number of weights is 126,794. The loss function used is cross entropy.

Other training settings are the same as for the Two Circles and Camelyon datasets reported in the main paper.

Table 1: *Summary of the thresholds resulting in the highest accuracy when training the corresponding networks from scratch. Lower thresholds mean that it is beneficial to start importance sampling late in the training, while a high threshold means that it is beneficial to start importance sampling earlier.*

Dataset	Best Threshold
Two Circles	0.66
MNIST	0.18
CIFAR-10	0.34
Camelyon	0.74

2.3. Optimal starting criteria

The threshold is used to formulate the starting criteria for when to switch from uniform sampling to importance based sampling. In Figure 1, the results when using different selections of thresholds are presented. Results for Two Circles and Camelyon are also included to facilitate comparisons between the four datasets.

The thresholds generating the best accuracy for each dataset are summarized in Table 1.

2.4. Importance Sampling Strategy Comparison

Based on the analysis of when to start importance sampling, we used the best thresholds from Table 1 to evaluate the performance of the different techniques. The result of this evaluation can be seen in Figure 2. For the Two Circles (top row) and MNIST (second row), we can see that the *highest loss* and *highest loss CDF* perform best. The *gradient norm* sampling strategy is too slow to compute, and the additional computations needed for the CDF-based method

† Authors contributed equally to this work

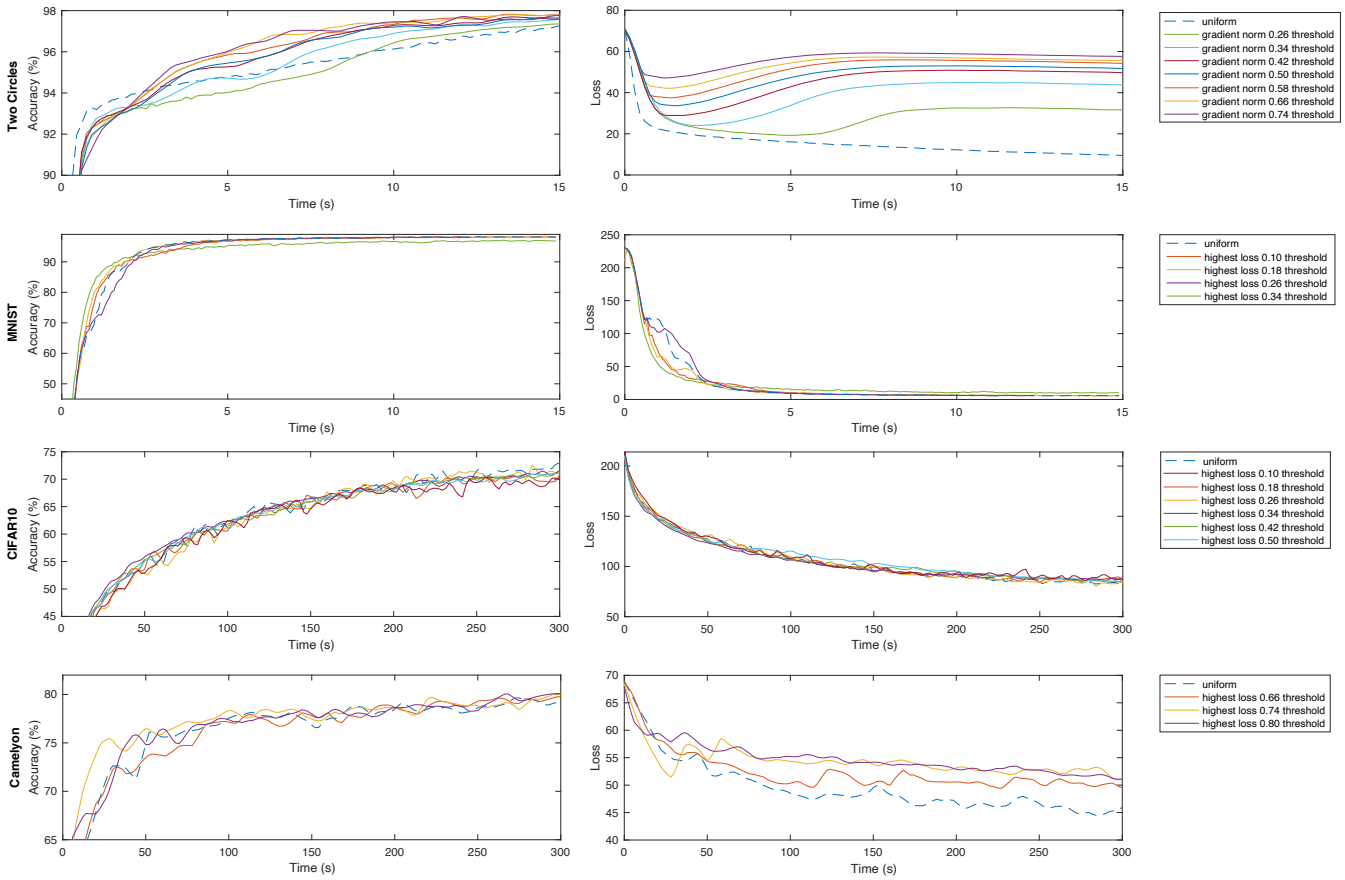

Figure 1: *Test accuracy and training loss graph depicting the performance for different thresholds determining when to start importance sampling on different dataset over time. Results for Two Circles and Camelyon are included for facilitating comparisons between datasets. Lower threshold are advantageous for less complex datasets, such as Two Circles and MNIST, while higher threshold yields superior performance on more complex data, such as CIFAR-10 and Camelyon.*

are too costly to introduce benefits for these small models and data sets.

None of the importance sampling methods outperform uniform sampling for CIFAR-10 (third row), but here it can be seen that the CDF-based importance sampling greatly improves the model accuracy compared to previous work that chose samples based on the highest loss or gradient norm.

For the significantly larger model trained on the Camelyon dataset (bottom row), we can see that the CDF-based sampling result in significantly higher accuracy compared to not using the CDF-based sampling for the *highest loss* and *gradient norm*. The *gradient norm CDF* is best at the beginning of the training but is later overtaken by the *highest loss CDF* strategy.

We also performed experiments based on batch number instead of measuring time in seconds. These experiments show which method that is best in case computational requirements are not considered. As such, they are mostly of theoretical interest unless a different implementation improves the importance computation time. Here, the *gradient norm* strategy produced similar model accuracy across batch number compared to *highest loss*. Thus, we can conclude that *gradient norm* is only worse due to its higher computational cost.

2.5. Important Training Samples

The most and least important samples for the training have been evaluated for MNIST (Figure 3) and CIFAR-10 (Figure 4). As can be seen from these figures, the most important samples for the training tend to be more difficult and more unique compared to the least important samples, which tend to be easy to classify. Important training samples are generally the ones with with unique features, making them stand out from other samples of the same class. Examples include slanted numbers in MNIST (Figure 3) and a plane photographed top down with ground on the background in CIFAR-10 (Figure 4).

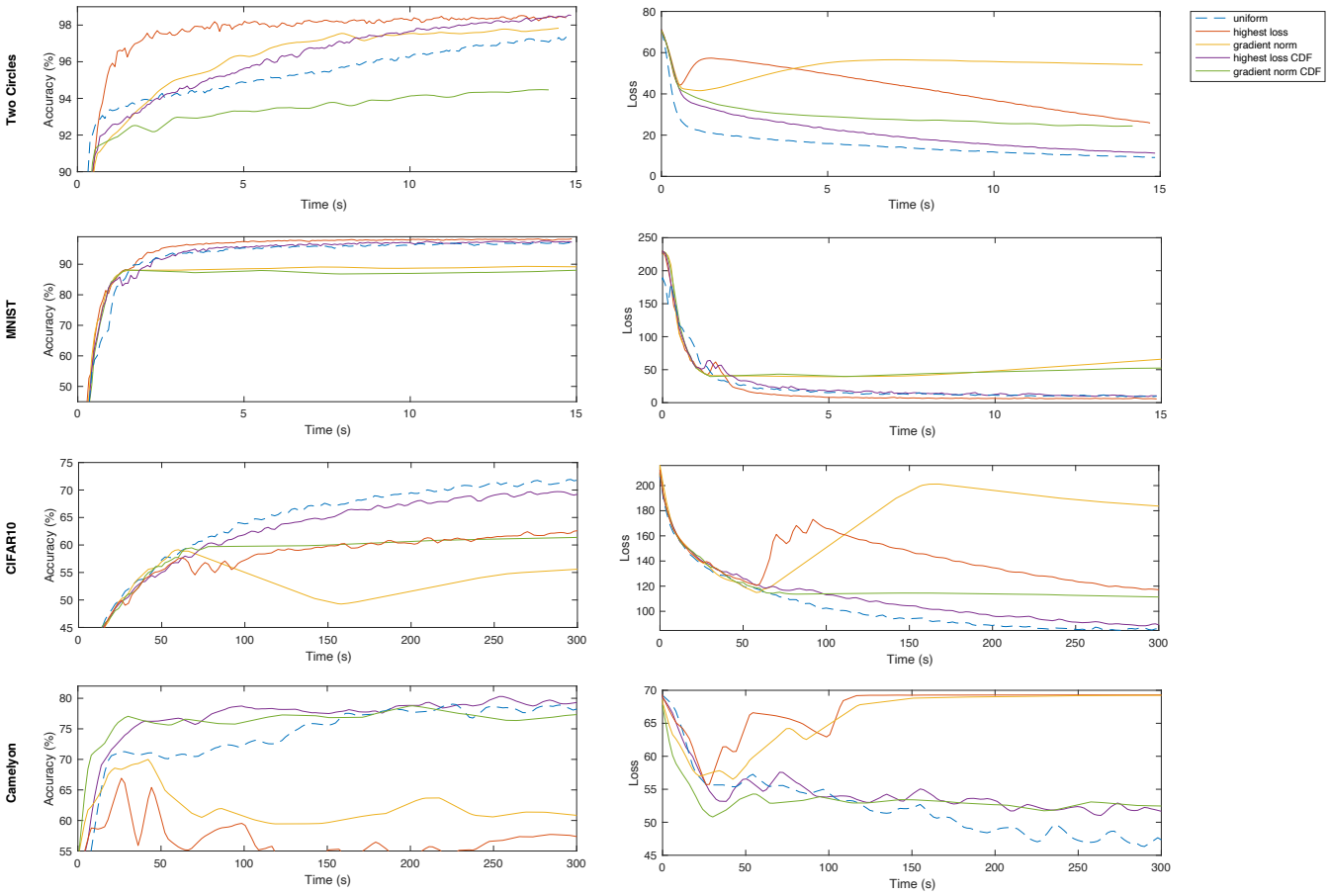

Figure 2: *Graphs depicting the test accuracy and training loss for different sampling methods over all datasets. A higher accuracy than the dashed line, representing uniform sampling, means better performance and efficiency. Results for Two Circles and Camelyon are included for facilitating comparisons between datasets.*

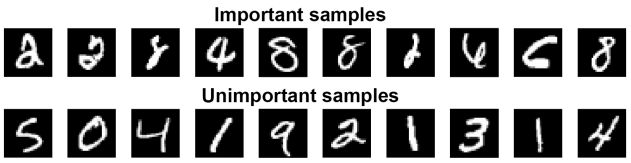

Figure 3: *The most and least important training samples from the MNIST dataset. Difficult to classify samples, e.g., skewed and uniquely shaped digits, are deemed important while easy samples, having clear strokes, are deemed unimportant.*

Figure 4: *The most and least important training samples from the CIFAR-10 dataset. Important samples primarily display cluttered backgrounds or off-centered objects. In contrast, the unimportant samples feature more isolated and centered objects.*

References

[KNH] KRIZHEVSKY A., NAIR V., HINTON G.: Cifar-10 (canadian institute for advanced research). URL: http://www.cs.toronto.edu/~kriz/cifar.html. 1

[LCB98] LECUN Y., CORTES C., BURGES C.: The MNIST database of handwritten digits, 1998. URL: http://yann.lecun.com/exdb/mnist/. 1