

Compact and Computationally Efficient Representation of Deep Neural Networks

Simon Wiedemann, Klaus-Robert Müller, *Member, IEEE*, and Wojciech Samek, *Member, IEEE*

Abstract—Dot product operations between matrices are at the heart of almost any field in science and technology. In many cases, they are the component that requires the highest computational resources during execution. For instance, deep neural networks such as VGG-16 require up to 15 giga-operations in order to perform the dot products present in a single forward pass, which results in significant energy consumption and thus limits their use in resource-limited environments, e.g., on embedded devices or smartphones. One common approach to reduce the complexity of the inference is to prune and quantize the weight matrices of the neural network and to efficiently represent them using sparse matrix data structures. However, since there is no guarantee that the weight matrices exhibit significant sparsity after quantization, the sparse format may be suboptimal. In this paper we present new efficient data structures for representing matrices with low entropy statistics and show that these formats are especially suitable for representing neural networks. Alike sparse matrix data structures, these formats exploit the statistical properties of the data in order to reduce the size and execution complexity. Moreover, we show that the proposed data structures can not only be regarded as a generalization of sparse formats, but are also more energy and time efficient under practically relevant assumptions. Finally, we test the storage requirements and execution performance of the proposed formats on compressed neural networks and compare them to dense and sparse representations. We experimentally show that we are able to attain up to x15 compression ratios, x1.7 speed ups and x20 energy savings when we *lossless* convert state-of-the-art networks such as AlexNet, VGG-16, ResNet152 and DenseNet into the new data structures.

Keywords—Neural network compression, computationally efficient deep learning, data structures, sparse matrices, lossless coding.

I. INTRODUCTION

The dot product operation between matrices constitutes one of the core operations in almost any field in science. Examples are the computation of approximate solutions of

This work was supported by the Fraunhofer Society through the MPI-FhG collaboration project “Theory & Practice for Reduced Learning Machines”. This research was also supported by the German Ministry for Education and Research as Berlin Big Data Center under Grant 01IS14013A and by the Institute for Information & Communications Technology Promotion and funded by the Korea government (MSIT) (No. 2017-0-01779 and No. 2017-0-00451).

S. Wiedemann and W. Samek are with Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany (e-mail: wojciech.samek@hhi.fraunhofer.de)

K.-R. Müller is with the Technische Universität Berlin, 10587 Berlin, Germany, with the Max Planck Institute for Informatics, 66123 Saarbrücken, Germany, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 136-713, South Korea (e-mail: klaus-robert.mueller@tu-berlin.de)

complex system behaviors in physics [1], iterative solvers in mathematics [2] and features in computer vision applications [3]. Also deep neural networks heavily rely on dot product operations in their inference [4], [5]; e.g., networks such as VGG-16 require up to 16 dot product operations, which results in 15 giga-operations for a single forward pass. Hence, lowering the algorithmic complexity of these operations and thus increasing their efficiency is of major interest for many modern applications. Since the complexity depends on the data structure used for representing the elements of the matrices, a great amount of research has focused on designing data structures and respective algorithms that can perform efficient dot product operations [6]–[8].

Of particular interest are the so called *sparse matrices*, a special type of matrices that have the property that many of their elements are zero valued. In principle, one can design efficient representations of sparse matrices by leveraging the prior assumption that most of their element values are zero and therefore, only store the non-zero entries of the matrix. Consequently, their storage requirements become of the order of the number of non-zero values. However, having an efficient representations with regard to storage requirement does neither imply that the dot product algorithm associated to that data structure will also be efficient. Hence, a great part of the research was focused on the design of data structures that have as well low complex dot product algorithms [8]–[11]. However, by assuming sparsity alone we are implicitly imposing a spike-and-slab prior¹ over the probability mass distribution of the elements of the matrix. If the actual distribution of the elements greatly differs from this assumption, then the data structures devised for sparse matrices become inefficient. Hence, sparsity may be considered a too constrained assumption for some applications of current interest, e.g., representation of neural networks.

In this work, we alleviate the shortcomings of sparse representations by considering a more relaxed prior over the distribution of the matrix elements. More precisely, we assume that this distribution has low Shannon entropy [12]. Mathematically, sparsity can be considered a subclass of the general family of low entropic distributions. In fact, sparsity measures the min-entropy of the element distribution, which is related to Shannon’s entropy measure through Renyi’s generalized entropy definition [13]. With this goal in mind, we ask the question:

“Can we devise efficient data structures under the implicit assumption that the entropy of the distribu-

¹That is, a delta function at 0 and a uniform distribution over the non-zero elements.

tion of the matrix elements is low?"

We want to stress once more that by efficiency we regard two related but distinct aspects

- 1) efficiency with regard to storage requirements
- 2) efficiency with regard to algorithmic complexity of the dot product associated to the representation

For the later, we focus on the number of elementary operations required in the algorithm, since they are related to the energy and time complexity of the algorithm. It is well known that the minimal bit-length of a data representation is bounded by the entropy of its distribution [12]. Hence, matrices with low entropic distributions automatically imply that we can design data structures that do not require high storage resources. In addition, as we will discuss in the next sections, low entropic distributions may also attain gains in efficiency if these data structures implicitly encode the distributive law of multiplications. By doing so, a great part of the algorithmic complexity of the dot product is reduced to the order of the number of distinct elements that appear per row in a matrix. This number is related to the entropy such, that it is likely to be small as long as the entropy of the matrix is low. Therefore, these data structures not only attain higher compression gains, but also require less total number of operations when performing the dot product.

Our contributions can be summarized as follows:

- We propose new highly efficient data structures that exploit on the improper prior that the matrix does not entail many distinct elements per row (i.e., low entropy).
- We provide a detailed analysis of the storage requirements and algorithmic complexity of performing the dot product associated to this data structures.
- We establish a relation between the known sparse and the proposed data structures. Namely, sparse matrices belong to the same family of low entropic distributions, however, they may be considered a more constrained subclass of them.
- We show through experiments that indeed, these data structures attain gains in efficiency on simulated as well as real-world data. In particular, we show that up to x15 compression ratios, x1.7 speed ups and x20 energy savings can be achieved when converting state-of-the-art neural networks into the proposed data structures.

In the following Section II we introduce the problem of efficient representation of neural networks and briefly review related literature. In Section III the proposed data structures are given. We demonstrate through a simple example that these data structures are able to: 1) achieve higher compression ratios than their respective dense and sparse counterparts and 2) reduce the algorithmic complexity of performing the dot product. Section IV analyses the storage and energy complexity of these novel data structures. Experimental evaluation is performed in Section V using simulations as well as state-of-the-art neural networks such as AlexNet, VGG-16, ResNet152 and DenseNet. Section VI concludes the paper with a discussion.

II. EFFICIENT REPRESENTATION OF NEURAL NETWORKS

Deep neural networks [14], [15] nowadays became the state-of-the-art in many fields of machine learning, such as in

computer vision, speech recognition, natural language processing [16]–[19], and have been progressively also used in the sciences, e.g. Physics [20], neuroscience [21], Chemistry [22], [23]. In their most basic form, they constitute a chain of affine transformations concatenated with a non linear function which is applied element-wise to the output. Hence, the goal is to learn the values of those transformation or weight matrices (i.e., parameters) such that the neural network performs its task particularly well. The procedure of calculating the output prediction of the network for a particular input is called *inference*. The computational cost of performing inference is dominated by computing the affine transformations (thus, the dot products between matrices). Since today's neural networks perform many dot product operations between large matrices, this greatly complicates their deployment onto resource constrained devices.

However, it has been extensively shown that most neural networks are overparameterized, meaning that there are many more parameters than actually needed for the tasks of interest [24]–[27]. This implies that these networks are highly inefficient with regard to the resources they require when performing inference. This fact motivated an entire research field of model compression [28]. One of the suggested approaches is to design methods that compress the weight elements of the neural network such, that their prediction accuracy is minimally affected. For instance, usual techniques include sparsification of the network [27], [29]–[31] or quantization of the weight elements [32], [33], or both simultaneously [26], [34], [35]. The advantage of these approaches is that after compression, one can greatly reduce both storage and execution complexity of inference by leveraging on the sparsity of the network, e.g., by representing the weight matrices in one of the sparse data structures and performing the dot products accordingly. However, it is not guaranteed that the distribution of the weight elements of the compressed matrices matches the sparse distribution assumption, which consequently would make the respective data structures inefficient.

Figure 1 demonstrates this discrepancy between the sparsity assumption and the real distribution of weight elements. It plots the distribution of the weight elements of the last classification layer of VGG-16 [36] (1000×4096 dimensional matrix), after having applied uniform quantization on the weight elements. We stress that the prediction accuracy and generalization of the network was not affected by this operation. As we can see, the distribution of the compressed layer does not satisfy the sparsity assumption, i.e., there is not one particular element (such as 0) that appears specially frequent in the matrix. The most frequent value is -0.008 and its frequency of appearance does not dominate over the others (about 4.2%). The entropy of the matrix is about $H = 5.5$, which indicates that highly likely not many distinct elements appear per row in the matrix (as a rough lower bound estimation one can think of $2^H \approx 45$ distinct elements per row, whenever the column dimension is large). Indeed, we notice that most of the entries are dominated by only 15 distinct values, which is 1.5% of the number of columns of the matrix.

Hence, designing efficient data structures for these type of matrices is of major interest in this application.

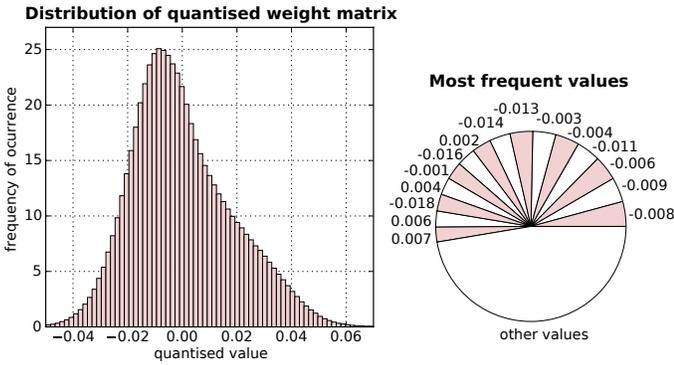


Fig. 1: Distribution of the weight matrix of the last layer of the VGG-16 neural network [36] after quantization. The respective matrix is 1000×4096 dimensional, transforming the 4096 last-layer features onto 1000 output classes. We applied an uniform quantizer over the range of values, with 2^7 quantization points, which resulted in no loss of accuracy on the classification task. Left: Probability mass distribution. Right: Frequency of appearance of the 15 most frequent values.

III. DATA STRUCTURES FOR MATRICES WITH LOW ENTROPY STATISTICS

In this section we introduce the proposed data structures and show that they implicitly encode the distributive law. Consider the following matrix

$$M = \begin{pmatrix} 0 & 3 & 0 & 2 & 4 & 0 & 0 & 2 & 3 & 4 & 0 & 4 \\ 4 & 4 & 0 & 0 & 0 & 4 & 0 & 0 & 4 & 4 & 0 & 4 \\ 4 & 0 & 3 & 4 & 0 & 0 & 0 & 4 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 & 4 & 0 & 3 & 4 & 4 & 0 & 0 \\ 0 & 4 & 4 & 0 & 0 & 4 & 0 & 4 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Now assume that we want to: 1) store this matrix with the minimum amount of bits and 2) perform the dot product with a vector $a \in \mathbb{R}^{12}$ with the minimum complexity.

A. Minimum storage

We firstly comment on the storage requirement of dense and sparse formats and then introduce two new formats which more effectively store matrix M .

Dense format: Arguably the simplest way to store the matrix M is in it's so called dense representation. That is, we store it's elements in a 5×12 long array (in addition to it's dimensions $m = 5$ and $n = 12$).

Sparse format: However, notice that more than 50% of the entries are $\bar{0}$. Hence, we may be able to attain a more compressed representation of this matrix if we store it in one of the well known sparse data structure, for instance, in the *Compressed Sparse Row* (or CSR in short) format. This particular format stores the values of the matrix in the following way:

- Scans the non-zero elements in row-major order (that is, from left to right, up to down) and stores them in an array (which we denote as W).
- Simultaneously, it stores the respective column indices in another array (which we call $colI$).
- Finally, it stores pointers that signalize when a new row starts (we denote this array as $rowPtr$).

Hence, our matrix M would take the form

$$\begin{aligned} W &: [3, 2, 4, 2, 3, 4, 4, 4, 4, 4, 4, 4, 3, \\ & \quad 4, 4, 2, 4, 4, 4, 3, 4, 4, 4, 4, 4] \\ colI &: [1, 3, 4, 7, 8, 9, 11, 0, 1, 5, 8, 9, 11, 0, \\ & \quad 2, 3, 7, 9, 3, 4, 5, 7, 8, 9, 1, 2, 5, 7] \\ rowPtr &: [0, 7, 13, 18, 24, 28] \end{aligned}$$

If we assume the same bit-size per element for all arrays, then the CSR data structure does not attain higher compression gains in spite of not saving the zero valued elements (62 entries vs. 60 that are being required by the dense data structure).

We can improve this by exploiting the low-entropy property of matrix M . In the following, we propose two new formats which realize this.

CER format: Firstly, notice that many elements in M share the same value. In fact, only the four values $\Omega = \{0, 4, 3, 2\}$ appear in the entire matrix. Hence, it appears reasonable to assume that data structures that repeatedly store these values (such as the dense or CSR structures) induce high redundancies in their representation. Therefore, we propose a data structure where we only store those values once. Secondly, notice that different elements appear more frequent than others, and their relative relation does not change throughout the rows of the matrix, and the ordering. Concretely, we have a set of unique elements $\Omega = \{0, 4, 3, 2\}$ which appear $P_{\#} = \{32, 21, 4, 3\}$ times respectively in the matrix, and we obtain the same relative order of highest to lowest frequent value throughout the rows of the matrix. Hence, we can design an efficient data structure which leverages on both properties in the following way:

- 1) Store unique elements present in the matrix in an array in frequency-major order (that is, from most to least frequent). We name this array Ω .
- 2) Store respectively the column indices in row-major order, excluding the first element (thus excluding the most frequent element). We denote it as $colI$.
- 3) Store pointers that signalize when the positions of the next new element in Ω starts. We name it ΩPtr . If a particular pointer in ΩPtr is the same as the previous one, this means that the current element is not present in the matrix and we jump to the next element.
- 4) Store pointers that signalize when a new row starts. We name it $rowPtr$.

Hence, this new data structure represents matrix M as

$$\begin{aligned}\Omega &: [0, 4, 3, 2] \\ colI &: [4, 9, 11, 1, 8, 3, 7, 0, 1, 5, 8, 9, 11, 0, \\ &\quad 3, 7, 2, 9, 3, 4, 5, 8, 9, 7, 1, 2, 5, 7] \\ \Omega Ptr &: [0, 3, 5, 7, 13, 16, 17, 18, 23, 24, 28] \\ rowPtr &: [0, 3, 4, 7, 9, 10]\end{aligned}$$

Notice that we can uniquely reconstruct M from this data structure. We refer to this data structure as the *Compressed Entropy Row* (or CER in short) data structure. One can verify that indeed, the CER data structure only requires 49 entries (instead of 60 or 62) attaining as such a compressed representation of the matrix M .

To our knowledge, this is the first format optimized for representing low entropy matrices.

CSER format: In some cases, it may well be that the probability distribution across rows in a matrix are not similar to each other. Hence, the second assumption in the CER data structure would not apply and we would only be left with the first one. That is, we only know that not many distinct elements appear per row in the matrix or, in other words, that many elements share the same value. The *compressed shared elements row* (or CSER in short) data structure is a slight extension to the previous CER representation. Here, we add an element pointer array, which signalizes which element in Ω the positions are referred to (we called it ΩI). Hence, the above matrix would then be represented as follows

$$\begin{aligned}\Omega &: [0, 2, 3, 4] \\ colI &: [4, 9, 11, 1, 8, 3, 7, 0, 1, 5, 8, 9, 11, 0, \\ &\quad 3, 7, 2, 9, 3, 4, 5, 8, 9, 7, 1, 2, 5, 7] \\ \Omega I &: [3, 2, 1, 3, 3, 2, 1, 3, 2, 3] \\ \Omega Ptr &: [0, 3, 5, 7, 13, 16, 17, 18, 23, 24, 28] \\ rowPtr &: [0, 3, 4, 7, 9, 10]\end{aligned}$$

Thus, for storing matrix M we require 59 entries, which is still a gain but not a significant one. Notice, that now the ordering of the elements in Ω is not important anymore.

The relationship between CSER, CER and CSR data structures is described in Section IV.

B. Dot product complexity

We just saw that we can attain gains with regard to compression if we represent the matrix in the CER and CSER data structures. However, we can also devise corresponding dot product algorithms that are more efficient than their dense and sparse counterparts. As an example, consider only the scalar product between the second row of matrix M with an arbitrary input vector $a = [a_1 \ a_2 \ \dots \ a_{12}]^T$. In principle, the difference in the algorithmic complexity arises because each data structure implicitly encodes a different expression of the

scalar product, namely

$$\begin{aligned}dense &: 4a_1 + 4a_2 + 0a_3 + 0a_4 + 0a_5 + 4a_6 \\ &\quad + 0a_7 + 0a_8 + 4a_9 + 4a_{10} + 0a_{11} + 4a_{12} \\ CSR &: 4a_1 + 4a_2 + 4a_6 + 4a_9 + 4a_{10} + 4a_{12} \\ CER/CSER &: 4(a_1 + a_2 + a_6 + a_9 + a_{10} + a_{12})\end{aligned}$$

For instance, the dot product algorithm associated to the dense format would calculate the above scalar product by

- 1) loading M and a .
- 2) calculating $4a_0 + 4a_1 + 0a_2 + 0a_3 + 0a_4 + 4a_5 + 0a_6 + 0a_7 + 4a_8 + 4a_9 + 0a_{10} + 4a_{11}$.

This requires 24 load (12 for the matrix elements and 12 for the input vector elements), 12 multiply, 11 add and 1 write operations (for writing the result into memory). We purposely omitted the accumulate operation which stores the intermediate values of the multiply-sum operations, since their cost can effectively be associated to the sum operation. Moreover, we only considered read/write operations from and into memory. Hence, this makes 48 operations in total.

In contrast, the dot product algorithm associated with the CSR representation would only multiply-add the non-zero entries. It does so by performing the following steps

- 1) Load the subset of $rowPtr$ respective to row 2. Thus, $rowPtr \rightarrow [7, 13]$.
- 2) Then, load the respective subset of non-zero elements and column indices. Thus, $W \rightarrow [4, 4, 4, 4, 4, 4]$ and $colI \rightarrow [0, 1, 5, 8, 9, 11]$.
- 3) Finally, load the subset of elements of a respective to the loaded subset of column indices and subsequently multiply-add them to the loaded subset of W . Thus, $a \rightarrow [a_0, a_1, a_5, a_8, a_{10}, a_{11}]$ and calculate $4a_0 + 4a_1 + 4a_5 + 4a_8 + 4a_9 + 4a_{11}$.

By executing this algorithm we would require 20 load operations (2 from the $rowPtr$ and 6 for the W , the $colI$ and the input vector respectively), 6 multiplications, 5 additions and 1 write. In total this dot product algorithm requires 32 operations.

However, we can still see that the above dot product algorithm is inefficient in this case since we constantly multiply by the same element 4. Instead, the dot product algorithm associated to, e.g., the CER data structure, would perform the following steps

- 1) Load the subset of $rowPtr$ respective to row 2. Thus, $rowPtr \rightarrow [3, 4]$.
- 2) Load the corresponding subset in ΩPtr . Thus, $\Omega Ptr \rightarrow [7, 13]$.
- 3) For each pair of elements in ΩPtr , load the respective subset in $colI$ and the element in Ω . Thus, $\Omega \rightarrow [4]$ and $colI \rightarrow [0, 1, 5, 8, 9, 11]$.
- 4) For each loaded subset of $colI$, perform the sum of the elements of a respective to the loaded $colI$. Thus, $a \rightarrow [a_0, a_1, a_5, a_8, a_{10}, a_{11}]$ and do $a_0 + a_1 + a_5 + a_8 + a_9 + a_{11} = z$.
- 5) Subsequently, multiply the sum with the respective element in Ω . Thus, compute $4z$.

A similar algorithm can be devised for the CSER data structure. One can find both pseudocodes in the appendix. The operations required by this algorithm are 17 load operations (2 from *rowPtr*, 2 from *ΩPtr*, 1 from Ω , 6 from *colI* and 6 from *a*), 1 multiplication, 5 additions and 1 write. In total this are 24 operations.

Hence, we have observed that for the matrix M , the CER (and CSER) data structure does not only achieve **higher compression rates**, but it also attains **gains in efficiency** with respect to the dot product operation.

In the next section we give a detailed analysis about the storage requirements needed by the data structures and also the efficiency of the dot product algorithm associated to them. This will help us identify when one type of data structure will attain higher gains than the others.

IV. AN ANALYSIS OF THE STORAGE AND ENERGY COMPLEXITY OF DATA STRUCTURES

Without loss of generality, in the following we assume that we aim to encode a particular matrix $M \in \Omega^{n \times m = N}$, where its elements $M_{ij} = \omega_k \in \Omega$ take values from a finite set of elements $\Omega = \{\omega_0, \omega_1, \dots, \omega_{K-1}\}$. Moreover, we assign to each element ω_k a probability mass value $p_k = \#(\omega_k)/N$, where $\#(\omega_k)$ counts the number of times the element ω_k appears in the matrix M . We denote the respective set of probability mass values $P_\Omega = \{p_0, p_1, \dots, p_{K-1}\}$. In addition, we assume that each element in Ω appears at least once in the matrix (thus, $p_k > 0$ for all $k = 0, \dots, K-1$) and that $\omega_0 = 0$ is the most frequent value in the matrix. We also denote with $\Omega^r \subset \Omega$ the subset of elements that appear at row $r \leq m$ (where $\omega_0 = 0 \in \Omega^r$ for all $r = 1, \dots, m$) and with P_Ω^r their respective frequencies. Finally, we order the elements in Ω and P_Ω in probability-major order, that is, $p_0 \geq p_1 \geq \dots \geq p_{K-1}$.

A. Measuring the energy efficiency of the dot product

This work proposes representations that are efficient with regard to storage requirements as well as their dot product algorithmic complexity. For the latter, we focus on the energy requirements, since we consider it as the most relevant measures for neural network compression. However, exactly measuring the energy of an algorithm may be unreliable since it depends on the software implementation and on the hardware the program is running on. Therefore, we will model the energy costs in a way that can easily be adapted across different software implementations as well as hardware architectures.

In the following we model a dot product algorithm by a computational graph, whose nodes can be labeled with one of 4 elementary operations, namely: 1) a *mul* or multiply operation which takes two numbers as input and outputs their multiplied value, 2) a *sum* or summation operation which takes two values as input and outputs their sum, 3) a *read* operation which reads a particular number from memory and 4) a *write* operation which writes a value into memory. Note, that we do not consider read/write operations from/into low level memory (like caches and registers) that store temporary runtime values, e.g., outputs from summation and/or multiplications, since

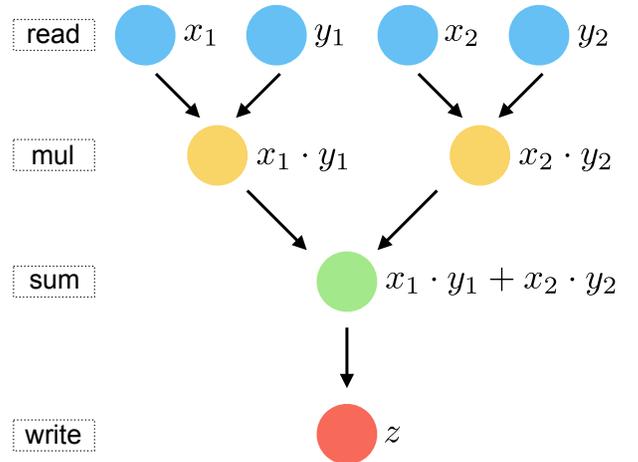


Fig. 2: Computational graph of a scalar product algorithm $X \cdot Y = z$ for two 2-dimensional input vectors X, Y . Any such algorithm can be described in terms of four elementary operations (*sum*, *mul*, *read*, *write*). These elementary operations are associated with functions $\sigma, \mu, \gamma, \delta$, which take a bit-size b as input and output the energy (and/or time) cost of performing that operation. Hence, assuming that all elements have same bit-size b , the total energy performance of the algorithm can be determined by calculating $E = 1\sigma(b) + 2\mu(b) + 4\gamma(b) + 1\delta(b)$.

their cost can be associated to those operations. Now, each of these nodes can be associated with an energy cost. Then, the total energy required for a particular dot product algorithm simply equals the total cost of the nodes in the graph.

However, the energy cost of each node depends on the hardware architecture and on the bit-size of the values involved in the operation. Hence, in order to make our model flexible with regard to different hardware architectures, we introduce four cost functions $\sigma, \mu, \gamma, \delta : \mathbb{N} \rightarrow \mathbb{R}$, which take as input a bit-size and output the energy cost of performing the operation associated to them²; σ is associated to the *sum* operation, μ to the *mul*, γ to the *read* and δ to the *write* operation.

Figure 2 shows the computational graph of a simple dot product algorithm for two 2-dimensional input vectors. This algorithm requires 4 *read* operations, 2 *mul*, 1 *sum* and 1 *write*. Assuming that the bit-size of all numbers is $b \in \mathbb{N}$, we can state that the energy cost of this dot product algorithm would be $E = 1\sigma(b) + 2\mu(b) + 4\gamma(b) + 1\delta(b)$. Note that similar energy models have been previously proposed [37], [38]. In the experimental section we validate the model by comparing it to real energy results measured by previous authors.

Considering this energy model we can now provide a detailed analysis of complexity of the CER and CSER data structure. However, we start with a brief analysis of the storage and energy requirements of the dense and sparse data structure in order to facilitate the comparison between them.

²The *sum* and *mul* operations take two numbers as input and they may have different bit-sizes. Hence in this case, we take the maximum of those as a reference for the bit-sizes involved in the operation.

B. Efficiency analysis of the dense and CSR formats

The dense data structure stores the matrix in an N -long array (where $N = m \times n$) using a constant bit-size b_Ω for each element. Therefore, it's effective per element storage requirement is

$$S_{\text{dense}} = b_\Omega \quad (1)$$

bits. The associated standard scalar product algorithm then has the following per element energy costs

$$E_{\text{dense}} = \sigma(b_o) + \mu(b_o) + \gamma(b_a) + \gamma(b_\Omega) + \frac{1}{n}\delta(b_o) \quad (2)$$

where b_a denotes the bit-size of the elements of the input vector $a \in \mathbb{R}^n$ and b_o the bit-size of the elements of the output vector. The cost (2) is derived from considering 1) loading the elements of the input vector $[\gamma(b_a)]$, 2) loading the elements of the matrix $[\gamma(b_\Omega)]$, 3) multiplying them $[\mu(b_o)]$, 4) summing the multiplications $[\sigma(b_o)]$, and 5) writing the result $[\delta(b_o)/n]$. We can see that both the storage and the dot product efficiency have a constant cost attached to them, despite the distribution of the elements of the matrix.

In contrast, the CSR data structure requires only

$$S_{\text{CSR}} = (1 - p_0)(b_\Omega + b_I) + \frac{1}{n}b_I \quad (3)$$

effective bits per element in order to represent the matrix, where b_I denotes the bits-size of the column indices. This comes from the fact that we need in total $N(1 - p_0)b_\Omega$ bits for representing the non-zero elements of the matrix, $N(1 - p_0)b_I$ bits for their respective column indices and mb_I bits for the row pointers. Moreover, it requires

$$E_{\text{CSR}} = (1 - p_0)(\sigma(b_o) + \mu(b_o) + \gamma(b_a) + \gamma(b_\Omega) + \gamma(b_I)) + \frac{1}{n}\gamma(b_I) + \frac{1}{n}\delta(b_o) \quad (4)$$

units of energy per matrix element in order to perform the dot product. The expression (4) was derived from 1) loading the non-zero element values $[(1 - p_0)\gamma(b_\Omega)]$, their respective indices $[(1 - p_0)\gamma(b_I) + \gamma(b_I)/n]$ and the respective elements of the input vector $[\gamma(b_a)]$, 2) multiplying and summing those elements $[\sigma(b_o) + \mu(b_o)]$ and then 3) writing the result into memory $[\delta(b_o)/n]$.

Different to the dense format, the efficiency of the CSR data structure increases as $p_0 \rightarrow 1$, thus, as the number of zero elements increases. Moreover, if the matrix-size is large enough, the storage requirement and the cost of performing a dot product becomes effectively 0 as $p_0 \rightarrow 1$.

For the ease of the analysis, we introduce the big \mathcal{O} notation for capturing terms that depend on the shape of the matrix. In addition, we denote the following set of operations

$$c_a = \sigma(b_a) + \gamma(b_a) + \gamma(b_I) \quad (5)$$

$$c_\Omega = \gamma(b_I) + \gamma(b_\Omega) + \mu(b_o) + \sigma(b_o) - \sigma(b_a) \quad (6)$$

c_a can be interpreted as the total effective cost of involving an element of the input vector in the dot product operation. Analogously c_Ω can be interpreted with regard to the elements

of the matrix. Hence, we can rewrite the above equations (2) and (4) as follows

$$E_{\text{dense}} = c_a + c_\Omega - 2\gamma(b_I) + \mathcal{O}(1/n) \quad (7)$$

$$E_{\text{CSR}} = (1 - p_0)(c_a + c_\Omega) + \mathcal{O}(1/n) \quad (8)$$

C. Efficiency analysis of the CER and CSER formats

Following a similar reasoning as above, we can state the following theorem

Theorem 1: Let $M \in \mathbb{R}^{m \times n}$ be a matrix. Let further $p_0 \in (0, 1)$ be the maximum likelihood estimate of the frequency of appearance of the zero element, and let $b_I \in \mathbb{N}$ be the bit-size of the numerical representation of a column or row index in the matrix. Then, the CER representation of M requires

$$S_{\text{CER}} = (1 - p_0)b_I + \frac{\bar{k} + \tilde{k}}{n}b_I + \mathcal{O}(1/n) + \mathcal{O}(1/N) \quad (9)$$

effective bits per matrix element, where \bar{k} denotes the average number of distinct elements that appear per row (excluding the most frequent value), \tilde{k} the average number of padded indices per row and $N = m \times n$ the total number of elements of the matrix. Moreover, the effective cost associated to the dot product with an input vector $a \in \mathbb{R}^n$ is

$$E_{\text{CER}} = (1 - p_0)c_a + \frac{\bar{k}}{n}c_\Omega + \frac{\tilde{k}}{n}\gamma(b_I) + \mathcal{O}(1/n) \quad (10)$$

per matrix element, where c_a and c_Ω are as in (5) and (6). Analogously, we can state

Theorem 2: Let M , p_0 , b_I , \bar{k} , c_a , c_Ω be as in theorem 1. Then, the CSER representation of M requires

$$S_{\text{CSER}} = (1 - p_0)b_I + \frac{2\bar{k}}{n}b_I + \mathcal{O}(1/n) + \mathcal{O}(1/N) \quad (11)$$

effective bits per matrix element, and the per element cost associated to the dot product with an input vector $a \in \mathbb{R}^n$ is

$$E_{\text{CSER}} = (1 - p_0)c_a + \frac{\bar{k}}{n}c_\Omega + \frac{\tilde{k}}{n}\gamma(b_I) + \mathcal{O}(1/n) \quad (12)$$

The proofs of theorems 1 and 2 are in the appendix. These theorems state that the efficiency of the data structures depends on the (p_0, \bar{k}) (sparsity - average number of distinct elements per row) values of the empirical distribution of the elements of the matrix. That is, these data structures are increasingly efficient for distributions that have high p_0 and low \bar{k} values. However, since the entropy measures the effective average number of distinct values that a random variable outputs³, both values are intrinsically related to it. In fact, from Renyi's generalized entropy definition [13] we know that $p_0 \geq 2^{-H}$. Moreover, the following properties are satisfied

- $\bar{k} \rightarrow \min\{K - 1, n\}$, as $H \rightarrow \log_2 K$ or $n \rightarrow \infty$, and

³From Shannon's source coding theorem [12] we know that the entropy H of a random variable gives the effective average number of bits that it outputs. Henceforth, we may interpret 2^H as the effective average number of distinct elements that a particular random variable outputs.

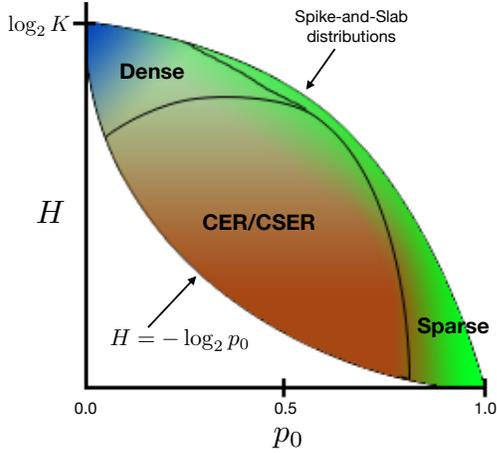


Fig. 3: Sketch of efficiency regions of the different data structures on the entropy-sparsity-plane (H denotes the entropy and p_0 the sparsity). A point in the plane corresponds to a distribution of the elements of a matrix, that has respective entropy-sparsity value. The intensity of the colors reflect the degree of the efficiency of the representations. More intense red regions indicate that the CER/CSER data structures are more efficient. Respectively, the colors blue and green indicate the degree of efficiency of the dense and sparse data structures. There are two lines that constrain the set of possible distributions. The bottom line corresponds to distributions whose entropy equal their respective min-entropy (that is, where $H = -\log_2 p_0$). The second line (at the most right) to the family of spike-and-slab distributions.

- $\bar{k} \rightarrow 0$, as $H \rightarrow 0$ or $n \rightarrow 1$.

Consequently, we can state the following corollary

Corollary 2.1: For a fixed set size of unique element $|\Omega| = K$, the storage requirements S as well as the cost of the dot product operation E of the CER and CSER representations satisfy

$$\begin{aligned} S, E &\leq \mathcal{O}(1 - 2^{-H}) + \mathcal{O}(K/n) + \mathcal{O}(1/N) \\ &= \mathcal{O}(1 - 2^{-H}) + \mathcal{O}(1/n) \end{aligned}$$

where p_0 , b_I , n and N are as in theorems 1 and 2, and H denotes the entropy of the matrix element distribution.

Thus, the efficiency of the CER and CSER data structures increase as the column size increases, or as the entropy decreases. Interestingly, when $n \rightarrow \infty$ both representations will converge to the same values, thus, will become equivalent. In addition, there will always exist a column size n where both formats are more efficient than the dense and sparse representations (see Fig. 5 where this trend is demonstrated experimentally).

D. Connection between CSR, CER and CSER

The CSR format is considered to be one of the most general sparse matrix representations, since it makes no further

assumptions regarding the empirical distribution of the matrix elements. Consequently, it implicitly assumes a spike-and-slab⁴ distribution on them, and are therefore accordingly optimized. However, spike-and-slab distributions are a particular class of low entropic (for sufficiently high sparsity levels p_0) distributions. In fact, spike-and-slab distributions have the highest entropy values compared to all other distributions that have same sparsity level. In contrast, as a consequence of corollary 2.1, the CER and CSER data structures relax this prior and can therefore efficiently represent the entire set of low entropic distributions. Hence, the CSR data structure may be interpreted as a more specialized version of the CER and CSER representations.

This may be more evident via the following example: consider the 1st row of the matrix example from section III

$$(0 \ 3 \ 0 \ 2 \ 4 \ 0 \ 0 \ 2 \ 3 \ 4 \ 0 \ 4)$$

The CSER data structure would represent the above row in the following manner

$$\begin{aligned} \Omega &:[0, 4, 3, 2] \\ colI &:[4, 9, 11, 1, 8, 3, 7] \\ \Omega I &:[1, 2, 3] \\ \Omega Ptr &:[0, 3, 5, 7] \\ rowPtr &:[0, 3] \end{aligned}$$

In comparison, the CER representation assumes that the ordering of the elements in ΩI is similar for all rows and therefore, it directly omits this array and implicitly encodes this information in the Ω array. Henceforth, the CER representation may be interpreted as a more explicit/specialized version of the CSER. The representation would then be

$$\begin{aligned} \Omega &:[0, 4, 3, 2] \\ colI &:[4, 9, 11, 1, 8, 3, 7] \\ \Omega Ptr &:[0, 3, 5, 7] \\ rowPtr &:[0, 3] \end{aligned}$$

Finally, since the CSR data structure assumes a uniform distribution over the non-zero elements, it omits the ΩPtr array since in such case all it's entries would redundantly be equal to 1. Henceforth, the respective representation would be

$$\begin{aligned} \Omega &:[3, 2, 4, 2, 3, 4, 4] \\ colI &:[1, 3, 4, 7, 8, 9, 11] \\ rowPtr &:[0, 7] \end{aligned}$$

Consequently, the CER and CSER representations will have superior performance for all those distributions that are not similar to the spike-and-slab distributions. Figure 3 displays a sketch of the regions on the entropy-sparsity plane where we expect the different data structures to be more efficient. The sketch shows that the efficiency of sparse data structures is high on the subset of distributions that are close to the right border line of the (H, p_0) -plane, thus, that are close to the family of

⁴That is, a spike at zero with probability p_0 and a uniform distribution over the non-zero elements.

TABLE I: Energy values (in pJ) of different elementary operations for a 45nm CMOS process [39]. We set the 8 bit floating point operations to be half the cost of a 16 bit operation, whereas we linearly interpolated the values in the case of the read and write operations.

Op	8 bits	16 bits	32 bits
float add	0.2	0.4	0.9
float mul	0.6	1.1	3.7
R/W (<8KB)	1.25	2.5	5.0
R/W (<32KB)	2.5	5.0	10.0
R/W (<1MB)	12.5	25.0	50.0
R/W (>1MB)	250.0	5000.0	1000.0

spike-and-slab distribution. In contrast, dense representations are increasingly efficient for high entropic distributions, hence, in the upper-left region. The CER and CSER data structures would then cover the rest of them. Figure 4 confirms this trend experimentally.

V. EXPERIMENTS

We applied the dense, CSR, CER and CSER representations on simulated matrices as well as on neural network weight matrices, and benchmarked their efficiency with regard to the following four criteria:

- 1) Storage requirements: We calculated the storage requirements according to equations (1), (3), (9) and (11).
- 2) Number of operations: We implemented the dot product algorithms associated to the four above data structures (pseudocodes of the CER and CSER formats can be seen in the appendix) and counted the number of elementary operations they require to perform a matrix-vector multiplication.
- 3) Time complexity: We timed each respective elementary operation and calculated the total time from the sum of those values.
- 4) Energy complexity: We estimated the respective energy cost by weighting each operation according to Table I. The total energy results consequently from the sum of those values. As for the case of the IO operations (read/write operations), their energy cost depend on the size of the memory the values reside on. Therefore, we calculated the total size of the array where a particular number is entailed and chose the respective maximum energy value. For instance, if a particular column index is stored using a 16 bit representation and the total size of the column index array is 30KB, then the respective read/write energy cost would be 5.0 pJ.

In addition, we used single precision floating point representations for the matrix elements and unsigned integer representations for the index and pointer arrays. For the later, we compressed the index-element-values to their minimum required bit-sizes, where we restricted them to be either 8, 16 or 32 bits.

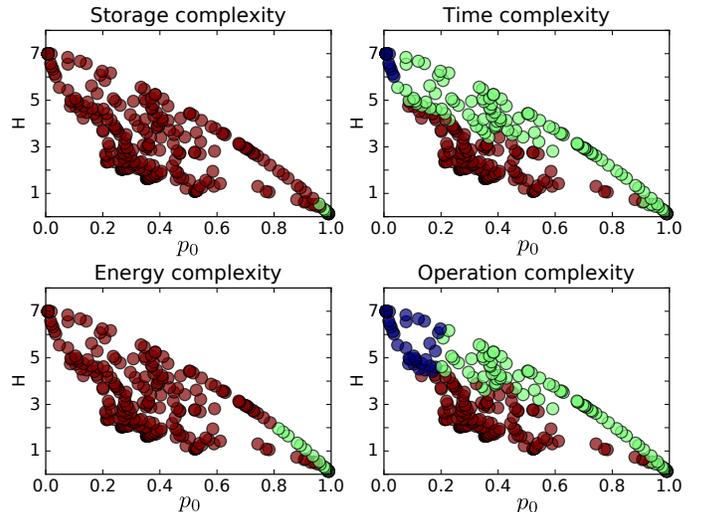


Fig. 4: The plots show the most efficient data structure at different points in the $H - p_0$ plane (to recall, H denotes the entropy of the matrix and p_0 the probability of the 0 value). We compare the dense data structure (blue), the CSR format (green) and the proposed data structures (red). The colors indicate which of the three categories was the most efficient at that point in the plane. The proposed data structures tend to be more efficient in the down left region of the plane. In contrast, sparse data structures tend to be more efficiency in the upper right corner, whereas dense structures in the upper left corner. For this experiment we employed a 100×100 matrix and calculated the average complexity over 10 matrix samples at each point. The size of the set of the elements was 2^7 .

A. Experiments on simulated matrices

As first experiments we aimed to confirm the theoretical trends described in Section IV.

1) *Efficiency on different regions of the entropy-sparsity plane*: Firstly, we argued that each distribution has a particular entropy-sparsity value, and that the superiority of the different data structures is manifested in different regions on that plane. Concretely, we expected the dense representation to be increasingly more efficient in the upper-left corner, the CSR on the bottom-right (and along the right border) and the CER and CSER on the rest.

Figure 4 shows the result of performing one such experiment. In particular, we randomly selected a point-distribution on the (H, p_0) -plane and sampled 10 different matrices from that distribution. Subsequently, we converted each matrix into the respective dense, CSR, CER and CSER representation, and benchmarked the performance with regard to the 4 different measures described above. We then averaged the results over these 10 different matrices. Finally, we compared the performances with each other and respectively color-coded the max result. That is, blue corresponds to the dense representation, green to the CSR and red to either the CER or CSER. As one can see, the result closely matches the expected behavior.

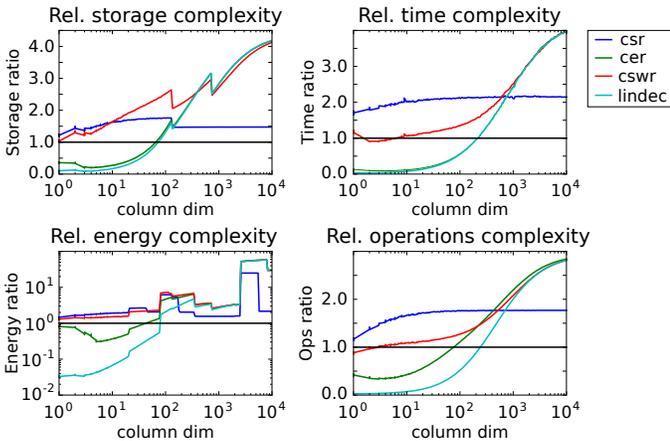


Fig. 5: Efficiency ratios compared to the dense data structure of the different data representations. We chose a matrix in the $H = 4.0 - p_0 = 0.55$ plane with fixed row size of 100, and averaged the values over 20 matrix samples. The size of the set of the elements was 2^7 . The proposed data structures tend to be more efficient as the column dimension of the matrix increases, and converge to the same value for $n \rightarrow \infty$.

2) *Efficiency as a function of the column size*: As second experiment, we study the asymptotic behavior of the data structures as we increase the column size of the matrices. From corollary 2.1 we expect that the CER and CSER data structures increase their efficiency as the number of columns in the matrix grows (thus, as $n \rightarrow \infty$), until they converge to the same point, outperforming the dense and sparse data structures. Figure 5 confirms this trend experimentally with regard to all four benchmarks. Here we chose a particular point-distribution on the (H, p_0) -plane and fixed the number of rows. Concretely, we chose $H = 4.0$, $p_0 = 0.55$ and $m = 100$ (the later is the row dimension), and measured the average complexity of the data structures as we increased the number of columns $n \rightarrow \infty$.

As a side note, the sharp changes in the plots are due to the sharp discontinuities in the values of table I. For instance, the sharp drops in storage ratios come from the change of the index bit-sizes, e.g., from $8 \rightarrow 16$ bits.

B. Compressed deep neural networks

As second set of experiments, we tested the efficiency of the proposed data structures on compressed deep neural networks. That is, we firstly lossy compressed the elements of the weight matrices of the networks, while ensuring that we negligible impact their prediction accuracy. As compression method we used an uniform quantizer over the range of weight values at each layer and subsequently rounded the values to their nearest quantization point. For all experiments of this section we quantized the weight elements down to 7 bits. Finally, we lossless converted the quantized weight matrices into the different data structures and tested their efficiency with regard to the four above mentioned benchmark criteria.

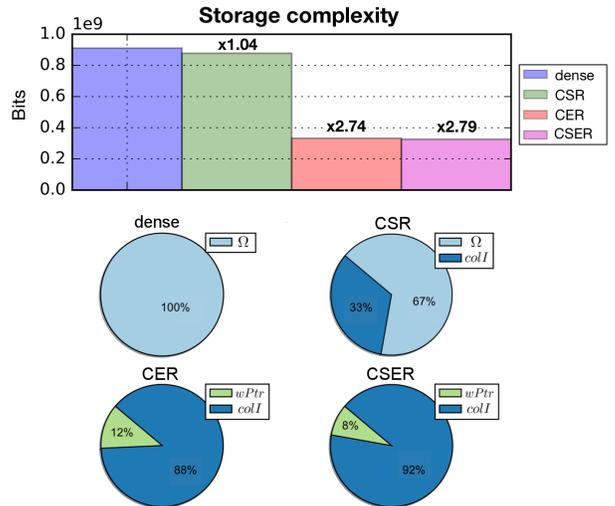


Fig. 6: Storage requirements of a compressed DenseNet [40] after converting it's weight matrices into the different data structures. The weights of the network layers were compressed down to 7 bits (resulting accuracy is 77.09%). The plots show the over the layers averaged result. Top chart: Compression ratio relative to the dense representation. Bottom charts: Contribution of the different parts of the data structures to the storage requirements. For the CER/CSER formats, most of the storage goes to the column indices.

TABLE II: Storage gains of different state-of-the-art neural networks after their weight matrices have been compressed down to 7 bits and, subsequently, converted into the different data structures. The performance gains are relative to the dense representation of the compressed weight matrices. The accuracy is measured with regard to the validation set (in parenthesis we show the accuracy of the uncompressed model).

Storage	Accuracy [%]	dense [MB]	CSR	CER	CSER
VGG16	68.51 (68.71)	553.43	x0.71	x2.11	x2.11
ResNet152	78.17 (78.25)	240.77	x0.76	x2.08	x2.10
DenseNet	77.09 (77.12)	114.72	x1.04	x2.74	x2.79

1) *Storage requirements*: Table II shows the gains in storage requirements of different state-of-the-art neural networks. Gains can be attained when storing the networks in CER or CSER formats. In particular, we achieve more than x2.5 savings on the DenseNet architecture, whereas in contrast the CSR data structure attains negligible gains. This is mainly attributed to the fact, that the dense and sparse representations store very inefficiently the weight element values of these networks. This is also reflected in Fig. 6, where one can see that most of the storage requirements for the dense and CSR representations is spend in storing the elements of the weight matrices Ω . In contrast, most of the storage cost for the CER and CSER data structures comes from storing the column indices $colI$, which is much lower than the actual

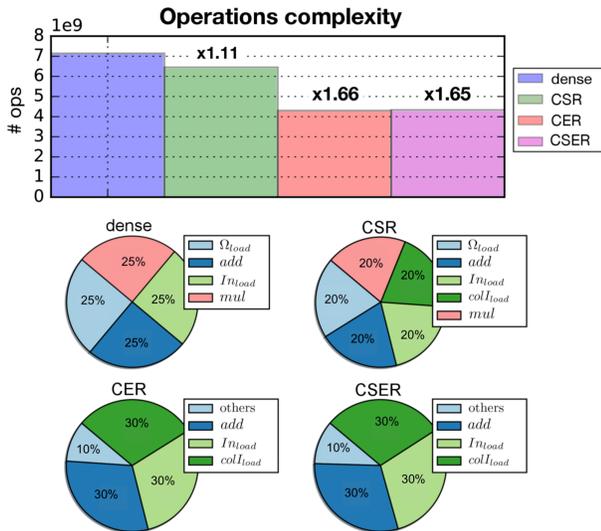


Fig. 7: Number of operations required to perform a dot product in the different formats for the experimental setup described in Fig. 6. The CER/CSER formats require less operations than the other formats, because 1) they do not need to perform as many multiplications and 2) they do not need to load as many matrix weight elements.

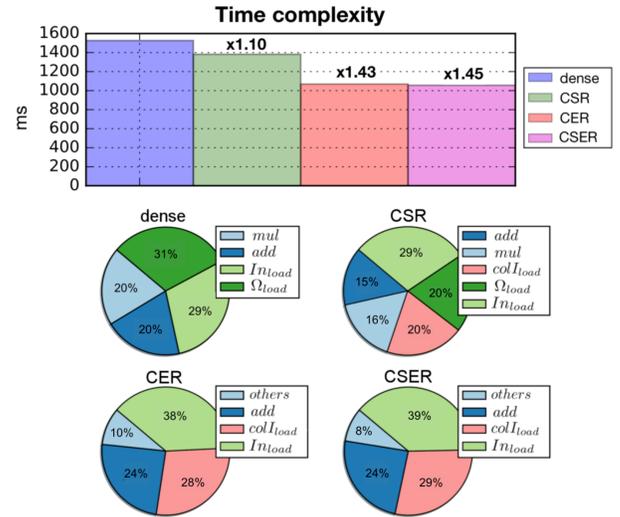


Fig. 8: Time cost of a dot product in the different formats for the experimental setup described in Fig. 6. The CER/CSER formats save time, because 1) they do not require to perform as many multiplications and 2) they do not spend as much time loading the matrix weight elements.

TABLE III: Gains attained with regard to the number of operations, time and energy cost needed for performing inference of different state-of-the-art neural networks. The experiment setting and table structure is the same as in Table II. The performance gains are relative to the dense representation of the compressed weight matrices.

	#ops [G]	dense	CSR	CER	CSER
VGG16	time [s]	15.08	x0.88	x1.40	x1.39
	energy [J]	3.37	x0.85	x1.27	x1.29
		2.70	x0.76	x2.37	x2.38
ResNet152	time [s]	10.08	x0.93	x1.42	x1.41
	energy [J]	2.00	x0.93	x1.30	x1.31
		1.92	x1.25	x3.73	x3.74
DenseNet	time [s]	7.14	x1.11	x1.66	x1.65
	energy [J]	1.53	x1.10	x1.43	x1.45
		0.51	x1.95	x6.40	x6.57

weight values.

2) *Number of operations*: Table III shows the savings attained with regard to number of elementary operations needed to perform a matrix-vector multiplication. As one can see, we can save up to 40% of the number of operations if we use the CER/CSER data structures on the DenseNet architecture. This is mainly due to the fact, that the dot product algorithm of the CER/CSER formats implicitly encode the distributive law of multiplications and consequently they require much less number of them. This is also reflected in Fig. 7, where

one can see that the CER/CSER dot product algorithms are mainly performing input load (In_{load}), column index load ($colI_{load}$) and addition (add) operations. In contrast, the dense and CSR dot product algorithms require an additional equal number of weight element load (Ω_{load}) and multiplication (mul) operations.

3) *Time cost*: In addition, Table III also shows that we attain speedups when performing the dot product in the new representation. Interestingly, Fig. 8 shows that most of the time is being consumed on IO's operations (that is, on $load$ operations). Consequently, the CER and CSER data structures attain speedups since they do not have to load as many weight elements. In addition, 20% and 16% of the time is spend in performing multiplications respectively in the dense and sparse representation. In contrast, this time cost is negligible for the CER and CSER representations.

4) *Energy cost*: Similarly, we see that most of the energy consumption is due to IOs operations (Fig. 9). Here the cost of loading an element may be up to 3 orders of magnitude higher than any other operations (see Table I) and henceforth, we obtain up to x6 energy savings when using the CER/CSER representations (see Table III).

Finally, Table IV and Fig. 10 further justify the observed gains. Namely, Table IV shows that the effective number of distinct elements per row is small relative to the networks effective column dimension. Moreover, Fig. 10 shows the distributions of the different layers of the networks on the entropy-sparsity plane where we see, that most of them lay in the regions where we expect the CER/CSER formats to be more efficient.

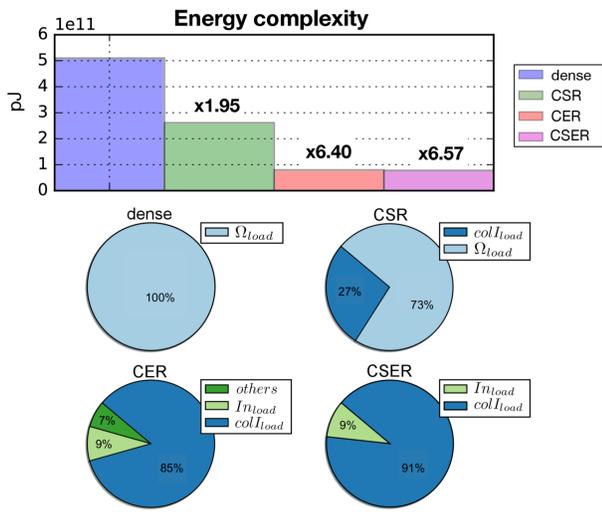


Fig. 9: Energy cost of a dot product in the different formats for the experimental setup described in Fig. 6. Performing loading operations consumes up to 3 orders more energy than sum and mult operations (see Table I). Since the CER/CSER formats need substantially less matrix weight element loading operations, they attain great energy saving compared to the dense and CSR formats.

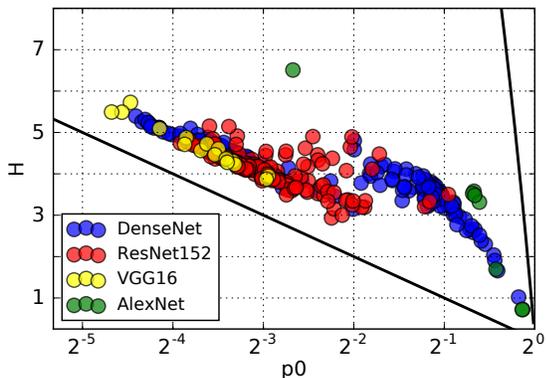


Fig. 10: Empirical distributions of the weight matrices of different neural network architectures after compression displayed on the entropy-sparsity plane. As we see, most of the layers lay in the region where the CER/CSER data structures to outperform the dense and sparse representation. The bottom and upper black line constrain the set of possible distributions.

C. Deep compression

Deep compression [26] is a technique for compressing neural networks that is able to attain high compression rates without loss of accuracy. It is able to do so by applying a three staged pipeline: 1) prune unimportant connections by employing algorithm [31], 2) cluster the non-pruned weight values and refine the cluster centers to the loss surface and 3) employ an entropy coder for storing the final weights. The

TABLE IV: Statistics of different neural network weight matrices taken over the entire network. p_0 denotes the effective sparsity level of the network, H stands for the effective entropy, \bar{k} represents the effective number of distinct elements per row, and n denotes the effective column dimension. We see that all neural networks have relatively low entropy, thus relatively low number of distinct elements compared to the very high column dimensionality.

	p_0	H	\bar{k}	n	\bar{k}/n
VGG16	0.07	4.8	55.80	10311.86	0.01
ResNet152	0.12	4.12	32.35	782.67	0.03
DenseNet	0.36	3.73	43.95	1326.93	0.03
AlexNet [26]	0.89	0.89	18.33	5767.85	0.01

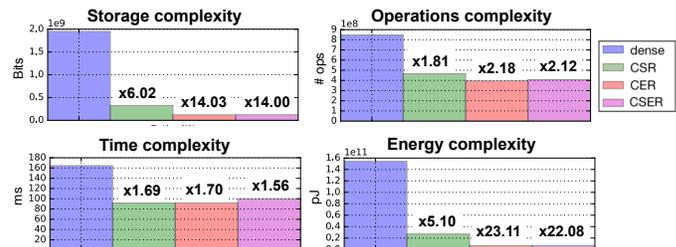


Fig. 11: Efficiency comparison of a compressed AlexNet [41] after converting its weight matrices into the different data structures. The network was compressed using the deep compression [26] technique. The plots show the over the layers averaged relative gains compared to the dense data structure.

first two stages allow to highly quantize the weight matrices of the networks without incurring significant loss of accuracy, whereas the third stage lossless converts the weight matrices into low-bit representation. However, one requires specialized hardware in order to efficiently exploit this final neural network representation during inference [42].

Henceforth, many authors benchmark the inference efficiency of highly compressed deep neural networks with regard to sparse representations when tested on standard hardware such as CPU's and/or GPU's [26], [35], [37]. However, we pay a high cost by doing so since sparse representations are highly redundant and inefficient for these type of networks. In contrast, the CER/CSER representation can be efficiently executed by conventional hardware and they are able to better exploit the redundancies entailed in low entropic neural networks. Hence, it is of high interest to benchmark their efficiency on highly compressed networks and compare them to their sparse (and dense) counterparts.

As experimental setup we chose the by the authors trained and quantized⁵ AlexNet architecture [41], where they were able to reduce the overall entropy of the network down to 0.89 without incurring any loss of accuracy. Figure 11 shows the gains in efficiency when the network layers are converted into the different data structures. We see, that the

⁵<https://github.com/songhan/Deep-Compression-AlexNet>

proposed data structures are able to surpass the dense and sparse data structures for all four benchmark criteria. Therefore, CER/CSER data structures are much less redundant and efficient representations of highly compressed neural network models. Interestingly, the CER/CSER data structures attain up to x20 energy savings, which is considerably higher than the sparse counterpart. We want to stress once more, that this gains apply directly, meaning that there is no need of specialized hardware in order to obtain them.

VI. CONCLUSION

We presented two new matrix representations, *compressed entropy row* (CER) and *compressed shared elements row* (CSER), that are able to attain high compression ratios and energy savings if the distribution of the matrix elements has low entropy. We showed on an extensive set of experiments that the CER/CSER data structures are more compact and computationally efficient representations of compressed state-of-the-art neural networks than dense and sparse formats. In particular, we attained up to x15 compression ratios and x20 energy savings by representing the weight matrices of an highly compressed AlexNet model in their CER/CSER forms.

By demonstrating the advantages of entropy-optimized data formats for representing neural networks, our work opens up new directions for future research, e.g., the exploration of entropy constrained regularization and quantization techniques for compressing deep neural networks. The combination of entropy constrained regularization and quantization and entropy-optimized data formats may push the limits of neural network compression even further and also be beneficial for applications such as federated or distributed learning [43], [44].



Simon Wiedemann received the M.Sc. degrees in applied mathematics from Technische Universität Berlin. He is currently with the Machine Learning Group, Fraunhofer Heinrich Hertz Institute, Berlin, Germany. His major research interests are in machine learning, neural networks and information theory.



Klaus-Robert Müller (M'12) has been a professor of computer science at Technische Universität Berlin since 2006; at the same time he is co-directing the Berlin Big Data Center. He studied physics in Karlsruhe from 1984 to 1989 and obtained his Ph.D. degree in computer science at Technische Universität Karlsruhe in 1992. After completing a postdoctoral position at GMD FIRST in Berlin, he was a research fellow at the University of Tokyo from 1994 to 1995. In 1995, he founded the Intelligent Data Analysis group at GMD-FIRST (later Fraunhofer FIRST) and directed it until 2008. From 1999 to 2006, he was a professor at the University of Potsdam. He was awarded the 1999 Olympus Prize by the German Pattern Recognition Society, DAGM, and, in 2006, he received the SEL Alcatel Communication Award, and, in 2014 he was granted the Science Prize of Berlin awarded by the Governing Mayor of Berlin, and, in 2017, he received the Vodafone Innovation Award. In 2012, he was elected to be a member of the German National Academy of Sciences-Leopoldina and in 2017 of the Berlin Brandenburg Academy of sciences, and in 2017 external scientific member of the Max Planck Society. His research interests are intelligent data analysis, machine learning, signal processing, and brain-computer interfaces.



Wojciech Samek (M'13) received a Diploma degree in computer science from Humboldt University of Berlin in 2010 and a Ph.D. degree in machine learning from the Technische Universität Berlin in 2014. In the same year he founded the Machine Learning Group at Fraunhofer Heinrich Hertz Institute which he currently directs. He is associated with the Berlin Big Data Center and was a Scholar of the German National Academic Foundation and a Ph.D. Fellow at the Bernstein Center for Computational Neuroscience Berlin. He was visiting Heriot-Watt University in Edinburgh and the University of Edinburgh from 2007 to 2008. In 2009 he was with the Intelligent Robotics Group at NASA Ames Research Center in Mountain View, CA. His research interests include interpretable machine learning, neural networks, signal processing and computer vision.

APPENDIX A

DETAILS ON NEURAL NETWORK EXPERIMENTS

A. Matrix preprocessing and convolutional layers

Before benchmarking the quantized weight matrices we applied the following preprocessing steps:

1) *Matrix decomposition*: After the quantization step it may well be that the 0 value is not included in the set of values and/or that it's not the most frequent value in the matrix. Therefore, we applied the following simple preprocessing steps: assume a particular quantized matrix $W_q \in \mathbb{R}^{m \times n}$, where each element $(W_q)_{ij} \in \Omega := \{\omega_0 = 0, \dots, \omega_{K-1}\}$ belong to a discrete set. Then, we decompose the matrix into the identity $W_q = (W_q - \omega_{\max} \mathbb{1}) + \omega_{\max} \mathbb{1} = \hat{W} + \omega_{\max} \mathbb{1}$, where $\mathbb{1}$ is the unit matrix whose elements are equal to 1 and ω_{\max} is the element that appears most frequently in the matrix. Consequently, \hat{W} is a matrix with 0 as it's most frequent element. Moreover, when performing the dot product with an input vector $x \in \mathbb{R}^n$, we only incur the additional cost of adding the constant value $c_{out} = \omega_{\max} \sum_i^n x_i$ to all the elements of the output vector. The cost of this additional operation is effectively of the order of n additions and 1 multiplication for the entire dot product operation, which is negligible as long as the number of rows is sufficiently large.

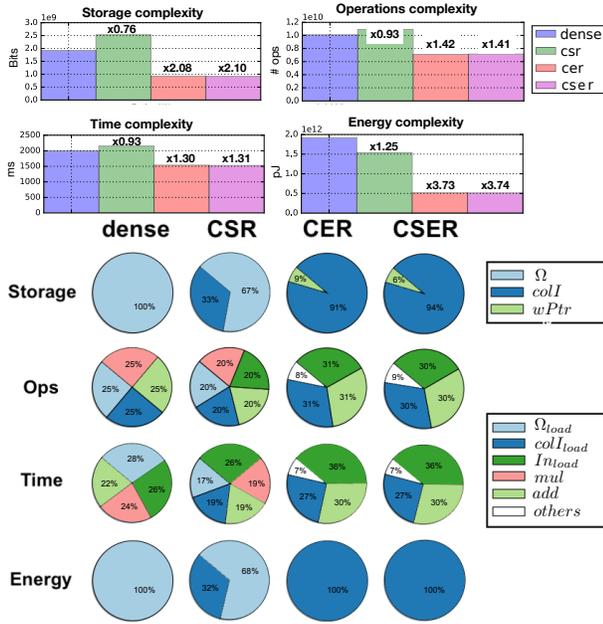


Fig. 12: Efficiency results from a compressed ResNet152. The experimental details are as described in section V-B

2) *Convolution layers*: A convolution operation can essentially be performed by a matrix-matrix dot product operation. The weight tensor containing the filter values would be represented as a $(F_n \times (n_{ch}m_Fn_F))$ -dimensional matrix, where F_n is the number of filters of the layer, n_{ch} the number of channels, and (m_F, n_F) the height/width of the filters. Hence, the convolution matrix would perform a dot product operation with an $((n_{ch}m_Fn_F) \times n_p)$ -dimensional matrix, that contains all the patches n_p of the input image as column vectors.

Hence, in our experiments, we reshaped the weight tensors of the convolutional layers into their respective matrix forms and tested their storage requirements and dot product complexity by performing a simple matrix-vector dot product, but weighted the results by the respective number of patches n_p that would have been used at each layer.

B. More results from experiments

Figures 12, 13 and 14 show our results for compressed ResNet152, VGG16 and AlexNet, respectively.

C. Dot product pseudocodes

Algorithms 1 and 2 show the pseudocodes of the dot product algorithm of the CER and CSER data structures.

APPENDIX B PROOF OF THEOREMS

1) *Theorem 1*: The CER data structure represents any matrix via 4 arrays, which respectively contain: $\Omega : K$, $colI : N - \#(0)$, $wPtr : \sum_{r=0}^m k_r + \tilde{k}_r$, $rowPtr : m$ entries, where K denotes the number of unique elements appearing in the

Algorithm 1 CER dot product

```

1: procedure DOTCER(A,X)
2:    $\Omega, colI, wPtr, rowPtr \leftarrow A$ 
3:    $N, L \leftarrow dim(X)$ 
4:    $Y = 0 \in \mathbb{R}^{M \times L}$ 
5:   for  $l \leq L$  do
6:      $r_{start} = 0$ 
7:      $w_{start} = 0$ 
8:     for  $r_{idx} = 1 < len(rowPtr)$  do
9:        $r_{end} \leftarrow rowPtr[r_{idx}]$ 
10:       $y = 0$ 
11:       $w_{count} = 1$ 
12:      for  $w_{idx} = r_{start} + 1 < r_{end} + 1$  do
13:         $w_{end} \leftarrow wPtr[w_{idx}]$ 
14:         $y' = 0$ 
15:        for  $i = w_{start} < w_{end}$  do
16:           $I \leftarrow colI[i]$ 
17:           $y' \leftarrow y' + X[I, l]$ 
18:          if  $w_{start} + 1 = w_{end}$  then
19:             $y \leftarrow y + y' * \Omega[w_{count}]$ 
20:           $w_{count} \leftarrow w_{count} + 1$ 
21:           $w_{start} \leftarrow w_{end}$ 
22:         $r_{start} \leftarrow r_{end}$ 
23:         $Y[m, k] \leftarrow y$ 
24:   return Y

```

Algorithm 2 CSER dot product

```

1: procedure DOTCSER(A,X)
2:    $\Omega, colI, wI, wPtr, rowPtr \leftarrow A$ 
3:    $N, L \leftarrow dim(X)$ 
4:    $Y = 0 \in \mathbb{R}^{M \times L}$ 
5:   for  $l \leq L$  do
6:      $r_{start} = 0$ 
7:      $w_{start} = 0$ 
8:      $w_{count} = 0$ 
9:     for  $r_{idx} = 1 < len(rowPtr)$  do
10:       $r_{end} \leftarrow rowPtr[r_{idx}]$ 
11:       $y = 0$ 
12:      for  $w_{idx} = r_{start} + 1 < r_{end} + 1$  do
13:         $w_{end} \leftarrow wPtr[w_{idx}]$ 
14:         $y' = 0$ 
15:        for  $i = w_{start} < w_{end}$  do
16:           $I \leftarrow colI[i]$ 
17:           $y' \leftarrow y' + X[I, l]$ 
18:           $y \leftarrow y + y' * \Omega[wI[w_{count}]]$ 
19:           $w_{count} \leftarrow w_{count} + 1$ 
20:           $w_{start} \leftarrow w_{end}$ 
21:         $r_{start} \leftarrow r_{end}$ 
22:         $Y[m, k] \leftarrow y$ 
23:   return Y

```

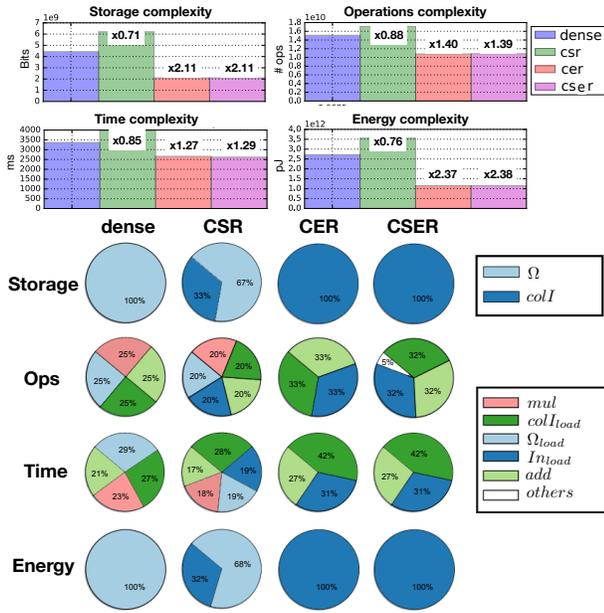


Fig. 13: Efficiency results from a compressed VGG16. The experimental details are as described in section V-B

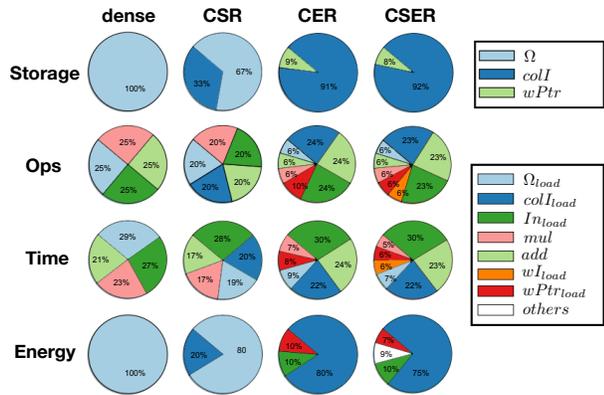


Fig. 14: Efficiency results of a compressed AlexNet. The experimental details are as described in section V-C

matrix, N the total number of elements, $\#(0)$ the total number of zero elements, m the row dimension and finally, \bar{k}_r the number of distinct elements that appeared at row r (excluding the 0) and \tilde{k}_r the number of redundant padding entries needed to communicate at row r .

Hence, by multiplying each array with the respective element bit-size and dividing by the total number of elements we get

$$\frac{Kb_\Omega}{N} + \left(1 - \frac{\#(0)}{N}\right)b_I + \frac{1}{N} \left(\sum_{r=0}^m \bar{k}_r + \tilde{k}_r\right)b_I + \frac{1}{n}b_I$$

where b_Ω and b_I are the bit-sizes of the matrix elements and the indices respectively. With $p_0 = \frac{\#(0)}{N}$ and $\bar{k} + \tilde{k} = \frac{1}{m} \sum_{r=0}^m \bar{k}_r + \tilde{k}_r$ we get equation 9.

The cost of the respective dot product algorithm can be estimated by calculating the cost of each line of algorithm 2. To recall, we denoted with $\sigma(b)$ the cost of performing a summation operation, which involved b bits. $\mu(b)$ the cost of a multiplication. $\gamma(b)$ the cost of a read and $\delta(b)$ of a write operation into memory. We further denoted with c_{oth} the cost of performing other types of operations. Moreover, assume an input vector (that is, $L = 1$), since the result can be trivially extended to input matrices of arbitrary size. Thus, algorithm 1 requires: from line 2) - 7) we assume a cost of c_{oth} , 8) mc_{oth} , 9) $m\gamma(b_I)$, 10) mc_{oth} , 11) mc_{oth} , 12) $m(\bar{k} + \tilde{k})c_{oth}$, 13) $m(\bar{k} + \tilde{k})\gamma(b_I)$, 14) $m(\bar{k} + \tilde{k})c_{oth}$, 15) $N(1 - p_0)c_{oth}$, 16) $N(1 - p_0)\gamma(b_I)$, 17) $N(1 - p_0)(\gamma(b_a) + \sigma(b_a))$, 18) $m\bar{k}(\gamma(b_\Omega) + \mu(b_o) + \sigma(b_o) - \sigma(b_a))$, 19) $m(\bar{k} + \tilde{k})c_{oth}$, 20) $m(\bar{k} + \tilde{k})c_{oth}$, 21) mc_{oth} , 22) $m\delta(b_o)$; where b_Ω , b_I and b_o are the bit-sizes of the matrix elements, the indices and output vector element respectively. Hence, adding up all above costs and replacing c_a and c_Ω as in equations (5) and (6), we can get the total cost of $c_{oth} + m(\gamma(b_I) + \delta(b_o) + 3c_{oth}) + m\bar{k}c_\Omega + m\tilde{k}(\gamma(b_I) + 4c_{oth}) + N(1 - p_0)(c_a + c_{oth})$. It is fair to assume that the cost c_{oth} is negligible compared to the rest for highly optimized algorithms. Indeed, figures 8 and 7 and 9 show that cost of these operations contribute very little to the total cost of the algorithm. Hence, we can assume the ideal cost of the algorithm to be equal to the above expression with $c_{oth} = 0$ (which corresponds to equation (10)).

□

2) *Theorem 2*: Analogously, we can follow the same line of arguments. Namely, each array in the CSER data structure contains: $\Omega : K$, $colI : N - \#(0)$, $wI : \sum_{r=0}^m \bar{k}_r$, $wPtr : \sum_{r=0}^m \bar{k}_r$, $rowPtr : m$ entries. Consequently, by adding those terms, multiplying by their bit-size and dividing by the total number of elements N we recover (11).

Each line of algorithm 2 induces a cost of: from line 2) - 8) we assume a cost of c_{oth} , 9) mc_{oth} , 10) $m\gamma(b_I)$, 11) mc_{oth} , 12) $m\bar{k}c_{oth}$, 13) $m\bar{k}\gamma(b_I)$, 14) $m\bar{k}c_{oth}$, 15) $N(1 - p_0)c_{oth}$, 16) $N(1 - p_0)\gamma(b_I)$, 17) $N(1 - p_0)(\gamma(b_a) + \sigma(b_a))$, 18) $m\bar{k}(\gamma(b_\Omega) + \gamma(b_I) + \mu(b_o) + \sigma(b_o) - \sigma(b_a))$, 19) $m\bar{k}c_{oth}$, 20) $m\bar{k}c_{oth}$, 21) mc_{oth} , 22) $m\delta(b_o)$.

Again, adding up all terms and replacing with c_a and c_Ω then we get the total cost of $c_{oth} + m(\gamma(b_I) + \delta(b_o) + 3c_{oth}) + m\bar{k}(c_\Omega + \gamma(b_I) + 4c_{oth}) + N(1 - p_0)(c_a + c_{oth})$ and with $c_{oth} = 0$ we recover equation (12). □

REFERENCES

- [1] R. H. Landau, J. Paez, and C. C. Bordeianu, *A Survey of Computational Physics: Introductory Computational Science*. Princeton, NJ, USA: Princeton University Press, 2008.
- [2] D. M. Young and R. T. Gregory, *A Survey of Numerical Mathematics*. New York, NY, USA: Dover Publications, Inc., 1988.
- [3] S. Krig, *Computer Vision Metrics: Survey, Taxonomy, and Analysis*, 1st ed. Berkely, CA, USA: Apress, 2014.
- [4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [5] Y. LeCun, L. Bottou, G. B. Orr, and K. Müller, "Efficient backprop;" in *Neural Networks: Tricks of the Trade - Second Edition*, Springer LNCS 7700, 2012, pp. 9–48.

- [6] S. Afroz, M. Tahaseen, F. Ahmed, K. S. Farshee, and M. N. Huda, "Survey on matrix multiplication algorithms," in *5th International Conference on Informatics, Electronics and Vision (ICIEV)*, 2016, pp. 151–155.
- [7] M. Bläser, "Fast matrix multiplication," *Theory of Computing, Graduate Surveys*, vol. 5, pp. 1–60, 2013.
- [8] I. S. Duff, "A survey of sparse matrix research," *Proceedings of the IEEE*, vol. 65, no. 4, pp. 500–535, 1977.
- [9] V. Karakasis, T. Gkountouvas, K. Kourtis, G. Goumas, and N. Koziris, "An extended compression format for the optimization of sparse matrix-vector multiplication," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 10, pp. 1930–1940, 2013.
- [10] J. King, T. Gilray, R. M. Kirby, and M. Might, "Dynamic-csr: A format for dynamic sparse-matrix updates," in *High Performance Computing - 31st International Conference, ISC High Performance 2016, Proceedings*, ser. LNCS, vol. 9697. Springer Verlag, 2016, pp. 61–80.
- [11] R. Yuster and U. Zwick, "Fast sparse matrix multiplication," in *Algorithms - ESA 2004*. Springer Berlin Heidelberg, 2004, pp. 604–615.
- [12] C. E. Shannon, "A mathematical theory of communication," *SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.
- [13] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 1961, pp. 547–561.
- [14] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [15] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.
- [16] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1026–1034.
- [17] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2018.
- [18] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very Deep Convolutional Neural Networks for Raw Waveforms," *arXiv preprint arXiv:1610.00087*, 2016.
- [19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *International Conference on Representation Learning (ICLR)*, 2015.
- [20] P. Baldi, P. Sadowski, and D. Whiteson, "Searching for exotic particles in high-energy physics with deep learning," *Nature Communications*, vol. 5, p. 4308, 2014.
- [21] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial eeg classification," *Journal of Neuroscience Methods*, vol. 274, pp. 141–145, 2016.
- [22] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," *Nature Communications*, vol. 8, p. 13890, 2017.
- [23] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, "Machine learning of accurate energy-conserving molecular force fields," *Science Advances*, vol. 3, no. 5, 2017.
- [24] M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. de Freitas, "Predicting parameters in deep learning," in *Advances in Neural Information Processing Systems*, 2013, pp. 2148–2156.
- [25] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *arXiv:1503.02531*, 2015.
- [26] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and Huffman coding," *arXiv:1510.00149*, 2015.
- [27] D. Molchanov, A. Ashukha, and D. Vetrov, "Variational dropout sparsifies deep neural networks," in *34th International Conference on Machine Learning*, 2017, pp. 2498–2507.
- [28] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," *arXiv:1710.09282*, 2017.
- [29] Y. L. Cun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in Neural Information Processing Systems 2*, 1990, pp. 598–605.
- [30] B. Hassibi, D. G. Stork, and G. J. Wolff, "Optimal brain surgeon and general network pruning," in *IEEE International Conference on Neural Networks*, 1993, pp. 293–299 vol.1.
- [31] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 1135–1143.
- [32] V. Vanhoucke, A. Senior, and M. Z. Mao, "Improving the speed of neural networks on cpus," in *NIPS'11 Deep Learning and Unsupervised Feature Learning Workshop*, 2011.
- [33] D. D. Lin, S. S. Talathi, and V. S. Annapureddy, "Fixed point quantization of deep convolutional networks," in *International Conference on Machine Learning*, 2016, pp. 2849–2858.
- [34] F. Li, B. Zhang, and B. Liu, "Ternary weight networks," *arXiv:1605.04711*, 2016.
- [35] C. Louizos, K. Ullrich, and M. Welling, "Bayesian Compression for Deep Learning," in *Advances in Neural Information Processing Systems*, 2017, pp. 3290–3300.
- [36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014.
- [37] T. Yang, Y. Chen, and V. Sze, "Designing energy-efficient convolutional neural networks using energy-aware pruning," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5687–5695.
- [38] Y. H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 127–138, 2017.
- [39] M. Horowitz, "1.1 computing's energy problem (and what we can do about it)," in *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, 2014, pp. 10–14.
- [40] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4700–4708.
- [41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012, pp. 1097–1105.
- [42] S. Han, X. Liu, H. Mao, J. Pu, A. Pedram, M. A. Horowitz, and W. J. Dally, "EIE: efficient inference engine on compressed deep neural network," in *ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, 2016, pp. 243–254.
- [43] H. B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.
- [44] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, "Sparse binary compression: Towards distributed deep learning with minimal communication," *arXiv:1805.08768*, 2018.