Experimental Results of Vectorized Posit-based DNNs on a Real ARM SVE High Performance Computing Machine

Marco Cococcioni[1], Federico Rossi[1], Emanuele Ruffaldi[2], and Sergio Saponara[1]

[1]University of Pisa, Dept. of Information Engineering
[2]MMI s.p.a., Calci, Pisa, Italy
{marco.cococcioni,sergio.saponara}@unipi.it
federico.rossi@ing.unipi.it
emanuele.ruffaldi@mmimicro.com

Abstract. With the pervasiveness of deep neural networks in scenarios that bring real-time requirements, there is the increasing need for optimized arithmetic on high performance architectures. In this paper we adopt two key visions: i) extensive use of vectorization to accelerate computation of deep neural network kernels; ii) adoption of the posit compressed arithmetic in order to reduce the memory transfers between the vector registers and the rest of the memory architecture. Finally, we present our first results on a real hardware implementation of the ARM Scalable Vector Extension.

Keywords: ARM SVE, Vectorization, alternative representation of reals, posit arithmetic, HPC

1 Introduction

Nowadays, Deep Neural Networks (DNNs) face new problems and challenges: on one hand, there is a need to reduce network design and computation complexity in order to better accomplish real-time tasks in resource-constrained devices. On the other hand, the trend is to address specific platform accelerators (for example, NVIDIA cuDNN for NVIDIA Graphics Processing Units (GPUs)) to significantly accelerate neural network processing in both the training and inference phases.

DNNs extensively use matrix multiplications, dot products and convolutions, highlighting the need for vectorization routines capable of increasing throughput for these operations. Although the use of GPUs in this field is important, high implementation costs and low-power requirements may prevent such components from being used. Several implementations of vector CPUs are available in most of the common processor architectures: i) Intel/AMD AVX2/SSE for x86 processors, ii) RISC-V "V" vector extension for the RISC-V architecture, iii) ARM SVE for the ARMv8 architecture. In [1–3] we were able to produce binaries that employ, respectively, ARM SVE and RISC-V vectorization. In particular, for the ARM SVE platform we were able to enable vectorization in two different tiers: one tier was the auto-vectorization approach that relies on automatic optimization from the compiler (for example, loop unrolling). The second tier involved was the use of intrinsic functions that allowed us to explicitly use ARM SVE instructions in the C++ code.

The idea behind vector extensions is to fit as much data as possible in the vector registers, that acts as very low latency memory for the vector processor. In order to increase the data we can fit in the registers and reduce the memory transfer, it is crucial to minimize the number of bits used to represent the weights of the DNNs. Several alternatives to the standard IEEE 754 32-bit floating point have been proposed: Google (Brain Float — BFloat16 — [4]), Intel (Flexpoint — FP16 — [5, 6]) have already suggested several concepts. BFloat8 [7] is also very interesting, adding the support for stochastic rounding.

The posit^{TM} format ([8–10]) is one of the most promising representations that deviates from the IEEE 754 standard. In machine learning, this kind has been shown to be a great drop-in replacement for 32-bit IEEE 754 floats, using only 16 bits [11–16]. Furthermore, it has been successfully used in low-precision inference down to 8-bit posit representation with minimal network inference accuracy degradation. Moreover, as explained in [12], this number system can be used to create quick, approximated, and efficient activation functions for neural networks such as the sigmoid function by simply using the already existing CPU integer arithmetic operations.

While in [1] we were not able to profile our code on a real hardware (we used the ARM Instruction Emulator for SVE (ARMIE)), in this paper we will instead evaluate the performance of the vectorized extension of the cppPosit C++ posit arithmetic library targeting a real hardware implementation of the ARM SVE architecture.

2 Posit number and cppPosit library

Posit numbers are represented by a fixed length format. The overall length (*nbits*) and exponent length (*esbits*) can be modified. The posit format can have a maximum of 4 fields as in Figure 1. Hereafter we describe the format fields:

- Sign field: 1 bit;
- Regime field: variable length, composed by a string of bits equal to 1 or 0 ended, respectively by a 0 or 1 bit;
- Exponent field: at most *esbits* bits (it can even be absent);
- Fraction field: variable length mantissa (it can even be absent too).

Given a posit$\langle nbits, esbits\rangle$, represented in 2's complement signed integer X and let e and f be exponent and fraction values, the real number r represented by X encoding is:

$$r = \begin{cases} 0, \text{ if } X = 0 \\ \text{NaN, if } X = -2^{(nbits-1)} \\ sign(X) \cdot useed^k \cdot 2^e \cdot (1+f), \text{ otherwise} \end{cases}$$

Where $useed = 2^{2^{esbits}}$ and k is strictly related to the regime length l and bitstring (b is the bit that composes the string of identical bits, e.g. in 00001 $b = 0$). If $b = 0$ the k is negative, otherwise the k is positive:

$$k = \begin{cases} -l, \text{ if } b = 0 \\ l - 1, \text{ otherwise} \end{cases}$$

In [17] we proved some interesting properties for the configuration ($esbits = 0$). Under this configuration, we could implement fast and approximated versions of common operations. We could evaluate these operations only using the arithmetic-logic unit (ALU) making them faster than the original ones computed using the FPU. These operations are the double and half operators ($2x$ and $x/2$), the inverse operator ($1/x$) and the one's complement ($1 - x$). In [1] we combined this idea with vectorization, obtaining several posit operations such as ELU and Tanh, exploiting the already existent vector integer operations in the ARM SVE vector environment.

We provide the software support for posit numbers through the cppPosit library, developed in Pisa and maintained by the authors of this work. We exploit templatization to configure the posit format. We classify posit operations into four different operational levels, identified with ($\mathcal{L}1$-$\mathcal{L}4$). Each level has increasing computational complexity (see [17]).

When in levels $\mathcal{L}3$-$\mathcal{L}4$ we need to use three different back-ends to accelerate posit operations that cannot be directly evaluated via ALU (waiting for proper posit hardware support):

- FPU back-end;
- Fixed back-end, exploiting big-integer support (64 or 128 bits) for operations;
- Tabulated back-end, generating lookup tables for most of the operations (suitable for Posit$\langle[8, 12], *\rangle$ due to table sizes).

3 The advantages of Vectorized CPUs

The newly introduced ARM SVE is a modern Single Instruction, Multiple Data (SIMD) for the 64-bit ARMv8 instruction set. It is intended as an evolution of the older ARM Neon vector instruction engine. The power of SVE lies in the Vector Length Agnostic (VLA) nature of the engine; indeed, there is no need to specify, at compilation time, the size of the vector registers. This dimension can be retrieved at run-time using a single assembly instruction. This design highly enhance portability of code across different ARM SVE platforms and revisions.

The VLA design is very similar to the one adopted RISC-V vector extension. The main differences between the RISC-V "V" extension and ARM SVE that we believe worht mentioning are:

- Maximum register size: while ARM SVE can only reach 2048-bits, RISC-V "V" can reach up to 16384-bits
- Register grouping: when dealing with different element sizes in the same vector loops, RISC-V can handle the wider element grouping registers so that it can be indexed as it was smaller (e.g. if we want to convert a vector of 16-bit posits to a vector of 32-bit floats).

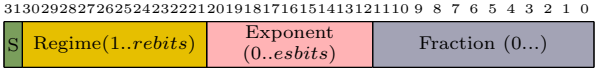

Fig. 1: Illustration of a posit$\langle 32, 9\rangle$ data type. Both the exponent and the fraction field can be absent, for specific configurations having a regime field particularly lengthy.

3.1 Vectorized CPUs and Deep Neural Networks

The most recurrent computations in deep neural networks are [1, 2, 17]:

- GEMM (general matrix-matrix multiplication) (training phase)
- matrix-vector multiplication (inference phase)
- matrix-matrix convolution product (both training and inference phases)
- vector-vector dot product (both training and inference phases, for computing the activations)
- non-linear activation function (both training and inference phases), computed over a vector of activations.

In this work we have used posit, since they allow to save memory for storing the weights. Moreover we and other authors have proved that posit16 is as accurate as 32-bit IEEE Floats for machine learning applications. In machine learning application even a posit8 can be accurate enough compared to 32-bit floats, thus saving 4x storage space (both on disk and, more importantly, on RAM and caches) with minimal accuracy loss.

4 Test-bench, methodologies and benchmarks

In [1] we tested ARM SVE capabilities using the ARM Instruction Emulator. We ran the emulator on a HiSilicon Hi1616 CPU with 32@2.4GHz ARMv8 Cortex-A72 cores. This emulator was able to trap all the illegal instruction interruptions coming from the execution of binaries compiled using the SVE instruction set extension. These instructions were then executed via software by the ARM Instruction Emulator. During *emulation* we were able to modify the vector register length from 128-bit to 2048-bit.

Instead, in this work we were able to use an actual hardware implementation of the ARM SVE architecture using the HPE Apollo80 machine available at University of Pisa. The Apollo80 is based on the ARMv8 A64FX core [18], the first commercial implementation of the ARM SVE architecture. In particular, the ARMv8 A64FX is the first processor to support the full feature set of the ARM SVE architecture without emulating any instruction.

This platform is particularly interesting since it will be employed in the European Processor Iniative framework [19], and it will be used as a base computing platform for the EUPEX and TEXTAROSSA EuroHPC projects.

In detail, this platform consists of 4 different blades equipped with 48 ARMv8 A64FX Cores with 512-bit vector registers, running, respectively, at 1.8 and 2.0 GHz. Each blade has access to 32GB of High Bandwidth Memory.

In order to evaluate SVE-related performance on this machine, we used following benchmarks: i) vectorized activation functions only using posits and integer vector instructions, ii) vectorized matrix-matrix multiplication and convolution.

Moreover, we employed posit numbers to compress and decompress data across the kernels, in order to reduce memory transfers by a factor 4 (with posit$\langle 8, 0\rangle$) or 2 (with posit$\langle 16, 0\rangle$). Also compression and decompression phases were implemented using vectorization, exploiting vector integer arithmetic.

Benchmarks were compiled using the `armclang++ 20.3` compiler, based on `LLVM 9.0.1` and executed on the aforementioned Apollo80 machine, running `CentOS Linux release 7.9.2009`.

5 Experimental results and discussion

Figure 2 shows the performance of the activation function benchmarks on the Apollo80 machine. The benchmarks consisted in the computation of sigmoid and extended linear unit (ELU) on 4096-wide vectors (even if the hardware only supports 512-bit). Each computation was repeated 100 times and the average computation time was reported.

As reported, the computation of the two activation functions using posit$\langle 8, 0\rangle$ benefits from the reduction in size of the format. This is because most of the steps of the activation function computation is performed using `int8_t` for posit$\langle 8, 0\rangle$ and `int16_t` for posit$\langle 16, 0\rangle$.

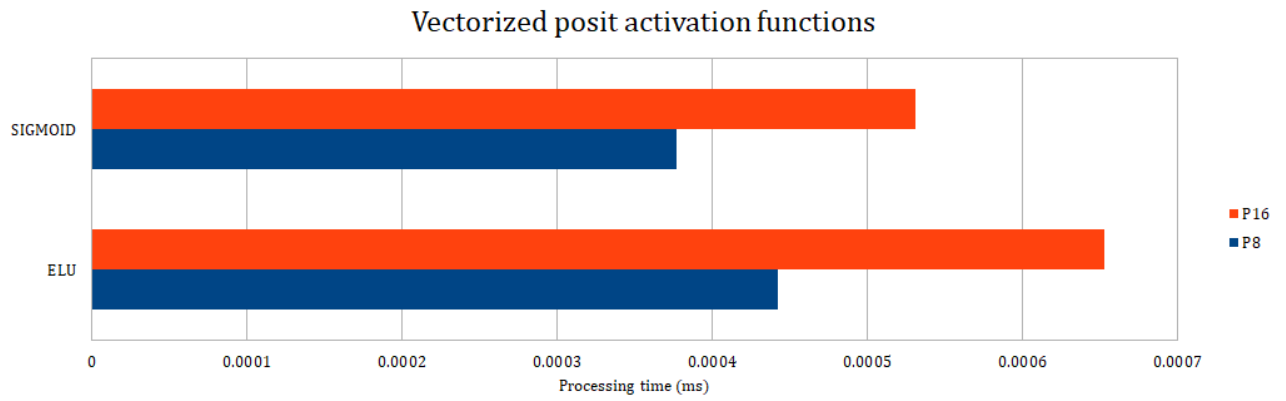

Fig. 2: Processing time of vectorized activation functions on a 4096-element vector with a 512-bit vector register length.

Figure 3 shows the performance of the kernel benchmarks, using posit$\langle 8,0\rangle$ and posit$\langle 16,0\rangle$. These benchmarks consisted in the computation of: i) dot product between vector of 4096 elements, ii) convolution on a 128×128 image with a 3×3 filter, iii) matrix-matrix multiplication between square matrices of 128×128 elements.

As reported, the benefit in reducing the information size is not as effective as in the previous case. The issue is that in this case, we could not use posits in every step of computation (of course ARM SVE lacks dedicated posit instructions). This means that we used posits just for data compression and decompression at the start and at the end of the kernels. For example, in the convolution kernel, we decompress the posit input to float using our vectorized routine, then we compute the convolution using native vectorized float support from ARM SVE and finally we compress the result back to posits.

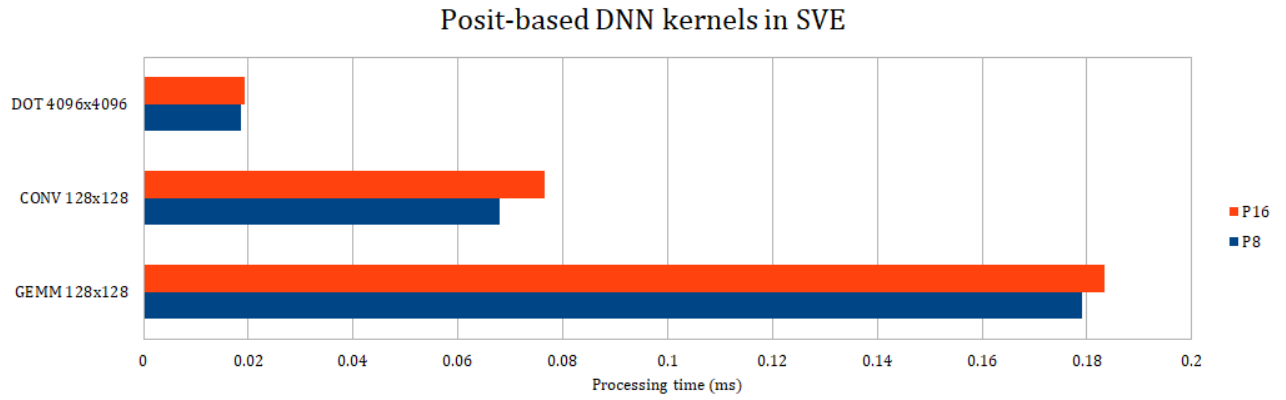

Fig. 3: Processing time of vectorized DNN kernels with a 512-bit vector register length (DOT: vector-vector dot product, CONV: matrix-matrix convolution, GEMM: General matrix matrix multiplication).

Figures 4 and 5 show the measured speedup from the emulated machine to the real hardware implementation. The speedup is computed as $t_{\text{HPE80}}/t_{\text{ARMIE}}$. Since we already proved that we can get better timing performance using posit$\langle 8,0\rangle$ instead of posit$\langle 16,0\rangle$, we reported the speedup relative to posit$\langle 8,0\rangle$ computations. As reported, the speedup spans from at least ~ 11, in the case of the GEMM function, up to ~ 1500 in the case of the ELU function.

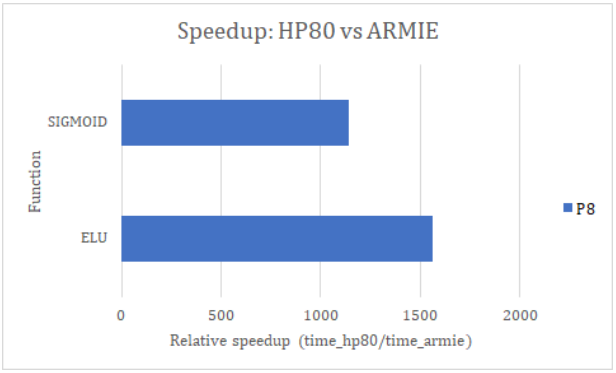

Fig. 4: Relative speedup of activation functions with 512-bit vector register length.

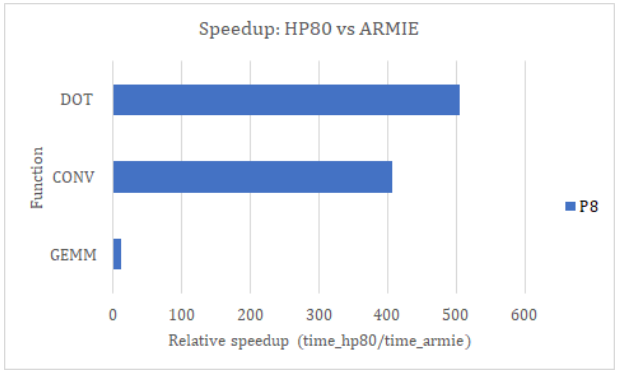

Fig. 5: Relative speedup of DNN kernels with a 512-bit vector register length.

6 Conclusions

In a previous work, we designed posit-based and vectorized operations on an ARM 64bit SVE emulator. The operations considered are the most time consuming ones in deep neural networks. In the present work we compared the impact of vectorization on a real machine. By using such machine, we were able to assess the true speedup due to vectorization, which turned out to be remarkable (with a speedup factor from 11× to 1500×, depending on the executed task). Future works will involve the combination of presented algorithms with MPI, to enable multi-processor or multi-node computation of deep neural networks (e.g. exploiting all the blades and cores of the HP80).

7 Acknowledgments

Work partially supported by H2020 projects (EPI grant no. 826647, `https://www.european-processor-initiative.eu/` and TEXTAROSSA grant no. 956831, `https://textarossa.eu/`) and partially by the Italian Ministry of Education and Research (MIUR) in the framework of the CrossLab project (Departments of Excellence). We thank the personnel of the Green DataCenter of the University of Pisa (`https://start.unipi.it/en/computingunipi`). In particular, we thank Prof. P. Ferragina, Dr. M. Davini and Dr S. Suin, for having provided us with the computational resources that have been used in the experimental section.

References

1. M. Cococcioni, F. Rossi, E. Ruffaldi, and S. Saponara, "Fast deep neural networks for image processing using posits and ARM scalable vector extension," *Journal of Real-Time Image Processing*, pp. 1–13, 2020.
2. ——, "Vectorizing posit operations on RISC-V for faster deep neural networks: experiments and comparison with ARM SVE," *Journal of Neural Computing and Applications*, vol. 33, pp. 10 575—10 585, 2021. [Online]. Available: https://doi.org/10.1007/s00521-021-05814-0
3. M. Cococcioni, F. Rossi, E. Ruffaldi, and S. Saponara, "Faster deep neural network image processing by using vectorized posit operations on a RISC-V processor," in *Real-Time Image Processing and Deep Learning 2021*, N. Kehtarnavaz and M. F. Carlsohn, Eds., vol. 11736, International Society for Optics and Photonics. SPIE, 2021, pp. 19 – 25. [Online]. Available: https://doi.org/10.1117/12.2586565
4. N. Burgess, J. Milanovic, N. Stephens, K. Monachopoulos, and D. Mansell, "Bfloat16 processing for neural networks," in *2019 IEEE 26th Symposium on Computer Arithmetic (ARITH)*, June 2019, pp. 88–91.
5. U. Köster, T. Webb, X. Wang, M. Nassar, A. K. Bansal, W. Constable, O. Elibol, S. Gray, S. Hall, L. Hornof *et al.*, "Flexpoint: An adaptive numerical format for efficient training of deep neural networks," in *In Proc. of teh 31st Conference on Neural Information Processing Systems (NIPS'17)*, 2017, pp. 1742–1752.
6. V. Popescu, M. Nassar, X. Wang, E. Tumer, and T. Webb, "Flexpoint: Predictive numerics for deep learning," in *In Proc. of the 25th IEEE Symposium on Computer Arithmetic (ARITH'18)*, June 2018, pp. 1–4.
7. N. Mellempudi, S. Srinivasan, D. Das, and B. Kaul, "Mixed precision training with 8-bit floating point," 05 2019.
8. J. L. Gustafson, *The End of Error: Unum Computing*. Chapman and Hall/CRC, 2015.
9. ——, "A radical approach to computation with real numbers," *Supercomputing Frontiers and Innovations*, vol. 3, no. 2, pp. 38–53, 2016.
10. J. L. Gustafson and I. T. Yonemoto, "Beating floating point at its own game: Posit arithmetic," *Supercomputing Frontiers and Innovations*, vol. 4, no. 2, pp. 71–86, 2017.
11. M. Cococcioni, F. Rossi, E. Ruffaldi, and S. Saponara, "Novel arithmetics to accelerate machine learning classifiers in autonomous driving applications," in *In Proc. of the 26th IEEE International Conference on Electronics, Circuits and Systems (ICECS'19)*, 2019, pp. 779–782.

12. M. Cococcioni, F. Rossi, E. Ruffaldi, and S. Saponara, "A fast approximation of the hyperbolic tangent when using posit numbers and its application to deep neural networks," *Int. Workshop on Applic. in Electronics Pervading Ind., Envir. and Society (ApplePies'19),* Lecture Notes in Electrical Engineering, vol. 627, 2019. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-37277-4_25
13. M. Cococcioni, E. Ruffaldi, and S. Saponara, "Exploiting posit arithmetic for deep neural networks in autonomous driving applications," in *In Proc. of the 2018 IEEE International Conference of Electrical and Electronic Technologies for Automotive (Automotive '18)*, 2018, pp. 1–6. [Online]. Available: https://doi.org/10.23919/EETA.2018.8493233
14. Z. Carmichael, H. F. Langroudi, C. Khazanov, J. Lillie, J. L. Gustafson, and D. Kudithipudi, "Deep positron: A deep neural network using the posit number system," in *2019 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2019, pp. 1421–1426.
15. S. H. Fatemi Langroudi, Z. Carmichael, J. Gustafson, and D. Kudithipudi, "Positnn framework: Tapered precision deep learning inference for the edge," 2019, pp. 53–59.
16. M. Cococcioni, F. Rossi, E. Ruffaldi, S. Saponara, and B. Dupont de Dinechin, "Novel arithmetics in deep neural networks signal processing for autonomous driving: Challenges and opportunities," *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 97–110, 2021. [Online]. Available: https://doi.org/10.1109/MSP.2020.2988436
17. M. Cococcioni, F. Rossi, E. Ruffaldi, and S. Saponara, "Fast approximations of activation functions in deep neural networks when using posit arithmetic," *Sensors*, vol. 20, no. 5, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/5/1515
18. "Fujitsu Processor A64FX," https://www.fujitsu.com/global/products/computing/servers/supercomputer/a64fx/, Accessed June 4th, 2021.
19. "European Processor Initiative, an H2020 project," https://www.european-processor-initiative.eu/, 2019-2021.