
Differentiable Sorting Networks for Scalable Sorting and Ranking Supervision

Felix Petersen¹ Christian Borgelt² Hilde Kuehne^{3,4} Oliver Deussen¹

Abstract

Sorting and ranking supervision is a method for training neural networks end-to-end based on ordering constraints. That is, the ground truth order of sets of samples is known, while their absolute values remain unsupervised. For that, we propose differentiable sorting networks by relaxing their pairwise conditional swap operations. To address the problems of vanishing gradients and extensive blurring that arise with larger numbers of layers, we propose mapping activations to regions with moderate gradients. We consider odd-even as well as bitonic sorting networks, which outperform existing relaxations of the sorting operation. We show that bitonic sorting networks can achieve stable training on large input sets of up to 1024 elements.

1. Introduction

Sorting and ranking as the ability to score elements by their relevance is an essential task in numerous applications. It can be used for choosing the best results to display by a search engine or organize data in memory. Starting in the 1950s, sorting networks have been presented to address the sorting task (Knuth, 1998). Sorting networks are sorting algorithms with a fixed execution structure, which makes them suitable for hardware implementations, e.g., as part of circuit designs. They are oblivious to the input, i.e., their execution structure is independent of the data to be sorted. As such hardware implementations are significantly faster than conventional multi-purpose hardware, they are of interest for sorting in high performance computing applications (Govindaraju et al., 2006). This motivated the optimization of sorting networks toward faster networks with fewer layers, which is a still-standing problem (Bidlo & Dobeš, 2019). Note that, although the name is similar, sorting networks are *not* neural networks that perform sorting.

Recently, the idea of end-to-end training of neural networks with sorting and ranking supervision by a continuous relaxation of the sorting and ranking functions has been presented by Grover et al. (2019). Sorting supervision means the ground truth order of some samples is known while their absolute values remain unsupervised. As the error has to be propagated in a meaningful way back to the neural network, it is necessary to use a continuous and continuously differentiable sorting function. Several such differentiable relaxations of the sorting and ranking functions have been introduced, e.g., by Adams & Zemel (2011), Grover et al. (2019), Cuturi et al. (2019), and Blondel et al. (2020). For example, they enable training a CNN based on ordering and ranking information instead of absolute ground truth values. As sorting a sequence of values requires finding the respective ranking order, we use the terms “sorting” and “ranking” interchangeably.

In this work, we propose to combine traditional sorting networks and differentiable sorting functions by presenting smooth differentiable sorting networks.

Sorting networks are conventionally non-differentiable as they use min and max operators for conditionally swapping elements. Thus, we relax these operators by building on the softmin and softmax operators. However, due to the nature of the sorting network, values with large as well as very small differences are compared in each layer. Comparing values with large differences causes vanishing gradients, while comparing values with very small differences can modify, i.e., blur, values as they are only partially swapped. This is because softmin and softmax are based on the logistic function which is saturated for large inputs but also returns a value close to the mean for inputs that are close to each other. Based on these observations, we propose an activation replacement trick, which avoids vanishing gradients as well as blurring. That is, we modify the distribution of the differences between compared values to avoid small differences close to 0 as well as large differences.

To validate the proposed idea and to show its generalization, we evaluate two sorting network architectures, the odd-even as well as the bitonic sorting network. The idea of odd-even sort is to iteratively compare adjacent elements and swap pairs that are in the wrong order. The method alternately compares all elements at odd and even indices with their

¹University of Konstanz, Germany ²University of Salzburg, Austria ³University of Frankfurt, Germany ⁴MIT-IBM Watson AI Lab. Correspondence to: Felix Petersen <felix.petersen@uni.kn>.

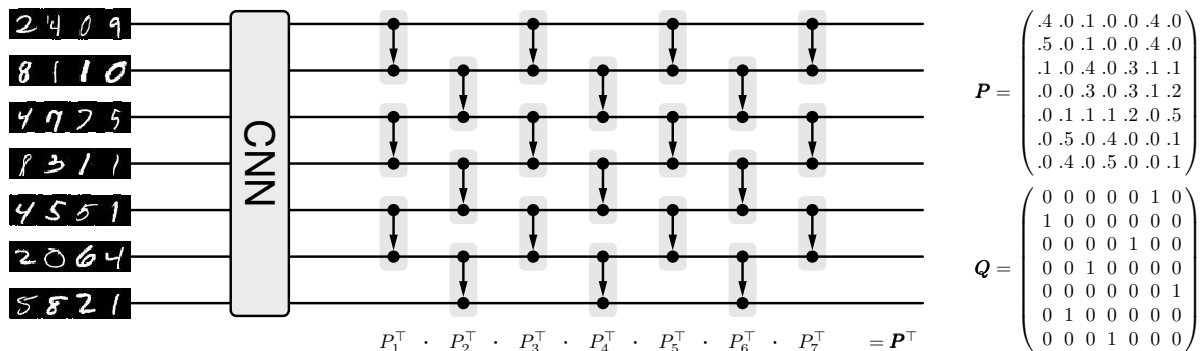


Figure 1: Overview of the system for training with sorting supervision. Left: input images are fed separately / independently into a Convolutional Neural Network (CNN) that maps them to scalar values. Center: the odd-even sorting network sorts the scalars by parallel conditional swap operations. Right: the sorting network produces a differentiable permutation matrix P which can then be compared to the ground truth permutation matrix Q using binary cross-entropy to produce the training loss. By propagating this error backward through the sorting network, we can train the CNN.

successors. To make sure that the smallest (or greatest) element will be propagated to its final position for any possible input of length n , we need n exchange layers. An odd-even network is displayed in Figure 1 (center). Odd-even networks can be seen as the most generic architectures, and are mainly suitable for small input sets as their number of layers directly depends on the number of elements to be sorted.

Bitonic sorting networks (Batcher, 1968) use bitonic sequences to sort based on the Divide-and-Conquer principle and allow sorting in only $\mathcal{O}(\log^2 n)$ parallel time. Bitonic sequences are twice monotonic sequences, i.e., they consist of a monotonically increasing and monotonically decreasing sequence. Bitonic sorting networks recursively combine pairs of monotonic sequences into bitonic sequences and then merge them into single monotonic sequences. Starting at single elements, they eventually end up with one sorted monotonic sequence. With the bitonic architecture, we can sort large numbers of input values as we only need $\frac{\log_2 n \cdot (\log_2 n + 1)}{2}$ layers to sort n inputs. As a consequence, the proposed architecture provides good accuracy even for large input sets and allows scaling up sorting and ranking supervision to large input sets of up to 1024 elements.

Following Grover et al. (2019) and Cuturi et al. (2019), we benchmark our continuous relaxation of the sorting function on the four-digit MNIST (LeCun et al., 2010) sorting supervision benchmark. To evaluate the performance in the context of a real-world application, we apply our continuous relaxation to the multi-digit images of the Street View House Number (SVHN) data set. We compare the performance of both sorting network architectures and evaluate their characteristics under different conditions. We show that both differentiable sorting network architectures outperform existing continuous relaxations of the sorting function on the four-digit MNIST sorting benchmark and also per-

form well on the more realistic SVHN benchmark. Further, we show that our model scales and achieves performance gains on larger sets of ordered elements and confirm this up to $n = 1024$ elements.

An overview of the overall architecture is shown in Figure 1.

In addition, we apply our method to top- k classification.

2. Related work

Sorting Networks The goal of research on sorting networks is to find optimal sorting networks, i.e., networks that can sort an input of n elements in as few layers of parallel swap operations as possible. Initial attempts to sorting networks required $\mathcal{O}(n)$ layers, each of which requires $\mathcal{O}(n)$ operations (examples are bubble and insertion sort (Knuth, 1998)). With parallel hardware, these sorting algorithms can be executed in $\mathcal{O}(n)$ time. Further research led to the discovery of the bitonic sorting network (aka. bitonic sorter) which requires only $\mathcal{O}(\log^2 n)$ layers (Knuth, 1998; Batcher, 1968). Using genetic and evolutionary algorithms, slightly better optimal sorting networks were found for specific n (Bidlo & Dobeš, 2019; Baddar & Batcher, 2012). However, these networks do not exhibit a simple, regular structure. Ajtai, Komlós, and Szemerédi (Ajtai et al., 1983) presented the AKS sorting network which can sort in $\mathcal{O}(\log n)$ parallel time, i.e., using only $\mathcal{O}(n \log n)$ operations. However, the complexity constants for the AKS algorithm are to date unknown and optimistic approximations assume that it is faster than bitonic sort if and only if $n \gg 10^{80}$. Today, sorting networks are still in use, e.g., for fast sorting implementations on GPU accelerated hardware as described by Govindaraju et al. (2006) and in hybrid systems as described by Gowanlock & Karsin (2019). Based on the bitonic sorting network, Lim & Wright (2016) propose a coordinate descent algorithm to solve hard permutation problems.

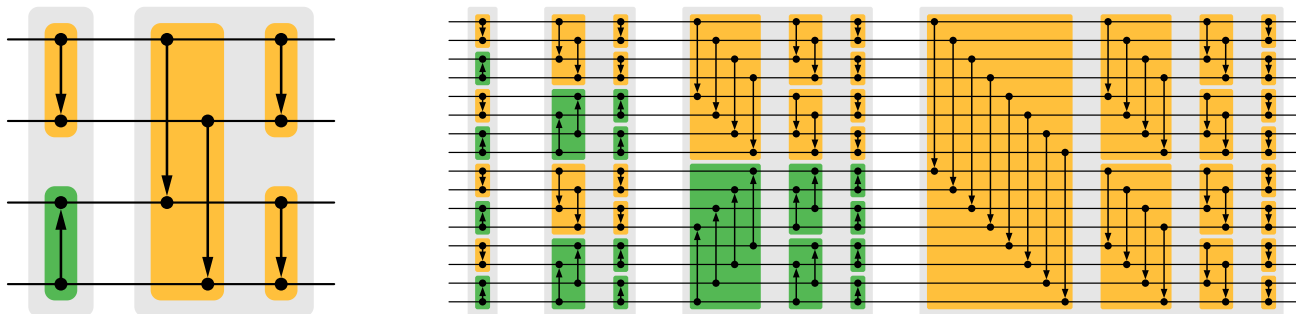


Figure 2: Bitonic sorting networks for 4 and 16 lanes, consisting of bitonic merge blocks (colored). Arrows pointing toward the maximum.

Neural Networks that Sort In the past, neural networks that sort have been proposed, e.g., by [Ceterchi & Tomescu \(2008\)](#), who proposed simulating sorting networks with spiking neural P systems. Spiking neural P systems are predecessors of current spiking networks, a form of computational models inspired by biological neurons. This was later adapted by [Metta & Kelemenova \(2015\)](#) for a spiking neural P system with anti-spikes and rules on synapses.

[Graves et al. \(2014\)](#) raised the idea of integrating sorting capabilities into neural networks in the context of Neural Turing Machines (NTM). The NTM architecture contains two basic components: a neural network controller based on an LSTM and a memory bank with an attention mechanism, both of which are differentiable. The authors use this architecture to sort sequences of binary vectors according to given priorities. [Vinyals et al. \(2016\)](#) address the problem of the order of input and output elements in LSTM sequence-to-sequence models by content-based attention. To show the effect of the proposed model, they apply it to the task of sorting numbers and formulate the task of sorting as an instance of the set2seq problem. [Mena et al. \(2018\)](#) introduce the Gumbel-Sinkhorn, a Sinkhorn-operator-based analog of the Gumbel-Softmax distribution for permutations. They evaluate the proposed approach, i.a., on the task of sorting up to 120 numbers. Note that these architectures learn to sort, while sorting networks and differentiable sorting functions sort provably correct. These methods allow sorting input values, as an alternative to classical sorting algorithms, but not training with sorting supervision because they are not differentiable.

Differentiable Sorting Closest to our work are differentiable sorting algorithms, which can be used to train neural networks based on sorting and ranking supervision.

[Adams & Zemel \(2011\)](#) propose relaxing permutation matrices to doubly-stochastic matrices based on marginals of distributions over permutation matrices. They apply their method to the LETOR learning-to-rank benchmark ([Liu, 2011](#)).

[Grover et al. \(2019\)](#) propose NeuralSort, a continuous relaxation of permutation matrices to the set of unimodal row-stochastic matrices via the Plackett-Luce family of distributions over permutations. For evaluation, they propose the benchmark of predicting the scalar value displayed on concatenated four-digit MNIST numbers. As supervision, they use the ranking of between 3 and 15 of those numbers. Additionally, they apply NeuralSort to differentiable quantile regression and k -nearest neighbors image classification.

Following this work, [Cuturi et al. \(2019\)](#) presented a method for smoothed ranking and sorting operators using optimal transport (OT). They use the idea that sorting can be achieved by minimizing the matching cost between elements and an auxiliary target of increasing values. That is, the smallest element is matched to the first value, the second smallest to the second value, etc. They make this differentiable by regularizing the OT problem with an entropic penalty and solving it by applying Sinkhorn iterations. Additionally, they devise a differentiable top- k operator for top- k supervised image classification. Based on this idea, [Xie et al. \(2020\)](#) have used OT and the differentiable top- k operator for k -nearest neighbors image classification and differentiable beam search.

Recently, [Blondel et al. \(2020\)](#) presented the idea of constructing differentiable sorting and ranking operators as projections onto the permutahedron, the convex hull of permutation matrices. They solve this by reducing it to isotonic optimization and make it differentiable by considering the Jacobians of the isotonic optimization and the projection. They apply their method to top- k supervised image classification, label ranking via a differentiable Spearman’s rank correlation coefficient, and robust regression via differentiable least trimmed squares.

3. Sorting Networks

In this section, we introduce two common sorting networks: the simple odd-even sorting network as well as the more complex, but also more efficient, bitonic sorting network.

3.1. Odd-Even Sorting Network

One of the simplest sorting networks is the fully connected odd-even sorting network. Here, neighboring elements are swapped if they are in the wrong order. As the name implies, this is done in a fashion alternating between comparing odd and even indexed elements with their successors. In detail, for sorting an input sequence $a_1 a_2 \dots a_n$, each layer updates the elements such that $a'_i = \min(a_i, a_{i+1})$ and $a'_{i+1} = \max(a_i, a_{i+1})$ for all odd or even indices i , respectively. Using n of such layers, a sequence of n elements is sorted as displayed in Figure 1 (center).

3.2. Bitonic Sorting Network

Second, we review the bitonic sorting network for sorting $n = 2^k$ elements where $k \in \mathbb{N}_+$. If desired, the sorting network can be extended to $n \in \mathbb{N}_+$ (Knuth, 1998).

The bitonic sorting networks builds on bitonic sequences: a sequence $(a_i)_{1 \leq i < n}$ is called bitonic if (after an appropriate circular shift) $a_1 \leq \dots \leq a_j \geq \dots \geq a_n$ for some j .

Following the Divide-and-Conquer principle, in analogy to merge sort, bitonic sort recursively splits the task of sorting a sequence into the tasks of sorting two subsequences of equal length, which are then combined into a bitonic sequence. Like merge sort, bitonic sort starts by merging individual elements, to obtain sorted lists of length 2 (first gray block in Figure 2). Pairs of these are then combined into bitonic sequences and then merged into monotonic sequences (second gray block in Figure 2). This proceeds, doubling the length of the sorted sequences with each (gray) block until the entire sequence is sorted. The difference to merge sort lies in the bitonic merge operation, which merges two sequences sorted in opposite order (i.e., a single bitonic sequence) into a single sorted (monotonic) sequence.

In Supplementary Material A, we give more details on the bitonic sorting network and sketch a proof why they work.

4. Differentiable Sorting Networks

To relax sorting networks, we need to relax the min and max operators, which are used as a basis for the swap operations in sorting networks. For that, we use softmin and softmax, which are convex combinations via the logistic sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. For two elements a_i, a_j , we define in accordance to softmin and softmax:

$$\text{softmin}(a_i, a_j) := \alpha_{ij} \cdot a_i + (1 - \alpha_{ij}) \cdot a_j \quad (1)$$

$$\text{softmax}(a_i, a_j) := (1 - \alpha_{ij}) \cdot a_i + \alpha_{ij} \cdot a_j \quad (2)$$

where

$$\alpha_{ij} := \sigma((a_j - a_i) \cdot s). \quad (3)$$

Here, s denotes a steepness hyperparameter such that for $s \rightarrow \infty$ the smooth operators converge to the discrete op-

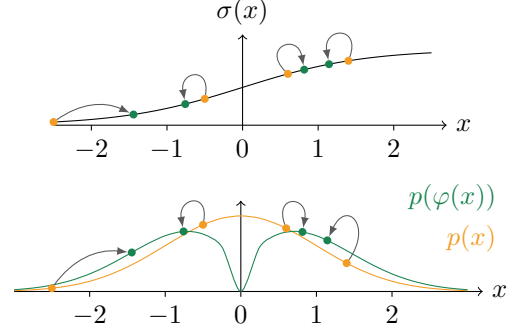


Figure 3: The Activation Replacement Trick. Top: on the logistic sigmoid function, the input values x (orange) are mapped to $\varphi(x)$ (green) and are thus closer to -1 and $+1$. Bottom: probability density functions of Gaussian distributed input values x (orange) and the distribution of replaced input values $\varphi(x)$ (green).

erators. As we show in the next section, it is necessary to extend this formulation by the activation replacement trick φ to avoid vanishing gradients and extensive blurring.

4.1. Activation Replacement Trick φ

Assuming that the inputs to a sorting network are normally distributed, there are many cases in which the differences of two values $|a_j - a_i|$ are very small as well as many cases in which the differences are very large. For the relaxation of sorting networks, this poses two problems:

If $|a_j - a_i|$ is close to 0, while we obtain large gradients, this also blurs the two values to a great extent, modifying them considerably. Thus, it is desirable to avoid $|a_j - a_i| \approx 0$.

On the other hand, if $|a_j - a_i|$ is large, vanishing gradients occur, which hinders training.

To counter these two problems at the same time, we propose the activation replacement trick. We transform the differences between two values to be potentially swapped (e.g., $x = (a_j - a_i)$) from a unimodal Gaussian distribution into a bimodal distribution, which has a low probability density around 0. To this end, we apply the transformation

$$\varphi : x \mapsto |x|^{1-\lambda} \cdot \text{sgn}(x) \quad (4)$$

to the differences x , where $\lambda \in [0, 1]$. φ pushes all input values (depending on the sign) toward -1 and $+1$, respectively. Thus, by applying φ before σ , we move the input values outside $[-1, +1]$ to positions at which they have a larger gradient, thus mitigating the problem of vanishing gradients. Simultaneously, we achieve a probability density of 0 at $|a_j - a_i| = 0$ (i.e., here $p(\varphi(0)) = 0$) as all values close to zero are mapped toward -1 and $+1$, respectively. This is displayed in Figure 3.

As we multiply by the steepness parameter s (Equation 3), we map the input to the sigmoid function toward $-s$ and $+s$, respectively. Thus, when replacing $\sigma(x \cdot s)$ by $\sigma(\varphi(x) \cdot s)$, we push the output values toward $\frac{1}{1+e^{-1 \cdot s}}$ or $\frac{1}{1+e^{1 \cdot s}}$. This increases the gradient $\frac{\partial \sigma(\varphi(x))}{\partial x}$ for large $\text{abs}(x)$ which are those values causing the vanishing gradients, addressing the problem of vanishing gradients. Further, for all $x \in (-1, +1)$ this pushes the output values away from $1/2$, addressing the problem of blurring of values.

Therefore, we extend our formulation of the relaxations of the min and max operators by defining

$$\alpha_{ij} := \sigma(\varphi(a_j - a_i) \cdot s). \quad (5)$$

Empirically, the activation replacement trick accelerates the training through our sorting network. We observe that, while sorting networks up to 21 layers (i.e., bitonic networks with $n \leq 64$) can operate with moderate steepness (i.e., $s \leq 15$) and without the activation replacement trick (i.e., $\lambda = 0$), for more layers, the activation replacement trick becomes necessary for good performance. Notably, the activation replacement trick also improves the performance for sorting networks with fewer layers. Further, the activation replacement trick allows training with smaller steepness s , which makes training more stable specifically for long sequences as it avoids exploding gradients.

Note that, in case of bitonic, in the first layer of the last merge block, $n/2$ elements in non-descending order are element-wise compared to $n/2$ elements in non-ascending order. Thus, in this layer, we compare the minimum of the first sequence to the maximum of the second sequence and vice versa. At the same time, we also compare the median of both sequences as well as values close to the median to each other. While we consider very large differences as well as very small differences in the same layer, the activation replacement trick achieves an equalization of the mixing behavior, reducing blurring and vanishing gradients.

4.2. Differentiable Permutation Matrices

For sorting and ranking supervision, i.e., training a neural network to predict scalars, where only the order of these scalars is known, we use the ground truth permutation matrix as supervision. Thus, to train an underlying neural network end-to-end through the differentiable sorting network, we need to return the underlying permutation matrix rather than the actual sorted scalar values. For that, we compute the permutation matrices for the swap operations for each layer as shown in Figure 1. Here, for all swap operations between any elements a_i and a_j that are to be ordered in non-descending order, the layer-wise permutation matrix is

$$P_{l,ii} = P_{l,jj} = \alpha_{ij} = \sigma(\varphi(a_j - a_i) \cdot s), \quad (6)$$

$$P_{l,ij} = P_{l,ji} = 1 - \alpha_{ij} = 1 - \sigma(\varphi(a_j - a_i) \cdot s) \quad (7)$$

where all other entries of P_l are set to 0. By multiplication, we compute the complete relaxed permutation matrix \mathbf{P} as

$$\mathbf{P} = P_n \cdot \dots \cdot P_2 \cdot P_1 = \left(\prod_{l=1}^n P_l^\top \right)^\top. \quad (8)$$

A column in the relaxed permutation matrix can be seen as a distribution over possible ranks for the corresponding input value. Given a ground truth permutation matrix \mathbf{Q} , we can define our column-wise cross entropy loss as


$$\mathcal{L} := \sum_{c=1}^n \left(\frac{1}{n} \text{CE}(\mathbf{P}_c, \mathbf{Q}_c) \right) \quad (9)$$

where \mathbf{P}_c and \mathbf{Q}_c denote the c th columns of \mathbf{P} and \mathbf{Q} , respectively. Note that, as the cross entropy loss is, by definition, computed element-wise, the column-wise cross entropy is equivalent to the row-wise cross entropy.

5. Experiments¹

5.1. Sorting and Ranking Supervision

We evaluate the proposed differentiable sorting networks on the four-digit MNIST sorting benchmark (Grover et al., 2019; Cuturi et al., 2019) as well as on the real-world SVHN data set.

MNIST For the four-digit MNIST sorting benchmark, MNIST digits are concatenated to four-digit numbers, e.g., . A CNN then predicts a scalar value corresponding to the value displayed in the four-digit image. For training, n of those four-digit images are separately processed by the CNN and then sorted by the relaxed sorting network as shown in Figure 1. Based on the permutation matrix produced by the sorting network and the ground truth ranking, the training objective is computed (Equation 9) and the CNN is updated. At test time, we forward single images of four-digit numbers from the test data set. For evaluation, the discrete rankings of the predicted values are compared to the rankings of their ground truth. Note that the n used for testing and evaluation can be independent of the n used for training because the n images are processed independently.

SVHN Since the multi-digit MNIST data set is an artificial data set, we also evaluate our technique on the SVHN data set (Netzer et al., 2011). This data set comprises house numbers collected from Google Street View and provides a larger variety wrt. different fonts and formats than the MNIST data set. We use the published ‘‘Format 1’’ and preprocess it as described by Goodfellow et al. (2013), cropping the centered multi-digit numbers with a boundary of


¹Our implementation is openly available at github.com/Felix-Petersen/diffsort.

Table 1: Results for the comparison to state-of-the-art (Grover et al., 2019; Cuturi et al., 2019) using the same network architectures averaged over 5 runs. The first three rows are duplicated from Cuturi et al. (2019). Metrics are (EM | EW | EM5).

Method	$n = 3$			$n = 5$			$n = 7$			$n = 9$			$n = 15$		
Stoch. NeuralSort	92.0	94.6		79.0	90.7	79.0	63.6	87.3		45.2	82.9		12.2	73.4	
Det. NeuralSort	91.9	94.5		77.7	90.1	77.7	61.0	86.2		43.4	82.4		9.7	71.6	
Optimal Transport	92.8	95.0		81.1	91.7	81.1	65.6	88.2		49.7	84.7		12.6	74.2	
Fast Sort & Rank	90.6	93.5	73.5	71.5	87.2	71.5	49.7	81.3	70.5	29.0	75.2	69.2	2.8	60.9	67.4
Odd-Even	95.2	96.7	86.1	86.3	93.8	86.3	75.4	91.2	86.4	64.3	89.0	86.7	35.4	83.7	87.6
	$n = 2$			$n = 4$			$n = 8$			$n = 16$			$n = 32$		
Odd-Even	98.1	98.1	84.3	90.5	94.9	85.5	63.6	87.9	83.6	31.7	82.8	87.3	1.7	69.1	86.7
Bitonic	98.1	98.1	84.0	91.4	95.3	86.7	70.6	90.3	86.9	30.5	81.7	86.6	2.7	67.3	85.4

Table 2: Results for training on the SVHN data set averaged over 5 runs. Metrics are (EM | EW | EM5).

Method	$n = 2$			$n = 4$			$n = 8$			$n = 16$			$n = 32$		
Det. NeuralSort	90.1	90.1	39.9	61.4	78.1	45.4	15.7	62.3	48.5	0.1	45.7	51.0	0.0	29.9	52.7
Optimal Transport	85.5	85.5	25.9	57.6	75.6	41.6	19.9	64.5	51.7	0.3	47.7	53.8	0.0	29.4	53.3
Fast Sort & Rank	93.4	93.4	57.6	58.0	75.8	41.5	8.6	52.7	34.4	0.3	36.5	41.6	0.0	14.0	27.5
Odd-Even	93.4	93.4	58.0	74.8	85.5	62.6	35.2	73.5	63.9	1.8	54.4	62.3	0.0	36.6	62.6
Bitonic	93.8	93.8	58.6	74.4	85.3	62.1	38.3	75.1	66.8	3.9	59.6	66.8	0.0	42.4	67.7

30%, resizing it to a resolution of 64×64 , and then selecting 54×54 pixels at a random location. As SVHN contains 1 – 5 digit numbers, we can avoid the concatenation and use the original images directly. Example images are . Otherwise, the experimental setup is as for the four-digit MNIST data set.

Network Architecture For the MNIST sorting task, we use the same convolutional neural network (CNN) architecture as Grover et al. (2019) and Cuturi et al. (2019) to allow for comparability. This architecture consists of two convolutional layers with a kernel size of 5×5 , 32 and 64 channels respectively, each followed by a ReLU and MaxPool layer; this is (after flattening) followed by a fully connected layer with a size of 64, a ReLU layer, and a fully connected output layer mapping to a scalar.

For the SVHN task, we use a network with four convolutional layers with a kernel size of 5×5 and (32, 64, 128, 256) filters, each followed by a ReLU and a max-pooling layer with stride 2×2 ; followed by a fully connected layer with size 64, a ReLU, and a layer with output size 1.

Evaluation Metrics For evaluation, discrete rankings based on the scalar predictions are computed and compared to the discrete ground truth rankings. As in previous works, we use the evaluation metrics of exact match (EM) of the predicted ranking, and fraction of element-wise correct ranks (EW) in the predicted ranking. For EM and EW, we follow Grover et al. (2019) and Cuturi et al. (2019), and use the

same n for training and evaluation. However, this can be a problem in the context of large input sets as these evaluation metrics become unreliable as n increases. For example, the difficulty of exact matches rises with the factorial of n , which is why they become too sparse to allow for valid conclusions for large n . To allow for a comparison of the performance independent of the number of elements n used for training, we also evaluate the models based on the EM accuracy for $n = 5$ (EM5). That is, the network can be trained with an arbitrary n , but the evaluation is done for $n = 5$. A table with respective standard deviations can be found in Supplementary Material C.

Training Settings We use the Adam optimizer (Kingma & Ba, 2015) with a learning rate of $10^{-3.5}$, and up to 10^6 steps of training. Furthermore, we set $\lambda = 0.25$ and use a steepness of two times the number of layers ($s = 2n$ for odd-even and $s = (\log_2 n)(1 + \log_2 n)$ for bitonic.) We use a constant batch size of 100 as in previous works unless denoted otherwise. Note that, although λ is chosen as a constant value for all n , a higher accuracy is possible when optimizing λ for each n separately.

5.1.1. Results

Comparison to State-of-the-Art (MNIST) We first compare our approach to the methods proposed by Grover et al. (2019) and Cuturi et al. (2019). Here, we follow the setting that the n used for evaluation is the same as the n used for training. The evaluation is shown in Table 1.

Table 3: Results for large n measured using the EM5 metric with fixed number of samples as well as a fixed batch size. Independent of the batch size, the model always performs better for larger n . Trained for 10^4 steps & averaged over 10 runs.

λ	0.25	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
n	32	32	64	128	256	512	1024	32	64	128	256	512	1024
batch size	128	128	64	32	16	8	4	4	4	4	4	4	4
$s = 30$	78.20	79.89	81.25	82.50	82.05	82.50	82.80	71.08	75.88	79.43	81.46	82.98	82.80
$s = 32.5$	76.98	79.62	81.66	80.15	81.87	82.64	81.63	72.31	75.59	79.71	81.36	82.99	81.63
$s = 35$	77.45	80.93	81.26	80.72	81.42	81.51	81.15	71.15	75.73	78.81	79.32	82.30	81.15
$s = 37.5$	76.40	80.02	80.05	81.50	80.05	82.67	80.07	70.69	75.80	79.11	80.64	82.70	80.07
$s = 40$	77.69	80.97	80.23	81.55	79.75	81.89	81.15	70.20	74.67	78.14	80.06	81.39	81.15
mean	77.35	80.29	80.89	81.28	81.03	82.24	81.36	71.09	75.53	79.04	80.57	82.47	81.36
best s	78.20	80.97	81.66	82.50	82.05	82.67	82.80	72.31	75.88	79.71	81.46	82.99	82.80
worst s	76.40	79.62	80.05	80.15	79.75	81.51	80.07	70.20	74.67	78.14	79.32	81.39	80.07

We report results for exact match, correct ranks, and EM5, respectively. For the odd-even architecture, we compare results for the original $n \in \{3, 5, 7, 9, 15\}$. Our approach outperforms current methods on all metrics and input set sizes. In addition, we extend the original benchmark set sizes by $n \in \{2, 4, 8, 16, 32\}$, allowing for the canonical version of the bitonic sorting network which requires input size of powers of 2. We apply $n \in \{2, 4, 8, 16, 32\}$ to the odd-even as well as the bitonic sorting network. In this direct comparison, we can see that the bitonic and the odd-even architectures perform similar. Notably, the EM and EW accuracies do not always correlate as can be seen for $n = 32$. Here, the EM accuracy is greater for the bitonic network and the EW accuracy is greater for the odd-even network. We attribute this to the odd-even network’s gradients causing swaps of neighbors while the bitonic network’s gradients provide a holistic approach favoring exact matches.

SVHN The results in Table 2 show that the real-world SVHN task is significantly harder than the MNIST task. On this data set, differentiable sorting networks are also better than current methods on all metrics and input set sizes. Here, the performances of odd-even and bitonic are similar. Notably, the EM5 accuracy is largest for the bitonic sorting network at $n = 32$, which demonstrates that the method benefits from longer input sets. Further, for $n \in \{8, 16, 32\}$, the bitonic sorting network marginally outperforms the odd-even sorting network on all metrics.

5.2. Large-Scale Sorting and Ranking Supervision

We are interested in the effect of training with larger input set sizes n . As the bitonic sorting network requires significantly fewer layers than odd-even and is (thus) faster, we use the bitonic sorting network for the scalability experiments. Here, we evaluate for $n = 2^k$, $k \in \{5, 6, 7, 8, 9, 10\}$ on the MNIST sorting benchmark, comparing the EM5 accuracy as shown in Table 3.

For this experiment, we consider steepness values of $s \in \{30, 32.5, 35, 37.5, 40\}$ and report the mean, best, and worst over all steepness values for each n . We set λ to 0.4 as this allows for stable training with $n > 128$. To keep the evaluation feasible, we reduce the number of steps during training to 10^4 , compared to the 10^6 iteration in Table 1. Again, we use the Adam optimizer with a learning rate of $10^{-3.5}$.

In the first two columns of Table 3, we show a head-to-head comparison with the setting in Table 1 with $\lambda = 0.25$ and $\lambda = 0.4$ for $n = 32$. Trained for 10^6 steps, the EM5 accuracy is 85.4%, while it is 78.2% after 10^4 steps. Increasing λ from 0.25 to 0.4 improves the EM5 accuracy from 78.2% to 80.97%.

This also demonstrates that already at this scale, a larger λ , i.e., a stronger activation replacement trick, can improve the overall accuracy of a bitonic sorting network compared to training with $\lambda = 0.25$.

As the size of training tuples n increases, this also increases the overall number of observed images during training. Therefore, in the left half of Table 3, we consider the accuracy for a constant total of observed images per iteration, i.e., for $n \times \text{batch size} = 4096$ (e.g., for $n = 32$ this results in a batch size of 128, while for $n = 1024$, the batch size is only 4). In the right half of Table 3, we consider a constant batch size of 4.

With increasing n , the accuracy of our model increases even for a constant number of observed images even though it has to operate on very small batch sizes. This shows that training with larger ordered sets results in better accuracy. This suggests that, if possible, larger n should be prioritized over larger batch sizes and that good results can be achieved by using the largest possible n for the available data to learn from all available information.

Table 4: Runtimes, memory requirements, and number of layers for sorting n elements. Runtimes reported for an Nvidia GTX 1070. We include NeuralSort (Grover et al., 2019), FastRank (Blondel et al., 2020), and OT Sort (Cuturi et al., 2019).

n	Differentiable Odd-Even Sort				Differentiable Bitonic Sort				NeuralSort		FastRank	OT Sort
	GPU	CPU	Memory	# Layers	GPU	CPU	Memory	# Layers	GPU	CPU	CPU	CPU
4	69 ns	1.9 μ s	1KB	4	52 ns	1.3 μ s	840B	3	145 ns	7.1 μ s	189 μs	1.0 ms
16	1.2 μs	54 μ s	42KB	16	759 ns	40 μ s	28KB	10	396 ns	11 μ s	215 μs	7.5 ms
32	7.4 μs	309 μ s	315KB	32	3.5 μs	159 μ s	152KB	15	969 ns	13 μ s	303 μs	17 ms
128	493 μs	19 ms	20.2MB	128	97 μs	5 ms	4.1MB	28	12 μs	177 μ s	834 μs	55 ms
1024	660 ms	31 s	4.9GB	1024	15 ms	1.7 s	549MB	55	1.2 ms	11 ms	4.8 ms	754 ms

5.3. Ablation Study and Hyperparameter Sensitivity

To assess the impact of the proposed activation replacement trick (ART), we evaluate both architectures with and without ART at $\lambda = 0.25$ in Table 5. The accuracy improves by using the ART for small as well as for large n . For large n , the activation replacement trick has a greater impact on the performance of both architectures. In Figure 4, we evaluate the sensitivity of the differentiable odd-even sorting network to the steepness hyperparameter s . For a broad range of s , the performance is stable. In Figure 5, we evaluate both differentiable sorting networks for varying ART intensities λ . Here, performance increases with larger λ (i.e., with a stronger ART). For $\lambda > 0.5$, the performance drops as φ converges to a discrete step function for $\lambda \rightarrow 1$.

5.4. Top- k Supervision

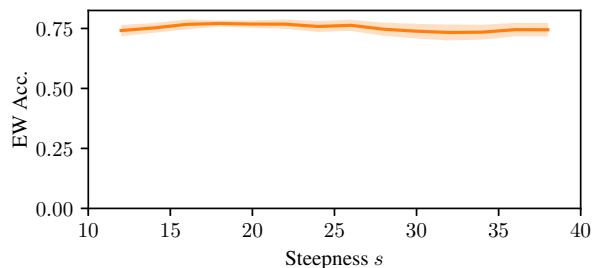
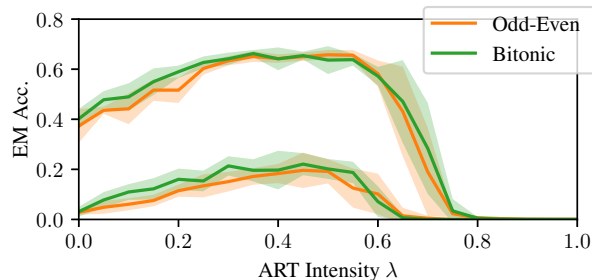
In addition to the sorting supervision task, we also benchmark our method on top- k supervision following Cuturi et al. (2019) and Blondel et al. (2020). Here, we train two models (ResNet18 and a vanilla CNN with 4 convolutional and 2 fully connected layers) on CIFAR-10 as well as CIFAR-100 and compare the results to training with the Softmax Cross-Entropy loss. Further details on the experimental setting can be found in Supplementary Material B.4. Following Cuturi et al. (2019) and Blondel et al. (2020), we focus on $k = 1$. We present the results for this in Table 6. Overall, Softmax Cross-Entropy and our differentiable top- k operator perform similar even in the 100 class classification problem.

5.5. Runtime and Memory Analysis

Finally, we report the runtime and memory consumption of differentiable sorting networks in Table 4. For GPU runtimes, we use a native CUDA implementation and measure the time and memory for sorting n input elements including forward and backward pass. For CPU runtimes, we use a PyTorch (Paszke et al., 2019) implementation. For a small number of input elements, the odd-even and bitonic sorting networks have around the same time and memory requirements, while for larger numbers of input elements, bitonic is much faster than odd-even.

 Table 5: Ablation Study: Evaluation of the ART ($\lambda = 0$ vs. $\lambda = 0.25$) for $n = 4$ and $n = 32$ on the MNIST and the SVHN data set. The displayed metric is EW.

Setting / λ	$n = 4$		$n = 32$	
	0	0.25	0	0.25
Odd-Even (MNIST)	94.5	94.9	61.5	69.1
Bitonic (MNIST)	93.6	95.3	62.8	67.3
Odd-Even (SVHN)	77.3	85.5	28.5	36.6
Bitonic (SVHN)	78.1	85.3	35.0	42.4

 Figure 4: Sensitivity of the odd-even sorting network to varying steepness s for $n = 16$.

 Figure 5: Comparing different ART strengths λ for $n = 8$ (top) and $n = 16$ (bottom). Training with $\lambda \leq 0.5$ is stable.

 Table 6: Top- k classification averaged over 10 runs.

Setting	Softmax CE	Diff. Top- k
CIFAR-10, Vanilla CNN	87.2%	88.0%
CIFAR-10, ResNet18	91.0%	90.9%
CIFAR-100, Vanilla CNN	58.2%	56.3%
CIFAR-100, ResNet18	61.9%	63.3%

The asymptotic runtime of differentiable odd-even sort is in $\mathcal{O}(n^3)$ and for bitonic sort the runtime is in $\mathcal{O}(n^2(\log n)^2)$. Note that, for this, the matrix multiplication in Equation 8 is a sparse matrix multiplication. We also report runtimes for other differentiable sorting and ranking methods. For large n , we empirically confirm that FastRank (Blondel et al., 2020) is the fastest method, i.e., because it produces only output ranks / sorted output values and not differentiable permutation matrices. Note that differentiable sorting networks also produce sorted output values. Computing only sorted output values is significantly faster than computing the full differentiable permutation matrices, however, for the effective cross-entropy training objective, differentiable permutation matrices are necessary.

6. Conclusion

In this work, we presented differentiable sorting networks for training based on sorting and ranking supervision. To this end, we approximated the discrete min and max operators necessary for pairwise swapping in traditional sorting network architectures with their respective differentiable softmin and softmax operators. We proposed an activation replacement trick to avoid the problems of vanishing gradients and well as blurred values. We showed that it is possible to robustly sort and rank even long sequences on large input sets of up to at least 1024 elements. In the future, we will investigate differentiable sorting networks for applications such as clustering and learning-to-rank.

Acknowledgment

The second author gratefully acknowledges the financial support from Land Salzburg within the WISS 2025 project IDA-Lab (20102-F1901166-KZP and 20204-WISS/225/197-2019).

References

- Adams, R. P. and Zemel, R. S. Ranking via sinkhorn propagation. *arXiv preprint arXiv:1106.1925*, 2011.
- Ajtai, M., Komlós, J., and Szemerédi, E. An $O(n \log n)$ sorting network. In *Proceedings of the Fifteenth Annual ACM Symposium on Theory of Computing*, 1983.
- Baddar, S. W. A.-H. and Batcher, K. E. *Designing sorting networks: A new paradigm*. Springer Science & Business Media, 2012.
- Batcher, K. E. Sorting networks and their applications. In *Proceedings of the April 30–May 2, 1968, spring joint computer conference*, pp. 307–314, 1968.
- Bidlo, M. and Dobeš, M. Evolutionary development of growing generic sorting networks by means of rewriting systems. *IEEE Transactions on Evolutionary Computation*, 2019.
- Blondel, M., Teboul, O., Berthet, Q., and Djolonga, J. Fast Differentiable Sorting and Ranking. In *International Conference on Machine Learning (ICML)*, 2020.
- Ceterchi, R. and Tomescu, A. I. Spiking neural p systems – a natural model for sorting networks. In *Proc. of the Sixth Brainstorming Week on Membrane Computing*, 2008.
- Cuturi, M., Teboul, O., and Vert, J.-P. Differentiable ranking and sorting using optimal transport. In *Proc. Neural Information Processing Systems (NIPS)*, 2019.
- Goodfellow, I. J., Bulatov, Y., Ibarz, J., Arnoud, S., and Shet, V. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013.
- Govindaraju, N. K., Gray, J., Kumar, R., and Manocha, D. Gputerasort: high performance graphics co-processor sorting for large database management. In *SIGMOD Conference*, 2006.
- Gowanlock, M. and Karsin, B. A hybrid cpu gpu approach for optimizing sorting throughput. *Parallel Computing*, 85, 02 2019.
- Graves, A., Wayne, G., and Danihelka, I. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.
- Grover, A., Wang, E., Zweig, A., and Ermon, S. Stochastic Optimization of Sorting Networks via Continuous Relaxations. In *International Conference on Learning Representations (ICLR)*, 2019.
- Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015.
- Kingma, D. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- Knuth, D. E. *The Art of Computer Programming, Volume 3: (2nd Ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., 1998.
- LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs*, 2, 2010. URL <http://yann.lecun.com/exdb/mnist>.
- Lim, C. H. and Wright, S. A box-constrained approach for hard permutation problems. In *International Conference on Machine Learning (ICML)*, 2016.
- Liu, T.-Y. Learning to rank for information retrieval. 2011.

- Mena, G., Belanger, D., Linderman, S., and Snoek, J. Learning latent permutations with gumbel-sinkhorn networks. In *International Conference on Learning Representations (ICLR)*, 2018.
- Metta, V. P. and Kelemenova, A. Sorting using spiking neural p systems with anti-spikes and rules on synapses. In *International Conference on Membrane Computing*, 2015.
- Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Proc. Neural Information Processing Systems (NIPS)*, pp. 8024–8035. 2019.
- Vinyals, O., Bengio, S., and Kudlur, M. Order matters: Sequence to sequence for sets. In Bengio, Y. and LeCun, Y. (eds.), *International Conference on Learning Representations (ICLR)*, 2016.
- Xie, Y., Dai, H., Chen, M., Dai, B., Zhao, T., Zha, H., Wei, W., and Pfister, T. Differentiable top-k with optimal transport. In *Proc. Neural Information Processing Systems (NIPS)*, 2020.