DeepCABAC: A Universal Compression Algorithm for Deep Neural Networks

Simon Wiedemann¹, Heiner Kirchhoffer¹, Stefan Matlage, Paul Haase, Arturo Marban, Talmaj Marine, David Neumann, Tung Nguyen, Ahmed Osman, Detlev Marpe*, Fellow, IEEE, Heiko Schwarz, Thomas Wiegand, Fellow, IEEE and Wojciech Samek*, Member, IEEE

Abstract—The field of video compression has developed some of the most sophisticated and efficient compression algorithms known in the literature, enabling very high compressibility for little loss of information. Whilst some of these techniques are domain specific, many of their underlying principles are universal in that they can be adapted and applied for compressing different types of data. In this work we present DeepCABAC, a compression algorithm for deep neural networks that is based on one of the state-of-the-art video coding techniques. Concretely, it applies a Context-based Adaptive Binary Arithmetic Coder (CABAC) to the network’s parameters, which was originally designed for the H.264/AVC video coding standard and became the state-of-the-art for lossless compression. Moreover, DeepCABAC employs a novel quantization scheme that minimizes the rate-distortion function while simultaneously taking the impact of quantization onto the accuracy of the network into account. Experimental results show that DeepCABAC consistently attains higher compression rates than previously proposed coding techniques for neural network compression. For instance, it is able to compress the VGG16 ImageNet model by x63.6 with no loss of accuracy, thus being able to represent the entire network with merely 8.7MB. The source code for encoding and decoding can be found at https://github.com/fraunhoferhhi/DeepCABAC.

Keywords—Neural Network Compression, Efficient Neural Network Representation, Source Coding, Rate-Distortion Quantization, Context-based Adaptive Binary Arithmetic Coding.

I. INTRODUCTION

It has been well established that deep neural networks excel at solving many complex machine learning tasks [1], [2]. Their relatively recent success can be attributed to three phenomena: 1) access to large amounts of data, 2) researchers having designed novel optimization algorithms and model architectures that allow to train very deep neural networks, 3) the increasing availability of compute resources [1]. In particular, the latter two allowed machine learning practitioners to equip neural networks with an ever-growing number of layers and, consequently, to consistently attain state-of-the-art results on a wide spectrum of complex machine learning tasks. However, this has triggered an exponential growth in the number of parameters these models entail over the past years [3]. Trivially, this implies that the models are becoming more and more complex in terms of memory. This can become very problematic since it does not only imply higher memory requirements, but also slower runtimes and high energy consumption [4]. In fact, IO operations can be up to three orders of magnitude more expensive than arithmetic operations. Moreover, [3] show that the memory-energy efficiency trends of most common hardware platforms are not able to keep up with the exponential growth of the neural networks’ sizes, thus expecting them to be more and more power hungry over time.

In addition, there has also been an increasing demand on deploying deep models to resource constrained devices such as mobile or wearable devices [5]–[7], as well as on training deep neural networks in a distributed setting such as in federated learning [8]–[10], since these approaches have direct advantages with regards to privacy, latency and efficiency issues. High memory complexity greatly complicates the applicability of neural networks for those use cases, in particular for the federated learning case since the parameters of the networks are transmitted through communication channels with limited capacity.

Model compression is one possible paradigm to solve this problem. Namely, by attempting to maximally compress the information contained in the network’s parameters we automatically leave only the bits that are necessary for solving the task. Thus, in principle, the memory complexity of deep models should only increase with the complexity of the learning task and not with its number of parameters¹. In addition, model compression has direct practical advantages such as reduced communication and compute cost [12]–[14]. In fact, the Moving Picture Expert Group (MPEG) of the International Organization of Standards (ISO) has recently issued a call on neural network compression [15], which stresses the relevance of the problem and the broad interest by the industry to find practical solutions.

A. Entropy coding in video compression

The topic of signal compression has been long studied and highly practical and efficient algorithms have been designed. State-of-the-art video compression schemes like H.265/HEVC [16] employ efficient entropy coding techniques that can also be used for compressing deep neural networks. Namely, the

¹To be more precise, the memory complexity of the model should only increase sublinearly with its number of parameters [11].
Context-based adaptive binary arithmetic coding (CABAC) engine [17] provides a very flexible interface for entropy coding that can be adapted to a wide range of applications. It is optimized to allow high throughput and a high compression ratio at the same time. In particular, the transform coefficient coding part of H.265/HEVC contains many interesting aspects that might be suitable for compressing deep neural network. Hence, it appears only natural to try to adapt current state-of-the-art compression techniques such as CABAC to deep neural networks and accordingly compress them.

B. Contributions

Our contributions can be summarized as follows:

1) We adapt CABAC for the task of neural network compression. To the best of our knowledge, we are the first to apply state-of-the-art coding techniques from video compression to deep neural networks.

2) We quantize the parameters of the networks by minimizing a generalized form of a rate-distortion function which takes the impact of quantization on the accuracy of the network into account.

3) In our experiments we show that DeepCABAC is able to attain very high compression ratios and that it consistently attains a higher compression performance than previously proposed coders.

C. Outline

In section II we start by reviewing some basic concepts from information theory, in particular from source coding theory. We also highlight the main difference between the classical source coding and the model compression paradigms in subsection II-D. Subsequently, we proceed by explaining DeepCABAC in section III. In section IV we provide a comprehensive review of the related work on neural network compression. Finally, we provide experimental results and a respective discussion in section V.

II. Source coding

Source coding is a subfield of information theory that studies the properties of so called codes. These are mappings that assign a binary representation and a reconstruction value to a given input element. Figure 1 depicts their most common structure. They are comprised of two parts, an encoder and a decoder. The encoder is a mapping that assigns a binary string of finite length \( b \) to an input element \( w \). In contrast, the decoder assigns a reconstruction value \( q \) to the corresponding binary representation. We will also sometimes refer to \( q \) as a quantization point. Furthermore, it is assumed that the output elements \( b \) and \( q \) of the code \( C \) are elements of finite countable sets, and that there is a one-to-one correspondence between them. Therefore, without loss of generality, we can decompose the encoder into a quantizer and a binarizer, where the former maps the input to an integer value \( Q(w) = i \in \mathbb{Z} \), and the latter maps the integers to their corresponding binary representation \( B(i) = b \). Analogously for the decoder. Naturally, it follows that the binarizer is always a bijective map, thus \((B^{-1} \circ B)(i) = i\).

We also distinguish between two types of codes, the so called lossless codes and lossy codes. They respectively correspond to the cases where \( Q \) is either bijective or not, thus, the latter implies that information is lost in the coding process. Therefore, we stress that the map \( Q^{-1} \) does not necessarily have to be the inverse of \( Q \).

After establishing the basic definition of codes, we will now formalize the source coding problem. In simple terms, source coding studies the problem of finding the code that maximally compresses a set of input samples, while maintaining the error between the input and reconstruction values under an error tolerance constraint.

Notice that the problem is probabilistic in its nature since it implicitly assumes that the decoder has no access to the element values being encoded. Moreover, the input values themselves may come from an unknown source distribution. Hence, we denote with \( P_{\text{Enc}}(w) \) the encoders probability model of \( w \), and with \( P_{\text{Dec}}(q) \) \((\equiv P_{\text{Dec}}(b) \equiv P_{\text{Dec}}(i)\)) the decoders probability model of \( q \) (or equivalently \( b \) and \( i \)). It is important to stress that both models do not have to coincide, thus \( P_{\text{Enc}}(Q(w)) = P_{\text{Enc}}(i) \neq P_{\text{Dec}}(i) \). Furthermore, we will assume that the encoder’s probability model follows the true underlying distribution of the input source, and therefore we will simply write \( P_{\text{Enc}}(w) \equiv P(w) \).

Thus, the source coding problem can be formulated more precisely as follows: let \( \mathbb{W} \subseteq \mathbb{R}^n \) be a given input set and let \( P(w) \) be the probability of an element \( w \in \mathbb{W} \) being sampled. Then, find a code \( C^* \) that

\[
C^* = \arg \min_C \mathbb{E}_{P(w)}[D(w, q) + \lambda L_C(b)] \tag{1}
\]
where \( b = (B \circ Q)(w) \), \( q = (Q^{-1} \circ Q)(w) \). \( D \) is some distance measure and \( L_C \) the length of the binary representation \( b \). We will sometimes refer to \( L_C(\cdot) \) as the code-length of a sample, and with \( D \) the distortion between \( w \) and \( q \). \( \mathbb{E}_P[\cdot] \) denotes expectations as taken by the probability distribution \( P \), \( \lambda \in \mathbb{R} \) is the Lagrange multiplier that controls the trade-off between the compression strength and the error incurred by it.

Minimization objectives of the form (1) are called rate-distortion objectives in the source coding literature. Namely, due to their sequential decomposition codes we want to recall an important property of joint probabil-

**A. Lossless coding**

Lossless coding implies that \( q = (Q^{-1} \circ Q)(w) = w \forall w \).

Thus, \( D(w,q) = 0 \ \forall w \in \mathbb{W} \) in (1) and the rate-distortion objective simplifies into finding a binarizer \( B^* \) that maximally compresses the input samples. Hence, throughout this section we will equate the general code \( C \) with the binarizer \( B \) and refer to it accordingly. Moreover, we will also assume that the decoder’s probability model equals the encoder’s, thus \( P_{\text{dec}} = P_{\text{enc}} \). In the next subsection we discuss the case when the latter property does not apply.

Information theory already makes concrete statements regarding the minimum information contained in a probability source. Namely, Shannon in its influential work [18] stated that the minimum information required to fully represent a sample \( w \) that has probability \( P(w) \) is of \(- \log_2 P(w)\) bits. Consequently, the entropy \( H_P(\mathbb{W}) = \sum_{w \in \mathbb{W}} -P(w) \log_2 P(w) \) states the minimum average number of bits required to represent any element \( w \in \mathbb{W} \subset \mathbb{R}^n \). This implies that

\[
H_P(\mathbb{W}) < \bar{L}_C(\mathbb{W})
\]

where \( \bar{L}_C(\mathbb{W}) = \sum_{w \in \mathbb{W}} P(w) L_C(w) \) is the average code-length that any code \( C \) assigns to each element \( w \in \mathbb{W} \). Eq. (2) is also referred to as the fundamental theorem of lossless coding.

Fortunately, from the source coding literature [19] we know of the existence of codes that are able to reach average code-length of up to only 1 bit of redundancy to the theoretical minimum. That is,

\[
\exists C : H_P(\mathbb{W}) < \bar{L}_C(\mathbb{W}) \leq H_P(\mathbb{W}) + 1
\]

Moreover, we even know how to build them.

Before we start discussing in more detail some of these codes we want to recall an important property of joint probability distributions. Namely, due to their sequential decomposition property, we can express the minimal information entailed in the output sample \( w \in \mathbb{R}^n \) of a joint probability distribution \( P(w) \) sequentially as

\[
- \log_2 P(w) = - \sum_{j=0}^{n-1} \log_2 P(w_j | w_{j-1}, \ldots, w_0)
\]

That is, we can always interpret a given input vector as an \( n \)-long random process and encode its outputs sequentially. As long as we know the respective conditional probability distributions, we can optimally encode the entire sequence. Respectively, we denote with \( w_j \) the scalar value of the \( j \)-th dimension of \( w \) (or equivalently \( j \)-th output of the random process). Also, we denote with \( \mathbb{W}_s \) the set of possible scalar inputs, where \( w_j \in \mathbb{W}_s \).

1) (scalar) Huffman coding: One optimal code is the well-known Huffman code [20]. It consists of building a binary tree such that each input sample \( w \) is associated with one of the leaves of the tree. Thus, each \( w \) can be associated with the sequence of binary decisions that traverse the tree from its root point. The main idea is then to build the tree in such a manner that shorter paths are associated to more probable samples \( w \). Huffman successfully proved that this code satisfies (3). We provide a pseudocode of the encoding and decoding process in the appendix (see algorithms 3, 1, and 2).

However, Huffman codes can be very inefficient in practice since the Huffman-tree grows very quickly for large input dimensions \( n \). Therefore, most often scalar Huffman codes are used instead. Scalar Huffman codes do only consider 1-dimensional inputs, and do accordingly encode each sample from the \( n \)-long random process. However, these codes are suboptimal in that they produce redundant binary representations and do therefore not satisfy (3). Concretely, they produce average code-lengths of

\[
H_P(\mathbb{W}_s) < \tilde{L}_{\text{SH}}(\mathbb{W}_s) \leq H_P(\mathbb{W}) + n
\]

where now \( P(w_j) \) is the probability of a scalar output \( w_j \) and \( \tilde{L}_{\text{SH}}(\cdot) \) is the average code-length produced by the scalar Huffman code. Moreover, they are limited to stationary processes since they do not take conditional dependencies into account, which could further reduce the average code-length.

2) Arithmetic coding: A concept that approaches the joint entropy \( H(\mathbb{W}) \) of eq. (3) in a practical and efficient manner is arithmetic coding. It consists of expressing a particular sequence of samples \( w_0, w_1, \ldots, w_{n-1} \) of an \( n \)-long random process as a so called coding interval. An overview of the idea is given in the following.

Let \( [L_j, L_j + R_j] \) be the coding interval before encoding symbol \( w_j \) and let \( L_0 = 0 \) and \( R_0 = 1 \). Encoding of a symbol
\( w_j \) corresponds to deriving a coding interval \([L_{j+1}, L_{j+1} + R_{j+1}]\) from the previous interval \([L_j, L_j + R_j]\) as follows. Subdivide \([L_j, L_j + R_j]\) into one subinterval for each element \( w_j \) of \( \mathbb{W} \) so that the interval width is given as

\[
R_j \cdot P(w_j | w_{j-1}, w_{j-2}, \ldots, w_0)
\]

for a given sequence of (already sampled) values \( w_{j-1}, w_{j-2}, \ldots, w_0 \), and arrange the subintervals so that they are non-overlapping and adjacent. The subinterval associated with the sample \( w_j \) to be encoded becomes the new coding interval \([L_{j+1}, L_{j+1} + R_{j+1}]\). Encoding of \( n \) symbols yields the coding interval \([L_n, L_n + R_n]\) and the sequence of symbols \( w_0, w_1, \ldots, w_{n-1} \) can be reconstructed (in the decoder) when an arbitrary value inside of this coding interval is known. Figure 2 exemplifies this procedure for a binary random process. Interestingly, the width of the coding interval \( R_n = P(w_0, w_1, \ldots, w_{n-1}) \) equals the probability of sequence \( w_0, w_1, \ldots, w_{n-1} \). As the minimum achievable code length for encoding of the \( n \) symbols is known to be \( -\log_2(R_n) \), the location of interval \([L_n, L_n + R_n]\) needs to be signaled to the decoder in a way so that the number of written bits gets as close to \( -\log_2(R_n) \) as possible. The basic encoding principle is as follows. Derive an integer \( k \) so that

\[
2^k \leq \frac{R_n}{2} < 2^{1-k}
\]

holds. Subdivide the unit interval \([0, 1]\) into \( 2^k \) (adjacent and non-overlapping) subintervals \([q2^k, (q+1)2^k]\) of width \( 2^k \). Equation (4) guarantees that one of the intervals \([q2^k, (q+1)2^k]\) is fully contained in the coding interval (regardless of the exact location \( L_n \) of interval \([L_n, L_n + R_n]\)) and if the decoder knows this interval, it can unambiguously identify \([L_n, L_n + R_n]\). Consequently, the index \( q \) identifying this interval is written to the bitstream using \( k \) bits. Equation (4) can be rewritten as

\[
k < -\log_2(R_n) + 2
\]

which shows that the ideal arithmetic coder only requires up to two bits more than the minimum possible code length for a sequence of length \( n \).

### B. Universal coding

In the previous subsection we learned that there exist codes that are able to produce binary representations of (almost) minimal redundancy (e.g. arithmetic codes). However, recall that the decoder has to know the joint probability distribution of the input source in order to build the most optimal binary representation. However, in most practical situations the decoder has no prior knowledge about it. Hence, in such cases, we have to rely on so called universal codes. They basically apply the following principle: 1) start with a general, data-independent probability model \( P_{\text{Dec}} \), 2) update the model upon seeing incoming samples, 3) encode the input samples with regards to the updated probability model.

Thus, the theoretical minimum of universal codes is upper bounded by the decoder’s probability estimate. Concretely, let

\[
P_{\text{Dec}} \text{ be the decoder’s estimate of the input’s probability model, then the minimum average code-length that can be achieved is }
\]

\[
\bar{L}_C(\mathbb{W}) = H_{P,P_{\text{Dec}}}(\mathbb{W}) = H_P(\mathbb{W}) + D_{KL}(P || P_{\text{Dec}})
\]

with \( H_{P,P_{\text{Dec}}}(\mathbb{W}) = -\sum_{w \in \mathbb{W}} P(w) \log_2 P_{\text{Dec}}(w) \) being the cross-entropy and \( D_{KL} \) the Kullback-Leibler divergence. Hence, a lossless code can only create binary representations with minimal redundancies if its decoder’s probability model is the same as the input sources. In other words, the better its estimate is, the better it can encode the input samples.

An example of a universal lossless code is the so called two-part Huffman code. Given a set of samples to be encoded, it firstly makes an estimate of their empirical probability mass distribution (EPMD) and, subsequently, it encodes the samples with regards to it. However, it has the natural caveat that the estimate needs to be encoded as well, which may add a significant number of bits in many practical situations. Moreover, as we already discussed in the previous subsection, Huffman codes also come with a series of undesired properties that make it very inefficient for cases where fast adaptability and coding efficiency is required [19].

In general, a universal lossless code should have the following desiderata:

- **Universality**: The code should have a mechanism that allows it to adapt its probability model to a wide range of different types of input distributions, in a sample-efficient manner.
- **Minimal redundancy**: The code should produce binary representations of minimal redundancy with regards to its probability estimate.
- **High efficiency**: The code should have high coding efficiency, meaning, that encoding/decoding should have high throughput.

1) **CABAC**: Context-based Adaptive Binary Arithmetic Coding is a form of universal lossless coding that fulfils all of the above properties, in that it offers a high degree of adaptation, optimal code-lengths, and a highly efficient implementation. It was originally designed for the video compression standard.
H.264/AVC [17], but it is also an integral part of its successor H.265/HEVC. It is well known to attain higher compression performance as well as higher throughput as compared to other entropy coding methods [21]. In short, it encodes each input sample by following the three stages:

1) **Binarization:** Firstly, it predefines a series of binary decisions (also called bins) under which each unique input sample element (or symbol) will be uniquely identified. In other words, it builds a predefined binary decision tree where each leaf identifies a unique input value.

2) **Context-modeling:** Then, it assigns a binary probability model to each bin (also named context model) which is updated on-the-fly by the local statistics of the data. This enables CABAC to model a high degree of different source distributions.

3) **Arithmetic Coding:** Finally, it employs an arithmetic coder in order to optimally and efficiently code each bin, based on the respective context model.

Moreover, its able to take into account, since the decoder is able to analogously update its context models upon sequentially decoding the input samples. Codes that have this property are called backward-adaptive codes. Moreover, its able to take local correlations into account, since the context models are updated by the local statistics of the incoming input data.

**C. Lossy Coding**

In contrast to lossless coding, information is lost in the lossy coding process. This implies that the quantizer $Q$ is non invertible, and therefore $\exists w : D(w,q) \neq 0$. An example of a distortion measure may be the mean-squared error $D(w,q) = ||w-q||^2_2$, but we stress that other measures can be considered as well (which will become apparent in section III).

The infimum of the rate-distortion objective (1) $\forall \lambda$ is referred to as the *rate-distortion function* in the source coding literature [19], and it represents the fundamental bound on the performance of lossy source coding algorithms. However, as we have already discussed above, finding the most optimal code that follows the rate-distortion function is most often NP-hard, and can be calculated only for very few types and/or special cases of input sources. Therefore, in practice, we relax the problem until we formalize an objective that we can solve in a feasible manner.

Firstly, we fix the binarization map $B$ by selecting a particular (universal) lossless code and condition the minimization of (1) on it. That is, now we only ask for the quantizer $Q$, along with its reconstruction values $Q^{-1}$, that minimize the respective rate-distortion objective. Secondly, we always assume that we encode an $n$-long 1-dimensional random process. Then, objective (1) simplifies to: given a lossless code $(B, B^{-1})$, find $(Q, Q^{-1})^*$ that

$$
(Q, Q^{-1})^* = \arg \min_{(Q, Q^{-1})} \mathbb{E}_{P(w_j)}[D(w_j, q_j) + \lambda L_Q(b_j)],
$$

$\forall j \in \{0, \ldots, n-1\}$, where $q_j \in Q_s := \{q_0, q_1, \ldots, q_{K-1}\} \subset \mathbb{R}$ and $K = |Q_s| < |\mathcal{W}| = n$.

For instance, if we choose $B$ such that it assigns a binary representation of fixed-length to all $w_j$, then the minimizer of (6) can be found by applying the K-Means algorithm.

The minimizers of (6) are called scalar quantizers, since they measure the distortion independently for each input sample. In contrast, vector quantizers are those that result from minimizing (6) when grouping a sequence of input samples together and measuring the distortion in the respective vector space. It is well known that the inum of scalar codes are fundamentally more redundant than vector quantizers. Nevertheless, due to the associated complexity of vector quantizers, it is more common to apply scalar quantizers in practice. Moreover, the inherent redundancy of scalar quantizers is mostly negligible for most practical applications [19].

We also want to stress that although the distortion in (6) is measured independently for each sample, the binarization $b_j$ (and consequently the respective code-length) of each sample can still depend on the other samples by taking correlations into account.

1) **Scalar Lloyd algorithm:** An example of an algorithm that finds a local optimum is the Lloyd algorithm. It approximates the average code-length of the quantized samples $q_j = (Q^{-1} \circ Q)(w_j)$ with the entropy of their empirical probability mass distribution (EPMD). Thus, it substitutes the code-length in (6) by $L_C(b_j) = -\log_2 p_{\text{EPMD}}(q_j)$ and applies a greedy algorithm in order to find the most optimal quantizer $Q$ and quantization points $Q^{-1}$ that minimize the respective objective. A pseudocode can be seen in the appendix (see algorithm 4).

2) **CABAC-based RD-quantization:** If we are given a set of quantization points $Q_s$ and select CABAC as our posterior universal lossless code, then we can trivially minimize (6) by sequentially quantizing the input samples. In the video coding standards, the set of quantization points are predefined by the particular choice of quantization strength $\lambda$ [16]. However, in the context of neural network compression we do not know of a good relationship between the quantization strength and the set of quantization points. In the next section III we describe how we tackled this problem.

**D. Model compression vs. source coding**

So far, we have reviewed some fundamental results of source coding theory. However, in this work, we are rather interested in the general topic of *model compression*. There is a fundamental difference between both paradigms. Namely, now we are more interested in the predictive performance of the resulting quantized model rather than the distance between the quantized and original parameters. Figure 4 highlights this distinction. We will now formalize the general model compression paradigm for the supervised learning setting. However, the problem can be analogously formulated for other learning tasks.

Firstly, we assume that we are given only one model sample with $n$ real-valued weight parameters (thus, here the input space is equivalent to the one discussed above). In addition,
the source coding literature in order to minimize the desired objective.

1) Relaxation of the model compression problem into a source coding problem: We may further assume that the given unquantized model has been pre-trained on the desired task and that it reaches satisfactory accuracies. Then, in such cases, it is reasonable to replace the first term in (7) by the KL-Divergence between the unquantized model \( P(y'|x, w) \) and the respective quantized model \( P(y''|x, q) \). That is, now we want to quantize our model such that its output distribution does not differ too much from its original version.

Furthermore, if we now assume that the output distributions do not differ too much from each other, then we can approximate the KL-Divergence with the Fisher Information Matrix (FIM). Concretely,

\[
\mathbb{E}_P \left[ D_{KL}(y''||y') \right] = \delta w F \delta w^T + \mathcal{O}(\delta w^2)
\]

with \( \delta w = q - w \) and \( F := \mathbb{E}_{P(x,w)} \left[ \delta \log P(y'|x,w)(\delta \log P(y'|x,w))^T \right] \).

Then, by substituting (8) in (7) we get the following minimization objective

\[
(Q, Q^{-1})^* = \arg \min_{(Q, Q^{-1})} \sum_{(x,y) \in \mathbb{D}} \mathcal{L}(y, y'') + \lambda L_Q(b)
\]

Objective (9) now follows the same paradigm as the usual source coding problem. However, with the peculiarity that now \( D(w, q) \) (approximately) measures the distortion of \( w \) and \( q \) in the space of output distributions instead the Euclidian space. The advantage of the rate-distortion objective (9) is that, after the FIM has been calculated, it can be solved by applying common techniques from the source coding literature, such as the scalar Lloyd algorithm.

However, minimizing (9) as well as estimating the FIM for deep neural networks usually requires considerable computational resources, and is most often infeasible for practical applications. Therefore, we further approximate it by only its diagonal elements (FIM-diagonals), which can be efficiently estimated (see appendix). As a result, (9) simplifies into

\[
(Q, Q^{-1})^* = \arg \min_{(Q, Q^{-1})} \sum_{i=1}^{n} F_i(q_i - w_i)^2 + \lambda L_Q(b)
\]

In the next section we will give a thorough description of our proposed coder. Its design complies with all desired properties that a coder for neural network compression should have.

III. DEEPCABAC

In light of the discussion in the previous section, we can highlight a set of desiderata that a coder for neural network compression should have.

- **Minimal redundancy**: State-of-the-art deep neural networks usually contain millions of parameters. Thus, any type of redundancy in the weight parameters may imply several additional MB being stored. Hence, the code should output a binary representation with minimal redundancy per weight element.


Fig. 5: Sketch of the DeepCABAC compression procedure. Firstly, DeepCABAC scans the weight parameters of each layer of the network in row-major order. Then, it selects a particular hyperparameter $\beta$ that will define the set of quantization points. Subsequently, it applies a quantizer on to the weight values that minimizes the respective weighted rate-distortion function (11). Then, it compresses the quantized parameters by applying our adapted version of CABAC. Finally, it reconstructs the network and measures the respective accuracy of it. The process is repeated for different hyperparameters $\beta$ until the desired trade-off between accuracy and bit-size of the network is achieved.

- **Universality:** The code should be applicable to any type of incoming neural networks, without having to know their distribution a priori. Hence, the code should entail a mechanism that allows it to adapt to a rich number of possible parameter distributions.

- **High coding efficiency:** The computational complexity of encoding/decoding should be minimal. In particular, the throughput of the decoder should be very high if performing inference on the compressed representation is desired.

- **Configurable error vs. compression strength:** The coder should have a hyperparameter that controls the trade-off between the compression strength and the incurred prediction error.

- **High data efficiency:** Minimizing (7) implies access to data. Hence, it is desirable that the coder finds a (local) solution with the least amount of data samples possible.

### A. DeepCABAC’s coding procedure

We propose a coding algorithm that satisfies all the above properties. We named it DeepCABAC, since it’s based on applying CABAC on the networks quantized weight parameters. Figure 5 shows the respective compression scheme. It performs the following steps:

1) it extracts the weight parameters of the neural networks layer-by-layer in row-major order$^2$.

2) Then, it selects a particular value $\beta$ which defines the set of quantization points.

3) Subsequently, it quantizes the weight values by minimizing a weighted rate-distortion function, which implicitly takes the impact of quantization on the accuracy of the network into account.

4) Then, it compresses them by applying our adapted version of CABAC.

5) Finally, it reconstructs the network and evaluates the prediction performance of the quantized model.

6) The process is repeated for a set of hyperparameters $\beta$, until the desired accuracy-vs-size trade-off is achieved.

This approach has several advantageous properties. Firstly, it applies CABAC to the quantized parameters and therefore we ensure that the code satisfies the desiderata 1-3. Secondly, by conducting the compression for a set of hyperparameters for the quantizer we can select the desired pareto-optimal solutions of the accuracy vs. bit-size plane, thus satisfying property 4. Finally, since only one evaluation of the model is required in the process, a significantly lower amount of data samples is required for the compression process than usually employed for training.

In the following we will explain in more detail the different components of DeepCABAC.

### B. Lossless coder of DeepCABAC

Consider the weight distribution of the last fully-connected layer of the trained VGG16 model displayed in figure 6. As we can see, there is one peak near 0 and the distribution is asymmetric and monotonically decreasing on both sides. In our experience, all layers we have studied so far have weight distributions with similar properties. Hence, in order to accommodate to this type of distributions, we adopted the following binarization procedure.

Given a quantized weight tensor in its matrix form$^3$, DeepCABAC scans the weight elements in row-major order and binarizes them as follows:

| Exp.-Golomb | SigFlag | SignFlag | AbsGr(\alpha)Flags | Unary | FL |
|-------------|---------|----------|---------------------|-------|----|

1) The first bit, $\text{sigFlag}$, determines if the weight element is a significant element or not. That is, it indicates if the weight value is 0 or not. This bit is then encoded using a binary arithmetic coder, according to its respective context model (color-coded in grey). The context model is initially set to 0.5 (thus, 50% probability that a weight element is 0 or not), but will automatically be updated to the local statistics of the weight parameters as DeepCABAC encodes more elements.

2) Then, if the element is not 0, the sign bit or $\text{signFlag}$ is analogously encoded, according to its respective context model.

$^2$Thus, it assumes a matrix form where the parameters are scanned from left-to-right, top-to-bottom.

$^3$For fully-connected layers this is trivial. For convolutional layers we converted them into their respective matrix form according to [22].
3) Subsequently, a series of bits are analogously encoded, which determine if the element is greater than 1, 2, ..., $n \in \mathbb{N}$ (hence AbsGr$(n)$Flag$s$). The number $n$ becomes a hyperparameter for the encoder.

4) Finally, the remainder is encoded using an Exponential-Golomb code [23], where each bit of the unary part are also encoded relative to their context-models. Only the fixed-length part of the code is not encoded using a context-model (color-coded in blue).

For instance, assume that $n = 1$, then the integer $-4$ would be represented as 111101, or the 7 as 10111010. Figure 7 depicts an example scheme of the binarization procedure.

The first three parts of the binarization scheme empower CABAC to adapt its probability estimates to any shape distribution around the value 0 and, therefore, to encode the most frequent values with minimal redundancy. For the remainder values, we opted for the Exponential-Golomb code since it automatically assigns smaller code-lengths to smaller integer values. However, in order to further enhance its adaptability, we also encode its unary part with the help of context models. We left the fixed-length part of the Golomb code without context models, meaning that we approximate the distribution of those values by a uniform distribution (see figure 6). We argue that this is reasonable since usually the distribution of large numbers become more and more flat, and it comes with the direct benefit of increasing the efficiency of the coder.

C. Lossy coder of DeepCABAC

After establishing CABAC as our choice of universal lossless code, now we aim to find the optimal quantizer that minimizes the objective stated in (7) (section II-D). To recall, this involves the optimization of two components (see figure 4):

- **Assignment**: finding the quantizer $Q$ that assigns the optimal set of integers to each weight parameter
- **Quant. points**: finding the optimal quantization points $q_j = (Q^{-1} \circ Q)(w_j)$.

Since neural networks usually rely on scalable, gradient-based minimization techniques in order to optimize their loss function, finding the quantizers that solve (7) becomes infeasible in most cases since $Q$ is a non-differentiable map. Therefore, we opted for a simpler approach.

Firstly, we decouple the assignment map $Q$ and the quantization points $Q^{-1}$ from each other and optimize them independently. The quantization points then become hyperparameters for the quantizer, and their values are selected such that they minimize the loss function directly. This separation between $Q$ and $Q^{-1}$ was empirically motivated, since we discovered that the networks performance is significantly more sensitive to the choice of $Q^{-1}$ than to the assignment $Q$. We discuss this in more detail in the experimental section.

1) **The quantization points**: Since finding the correct map $Q^{-1}$ for a large number of points can be very complex, we constrain them to be equidistant to each other with a specific step-size $\Delta$. That is, each point $q_k$ can be rewritten as to be $q_k = \Delta l_k$ with $l_k \in \mathbb{Z}$. This does not only considerably

\begin{figure}
\centering
\includegraphics[width=\textwidth]{distribution.pdf}
\caption{Distribution of the weight matrix of the last layer of VGG-16 (trained on ImageNet) after uniform quantization over the range of values. In red is CABAC’s (possible) estimation of the distribution. The first $n + 2$ bits allow to adapt to any type of shape around 0 since they are encoded with regards to a context model. The remainder can only approximate the shape by a step-like distribution, since they are encoded with an Exponential-Golomb where the fixed-length parts are encoded without a context model.}
\end{figure}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{binarization.pdf}
\caption{DeepCABAC binarization of neural networks. It encodes each weight element by performing the following steps: 1) encodes a bit named signflag which determines if the weight is a significant element or not (in other words, if its 0 or not). 2) If it’s not 0, then the sign bit, signflag, is encoded. 3) Subsequently, a series of bits are encoded, which indicate if the weight value is greater than equal than 1, 2, ..., $n \in \mathbb{N}$ (the so-called AbsGr($n$)Flag). 4) Finally, the remainder is encoded. The gray bits (also named regular bins) represent bits that are encoded using an arithmetic coder according to a context model. The other bits, the so-called bypass bins, are encoded in fixed-point form. For instance, in the above diagram $n = 1$, and therefore $1 \rightarrow 100$, $-4 \rightarrow 111101$ or $7 \rightarrow 1011010$.}
\end{figure}

\footnote{To recall, the Exponential-Golomb code encodes a positive integer $2^k < i \leq 2^{k+1}$ by firstly encoding the exponent $k$ using an unary code and subsequently the remainder $r = i - 2^k$ in fixed-point representation.}
simplify the problem, but it does also encourage fixed-point representations which can be exploited in order to perform inference with lower complexity [24], [25].

2) The assignment: Hence, the quantizer has two configurable hyperparameters \( \beta = (\Delta, \lambda) \), the former defining the set of quantization points and the latter the quantization strength. Once a particular tuple is given, the quantizer \( Q_\beta \) will then assign each weight parameter \( w_i \) to its corresponding quantization point by minimizing the weighted rate-distortion function

\[
Q_\beta(w_i) = k^* = \arg \min_k F_i(w_i - q_k)^2 + \lambda L_{ik}
\]

\( \forall i \in \{0, ..., n - 1\} \), where \( L_{ik} \) is the code-length of the quantization point \( q_k \) at the weight \( w_i \) as estimated by CABAC.

As previously mentioned, we perform a grid-search algorithm over the hyperparameters \( \Delta \) and \( \lambda \) in order to find the quantizer configuration that achieves the desired accuracy vs. bit-size trade-off. However, for that we need to define a predefined set of hyperparameter candidates to look for, in particular for the step-sizes \( \Delta \). In this work we considered two approaches for finding the set of step-sizes, which we denote as DeepCABAC-version1 (DC-v1) and DeepCABAC-version2 (DC-v2).

3) DeepCABAC-version1 (DC-v1): In DC-v1 we firstly estimate the diagonals of the FIM by applying scalable Bayesian techniques. Concretely, we parametrize the network with a fully-factorized Gaussian posterior and minimize the variational inference with lower complexity [24], [25].

After the FIM-diagonals have been estimated, we define the set of considered step-sizes as follows:

\[
q_k = \Delta I_k, \quad \Delta = \frac{2|w_{\max}|}{2|w_{\max}| + S}, \quad S, I_k \in \mathbb{Z}
\]

where \( \sigma_{\min} \) is the smallest standard deviation and \( w_{\max} \) the parameter with highest magnitude value. \( S \) is then the quantizers hyperparameter, which controls the “coarseness” of the quantization points. By selecting \( \Delta \) in such a manner we ensure that the quantization points lie within the range of the standard deviation of each weight parameter, in particular for values \( S \geq 0 \). Hence, we selected \( S \) to be \( S \in \{0, 1, ..., 256\} \).

One advantage of this approach is that we can have one global hyperparameter \( S \) for the entire network, but each layer will still attain a different value for its step-size if we select one \( \sigma_{\min} \) per layer. Thus, with this approach we can adapt the step-size to the layer’s sensitivity to perturbations. Moreover, the quantization will also take the sensitivity of each single parameter into account.

4) DeepCABAC-version2 (DC-v2): Estimating the diagonals of the FIM can still be computationally expensive since it requires the application of the backpropagation algorithm for several iterations in order to minimize the variational objective. Moreover, it only offers an approximation of the robustness of each parameter, and can therefore sometimes be misleading and limit the potential compression gains that can be achieved. Therefore, due to simplicity and complexity reasons, we also considered to directly try to find a good set of candidates \( \Delta \in \{\Delta_0, ..., \Delta_{m-1}\} \). We do so by applying a first round of the grid-search algorithm while applying a nearest-neighbor quantization scheme (that is, for \( \lambda = 0 \)). This allows us to identify the range of step-sizes that do not considerably harm the networks accuracy when applying the simplest quantization procedure. Then, we quantize the parameters as in eq. (11), but without the diagonals of the FIM (thus, \( F_j = 1 \forall j \)).

Under a limited computational budget, this approach has the advantage that we can directly search for a more optimal set of step-sizes \( \Delta \) since we spare the computational complexity of having to estimate the FIM-diagonals. However, since DC-v2 considers only one global step-size for the entire network, it cannot adapt to the different sensitivities of the layers.

IV. RELATED WORK

There has been a plethora of work focusing on the topic of neural network compression [12], [13]. In some way or another, all of them try to partially solve the general model compression objective (7) (section II-D). In the following we will thoroughly examine the currently proposed approaches and discuss some of their advantages and disadvantages.

A. Lossy neural network compression

Some of the insofar proposed approaches include:

1) Trained scalar quantization: These are methods that aim to minimize the model compression objective (7) by applying training algorithms that learn the optimal quantization map and reconstruction values. Inter alia, this includes sparsification methods [26], [28]–[33] which try to minimize the \( L_0 \)-norm of the network’s parameters. Others attempted to find optimal binary or ternary weighted networks [34]–[37], or a more general set of (locally) optimal quantizers [38]–[42].

Although these methods are able to attain very high compression ratios, they are also very computationally expensive in that several training iterations over a large training set have to be performed in order to find the optimal quantized network. In contrast, DeepCABAC does not require any retraining, nor access to the full training dataset in order to be applicable.

2) Non-trained scalar quantization methods: Another line of work has focused on implicitly minimizing (7). They also rely on distance measures for quantizing the network’s parameters [43]–[46]. In fact, these methods can be seen as special cases of (10), in that they either use different approximations of the FIM-diagonals or apply other minimization algorithms. To the best of our knowledge, mainly two quantizers are widely applied by the community, either the scalar uniform quantizers or the weighted scalar Lloyd algorithm. Usually, the former basically consist on uniformly spreading \( K \in \mathbb{N} \) quantization
points over the range of parameter values and then applying nearest-neighbor quantization on to them [38], [40], [45]. The latter consists of applying the scalar Lloyd algorithm in order to find the most optimal quantizer that minimizes a weighted rate-distortion objective (10). In particular, [45] considers the diagonal elements of the empirical average of the Hessian of the loss function, which has a close connection to the FIM-diagonals (see appendix for a comprehensive discussion).

Applying quantization methods that do not rely on retraining are significantly less computationally expensive. But their compression gains are heavily limited by the networks un-quantized parameter distribution, since they rely on a distance measure for quantization. Moreover, as already mentioned, most of these methods do only implicitly take the impact on to the accuracy of the network into account. In contrast, DeepCABAC does explicitly take the accuracy of the network into account since its hyperparameters are optimized with regards to it.

B. Lossless neural network compression

In the field of lossless network compression, we are given an already quantized model and we want to apply a universal lossless code to its parameters in order to maximally compress it. Hence, this setting is entirely equivalent to the usual lossless source coding setting discussed in sections II-A and II-B, and therefore all of its theorems and results can be applied in a straight-forward manner. Nevertheless, most of the previous work did not apply state-of-the-art universal lossless compression algorithms to them. Instead, these are some of the most commonly used:

1) Fixed-length numerical representations: These methods reduce the bit-length representation of the parameter values after quantization [34]–[37], [47]–[52]. They usually have the advantage of immediately reducing the complexity for performing inference, however, at the expense of having a highly redundant network representation.

2) Scalar Huffman code: Others applied the scalar Huffman code on to quantized neural networks [45], [46]. However, as we have already discussed in section II-A, this code has several disadvantages compared to other state-of-the-art lossless codes such as arithmetic codes. Probably the most prominent one is that this code is suboptimal in that it incurs up to 1 bit of redundancy per parameter being encoded. This can be quite significant for large networks with millions of parameters. For instance, VGG16 [53] contains 138 million parameters, meaning that the binary representation of any quantized version of it may have about 17MB of redundancy if we encode it using the scalar Huffman code.

3) Compressed matrix representations: Most of the literature that sparsifies the networks parameters aim to convert the resulting networks into a compressed sparse matrix representation, e.g., the Compressed Sparse Row (CSR) representation. These matrix data structures do not only offer compression gains, but also an efficient execution of the associated dot product algorithm [54]. Similarly, [14] proposed two novel matrix representation, the Compressed Entropy Row (CER) and Compressed Shared Elements Row (CSER) representations, that are provably more optimal than the CSR with regards to both, compression and execution efficiency when the networks parameters have low entropy statistics.

However, these matrix representations are also redundant in that they do not approach the reachable entropy limit (3) (section II-A). [38] attempted to extract some of the redundancies entailed in the CSR representations by applying a scalar Huffman code to its numerical arrays. However, this has again the same limitations that come by applying the scalar Huffman code.

C. Compression pipelines/frameworks

Among all different proposed approaches for deep neural network compression there is one paradigm that stands out in that very high compression gain can be achieved with it [26], [38], [40], [42], [55]. Namely, it consists on applying four different compression stages:

1) Sparsification: Firstly, the networks are maximally sparsified by applying a trained sparsification technique.

2) Quantization: Then, the non-zero elements are quantized by applying one of the non-trained quantization techniques.

3) Fine-tuning: Subsequently, the quantization points are fine-tuned in order to recover the accuracy loss incurred by the quantization procedure.

4) Lossless compression: Finally, the quantized values are encoded using a lossless coding algorithm.

Hence, DeepCABAC is designed to enhance points 2 and 4. As we will see in the next section, DeepCABAC is able to considerably boost the attainable compression gains, surpassing previously proposed methods for steps 2 and 4.

V. Experiments

In this section we benchmark DeepCABAC and compare it to other compression algorithms. We also design further experiments with the purpose to shed light on the effectiveness of its different components.

A. General compression benchmark

Here we benchmark the maximum compression gains attained by applying DeepCABAC. In order to assess its universality, we applied it to a wide set of pretrained network architectures, trained on different data sets. Concretely, we used the VGG16, ResNet50 and MobileNet-v1 architectures, trained on the ImageNet dataset, a smaller version of the VGG16 architecture trained on the CIFAR10 dataset, which we denote as Small-VGG16, and the LeNet-300-100 and LeNet5 trained on MNIST.

In addition, we also applied DeepCABAC to pre-sparsified versions of these networks. For that, we employed the variational sparsification algorithm [26] to all networks, except

3Although a better compression ratio was attained, we were not able to get an accuracy in the ±0.5 percentage point range of the original accuracy. Therefore, this result shall not be considered as the best result.

4http://torch.ch/blog/2015/07/30/cifar.html
TABLE I: Compression ratios achieved at no loss of accuracy when applying different coding methods. DC-v1 & DC-v2 denote the two versions of DeepCABAC, whereas Lloyd denotes the weighted Lloyd algorithm and uniform the nearest-neighbor quantization scheme. For the latter two, we report the best compression results attained after applying scalar Huffman, CSR-Huffman [38] and the bzip2 lossless coding algorithms on to the quantized networks. In parenthesis are the resulting top-1 accuracies, and the sparsity ratios achieved as measured \( \frac{|w| - |\hat{w}|}{|w|} \).

| Models (Spars. [%]) | Org. acc. (top1 [%]) | Org. size | DC-v1 [%] | DC-v2 [%] | Lloyd [%] | Uniform [%] |
|---------------------|----------------------|----------|-----------|-----------|-----------|-------------|
| **NON-SPARSE**      |                      |          |           |           |           |             |
| VGG16 69.94         | 553.43 MB            |          | 5.84      | 3.96      | 7.74      | 17.37       |
|                     |                      |          | (69.44)   | (69.54)   | (69.50)   | (69.90)     |
| ResNet50 74.98      | 102.23 MB            |          | 10.14     | 10.14     | 13.04     | 15.58       |
|                     |                      |          | (74.40)   | (74.51)   | (74.74)   | (74.64)     |
| MobileNet-v1        | 70.69 MB             |          | 21.40     | 22.08     | 15.00     | 24.23       |
|                     |                      |          | (70.21)   | (70.21)   | (68.10)   | (70.10)     |
| Small-VGG16         | 91.54 MB             |          | 6.35      | 5.88      | 9.98      | 16.18       |
|                     |                      |          | (91.11)   | (91.13)   | (91.59)   | (91.53)     |
| LeNet5 99.46        | 1722 KB              |          | 3.77      | 2.52      | 3.96      | 20.60       |
|                     |                      |          | (99.23)   | (99.12)   | (98.96)   | (99.45)     |
| LeNet-300-100       | 98.32 KB             |          | 8.61      | 5.87      | 8.07      | 15.01       |
|                     |                      |          | (98.04)   | (98.00)   | (97.92)   | (98.30)     |
| **SPARSE**          |                      |          |           |           |           |             |
| VGG16 (9.85)        | 69.43 MB             |          | 1.58      | 1.67      | 1.72      | 2.77        |
|                     |                      |          | (69.43)   | (69.04)   | (69.01)   | (69.42)     |
| ResNet50 (25.40)    | 74.09 MB             |          | 5.45      | 5.14      | 5.61      | 6.68        |
|                     |                      |          | (73.73)   | (73.65)   | (73.73)   | (73.98)     |
| MobileNet-v1 (50.73)| 66.18 MB             |          | 13.29     | 12.89     | 11.16     | 14.78       |
|                     |                      |          | (66.01)   | (66.02)   | (65.63)   | (65.71)     |
| Small-VGG16 (7.57)  | 91.35 MB             |          | 1.90      | 1.95      | 2.08      | 2.84        |
|                     |                      |          | (91.03)   | (91.06)   | (91.10)   | (91.20)     |
| LeNet5 (1.90)       | 1722 KB              |          | 0.88      | 0.87      | 1.09      | 3.01        |
|                     |                      |          | (99.14)   | (99.02)   | (99.25)   | (99.22)     |
| LeNet-300-100 (9.05)| 98.29 KB             |          | 2.26      | 2.20      | 1.69      | 4.17        |
|                     |                      |          | (98.00)   | (98.00)   | (97.76)   | (98.36)     |

for the VGG16 and ResNet50 due to the high computational complexity demanded by the method. The advantage of employing [26] is that we obtain the variances of each weight parameters as a byproduct of the methods output, thus being able to directly apply DC-v1 after the sparsification process finished. In the cases of the VGG16 and ResNet50 networks, we firstly applied the iterative sparsification algorithm [30] on them and subsequently estimated their FIM-diagonals by minimizing the same variational objective proposed in [26] (see appendix for a more in comprehensive explanation).

We compare the two versions of DeepCABAC, DC-v1 & DC-v2, against two previously proposed quantization schemes. Namely, similarly to [38], [45], [46], we applied the nearest-neighbor quantization scheme on to the networks. In addition, we also applied the weighted Lloyd algorithm as proposed by [43], [45], [46]. As possible lossless compression candidates, we considered the scalar Huffman code, the code proposed by [38] which we denote CSR-Huffman, and the bzip2 [56] algorithm. See appendix for a more detailed explanation of the respective implementations.

Table I shows the results. As one can see, DeepCABAC is able to attain higher compression gains on most networks as compared to the previously proposed coders. It is able to compress the pretrained by \( \times 18.9 \) and the sparsified models by \( \times 50.6 \) on average. In contrast, the Lloyd algorithm compresses the models by \( \times 13.6 \) and \( \times 47.3 \) on average, whereas uniform quantization only achieves \( \times 5.7 \) and \( \times 25.0 \) compression gains.

| step-sizes (top1 acc.) | DC-v1 | DC-v2 | Lloyd | Uniform |
|------------------------|-------|-------|-------|---------|
| NON-SPARSE             |       |       |       |         |
| 0.032 (90.35)          | 1.48  | 1.48  | 1.79  | 1.60    |
| 0.016 (91.13)          | 2.21  | 2.20  | 2.29  | 2.40    |
| 0.001 (91.55)          | 4.27  | 4.80  | 2.34  | 5.61    |
| SPARSE (5.75%)         |       |       |       |         |
| 0.032 (90.22)          | 0.47  | 0.47  | 0.52  | 0.48    |
| 0.016 (91.06)          | 0.59  | 0.58  | 0.62  | 0.60    |
| 0.001 (91.17)          | 0.91  | 1.00  | 0.74  | 1.00    |

B. Assignment vs. quantization points

To recall, lossy quantization involves two types of mappings, the quantization map \( Q \) where input values are assigned to integers, and the reconstruction map \( Q^{-1} \) which assigns a quantization point to each integer. Hence, the following experiment aims to assess the effectiveness of these components individually.
For that, we selected a predefined set of step-sizes and subsequently quantized the parameters according to different quantization schemes. In this way we can attain insights into the compression gains attained only by the influence of the quantization map.

Table II shows the average bit-sizes per parameter attained by applying different quantizers with the three-given step-sizes to the Small-VGG16 model. In order to decouple the lossless part from the quantization, the bit-sizes are calculated with regards to the entropy of the empirical probability mass distribution (EPMD) of the quantized models in the case of the Lloyd and uniform algorithms, since it marks the theoretical minimum for lossless codes that do not take correlations between the parameters into account. In contrast, since DeepCABAC’s quantizer is optimized explicitly under CABAC’s lossless coder, we calculated the average bit-size with regards to the total bit-size of the model as outputted by CABAC. We also want to stress that we chose the networks that resulted in having equal accuracies, thus, within the ±0.1 percentage point range from the accuracy attained after applying a uniform quantizer.

We attain many insights from table II. Firstly, notice how DeepCABAC’s performance is very sensitive to the particular choice of the step-size. This is due to fact that, usually, the best compression performances are attained for small compression strengths λ ≈ 0 at high accuracies. Thus, DeepCABAC’s quantization map behaves similarly to uniform quantization in those cases, and therefore it becomes sensitive to the particular choice of the step-size. Nevertheless, notice how DeepCABAC still always attains better compression performance than uniform quantization.

Secondly, notice how for small step-sizes DC-v1 outperforms DC-v2 and thus, makes better rate-distortion decisions. We attribute this to the property that DC-v1 takes the “robustness” to perturbations of each element during quantization into account. As we have already discussed in sections II-D and III (and in more detail in the appendix), the latter can be interpreted as minimizing an approximation to the desired MDL-loss function. However, since it is only an approximation, this only applies for small step-sizes and becomes more inaccurate for larger ones. Indeed, table II shows how DC-v2 attains similar results as DC-v1 as the step-size is increased, implying that the particular expression of the RD-function becomes more and more irrelevant.

These insights motivated the design of DC-v2 in the first place, since it is able to explore a larger set of step-sizes for the best accuracy vs. bit-size trade-offs. Indeed, as table I from the previous experiment shows, DC-v2 attains similar or even higher compression gains than DC-v1, in particular in the case of pretrained networks.

| Quantizers → | Uniform | Lloyd | DC-v2 |
|--------------|---------|-------|-------|
| scalar-Huffman | 5.18    | 3.19  | 2.33  |
| bzip2        | 5.22    | 3.22  | 2.42  |
| CABAC        | 4.77    | 2.74  | 2.07  |
| H            | 5.09    | 2.91  | 2.20  |
| SPARSE       | 2.58    | 1.67  | 1.80  |

C. Lossless coding

In our last experiment we aimed to assess the efficiency of different universal lossless coders. For that, we quantized the Small-VGG network using three different quantizers, and subsequently compressed each of them using different universal lossless coders. More concretely, we quantized the model by applying DC-v2, the weighted Lloyd algorithm and the nearest-neighbor quantizer. We then applied the scalar Huffman code, the CSR-Huffman code [38], the bzip2 coder, and by CABAC. The second last row denotes the entropy of the EPMD.

VI. Conclusion

In this work we proposed a novel compression algorithm for deep neural networks called DeepCABAC, that is based on applying a Context-based Adaptive Binary Arithmetic Coder (CABAC) to the networks parameters, which is the state-of-the-art universal lossless coder employed in the H.264/HEVC and H.265/HEVC video coding standards. DeepCABAC also incorporates a novel quantization scheme that explicitly minimizes the accuracy vs. bit-size trade-off, without relying...
on expensive retraining or access to large amounts of data. Experiments showed that it can compress pretrained neural networks by x18.9 on average, and their sparsified versions by x50.6, consistently attaining higher compression performance than previously proposed coding techniques with similar characteristics. Moreover, DeepCABAC is able to capture correlations between the network’s parameters, as such being able to compress the networks parameters beyond the entropy limit of codes that only assume a stationary distribution.

As future work we will investigate the impact of compression on the neural network’s problem-solving strategies [57] and apply DeepCABAC in distributed training scenarios, where communication overhead of the networks update parameters is critical for the overall training efficiency.

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. E. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, 2015.
[2] J. Schmidhuber, “Deep learning in neural networks: An overview,” Neural Networks, vol. 61, pp. 85–117, 2015.
[3] X. Xiaowei, D. Yukun, H. S. Xiaobo, N. Michael, C. Jason, H. Yu, and S. Yiju, “Scaling for edge inference of deep neural networks,” Nature Electronics, vol. 1, pp. 216–222, Apr 2018.
[4] 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), Feb 2014, pp. 10–14.
[5] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” arXiv:1704.04861, 2017.
[6] K. Ota, M. S. Dao, V. Mezaris, and F. G. B. D. Natale, “Deep learning for mobile multimedia: A survey,” ACM Trans. Multimedia Comput. Commun. Appl., vol. 13, no. 3, pp. 34:1–34:22, Jun. 2017.
[7] C. Zhang, P. Patras, and H. Haddadi, “Deep learning in mobile and wireless networking: A survey,” IEEE Communications Surveys Tutorials, pp. 1–1, 2019.
[8] H. B. McMahan, E. Moore, D. Ramage, and B. A. y Arcas, “Federated learning of deep neural networks using model averaging,” arXiv:1602.05629, 2016.
[9] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, “Sparse binary compression: Towards distributed deep learning with minimal communication,” in IEEE International Joint Conference on Neural Networks (IJCNN), 2019.
[10] F. Sattler, S. Wiedemann, K.-R. Müller, and W. Samek, “Robust and communication-efficiency federated learning from non-iid data,” arXiv:1903.02891, 2019.
[11] P. Grünwald and J. Rissanen, The Minimum Description Length Principle, ser. Adaptive computation and machine learning. MIT Press, 2007.
[12] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, “A survey of model compression and acceleration for deep neural networks,” arXiv:1710.09282, 2017.
[13] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, “Model compression and acceleration for deep neural networks: The principles, progress, and challenges,” IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 126–136, Jan 2018.
[14] S. Wiedemann, K.-R. Müller, and W. Samek, “Compact and computationally efficient representation of deep neural networks,” IEEE Transactions on Neural Networks and Learning Systems, 2019.
[15] MPEG Requirements, “Updated call for proposals on neural network compression,” Moving Picture Experts Group (MPEG), Marrakech, MA, CIP, Jan. 2019.
[16] V. Sze, M. Budagavi, and G. J. Sullivan, High Efficiency Video Coding: Algorithms and Architectures. Springer, 2014.
[17] D. Marpe, H. Schwarz, and T. Wiegand, “Context-based adaptive binary arithmetic coding in the h.264/avc video compression standard,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 620–636, July 2003.
[18] C. E. Shannon, “A mathematical theory of communication,” SIGMOBILE Mobile Computing and Communications Review, vol. 5, no. 1, pp. 3–55, 2001.
[19] T. Wiegand and H. Schwarz, “Source coding: Part 1 of fundamentals of source and video coding,” Found. Trends Signal Process., vol. 4, no. 1–2, pp. 1–222, 2011.
[20] D. A. Huffman, “A method for the construction of minimum-redundancy codes,” Proceedings of the IRE, vol. 40, no. 9, pp. 1098–1101, Sep. 1952.
[21] D. Marpe and T. Wiegand, “A highly efficient multiplication-free binary arithmetic coder and its application in video coding,” in Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429), vol. 2, Sep. 2003, pp. 263–266.
[22] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanese, and E. Shelhamer, “Cudnn: Efficient primitives for deep learning,” arXiv:1410.0759, 2014.
[23] J. Teuhola, “A compression method for clustered bit-vectors,” Information Processing Letters, vol. 7, no. 6, pp. 308–311, 1978.
[24] “QNNPACK open source library for optimized mobile deep learning,” https://github.com/pytorch/QNNPACK, accessed: 28.02.2019.
[25] “TensorFlow Lite,” https://www.tensorflow.org/lite, accessed: 28.02.2019.
[26] D. Molchanov, A. Ashukha, and D. Vetrov, “Variational dropout sparsifies deep neural networks,” in International Conference on Machine Learning (ICML), 2017, pp. 2498–2507.
[27] A. Achille, M. Rovère, and S. Soatto, “Critical learning periods in deep neural networks,” arXiv:1711.08856, 2017.
[28] 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.
[29] Y. L. Cun, J. S. Denker, and S. A. Solla, “Optimal brain damage,” in Advances in Neural Information Processing Systems (NIPS), 1990, pp. 598–605.
[30] 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 (NIPS), 2015, pp. 1135–1143.
[31] Y. Guo, A. Yao, and Y. Chen, “Dynamic network surgery for efficient dnn,” arXiv:1608.04493, 2016.
[32] C. Louizos, M. Welling, and D. P. Kingma, “Learning Sparse Neural Networks through L_0 Regularization,” arXiv:1712.01312, 2017.
[33] E. Tartaglione, S. Lepschy, A. Fiandratti, and G. Francini, “Learning sparse neural networks via sensitivity-driven regularization,” in Advances in Neural Information Processing Systems (NIPS), 2018, pp. 3878–3888.
[34] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, “Xnor-net: Imagenet classification using binary convolutional neural networks,” arXiv:1603.05279, 2016.
[35] M. Courbariaux and Y. Bengio, “Binarynet: Training deep neural networks with weights and activations constrained to +1 or -1,” arXiv:1602.02830, 2016.
[36] C. Zhu, S. Han, H. Mao, and W. J. Dally, “Trained ternary quantization,” arXiv:1603.05279, 2016.
[37] M. Courbariaux and Y. Bengio, “Binarynet: Training deep neural networks with weights and activations constrained to +1 or -1,” arXiv:1602.02830, 2016.
[38] C. Zhu, S. Han, H. Mao, and W. J. Dally, “Trained ternary quantization,” arXiv:1612.01064, 2016.
[39] F. Li, B. Zhang, and B. Liu, “Ternary weight networks,” arXiv:1605.04711, 2016.
[40] 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.
[39] K. Ullrich, E. Meeds, and M. Welling, “Soft Weight-Sharing for Neural Network Compression,” arXiv:1702.04008, 2017.
[40] C. Louizos, K. Ullrich, and M. Welling, “Bayesian Compression for Deep Learning,” in Advances in Neural Information Processing Systems (NIPS), 2017, pp. 3288–3298.
[41] J. Achterhold, J. M. Köhler, A. Schmeink, and T. Genewein, “Variational network quantization,” in International Conference on Representation Learning (ICLR), 2018.
[42] S. Wiedemann, A. Marbán, K.-R. Müller, and W. Samek, “Entropy-constrained training of deep neural networks,” in IEEE International Joint Conference on Neural Networks (IJCNN), 2019.
[43] E. Park, J. Ahn, and S. Yoo, “Weighted-entropy-based quantization for deep neural networks,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, pp. 7197–7205.
[44] M. Tu, V. Berisha, Y. Cao, and J. S. Seo, “Reducing the model order of deep neural networks using information theory,” in 2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2016, pp. 93–98.
[45] Y. Choi, M. El-Khamy, and J. Lee, “Towards the limit of network quantization,” arXiv:1612.01543, 2016.
[46] Y. Choi, M. El-Khamy, and J. Lee, “Universal deep neural network compression,” arXiv:1802.02271, 2018.
[47] N. Mellempudi, A. Kundu, D. Das, D. Mudigere, and B. Kaul, “Mixed low-precision deep learning inference using dynamic fixed point,” arXiv:1701.08978, 2017.
[48] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, “Deep learning with limited numerical precision,” arXiv:1502.02551, 2015.
[49] X. Chen, X. Hu, N. Xu, H. Zhou, and and, “Fxpnet: Training deep convolutional neural network in fixed-point representation,” in IEEE International Joint Conference on Neural Networks (IJCNN), 2017.
[50] A. Zhou, A. Yao, Y. Guo, L. Xu, and Y. Chen, “Incremental network quantization: Towards lossless cnns with low-precision weights,” arXiv:1702.03044, 2017.
[51] R. Banner, I. Hubara, E. Hoffer, and D. Soudry, “Scalable methods for 8-bit training of neural networks,” in Advances in Neural Information Processing Systems (NIPS), 2018, pp. 5145–5153.
[52] N. Wang, J. Choi, D. Brand, C. Chen, and K. Gopalakrishnan, “Training deep neural networks with 8-bit floating point numbers,” arXiv:1812.08011, 2018.
[53] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv:1409.1556, 2014.
[54] I. S. Duff, “A survey of sparse matrix research,” Proceedings of the IEEE, vol. 65, no. 4, pp. 500–535, 1977.
[55] M. Federici, K. Ullrich, and M. Welling, “Improved Bayesian Compression,” arXiv:1711.06494, 2017.
[56] J. Seward, “bzip2,” http://www.bzip.org, 1998, accessed: 09.06.2019.
[57] S. Lapuschkin, S. Waldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, “Unmasking clever hans predictors and assessing what machines really learn,” Nature Communications, vol. 10, p. 1096, 2019.
[58] D. P. Kingma, T. Salimans, and M. Welling, “Variational dropout and the local reparameterization trick,” in Advances in Neural Information Processing Systems (NIPS), 2015, pp. 2575–2583.
[59] J. Martens, “New perspectives on the natural gradient method,” arXiv:1412.1193, 2014.
TABLE IV: The number of clusters per layer employed for the experiment described in section V-A.

| Model              | Clusters |
|--------------------|----------|
| LeNet-300-100      | 256      |
| LeNet-300-100 Sparse | 32       |
| LeNet5             | 256      |
| LeNet5 Sparse      | 32       |
| Small-VGG16        | 256      |
| Small-VGG16 Sparse | 128      |
| ResNet50           | 256      |
| ResNet50 Sparse    | 256      |
| VGG16              | 256      |
| VGG16 Sparse       | 256      |
| MobileNet v1       | 1,024    |
| MobileNet v1 Sparse | 1,024    |

TABLE V: The number of clusters for the whole network as well as the λ values used in experiment V-A

| Model              | λ   | Clusters |
|--------------------|-----|----------|
| LeNet-300-100      | 0.0144 | 256      |
| LeNet-300-100 Sparse | 1.0503 | 256      |
| LeNet5             | 0.1222 | 256      |
| LeNet5 Sparse      | 0.4   | 256      |
| Small-VGG16        | 0.2105 | 256      |
| Small-VGG16 Sparse | 0.2368 | 256      |
| ResNet50           | 0.05  | 256      |
| ResNet50 Sparse    | 0.0105 | 256      |
| VGG16              | 0.0063 | 256      |
| VGG16 Sparse       | 0.0474 | 256      |
| MobileNet v1       | 0.9474 | 10,240   |
| MobileNet v1 Sparse | 0.9474 | 3,072    |

APPENDIX A

EXPERIMENT DETAILS

A. Uniform quantization

Uniform quantization is essentially one step of the weighted Lloyd algorithm with no importance measure, λ = 0, and no cluster center update. One major difference between uniform quantization and the weighted Lloyd algorithm is, that in the weighted Lloyd algorithm, the neural network is quantized as a whole, while in uniform quantization, the neural network is quantized layer-wise (see algorithm 5).

For the experiment described in section V-A, the number of clusters were determined by first starting out with 256 clusters for un-sparsified networks and with 32 clusters for sparsified networks. Then the networks were quantized and evaluated. If the accuracy was not within a range of ±0.5 percentage points as compared to the original accuracy, then the number of clusters was doubled until the accuracy was within that range.

The quantized models were then compressed using scalar Huffman coding and bzip2. Additionally, the sparse models were compressed using CSR-Huffman coding. Since additional parameters such as biases were not quantized, their original size was added to the compressed size of all compression methods.

B. Lloyd’s hyperparameter selection

To determine the optimal number of clusters and optimal values for λ, in the beginning, the number of clusters were fixed to 256. The ranges for λ were determined iteratively under the assumption that the accuracy decreases roughly monotonic with an increasing λ. At first, 20 experiments with a λ value between 0.0 and 1.0 were started and evaluated to establish a rough range of λ in which the accuracies are within a range of ±0.5 percentage points as compared to its original accuracy. In one case (LeNet-300-100 Sparse) all experiments yielded accuracies within that range. In that case another 20 experiments were conducted with λ values between 1.0 and 2.0. In the case of MobileNet v1 and its sparse counterpart, even a value of 0.0 for λ produced accuracies below the ±0.5 percentage point threshold. In both cases the number of clusters was doubled, then 20 experiments with λ in the range of 0.0 to 1.0 were conducted. This process was repeated until the accuracies were within the threshold7. Then for all networks two adjacent values of λ were selected where the accuracy lies within the range and that produced the smallest entropies. Again, 20 experiments with λ values between these selected λs were conducted. This process was repeated until there were no longer any gains in the entropies. Typically, only two rounds were enough to find no further improvement.

C. DeepCABAC’s hyperparameter selection

In all experiments we set the AbsGr(n)-Flag to 10.

D. DC-v1

For the experiment in section V-A, we searched through the following set of hyperparameters:

\[ S = [0.0, 8.0, 16.0, 32.0, 64.0, 96.0, 128.0, 160.0, 172.0, 192.0, 256.0] \]

\[ \lambda = 0.0001 \cdot 2^{\left(\log_2 \left(10^7 i \pi \right)\right)}, \forall i \in \{0, ..., 99\} \]

E. DC-v2

For the experiment in section V-A, we searched through the following set of hyperparameters:

\[ \lambda = \frac{0.02}{i + 0.01}, \forall i \in \{0, ..., 20\} \]

\[ \Delta_1 = 0.001 \cdot 2^{\left(\log_2 \left(\frac{0.15}{\pi} \right)\right)}, \forall i \in \{0, ..., 70\} \]

\[ \Delta_2 = 0.064 \cdot 2^{\left(\log_2 \left(\frac{0.128}{\pi} \right)\right)}, \forall i \in \{0, ..., 30\} \]

7Please note that, although the number of clusters was drastically increased for MobileNet v1, we were not able to achieve accuracies within the target threshold.
After convergence, it results in the following solution
and-slab distribution over the quantization points, then the identity map), and consider a decoder that assumes a spike-

\[ (\mu, \sigma)^* = \min_{(\mu, \sigma)} \mathbb{E}_{P(w|\mu, \sigma)}[\mathcal{L}(y, y')] + \beta D_{KL}(P(w|\mu, \sigma)||P(w)) \]  

(13)

with \( y' \sim P(y|x, w) \) being the output samples of the neural network, \( (x, y) \) the data samples, and \( (\mu, \sigma) \) the mean and standard deviations of all the networks parameters, and \( \beta \in \mathbb{R} \) the Lagrange-multiplier. As the KL-Divergence cannot be calculated analytically, they proposed to approximate it by

\[ D_{KL}(P(w|\mu, \sigma)||P(w)) \approx \sum_i k_i \text{sgm}(k_2 + k_3 \log \alpha_i) - \frac{1}{2} \log \left( \frac{\alpha_i + 1}{\alpha_i} \right) \]  

(14)

with \( \text{sgm}(\cdot) \) being the sigmoid function, \( \alpha_i = \sigma_i^2/\mu_i^2 \) the inverse of the signal-to-noise ratio of the parameters, \( k_1 = 0.63576, k_2 = 1.87320 \) and \( k_3 = 1.48695 \). Then, they minimize (13) by applying scalable sampling techniques proposed by [58].

A. Connection between pruning and quantization

As a result of minimizing (13) we get a mean and standard deviation for each parameter of the network. In our work, we interpreted the former as their (new) value and the latter as a measure of their “robustness” against perturbations. Indeed, the authors suggested to prune away (set to 0) parameters with a

\[ \text{signal-to-noise ratio, and hessian of the loss function, and, respectively, with the FIM-diagonals.} \]

B. Connection between variances, Hessian, and FIM-diagonals

Firstly, as thoroughly discussed in [59] and mentioned in [27], it is important to recall that the FIM is a semi-positive approximation of the Hessian of the loss function. Hence, similar to [27], we can derive a more rigorous connection between the estimated variances from minimizing (13), the FIM-diagonals and the Hessian. Namely, assuming that the variational loss function can be approximated by its second order expansion around the weight configuration \( w \), we get the following expression

\[ \mathbb{E}_{P(w|\mu, \sigma)}[\mathcal{L}(y, y')] \approx \mathcal{L}(w) + \frac{1}{2} \text{tr}[\sigma H(w)] \]

with \( \mathcal{L}(w) \) being the loss value at \( w \) and \( \text{tr}[] \) the trace. Hence, if we substitute this expression into (13) and take the derivative with respect to \( \sigma_i \) we attain

\[ H_i = \frac{1}{\sigma_i^2} K(\alpha_i^{-1}) \]

with \( K(\alpha_i^{-1}) \in (0, 1) \) being (approximately) a monotonically increasing function of the signal-to-noise ratio of the parameter. Hence, there is a direct connection between the variances, signal-to-noise ratio, and hessian of the loss function and, consequently, with the FIM-diagonals.

C. Hessian-weighted vs. variance-weighted quantization

[45] suggested a Hessian-weighted Lloyd algorithm for quantizing the neural networks parameters, where the diagonals of the empirical Hessian are taken as weights in the algorithm. As we have already discussed above, these coefficients are closely connected to the FIM-diagonals, and are thus also theoretically well motivated. However, we experienced
the algorithm to be less stable in practice when we used the
Hessian-diagonals instead of the variances. Figure 8 shows the
rate-accuracy curves when we quantized the LeNet5 model
with both alternatives. As we can see, the curves of the
variances are more stable, even achieving better compression
results than the Hessian-weighted variant.

APPENDIX C
PSEUDOCODES

Algorithm 1 Encoding a message using a Huffman code.
1: Input: A message \( M := (s_1, s_2, \ldots, s_n) \) of length \( n \), over
alphabet \( \Sigma \) and a Huffman code \( H : \Sigma \rightarrow \{0, 1\}^+ \)
2: Output: A sequence of bits \( E \in \{0, 1\}^+ \) containing the
encoded symbol sequence
3: Let \( E \leftarrow \epsilon \)
4: for all \( s \in M \) do
5: \( E \leftarrow E \odot H_s \)
6: return \( E \)

Algorithm 2 Decoding a message using a Huffman code.
1: Input: A sequence of bits \( E \in \{0, 1\}^+ \) to decode and a
Huffman code \( H : \Sigma \rightarrow \{0, 1\}^+ \)
2: Output: A message \( M := (s_1, s_2, \ldots, s_n) \) of length \( n \), over alphabet \( \Sigma \) representing the decoded message
3: Let \( M \leftarrow \epsilon \)
4: while \( |E| > 0 \) do
5: for all \( \sigma \in \Sigma \) do
6: if \( \exists x \in \Sigma^* : E = H_\sigma \odot x \) then
7: \( M \leftarrow M \odot \sigma \)
8: \( E \leftarrow (E[\epsilon_x], \ldots, E[|E| - |E_\sigma|]) \)
9: return \( M \)

Algorithm 3 Generating a Huffman code for a message
1: Input: A message \( M := (s_1, s_2, \ldots, s_n) \) of length \( n \), over
alphabet \( \Sigma \)
2: Output: A Huffman code, which maps symbols to code
words: \( H : \Sigma \rightarrow \{0, 1\}^+ \)
3: for all \( \sigma \in \Sigma \) do
4: \( F_\sigma \leftarrow \frac{|\{s \in M | s = \sigma\}|}{|M|} \) \quad \text{Calculate symbol frequency}
5: Let \( Q \) be a priority queue \( \text{Dequeuing from } Q \text{ yields smallest element} \)
6: for all \( s \in M \) do
7: if \( F_s > 0 \) then
8: \( \text{ENQUEUE}((s, F_s), Q) \)
9: \( |Q| > 1 \) do \quad \text{Build Huffman tree}
10: \( (n_l, F_l) \leftarrow \text{DEQUEUE}(Q) \)
11: \( (n_r, F_r) \leftarrow \text{DEQUEUE}(Q) \)
12: \( \text{ENQUEUE}((n_l, n_r), F_l + F_r), Q) \)
13: \( (n, F_n) \leftarrow \text{DEQUEUE}(Q) \)
14: return \( \text{ASSIGNCODEWORDS}(n, \epsilon) \)
15: procedure \( \text{ASSIGNCODEWORDS}(n, c) \)
16: if \( n \) is leaf node then \quad \text{Assign code word to leaf nodes}
17: \( H_n \leftarrow c \)
18: else \quad \text{Traverse tree recursively}
19: \( H \leftarrow H \cup \text{ASSIGNCODEWORDS}(n_l, c \odot 0) \)
20: \( H \leftarrow H \cup \text{ASSIGNCODEWORDS}(n_r, c \odot 1) \)
21: return \( H \)
Algorithm 4 Weighted Lloyd’s algorithm
1: Input: A set of clusters $C = \{c_1, c_2, \ldots, c_k\}$, a set of neural network parameters $W = \{w_1, w_2, \ldots, w_n\}$, an importance measure for each neural network parameter $F = \{f_1, f_2, \ldots, f_n\}$, and a Lagrangian multiplier $\lambda$
2: Output: A set of cluster assignments $C_j = \{w_1, w_2, \ldots, w_i\} \forall j \in \{1, \ldots, k\}$ that minimize the Lagrangian loss function $J_\lambda$
3: for $j = 1 \rightarrow k$ do  \hspace{1em} ▷ Initialize cluster probabilities
4: $P_j \leftarrow \frac{1}{k}$
5: repeat
6: for $j = 1 \rightarrow k$ do  \hspace{1em} ▷ Reset the clusters
7: $C_j \leftarrow \emptyset$
8: for $i = 1 \rightarrow n$ do  \hspace{1em} ▷ Assignment step
9: $j = \text{argmin}_j (F_i(W_i - C_j)^2 - \lambda \log_2(P_j))$
10: $C_j \leftarrow C_j \cup \{W_i\}$
11: for $j = 1 \rightarrow k$ do  \hspace{1em} ▷ Update step
12: $C_j \leftarrow \frac{\sum_{w_i \in C_j} F_i w_i}{\sum_{w_i \in C_j} F_i}$
13: $P_j \leftarrow \frac{|C_j|}{n}$
14: $j = \text{argmin}_j |C_j|$  \hspace{1em} ▷ Enforce 0-cluster
15: $C_j \leftarrow 0$
16: $J_\lambda = \sum_{j=1}^k \sum_{w_i \in C_j} F_i (W_i - C_j)^2 - \lambda \log_2(P_j)$  \hspace{1em} ▷ Calculate loss
17: until convergence, e.g. $J_\lambda$ decreases below some threshold
18: return $C$

Algorithm 5 Uniform quantization algorithm
1: Input: A set of $n$ neural network layers $W_l = \{w_1, w_2, \ldots, w_n\} \forall l \in \{1, \ldots, n\}$ and a set of cluster centers for each layer of the neural network: $C_l = \{c_1, c_2, \ldots, c_k\} \forall l \in \{1, \ldots, n\}$
2: Output: A set of cluster assignments $C_l = \{w_1, w_2, \ldots, w_i\} \forall l \in \{1, \ldots, n\}$
3: for $l = 1 \rightarrow n$ do
4: for $i = 1 \rightarrow n_l$ do
5: $j = \text{argmin} (W_{li} - C_{lj})^2$
6: $C_{lj} \leftarrow C_{lj} \cup \{W_{li}\}$
7: return $C$