



HAL
open science

A Framework for Semi-Automatic Precision and Accuracy Analysis for Fast and Rigorous Deep Learning

Christoph Lauter, Anastasia Volkova

► **To cite this version:**

Christoph Lauter, Anastasia Volkova. A Framework for Semi-Automatic Precision and Accuracy Analysis for Fast and Rigorous Deep Learning. IEEE Symposium on Computer Arithmetic (ARITH), Jun 2020, Portland, United States. 10.1109/arith48897.2020.00023 . hal-02473300

HAL Id: hal-02473300

<https://hal.science/hal-02473300v1>

Submitted on 10 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Framework for Semi-Automatic Precision and Accuracy Analysis for Fast and Rigorous Deep Learning

Christoph Lauter

*Department of Computer Science - College of Engineering
University of Alaska Anchorage (UAA)
3211 Providence Dr Anchorage, AK, 99508
christoph.lauter@christoph-lauter.org
ORCID: 0000-0001-7335-8220*

Anastasia Volkova

*Université de Nantes
CNRS, LS2N F-44000 Nantes, France
anastasia.volkova@univ-nantes.fr
ORCID: 0000-0002-0702-5652*

Abstract—Deep Neural Networks (DNN) represent a performance-hungry application. Floating-Point (FP) and custom floating-point-like arithmetic satisfies this hunger. While there is need for speed, inference in DNNs does not seem to have any need for precision. Many papers experimentally observe that DNNs can successfully run at almost ridiculously low precision.

The aim of this paper is two-fold: first, to shed some theoretical light upon why a DNN’s FP accuracy stays high for low FP precision. We observe that the loss of relative accuracy in the convolutional steps is recovered by the activation layers, which are extremely well-conditioned. We give an interpretation for the link between precision and accuracy in DNNs.

Second, the paper presents a software framework for semi-automatic FP error analysis for the inference phase of deep-learning. Compatible with common Tensorflow/Keras models, it leverages the frugally-deep Python/C++ library to transform a neural network into C++ code in order to analyze the network’s need for precision. This rigorous analysis is based on Interval and Affine arithmetics to compute absolute and relative error bounds for a DNN. We demonstrate our tool with several examples.

Keywords—*deep learning, floating-point arithmetic, error analysis, interval arithmetic, affine arithmetic*

I. INTRODUCTION

The area of Deep Learning (DL) [1], and in particular learning approaches based on Deep Neural Networks (DNNs), has seen some remarkable advances in the past decade. Neural networks are computing systems that can be seen as collections of connected nodes that are called neurons and are typically organized into sequentially inter-connected layers. Each layer performs an affine transformation defined by the layer’s parameters (weights and bias), followed by a non-linear transformation called activation. DNNs can “learn” to perform specific tasks by training on examples and then infer the results for new input data. For example, when given enough training samples, a classification network can learn the values of weights and biases for each neuron such that given a new image it can distinguish cats from dogs on images.

Many key algorithmic ideas underlying DNNs go back to as far as late 1960s [2] with a reinstated interest in 1990s [3]. The huge potential of DNNs is to solve complex problems while

accepting input data in a raw and even heterogeneous form faced practical difficulties: hardware was not powerful enough, allowing only small-sized examples. It is not until the 2000s that some realistic problems could be solved by DNNs. The best example is the Imagenet [4] image classification problem. Then a huge wave of results in applying DL to real-world problems in the areas of computer vision, speech recognition, language-understanding and navigation in autonomous driving through reinforcement learning followed. Both increasingly large datasets and increasingly complex models were critical for the success. For example, to recognize handwritten digits using MNIST network neural network is defined using around 0.7 million parameters; the MobileNet network for the Imagenet problem requires around 27 million of parameters, while the BERT network goes up to 345 million parameters to solve Natural Language Processing problems.

Deep Learning became an extremely computationally-hungry application in the end of the Moore’s Law era [5], when again performance improvement in CPUs and even GPUs is not enough. On the computer arithmetic level, performance can be improved by reducing the bit-widths of data and operators, which results in smaller hardware area and memory-access times, and faster computations. However, a tradeoff must be found in order to keep the DNN inference accurate.

While DNN training is usually performed in Floating-Point (FP) arithmetic using uniform float32 or mixed float32/half [6] precision, inference can be performed in smaller formats, or even in Fixed-Point arithmetic. There are several new low-precision FP formats that have been suggested by the major hardware manufacturers: bfloat16 (Intel [7], ARM [8]), DLfloat (IBM) [9], MSFP8-11 (Microsoft [10]) and the dedicated hardware is on its way. Obviously, FPGAs and ASICs offer even richer design space.

Existing literature provides a large body of research on post-training quantization of DNN’s parameters, typically down to 8-bit (Google’s TPU) and 4-bit integers [11], or even down to 1 bit, which results in Binary Neural Networks [12], [13]. Models can even be trained directly to have low-precision representation of weights and biases [14]–[16].

However, in the existing literature, the impact of rounding

errors due to the precision of the underlying arithmetic has been, to the best of our knowledge, surprisingly missing. Perhaps negligible for close-to-float32 precisions, the arithmetic rounding errors in low-precision implementations can potentially grow and impact the network’s classification choice. The experimental studies [7], [8], [17] make us think that DNNs can nevertheless maintain high inference accuracy even with low-precision FP arithmetic. The first contribution of this paper is to shed some theoretical light on why this is the case by having a computer arithmetic look at DNN layers.

A typical study of the impact of rounding errors in DNNs is based on a comparison with a reference output on a (moderate) set of testing data. More formal approaches do exist, to analyze the robustness of DNNs with respect to perturbation of input parameters, e.g. the SafeAI project¹ [18] based on abstract interpretation or SMT [19]. However, these tools do not account for FP rounding errors in their analyses. Our second contribution is to provide an interpretation of the impact of a precision choice upon the accuracy of a DNN.

Finally, we present a semi-automatic software framework for an automatic precision and accuracy analysis tool. Our versatile tool has a front-end accepting DNN models from common design frameworks such as Tensorflow/Keras, and is based on a generic error analysis technique, parametrizable by the target FP precision. The latter feature permits to analyze the behavior of DNNs upon a variety of target FP formats. In order to provide both relative and absolute error bounds, we introduce a combination of Affine and Interval arithmetics.

We start by recalling to the reader the basic notions of Deep Neural Networks in Section II. Then, in Sections III and IV we define our arithmetic toolkit and present a theoretical look at the numerical computations within DNN’s layers, respectively. We follow by the description of the software tool and numerical experiments in Section V before concluding and discussing future research.

II. DEEP NEURAL NETWORKS

The basic computational units of DNNs are neurons, that can be seen as nodes parameterized by a weight $w \in \mathbb{R}$ and a bias $b \in \mathbb{R}$. Given an input $x \in \mathbb{R}$, it computes an output $y = g(w \cdot x + b)$, where $g : \mathbb{R} \rightarrow \mathbb{R}$ is a non-linear activation function. Typically, the inputs of a neural Network are assembled into a vector (e.g. a 32x32 pixel gray-scale image is flattened into a 1024x1 vector), hence the per-layer computations are dot products. In a general case, DNNs operate on N -dimensional tensors.

Conceptually, DNN layers are divided into the input layer, the output layer and the intermediate, so-called hidden, layers as illustrated in Fig. 1. A trained DNN model is defined by its topology (number/type of layers) and the learned parameters (weights and biases). Even though classically a network layer was comprised of a linear computation (e.g. dot product) and a non-linear activation function, modern literature often speaks of “activation layer” as an independent entity. Following this trend, we assume that computational layers are interleaved with the activation ones.

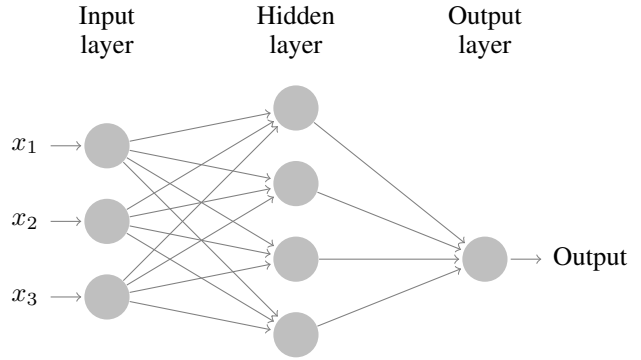


Fig. 1. A DNN with 1 hidden layer. Each layer is parametrized by a weight matrix, a bias vector and an activation function.

Activation layers: There exist a variety of non-linear activation functions, having different properties, e.g. bounded/unbounded, monotonic, continuously differentiable, etc. Some of the most common activations over vectors are the following ones:

- Sigmoid (sigmoid): computes

$$\sigma(x_i) = \frac{1}{1 + e^{-x_i}} \quad (1)$$

- Hyperbolic Tangent (tanh): simply applies $\tanh(x_i)$
- Rectified Linear Unit (ReLU):

$$\text{ReLU}(x_i) = \max\{x_i, 0\}, \quad i = 1, \dots, n. \quad (2)$$

If the input x is an N -dimensional tensor, all of the above functions are applied element-wise.

- Softmax activation (Softmax): normalizes x into a probability distribution over output classes. It is evaluated via

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}, \quad i = 1, \dots, n. \quad (3)$$

If the input x is an N -dimensional tensor, the Softmax activation is applied along each axis separately.

It should be noted that all of the above functions are bounded, yielding output values in $[0, 1]$, except for the ReLU, which results in $[0; \bar{X}]$, i.e. it maintains an upper bound \bar{X} on the layer’s input while clipping negative values.

Computational layers: Some of the most common layer types are:

- Dense layer (Dense): typically accepting an input data vector $x \in \mathbb{R}^n$ and parameterized by the weights matrix $A \in \mathbb{R}^{m \times n}$ and a bias $b \in \mathbb{R}^m$. The output is compute via $y = A \cdot x + b \in \mathbb{R}^m$. For multidimensional inputs, this operator extends into a tensor product.

- Convolution² layer (Conv): its purpose is to extract the feature maps out of data represented as multi-dimensional arrays through a linear transformation. This layer is parameterized by a convolution kernel that is convolved with the layer input to produce a tensor of outputs. The guide [20] offers a comprehensive description of the convolutional arithmetic for deep learning. For example, an image is given as a 3D

¹<http://safeai.ethz.ch>

²Should not be confused with a mathematical definition of convolution.

tensor (rows, cols, axes), where axes provides the number of color channels. Convolution kernel is also a 3D tensor of the size (kernel_width, kernel_height, axes). Convolution operation consists of the kernel sliding across the input data when at each location, the product between each element of the kernel and the input is computed; consequently the results are summed up to obtain a scalar output in the current location. The basic arithmetic operation in the convolution layers is again, the dot product.

- Pooling layer (Pool) : Pooling operations reduce the size of feature maps by using some function to summarize sub-regions, such as taking the average or the maximum value. The basic arithmetic operation in this layer is summation and/or max function.

- Batch Normalization layer (BatchNormalization): introduced in [21], this technique is used to normalize the input of the layer and only then apply a dot-product. The idea is to divide the input data into mini-batches and first perform per-batch normalization. Let m be a size of a mini-batch B and $x \in \mathbb{R}^{m \times d}$ be a d -dimensional input for the batch normalization layer. Then, the normalization is applied for each dimension separately via:

$$\tilde{x}_i^{(k)} = \frac{x_i^{(k)} - \mu_B^{(k)}}{\sqrt{\sigma_B^{(k)^2} + \epsilon}}, \quad k = 1, \dots, d, \quad i = 1, \dots, m, \quad (4)$$

where the mean $\mu_B = \frac{1}{m} \sum_{i=1}^m x_i \in \mathbb{R}^k$ and variance $\sigma_B^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_B)^2 \in \mathbb{R}^k$ are computed per mini-batch B and $\epsilon > 0$ is a small parameter. After normalization, the input vector is transformed as follows : $y_i^{(k)} = \gamma^{(k)} \tilde{x}_i^{(k)} + \beta^{(k)}$, where γ and β are d -dimensional vectors of parameters that are learned during the optimization.

The continuously updating list of computational layers can be found in the documentation for the Tensorflow/Keras.

III. ARITHMETIC TOOLKIT

FP operations, such as addition, subtraction, multiplication or more complex functions like \exp, \tanh , can necessarily not be exact: floating-point representation uses finite memory. Rounding is hence necessary after almost every operation. These roundings induce error into the computation, affecting the final result [22]. For a DNN, this means e.g. that the output class and the attached confidence probability are affected by rounding error. When confidence is low, that error might have even toggled the class the DNN output. In order to make DNNs rigorous, the overall FP rounding error affecting the results must be analyzed.

The error due to solely one FP operation mainly depends on the *precision* k of the FP format used. For binary FP formats – which we shall focus on in this paper – precision expresses the number k of bits held in the format’s mantissa. For example, for the IEEE754-2019 binary32, $k = 24$ and for IEEE754-2019 binary64, $k = 53$. For an IEEE754-2019 FP operation with precision k that does not overflow nor underflow, i.e. exceed the format’s exponent range, the following holds for rounding-to-nearest [22]: let $u = 2^{-k+1}$. Then, for every FP input a, b , there exists $\epsilon \in [-1/2; 1/2]$ such that

$$a \odot b = (a \circ b) \cdot (1 + \epsilon u) \quad (5)$$

where \odot is the FP realization of operation $\circ \in \{+, -, \times, /\}$. A similar bound is available for unary operations such as $\sqrt{\cdot}, \exp, \tanh$. This representation of the FP rounding error is also called the first FP error model [23]. Remark that (5) holds independently of the value of u , i.e. independently of the chosen precision k . This model therefore allows code to be analyzed for a given precision so to tailor it for an application.

The difficulty in analyzing a given FP code’s *accuracy* by analyzing the total amount of error affecting a result value lies in the intricate ways the different elementary errors, which are given by eq. (5). For example, suppose two multiplication operations are execute one after the other. By applying eq. (5) twice, we already obtain:

$$a \otimes (b \otimes c) = (a \times b \times c) \cdot (1 + \epsilon_1 u + \epsilon_2 u + \epsilon_1 \epsilon_2 u^2).$$

A single scalar product of a DNN’s convolution layer with n inputs uses n multiplications and n additions, each of which will result in an error term $\epsilon_i u$ by application of eq. (5) but will also combine with all other error terms in the most intricate way. Manual analysis with this approach is hence completely intractable. Wilkinson and Higham therefore invented extended ways of analyzing the combination of elementary errors in FP code [23], [24]. However, their analysis requires human insight into the different algorithms used, such as addition of FP values, scalar products etc. When existing code is to be analyzed for accuracy automatically using software, Higham’s approach is hence not usable either.

A workable approach is found with *Affine Arithmetic* (AA), as developed e.g. by Putot [25]. Every FP quantity \hat{q} is annotated with a bound $\bar{\epsilon}$ that expresses the quantity’s error with respect to the mathematically ideal, but unknown quantity q in a similar manner as in eq. (5) above:

$$\hat{q} = q \cdot (1 + \epsilon u) \quad \text{with } |\epsilon| \leq \bar{\epsilon}. \quad (6)$$

If \hat{q} stems from an exact quantity q in a way that involves only one FP operation, $\bar{\epsilon}$ is set to $1/2$ according to (5). Otherwise, when \hat{q} is the result of combining two quantities \hat{r} and \hat{s} with an operation \odot , the error terms $\bar{\epsilon}_r$ and $\bar{\epsilon}_s$ as well as the error term $\bar{\epsilon}_{\odot}$ due to the rounding in the operation \odot are combined to yield one single new error term $\bar{\epsilon}$ for \hat{q} with respect to q . This combination is specific to each operation type. For example, for addition, $\odot = \oplus$ we obtain:

$$\begin{aligned} \hat{q} &= \hat{r} \oplus \hat{s} = (\hat{r} + \hat{s}) \cdot (1 + \epsilon_{\odot} u) \\ &= (r \cdot (1 + \epsilon_r u) + s \cdot (1 + \epsilon_s u)) \cdot (1 + \epsilon_{\odot} u) \\ &= (r + s) \cdot \left(1 + \epsilon_r \frac{r}{r+s} u + \epsilon_s \frac{s}{r+s} u \right) \cdot (1 + \epsilon_{\odot} u) \\ &= q \cdot (1 + \epsilon u), \end{aligned} \quad (7)$$

with

$$\epsilon = \frac{\left(1 + \epsilon_r \frac{r}{r+s} u + \epsilon_s \frac{s}{r+s} u \right) \cdot (1 + \epsilon_{\odot} u) - 1}{u}. \quad (8)$$

To annotate \hat{q} with $\bar{\epsilon}$ bounding that ϵ in (8), we need to

- have access to the annotations ϵ_r and ϵ_s of the operands,
- know a bound on ϵ_{\odot} , which is provided by (5) and
- to be able to bound the error amplification (or attenuation) quantities $\alpha_r = \frac{r}{r+s}$ and $\alpha_s = \frac{s}{r+s}$.

For the latter task, to bound α_r and α_s , we use Interval Arithmetic (IA), as we shall explain below. We have detailed only the case of FP addition, $\odot = \oplus$. Similar error combination rules exist for the other operations that occur in DNNs, such as subtraction, multiplication, division, square root, exp, tanh.

However, the AA approach above is based on a relative error model: the term εu describes the relative error of the FP quantity \hat{q} with respect to the ideal, unknown quantity q . Such a relative error bound does not always exist, in which case relative AA breaks down. An easy-to-understand example is when the result of an FP addition $\hat{r} \oplus \hat{s}$ cancels out completely, amplifying the incoming errors $\varepsilon_r u$ and $\varepsilon_s u$ by an infinite amount. Typically, in this case, the quantities α_r and α_s do not stay bounded as their denominator becomes zero.

A solution to this issue is to use AA not with a relative error term but with an absolute error term. An FP quantity \hat{q} is labeled with an absolute error bound $\bar{\delta}$ such that

$$\hat{q} = q + \delta u \quad \text{with } |\delta| \leq \bar{\delta}. \quad (9)$$

For FP addition, the absolute error terms just add up, plus an additional absolute error term, which can easily be deduced out of the relative term given by (5), by multiplying the relative error bound by an upper bound on the exact result's absolute value. Such an upper bound is easily computed using IA

As a matter of course, due to the “there’s no free lunch”-rule, the ease of use of absolute AA for addition comes with issues for other operators like division, where amplification terms similar to α_r and α_s become unbounded and hurt absolute AA the same way as addition hurts relative AA.

Our solution to this issue is to maintain both absolute and relative error “bounds” $\bar{\delta}$ and $\bar{\varepsilon}$ for each quantity \hat{q} and to let them become infinite whenever no such bound exists. Operators like addition, multiplication, division, square root, exp etc. try to propagate both the absolute and the relative error bounds whenever possible, using the information in both bounds when appropriate. This is, addition and subtraction, which may cancel, propagate the absolute error bound and may yield an infinite relative error bound. Multiplication, division and square root start off the relative error bounds and propagate those. Exponential propagates the entering absolute error bound as a relative error bound as in

$$e^{q+\delta u} = e^q \cdot \left(1 + \frac{e^{\delta u} - 1}{u} u \right).$$

Logarithm does the inverse, transforming a relative error bound into an absolute one. The function tanh, used a lot in DNNs activation layers, can propagate the absolute error with no amplification factor and may propagate the relative error bound $\bar{\varepsilon}$ with a small amplification factor of 2.63 whenever $\bar{\varepsilon} u \leq 1/4$. The details of this combined absolute and relative AA (CAA) go beyond the scope of this paper but we may state:

$$\begin{aligned} \tanh(q + \delta u) &= \tanh(q) + \delta' u \quad \text{with } \bar{\delta}' = \bar{\delta}, \\ \tanh(q \cdot (1 + \varepsilon u)) &= \tanh(q) \cdot (1 + \varepsilon' u) \\ &\quad \text{with } \bar{\varepsilon}' = 2.63 \bar{\varepsilon} \text{ if } \bar{\varepsilon} u \leq 1/4. \end{aligned}$$

Whenever possible, the proposed CAA improves the one bound –absolute or relative– of a quantity using the other. For example, it is often possible to deduce tight relative error

bounds out of absolute when a quantity can be shown never to be zero. Likewise, an absolute error bound is readily deduced from the relative one and an upper bound on the quantity.

In order to do so and, as explained above, to be able to combine the operands’ error terms and to bound absolute elementary errors using eq. (5), bounds for the quantities occurring in a computation must be known. We compute these bounds using *Interval Arithmetic* (IA) [26]. For IA, each quantity is replaced by an interval the quantity can be shown to lie in. Each IA operator working on intervals produces an interval that surely encompasses all possible images of the operation and operands in the operand intervals. All roundings are performed in such a way –viz. outwards– that this enclosure property is satisfied even in the presence of roundings [26].

Both IA and (absolute, relative or combined) AA are plagued by a phenomenon called the *decorrelation effect* [25]. Consider the following code snippet:

$$\begin{aligned} y &= x; \\ z &= x - y; \end{aligned}$$

While mathematically, z will always be zero, as $y = x$, and while even IEEE754-2019 FP code ensures that z will be zero due to full cancellation of all bits of x and its copy y , IA and AA will have no global understanding that x and y are correlated, and –actually– equal. So assuming x to be bounded by the interval $[-1; 1]$, z will evaluate to the interval $[-2; 2]$ instead of the interval $[0; 0]$. For CAA, the issue will be that the relative error bound $\bar{\varepsilon}$ on \hat{z} instead of becoming zero as the errors on \hat{x} and \hat{y} cancel out will become infinite due to the detected “catastrophic” cancellation. The absolute error bound $\bar{\delta}$ for \hat{z} will not become zero but the double of the one on \hat{x} . For all kind of arithmetic techniques, such as IA or CAA, which only label quantities in code with interval bounds or absolute and relative error bounds but which do not gain any global understanding of the code, the decorrelation effect has no simple solution. It depends on the application whether or not the decorrelation effect occurs and whether or not its consequences are bearable or not for that type of application. As we shall see in more detail in Section IV, in code for DNNs, the decorrelation effect does occur, typically in precisely the way illustrated with the code sequence above. It does not occur in its even more intractable appearances, like in cases when $y = \sin x$ and $z = x - y$, where y and x are correlated for small x , as the Taylor series of sin starts with $x - 1/6 x^3$.

For the easy case when two variables in code correlate because they are copies on of the other –as in the example code illustrated above– a simple solution to overcome the decorrelation effect in CAA (and IA) exists: all FP quantities analyzed by CAA are labeled with a unique identifier that relates to their moment of creation in the execution of the program. This identifier is never repeated for any other FP quantity but for assignment, where the identifier does get copied. Subtraction (and division) operations in CAA can then start by checking whether the identifier of both operands happens to be the same. If it does, both operands are correlated as they are copies one of each other. Interval bounds of $[0; 0]$ and CAA error bounds of $\bar{\delta} = 0$ and $\bar{\varepsilon} = 0$ can then be returned. This solution is crude but addresses all simple decorrelation cases found in DNNs code.

Yet another issue with code analysis with CAA and IA comes in the form of **if** statements depending on FP variables –which are to be analyzed– and, generally, *control flow* depending on FP variables [25]. FP code might take the one or the other branch by evaluating a comparison like $x < y$ to a boolean, which, of course, might be falsed due to the errors on x and y . In contrast, CAA and IA, which replace x and y by whole classes (intervals resp. abstract approximate quantities with bounded error measured in units of u), cannot even evaluate the expression to a unique boolean in cases where the intervals for x and y intersect or where the errors make the boolean answer not unique. Some approaches for this control flow issue have been proposed [27].

Fortunately, in code for DNNs, this issue is virtually absent. Code for DNNs, in inference mode, does not contain control flow in the form of loops that depend on FP values. In other words, no iterative FP techniques are used. All control flow for loops comes from the DNN’s configuration and the respective dimensions of the manipulated tensors. As we shall see in more detail in Section IV, the only **if** statements in DNNs that depend on FP values are encountered in activation layers such as pooling or softmax layers. This is due to the very nature of DNNs: in order to make training possible, from a bird-view perspective, DNNs need to represent (non-linear) functions that are differentiable. Branches would necessarily introduce discontinuities of that derivative. In the concerned code sequences, the **if** serve the only purpose of computing minima and maxima on vectors of FP values, hence does not influence the output directly. We solve the control flow issue in a similar way as the decorrelation effect: the point is to provide the CAA and IA arithmetics, which, again, are concerned with local effects, with just enough global insight on the program’s logic. For instance, the quantities analyzed with CAA and IA can be labeled with bounds given in the form of other CAA+IA quantities that are minimum or maximum bounds for them. Subsequent CAA or IA operations, like subtraction, can then exploit the fact that if for example a quantity \hat{q} is upper-bounded by \hat{M} , i.e. $\hat{q} \leq \hat{M}$, the result of the subtraction $\hat{q} - \hat{M}$ will always be bounded by $\hat{q} - \hat{M} \leq 0$.

As a matter of course, DNNs that perform classification tasks do contain **if** statements that depend on FP values. These statements are the one executed as the very last step, when the one-hot output of a softmax layer [1] gets translated into the predicted numerical integer class. This code boils down to computing the integer *argmax* index for a vector of FP values of probabilities; the DNN picks the class that is the most probable. However, it is the very aim of this paper to analyze the effects of roundings in the FP arithmetic for the DNN’s inference on the output class, which we discuss in Section IV.

To wrap it up, our approach is to analyze DNNs for FP rounding errors using CAA and IA, where the FP quantity in the DNN to be analyzed gets replaced by an arithmetical object containing the following entries:

- a unique ID of the quantity, in the form of an integer,
- the FP value in the IEEE754-2019 (or any other) FP format that would be used if the DNNs were implemented without this enhanced CAA+IA arithmetic,
- an interval holding the actual error of the latter FP value,

for reference purposes,

- an absolute error bound $\bar{\delta} \in \mathbb{R}^+ \cup \{+\infty\}$, for this quantity, in units of u ,
 - a relative error bound $\bar{\varepsilon} \in \mathbb{R}^+ \cup \{+\infty\}$, for this quantity, in units of u ,
 - an interval safely enclosing all possible values for this quantity if no FP rounding error occurred,
 - an interval safely enclosing all possible values for this quantity, as it is evaluated with rounding FP arithmetic and,
 - optionally, a lower and an upper bound for this quantity.
- These bounds are given in the form of arithmetical objects of the same nature.

All operators required for DNNs, starting with assignment, going over computational operators like $+$, $-$, \times , $/$, $\sqrt{\quad}$ to functions like \exp , \log and \tanh are overloaded to work on such CAA+IA arithmetical objects, propagating all entries as described above. As a result, a DNNs run on an example input, widened with interval bounds for the inputs’ ranges, provides an output in these arithmetical objects, from which errors on probabilities etc. can be read off. As the absolute and relative error bounds are expressed in units of u , that same output can be used to tailor a DNN’s FP precision to just the right amount of tolerable final error. We shall describe the use of this arithmetic just below, in Section IV. As for the technical realization of this enhanced CAA+IA arithmetic in an actual software tool, we refer the reader to Section V.

IV. COMPUTER ARITHMETIC LOOK AT DNNs

As we shall see in the next Section V, the enhanced CAA arithmetic we just described is able to automatically analyze given FP code for DNNs and to come up with absolute or relative error bounds, expressed in units of u , that are pretty tight and suit their purpose. However, as useful as this automatic analysis might be for application programmers, we wanted to ensure its tightness and validity from a more theoretical standpoint. This Section strives at providing this insight. For the sake of brevity, we shall focus on DNNs for classification problems. The analysis is similar for other types of problems. We will present a concrete example of a DNN for a non-classification problem in Section V.

DNNs for classification problems transform high-dimensional input data into an output vector that is a one-hot representation of the class detected for the input’s class. This is, the output vector has as many entries as there are classes, and the i -th entry of that vector contains a probabilistic estimate of the confidence of the DNN the input is in the i -th class. That estimate is expressed as a probability; all entries are hence between 0 and 1 and sum up to 1. Post-processing after a DNN picks the class for which the confidence estimate is highest, computing the *argmax* on the output vector. The index of this output class can then be translated into e.g. a textual representation of the class, such as “*Cat*” or “*Dog*”.

In the case when the maximum confidence estimate is at 50% and the second-to-maximum confidence estimate is also at 50%, the slightest change to these output values will of course make the DNN commit a misclassification, outputting e.g. “*Dog*” when the input represents a “*Cat*”. Such a slight change may stem from FP rounding errors. For DNN input data where maximum confidence is at 50%, no FP arithmetic –but

exact arithmetic—exists avoiding misclassifications. However, when external knowledge on the DNN exists that guarantees that, on all possible inputs³, the DNN will output a one-hot vector with a top-1 value $p^* > 0.5$, it can be guaranteed that the second-to-maximum, top-2, value will be $p^\dagger < 1 - p^*$, leaving a margin of $1/2 (p^* - p^\dagger) > p^* - 1/2$ for each of the maximum and second-to-maximum entry to be affected by FP rounding error. Gaining this external knowledge is beyond the scope of this article, but approaches like SafeAI [18] seem to be able to provide it. This external minimum bound may also just be specified, accepting a certain percentage of misclassifications.

From a computer arithmetic perspective, we may hence assume that there is an absolute FP error margin $\mu = \frac{p^* - 1/2}{2}$ available for each element in the output vector, where $\frac{p^*}{2} > 0.5$ is the minimum bound established with external knowledge. Similarly, we may assume a relative FP error margin $\nu = \frac{2p^* - 1}{2p^* + 1}$ available. Our job is to rigorously ensure that no misclassification may occur given that error margin. Hence we need to choose FP precision k , resp. $u = 2^{-k+1}$ in such a way that we can guarantee that the DNN's inference accuracy is enough so that the FP rounding error does not exceed the margins. We may hence start reclaiming the DNN's FP algorithm from its end with that margin as some kind of FP error budget to be burnt for FP roundings.

As their last layer, most classification DNNs have a Softmax layer, as it was defined in Section II. We must hence analyze the FP error in output of a Softmax layer. This analysis will also serve as an illustration of error analysis for the different layers; we analyzed all layers, for the sake of brevity, we shall only report on the Softmax layer. The error in output of a Softmax layer has two sources: (1) the FP rounding errors committed during the layer's evaluation and (2) the errors present in input to the layer, propagated in an amplified or attenuated manner by the layer. The analysis of the first kind of errors, the rounding errors, is trivial, as the Softmax function required just the evaluation of exponentials, a division (often implemented as a division of a logarithm) and the summation of positive values, obtained by exponentiating the input. We shall hence not address this point any further.

For the propagated error of a Softmax layer, the following analysis can be performed; herein, \hat{x}_i are the elements of the computed input vector affected by an absolute error δ_i , approximating the unknown, ideal x_i . The \hat{y}_i are the output vector elements, approximating the unknown, ideal y_i .

$$\begin{aligned} \hat{y}_i &= \frac{e^{x_i + \delta_i}}{\sum_k e^{x_k + \delta_i}} \\ &= \frac{e^{x_i}}{\sum_k e^{x_k} \cdot \left(1 + \frac{\sum_k e^{x_k} \cdot (e^{\delta_k - \delta_i} - 1)}{\sum_k e^{x_k}}\right)} \\ &= y_i \cdot \left(1 + \frac{1}{1 + \eta_i} - 1\right) = y_i \cdot (1 + \varepsilon_i) \quad (10) \end{aligned}$$

with

$$\eta_i = \sum_k \frac{e^{x_k}}{\sum_j e^{x_j}} \cdot (e^{\delta_k - \delta_i} - 1).$$

This quantity is easily bounded with

$$\begin{aligned} |\eta_i| &\leq \sum_k \frac{e^{x_k}}{\sum_j e^{x_j}} \cdot \max_t |e^{\delta_t - \delta_i} - 1| \\ &\leq \max_k |e^{\delta_k - \delta_i} - 1|. \end{aligned}$$

With some mild assumptions on bounds for the δ_k and η_i , it can further be shown that the relative error ε_i affecting \hat{y}_i is bounded by

$$|\varepsilon_i| \leq 11/2 \max_k |\delta_k|, \quad (11)$$

essentially by taking the Taylor development of $e^x - 1$.

This analysis therefore shows us the following: the Softmax layer transforms the absolute error in its input into a relative error of approximately⁴ the same amount in output. Our margin of $\nu = \frac{2p^* - 1}{2p^* + 1}$ becomes hence an absolute error margin in input of the Softmax layer. This input of the Softmax layer is in general the output of a convolutional layer, where the arithmetical difficulty is in a summation, which lives very well when it just needs to satisfy an absolute error bound. Amazingly, the bound given above does not at all depend on the number of elements of the vectors x and y .

In order to illustrate this stability argument more intuitively, let us give a numerical example: let $p^* = 0.60$, i.e. the classifying DNN shows at least 60% confidence for the best output class. Then $\nu > 0.0909 > 2^{-3.45}$, meaning that FP results with about 3.45 valid bits are sufficient. Then a maximum element-wise absolute error of $\frac{0.0909}{5.5} > 1.65 \cdot 10^{-2}$ is still tolerated on the input of the softlayer. This means for a convolution or dot-product, i.e. summation, fixed-point arithmetic with a quantization unit of $1.65 \cdot 10^{-2}$, i.e. about 2^{-6} is enough. FP arithmetic can only do better, its precision is at least these $6+g$ bits, provided the inputs to the summation are bounded around 2^g .

The boundedness of the values manipulated by DNNs is something we have already stated in Section II. Most activation layers bound their outputs to $[0; 1]$, equivalent to a bound $2^g = 2^0$. Convolutional or fully-connected layers also exhibit pretty small bounds 2^g , which are easily established, and, by the way, perfectly bounded with IA [28], considering the boundedness of the DNN's coefficients, the relatively small dimensions of the manipulated vectors, matrices and tensors and the boundedness of the preceding input.

It is hence all but surprising to observe that

- DNN inference behaves very well for FP arithmetic with low precision,
- DNN inference behaves very well for FP arithmetic with low exponent range, as fixed-point arithmetic already provided enough accuracy for the subsequent layers, such as softmax,
- and analysis with CAA that exhibits small FP error bound in output does provide tight and sensible results.

³For a reasonable definition of what a *possible* input is.

⁴i.e. 5.5 times larger

	max absolute error in u	max relative error in u	analysis time	required precision to prevent misclassification with $\underline{p}^* = 0.60$
Digits	1.1u	3.4u	12s per class	$k = 8$
MobileNet	22.4u	11.5u	4.2h per class	$k = 8$
Pendulum	1.7u	–	100ms	–

TABLE I. NUMERICAL RESULTS FOR EXPERIMENTS WITH $u \leq 2^{-7}$.

V. CAA-BASED FP ERROR ANALYSIS AND EXPERIMENTAL RESULTS

We implemented our semi-automatic FP accuracy analysis tool building upon a combination of existing software packages for DNNs, such as `frugally-deep`⁵, which we patched pretty heavily. We coupled these packages with a C++ implementation of the enhanced CAA that we have described in Section III. This C++ implementation of CAA was written from scratch. The implementation is currently based on IA provided by MPFI 1.5.3, which is itself based on MPFR 4.0.2 on top of GMP 6.1.2 [29]–[31]. However, we wrapped MPFI in a C++ façade class in order to facilitate transition to other IA libraries later. We use `g++` version 8.3.0 to compile our code. The `frugally-deep` library we are using requires the use of C++ in its C++20 version [32]. Our contribution in terms of code consists in C++ classes to implement CAA as we have described it and in the patches required to allow for binding and use of that CAA arithmetic instead of plain IEEE754-2019 arithmetic in `frugally-deep`. Our workflow runs only with our version of `frugally-deep`, not with a stock version. Thanks to `frugally-deep`, our semi-automatic FP accuracy analysis tool is compatible with almost all DNNs as they are designed and trained with TensorFlow/Keras [33], [34]. The `frugally-deep` package first converts DNN models to JSON files, and then provides C++ header classes that allow loading of JSON files as object graphs that can be evaluated on the input data. The `frugally-deep` library leverages several other C++ libraries for this task, the most prominent of which is Eigen [35] that permits binding of custom arithmetic.

Our CAA class structure consists of three classes: a façade class for the front-end binding with `frugally-deep`; a class that actually implements the CAA arithmetic and overloads all necessary arithmetic operations; and a back-end wrapper for IA. We did not use existing MPFI wrappers in C++, in order to have the possibility to exploit the performance advantage of new C++20 features, like move constructors and move assignment operators.

Our workflow runs as follows: using `frugally-deep` we construct a C++ program to load a DNN model designed in TensorFlow/Keras, as well as the input data, expressed with CAA objects. The bounds on these data are trivial in most cases, e.g. image data gets annotated with 8-bit unsigned values in $[0; 255]$. We run the resulting program for all possible classes to cover all possible control flows. And this can be done only for one representative of the class, no additional tests are required. The program outputs the inference result and the absolute and relative error on it. The error bounds are all given in units of u , an upper bound on which is user-configurable. The output error bounds can then be used to tailor the DNNs actual FP arithmetic, by applying the theory we described in

Section IV, determining the value of u such that the required accuracy bounds are still met.

We demonstrate our tool on several examples of DNNs and give some results in Table I.

a) Digits: We built a simple DNN for the recognition of hand-written digits and trained it on the MNIST dataset [36]. This model requires around 0.7 million parameters and consists of three Dense, two ReLU and a Softmax layer. As input, it takes 28×28 gray-scale images (i.e. a flattened vector of length 784) and has a 10-dimensional output vector whose i^{th} element indicates the probability that the input image is the digit i . Table I illustrates the results of analysis, where the maximum absolute and relative errors denote the maximum errors over all possible classes. We also observed that on the top-1 choice, the relative error bounds are quite tight, while on the other elements the relative error looks less good. However, the bound (11) still holds and in any case, the absolute error stays low. Our analysis shows that the network can safely run with 7-bit precision FP.

b) MobileNet: We used a Keras pre-trained model for this considerably bigger network for Imagenet classification. MobileNet requires around 27 million parameters. The complete architecture can be found in Keras documentation, we will only state here that it is a Convolutional Neural Network with 28 Conv layers, 27 BatchNorm layers followed by ReLU activations and a Softmax layer that classifies 224×224 RGB images over 1000 classes. This challenging example revealed a performance bottleneck in our tool. To analyze the model over one class it took the tool around 4 hours on a conventional laptop. Our performance analysis determined that most of the analysis time was dedicated to the memory allocation process somewhere deep in MPFI. Regardless of the analysis time, the tool successfully illustrated that even for large-scale models our analysis techniques compute tight error bounds.

c) Pendulum: This small neural network model comes from the context of reinforcement learning applied for autonomous control [19] and aims at approximating a Lyapunov function for a non-linear controller. The article [19] proposes a new methodology for certified approximation of neural controllers based on SAT theory. This example is interesting from the formal verification point of view: our bound on the absolute error can be effortlessly incorporated into the existing verification procedure. This network has two Dense layers and two tanh activations. It takes a $2D$ coordinate vector as input and, as in [19], we tested it for the interval $[-6; 6]$. Our tool provided an absolute error bound in a fraction of a second. A relative error bound does not exist since the output interval contains zero.

VI. CONCLUSION AND PERSPECTIVES

With this work we proposed a semi-automated way to *bound* and *interpret* the impact of rounding errors due to the

⁵<https://github.com/Dobiasd/frugally-deep>

precision choice for inference in generic DNNs. We presented a software tool that, thanks to frugally-deep library, can receive any TensorFlow/Keras model in the front-end. We support the most common activation and computational layers.

The back-end that we developed automatically computes and propagates rounding errors through the computations. For this, we have introduced a combination of Affine and Interval Arithmetics called CAA. This new construction permitted us to compute both relative and absolute error bounds. Our implementation of CAA is based on rigorous error analysis for arithmetic operations, as well as for all the necessary elementary functions for activation layers (e.g. tanh, exp). We enhance CAA with just enough global insight on the program’s logic in order to fight the decorrelation effect and take care of the control flow depending on FP variables. Software implementation of the arithmetic is done in C++ in a generic way.

We offer a computer arithmetic look at the computations with the DNNs. When analyzing the computations within activation layers, we establish that activation functions are actually transforming absolute errors on their inputs into relative errors in the output. Which means that even if the computational layers yield relatively imprecise results, activation layers will recover decent relative errors, as long as inputs are bounded (which is basically always the case thanks to bounded activation functions and normalizations).

Finally, we offered the first, to the best of our knowledge, interpretation of the impact of precision choice upon the top-1 accuracy of classification networks. If we reason that the goal is to preserve the top-1 choice w.r.t. the reference model, we can establish the minimum required precision as a function of the deduced error bound and the distance between the top-1 and top-2 choice.

This reasoning reflects perfectly the fact that as long as the model was well-trained for some classification problem and can clearly distinguish between classes, then the network is extremely robust to low-precision evaluation.

We identify several axes of future improvements to this work. The first improvement will be to improve the tool’s performance, which so far does not scale up to models having tens of millions parameters. We identified the performance bottleneck to be the memory management in MPFI. The solution would be to replace the the underlying IA implementation by a faster one. In addition to that, the manually-coded error analysis is of course error-prone, from the formal verification point of view. Another limitation of the proposed tools is that it analyzes one implementation produced by the frugally-deep library. In order to support other implementations, e.g. using Kahan summation instead of a straightforward one, a corresponding code generation phase needs to be added. The second improvement concerns mixed-precision implementations, as proposed by NVIDIA [6], which can be achieved by removing the global u and parameterizing the error analysis with the input/output precision. To go further, we would like to combine our results with a static analysis and mixed-precision tuning tool like Daisy [37] to accelerate specific parts of DNN models. Extension towards the training of DNNs is non-trivial and requires an analysis of gradient descent algorithms. Finally, [38] proposes to relate the classification capacity of DNNs

with the geometry of the object manifolds issued after each layer. By combining our error analysis with the quantitative “separability” measure from [38], we hope to back up our interpretation of the relation between precision and accuracy for DNNs with a solid theoretic-geometrical basis.

REFERENCES

- [1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.
- [2] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA, USA: MIT Press, 1969.
- [3] G. Tesauro, “Td-gammon, a self-teaching backgammon program, achieves master-level play,” *Neural Comput.*, vol. 6, no. 2, p. 215–219, Mar. 1994.
- [4] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 248–255.
- [5] J. L. Hennessy and D. A. Patterson, *Computer Architecture, Sixth Edition: A Quantitative Approach*, 6th ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2017.
- [6] P. Micikevicius, S. Narang *et al.*, “Mixed precision training,” *CoRR*, vol. abs/1710.03740, 2017.
- [7] G. Henry, P. T. P. Tang, and A. Heinecke, “Leveraging the bfloat16 artificial intelligence datatype for higher-precision computations,” in *IEEE 26th Symposium on Computer Arithmetic (ARITH)*, 2019.
- [8] N. Burgess, J. Milanovic, N. Stephens, K. Monachopoulos, and D. Mansell, “Bfloat16 processing for neural networks,” in *IEEE 26th Symposium on Computer Arithmetic (ARITH)*, 2019.
- [9] A. Agrawal, S. M. Mueller, B. M. Fleischer, X. Sun, N. Wang, J. Choi, and K. Gopalakrishnan, “Dfloat: A 16-b floating point format designed for deep learning training and inference,” in *IEEE 26th Symposium on Computer Arithmetic (ARITH)*, 2019.
- [10] (2020) Microsoft brainwave project. [Online]. Available: <https://www.microsoft.com/en-us/research/project/project-brainwave/>
- [11] R. Banner, Y. Nahshan, and D. Soudry, “Post training 4-bit quantization of convolutional networks for rapid-deployment,” in *Advances in Neural Information Processing Systems 32, NeurIPS*, 2019.
- [12] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized neural networks,” in *Advances in Neural Information Processing Systems 29, NeurIPS*, 2016.
- [13] —, “Quantized neural networks: Training neural networks with low precision weights and activations,” *CoRR*, vol. abs/1609.07061, 2016.
- [14] E. Kravchik, F. Yang, P. Kisilev, and Y. Choukroun, “Low-bit quantization of neural networks for efficient inference,” in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2019.
- [15] Y. Choi, M. El-Khamy, and J. Lee, “Learning low precision deep neural networks through regularization,” *CoRR*, vol. abs/1809.00095, 2018.
- [16] M. Courbariaux and Y. Bengio, “Binarynet: Training deep neural networks with weights and activations constrained to +1 or -1,” *CoRR*, vol. abs/1602.02830, 2016.
- [17] D. D. Kalamkar, D. Mudigere *et al.*, “A study of BFLOAT16 for deep learning training,” *CoRR*, vol. abs/1905.12322, 2019.
- [18] T. Gehr, M. Mirman, D. Drachler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev, “Ai2: Safety and robustness certification of neural networks with abstract interpretation,” in *2018 IEEE Symposium on Security and Privacy (SP)*, 2018.
- [19] Y. Chang, N. Roohi, and S. Gao, “Neural lyapunov control,” in *NeurIPS*, 2019.
- [20] V. Dumoulin and F. Visin, “A guide to convolution arithmetic for deep learning,” 2016.
- [21] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *CoRR*, vol. abs/1502.03167, 2015.
- [22] J.-M. Muller *et al.*, *Handbook of Floating-Point Arithmetic*, 2nd ed. Birkhäuser Basel, 2018.
- [23] N. J. Higham, *Accuracy and stability of numerical algorithms (2 ed.)*. SIAM, 2002.

- [24] J. H. Wilkinson, "Error analysis of floating-point computation," *Numer. Math.*, vol. 2, no. 1, p. 319–340, Dec. 1960.
- [25] E. Goubault and S. Putot, "Static analysis of finite precision computations," in *Proceedings of the 12th International Conference on Verification, Model Checking, and Abstract Interpretation*, ser. VMCAI'11. Berlin, Heidelberg: Springer-Verlag, 2011, p. 232–247.
- [26] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*. SIAM, 2009.
- [27] E. Goubault, "Static analysis by abstract interpretation of numerical programs and systems, and fluctuat," *LNCS*, vol. 7935, pp. 1–3, 2013.
- [28] S. Sergey, *Interval algebraic problems and their numerical solution (habilitation in Russian)*. Institute of Computational Technologies. Russian Academy of Sciences, 2000.
- [29] N. Revol and F. Rouillier, "Motivations for an arbitrary precision interval arithmetic and the MPFI library," *Reliable Computing*, vol. 11, no. 4, pp. 275–290, 2005.
- [30] L. Fousse, G. Hanrot, V. Lefèvre, P. Pélicier, and P. Zimmermann, "Mpfr: A multiple-precision binary floating-point library with correct rounding," *ACM Trans. Math. Softw.*, vol. 33, no. 2, p. 13–es, Jun. 2007.
- [31] T. Granlund and the GMP development team, *GNU MP: The GNU Multiple Precision Arithmetic Library*, 5th ed., 2012, <http://gmplib.org/>.
- [32] R. Smith, "Working draft, standard for programming language c++," <http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2020/n4849.pdf>, 2020-01-14.
- [33] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: <http://tensorflow.org/>
- [34] F. Chollet *et al.*, "Keras," <https://keras.io>, 2015.
- [35] G. Guennebaud *et al.*, "Eigen v3," <http://eigen.tuxfamily.org>, 2010.
- [36] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: <http://yann.lecun.com/exdb/mnist/>
- [37] E. Darulova *et al.*, "Daisy - framework for analysis and optimization of numerical programs," in *Tools and Algorithms for the Construction and Analysis of Systems*, 2018.
- [38] U. Cohen, S. Chung, D. D. Lee, and H. Sompolsky, "Separability and geometry of object manifolds in deep neural networks," *bioRxiv*, 2019.