Vector-Vector-Matrix Architecture: A Novel Hardware-Aware Framework for Low-Latency Inference in NLP Applications

Matthew Khoury[*,1], **Rumen Dangovski**[*,†,2], **Longwu Ou**[1],
Preslav Nakov[3], **Yichen Shen**[2], **Li Jing**[†,1]
(*) equal contribution, (†) work done at Lightelligence, Inc.
[1]Lightelligence, Inc.
{matthew.khoury, longwu.ou, yichen, li}@lightelligence.ai
[2]Massachusetts Institute of Technology
rumenrd@mit.edu
[3]Qatar Computing Research Institute, HBKU
pnakov@hbku.edu.qa

Abstract

Deep neural networks have become the standard approach to building reliable Natural Language Processing (NLP) applications, ranging from Neural Machine Translation (NMT) to dialogue systems. However, improving accuracy by increasing the model size requires a large number of hardware computations, which can slow down NLP applications significantly at inference time. To address this issue, we propose a novel vector-vector-matrix architecture (VVMA), which greatly reduces the latency at inference time for NMT. This architecture takes advantage of specialized hardware that has low-latency vector-vector operations and higher-latency vector-matrix operations. It also reduces the number of parameters and FLOPs for virtually all models that rely on efficient matrix multipliers without significantly impacting accuracy. We present empirical results suggesting that our framework can reduce the latency of sequence-to-sequence and Transformer models used for NMT by a factor of four. Finally, we show evidence suggesting that our VVMA extends to other domains, and we discuss novel hardware for its efficient use.

1 Introduction

Artificial neural networks have become increasingly popular over the last decade as they excel in tasks such as object detection and speech recognition (LeCun et al., 2015), which are becoming more commonplace with the use of self-driving cars and virtual assistants. The rapid development of deep neural networks has also made them the dominant approach for natural language processing (NLP) applications, ranging from neural machine translation (NMT) (Bahdanau et al., 2015; Klein et al., 2017; Wu et al., 2016) and text summarization (Rush et al., 2015; Nallapati et al., 2016; Liu et al., 2018) to virtual assistants such as Apple Siri, Amazon Alexa, and Google Home.

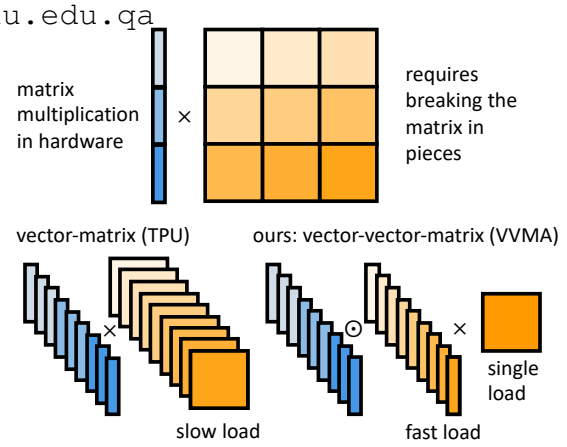

Figure 1: TPU vs. VVMA. Top: to multiply a vector by a matrix, the hardware tiles up the matrix. Bottom left: the TPU loads each piece. Bottom right: the VVMA loads a single piece (for broadcasting) and adds diagonals for element-wise multiplication, which is faster.

Unfortunately, neural networks are slow for training, inference and use due to their vast computational complexity. Several approaches have been proposed to address these issues including (*a*) quantization and pruning, (*b*) efficient models with less computational demand, and (*c*) specialized hardware accelerators (Sze et al., 2017). While direction (*a*) has been well-studied (LeCun et al., 1990; Han et al., 2016b,a; Guo, 2018; Quinn and Ballesteros, 2018), and can be considered complementary to (*b,c*), optimizing the combination of (*b*) and (*c*) has not been considered, to the best of our knowledge. Thus, here we propose a novel *vector-vector-matrix architecture* (VVMA) that compresses neural networks, while optimizing for hardware performance at inference time. Therefore, we optimize (*b*) and (*c*), without conflicting with (*a*), i.e., using quantization and pruning can potentially further boost the efficiency of our framework. Figure 1 illustrates this VVMA in contrast to a traditional vector-matrix architecture.

Moreover, the inherently sequential nature of many NLP tasks can increase the latency at inference time. Constrained by their memory bandwidth and footprint, modern accelerators rely on large batch sizes to avoid under-utilization. However, it is not always possible to increase the batch size if conclusions have to be inferred quickly, e.g., for real-time inference. For example, the matrix multiply unit of state-of-the-art accelerators, such as Google's Tensor Processing Unit (TPU), will "stall" when translating a single sentence, thus increasing the overall latency (Jouppi et al., 2017).

Our architecture can improve the TPU and other AI accelerators for *small-batch inference*. Thus, unlike other methods for compressing neural networks, the VVMA is designed to take advantage of the dataflow and the architecture of certain kinds of hardware accelerators such as the TPU.

Our contributions are as follows:

- We tailor an efficient model to state-of-the-art hardware accelerators.
- We provide an efficient vector-vector-matrix architecture (VVMA) framework for inference with small batch sizes.
- We use VVMAs to speed up inference in the computationally expensive Neural Machine Translation (NMT) task by a factor of four without losing much in terms of quality.
- We highlight promising applications of the VVMA in other deep learning domains and novel Artificial Intelligence (AI) accelerators.

The rest of this paper is organized as follows: In Section 2, we elaborate on directions (*b*) and (*c*), and we relate them to VVMAs. In Section 3, we motivate VVMAs as a faster improvement of the TPU's architecture and dataflow at inference time, and we then analyze our framework in its universality, including tips for efficient implementation of VVMAs. As a proof of concept, in Section 4 we demonstrate empirical inference speed-ups for NMT using Seq2seq-LSTM and Transformer models, which are both notorious for their computational complexity. We also show ablation studies and extensions to other tasks. In Section 5, we explore novel accelerators that can benefit from VVMAs. Finally, we offer concluding remarks in Section 6, and we point to possible directions for future work.

2 Background

Here, we look at efficient models from the software and the hardware side, and we discuss the advantages of merging them in a *co-design* manner. We further discuss the importance of wall-clock speed versus floating point operations and why from this perspective our weight sharing matrices will decrease inference rather than training time.

2.1 Efficient Models from the Software Side for Training and Inference

Efficient model architectures can decrease the complexity of neural networks. Some techniques to achieve this are described in (Chen et al., 2015; Zhang et al., 2018; Gao et al., 2018).

Zhang et al. (2018) added a new type of layer, a *channel shuffle layer*, to neural networks that use group convolution. By shuffling the data between layers, they reduced the number of parameters in the other layers while retaining similar accuracy. Gao et al. (2018) used a technique similar to group convolution, but applied it to recurrent neural networks. They used shuffling operations with a group recurrent neural network and showed improvements for NMT and text summarization.

Chen et al. (2015) compressed a weight matrix into a learned vector of weights. They used a hash function to map entries in the weight matrix to elements in the vector. As a result, many matrix entries share a single weight in the vector.

As Transformers are becoming the standard building block for NLP tasks, there is a growing effort to make them efficient, since their inference time scales as $O(N^2)$, where N is the number of input tokens. Child et al. (2019) proposed Sparse Transformers with $O(N\sqrt{N})$ complexity. Likewise, Sukhbaatar et al. (2019) developed Adaptive Attention Span and Kitaev et al. (2020) proposed Reformer using locality-sensitive hashing, and achieved $O(N \log N)$ complexity. See (Ganesh et al., 2020) for a broader overview.

In a similar fashion, our VVMA is an efficient model because it reduces the computational complexity at inference time without much decrease in performance. However, unlike the above models, VVMAs focus on the low levels of execution: the VVMA is an architecture that speeds up matrix multiplications. Thus, it is an efficient model that relates to hardware accelerators directly and it is universal, as matrix multiplication is the dominant computational factor for neural network inference.

2.2 Efficient Models from the Hardware Side

As we have mentioned above, successful NLP applications have been based on a variety of neural network models: recurrent and convolutional neural networks, memory-augmented networks, attention mechanism, Transformers, etc. These models were designed to solve numerous tasks ranging from language modeling and named entity recognition to NMT and other sequence modeling and sequence generation tasks. Most of the computation in such models is matrix multiplication both at inference and at training time, which is expensive. Therefore, specialized hardware accelerators for neural networks have been designed, focusing on making matrix multiplication efficient.

Note that the above techniques assume general hardware, i.e., they do not utilize the specific dataflow or architecture of an AI accelerator to improve efficiency. Yet, several such accelerators have been developed recently, e.g., the Horizon Robotics Brain Processing Unit, Graphcore Intelligence Processing Unit, NVIDIA Tensor Core, and Google Tensor Processing Unit (TPU).

A *matrix-matrix* architecture is a hardware unit that takes two matrices and multiplies them, e.g., NVIDIA Tensor Core. A *vector-matrix* architecture such as Google's TPU multiplies a vector and a matrix. As shown in Figure 1, the VVMA *vector-vector-matrix* architecture takes two vectors and a matrix, and it multiplies element-wise the first vector by the second vector, and then multiplies the resulting vector by the matrix.

Furthermore, VVMAs are optimized for certain AI accelerators, such as the TPU architecture. We specifically take advantage of the dataflow of the matrix multiply unit in the TPU, which is described in (Jouppi et al., 2017). This matrix multiply unit allows to re-use weights for multiple batches of data, while also using a systolic loop to perform matrix multiplication extremely fast. Therefore, we reduce the computational complexity of the "matrix" component in the TPU's vector-matrix unit, but we also maintain representational accuracy by inserting an extra "vector" part to get the vector-vector-matrix unit. By switching to this unit, we introduce a trade-off by increasing the efficiency of the model while decreasing its flexibility and generalization power. Likewise, we expect to have comparable accuracy to other compression techniques while also providing even faster performance at inference time.

2.3 Trade-Off between Flexibility and Efficiency at Inference Time

While every neural network requires a certain budget of floating point operations for a target computation, how fast such computations are in practice depends not on the size of this budget but rather on the number of wall clocks needed in order to cover all floating point operations. Thus, it is important to combine the software and the hardware advances in a co-design manner to optimize an efficient model for the correct metric: wall clocks.

Designed to optimize for the number of wall clocks, our VVMA introduces an extra vector component that maintains accuracy, but increases the computational complexity. We achieve this in part by optimizing our VVMA to specifically take advantage of the TPU architecture and dataflow. This creates a trade-off between flexibility and efficiency, e.g., the more we reuse weights, the more we have to compensate for the model accuracy.

Neural networks that are specifically designed to work in conjunction with certain AI accelerators will encounter a similar trade-off. That is, the more a neural network is tuned for efficiency, the less flexibility for change the model will have (Han et al., 2015). Nonetheless, we find regimes that suppress this trade-off and yield faster neural networks inference with VVMA. Thus, we believe that our VVMAs provide enough flexibility to be useful in a variety of existing neural architectures.

Training is the process of using (large) datasets to learn specific weights in neural networks. This process is usually very computationally expensive and can take days or months to complete. Once a neural network has finished training, the set of weights that were learned through the training process can remain fixed while making predictions. This process of using a fixed set of weights to make predictions with a neural network is called *inference* (Sze et al., 2017). Training can be done faster when parallelizing the process and increasing the amount of data fed into the network at a given time. This roughly translates to increasing the throughput of the training process. However, when performing inference on a single data point, the latency of making predictions seems to dominate the runtime (Jouppi et al., 2017). The VVMA we propose can be used specifically to decrease the latency of a neural network. Likewise, we expect this technique to be used to decrease inference time rather than to decrease training time.

3 Architecture

In this section, we present our approach to constructing a VVMA, including implementation details that are necessary to use VVMAs in practice.

3.1 Motivation

Google announced their first application-specific AI accelerator called the *Tensor Processing Unit* (TPU) in 2016. As described by Jouppi et al. (2017), the TPU uses a systolic loop to perform matrix multiplications (Jouppi et al., 2017), which are the most demanding computations in deep neural networks. Let W be an $n \times n$ weight matrix and x be an n-dimensional input vector. In order to perform Wx on the TPU, we must first break up W and x into $k \times k$ sections, where $k \times k$ is the size of the matrix multiply unit:

$$Wx = \begin{bmatrix} W_{1,1} & W_{1,2} & \cdots \\ W_{2,1} & W_{2,2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}. \quad (1)$$

Here, $W_{i,j}$ is a $k \times k$ block of W, and x_j is a k-dimensional block of x. Likewise, the TPU must load each block $W_{i,j}$ onto the matrix multiply unit before multiplying it by x_j. Loading a $k \times k$ block takes $O(k)$ clocks on the TPU. After loading a block $W_{i,j}$ onto the TPU, it takes $O(2k + t)$ clocks to multiply t k-dimensional vectors x_j by the matrix $W_{i,j}$. So, the total number of clocks to multiply t n-dimensional vectors x by W is

$$O\left(\frac{n^2}{k^2}(k + 2k + t)\right). \quad (2)$$

Note the large latency for single-batch inference, i.e., for $t = 1$. In order to decrease it, we tweak the weight matrix W, so that we only have to load a single $k \times k$ block M onto the matrix multiply unit. We then perform vector operations to each x_j in order to make up for the extra parameters that are lost by re-using the same $k \times k$ matrix M. Figure 2 shows an illustration of this process.

With this new procedure, the total number of clocks to multiply t n-dimensional vectors by the larger matrix is given by

$$O\left(k + 2k + \frac{n^2 t}{k^2}\right). \quad (3)$$

We can see that this new procedure significantly decreases the total number of clocks for single-batch inference with $t = 1$.

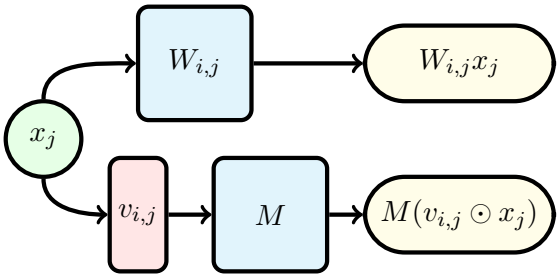

Figure 2: Illustration of how we can save time by sharing a weight matrix M. The top path shows the traditional dataflow, where each $W_{i,j}$ must be loaded onto the matrix multiply unit. The bottom path shows our approach, where M is loaded onto the matrix multiply unit only once. We then add a vector-vector operation $v_{i,j} \odot x_j$ before doing the matrix multiplication, where $\odot$ denotes element-wise multiplication.

3.2 Vector-Vector-Matrix Architecture

We construct the VVMA as follows. Let W be a large $n \times n$ weight matrix and let M be a smaller $k \times k$ weight matrix. First, we tile M into a larger matrix, so that its size is greater than or equal to the weight matrix W. Then, we multiply each copy of M by a unique diagonal matrix. Mathematically, we replace W with a structured matrix as shown below:

$$\begin{bmatrix} MD_{1,1} & MD_{1,2} & \cdots \\ MD_{2,1} & MD_{2,2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}, \quad (4)$$

where M is a shared $k \times k$ weight matrix and $D_{i,j}$ is a diagonal $k \times k$ weight matrix.

We use the diagonal matrices $D_{i,j}$ in order to introduce variation to each of the copies of M. We found that this is necessary for a VVMA to be able to effectively replace the original matrix W. Each of the entries in the matrix M is shared in multiple blocks of the new matrix, thus decreasing the total number of parameters compared to the original weight matrix W. Moreover, each of the entries of M as well as the entries in each diagonal matrix $D_{i,j}$ are learned as part of the training process.

Even though each entry $D_{i,j}$ is mathematically represented as a matrix in Equation 4, we can also see it as a k-dimensional vector $v_{i,j}$. We can then perform the matrix multiplication $D_{i,j}x$ as an element-wise multiplication $v_{i,j} \odot x_j$, as shown in Figure 2.

Task	Model	Architecture	# Params	BLEU	# Clocks	FLOPs
German-English	Seq2Seq-LSTM	Original	210.9M	22.42	322.1M	421.7M
		VVMA	115.4M	21.53	98.3M	230.9M
	Transformer	Original	61.4M	29.66	145.2M	122.6M
		VVMA	18.8M	23.32	42.2M	37.5M
English-German	Seq2Seq-LSTM	Original	210.9M	20.70	322.1M	421.7M
		VVMA	115.4M	18.90	98.3M	230.9M
	Transformer	Original	61.4M	24.57	145.2M	122.6M
		VVMA	18.8M	18.99	42.2M	37.5M
Vietnamese-English	Seq2Seq-LSTM	Original	32.3M	22.42	46.3M	64.6M
		VVMA	21.9M	20.86	21.9M	43.8M
English-Vietnamese	Seq2Seq-LSTM	Original	27.5M	25.34	34.8M	54.9M
		VVMA	17.1M	24.42	10.3M	34.1M

Table 1: Comparing the original Seq2seq-LSTM and Transformer models to such with VVMAs. Shown are the number of parameters, the BLEU score, and the estimated number of clock cycles and floating point operations.

3.3 Implementation Details

In order to implement equation 4 as a trainable matrix, we found that it was inefficient to actually construct the entire matrix representation. Instead, it was better to take advantage of broadcasting, which allows us to element-wise multiply tensors of different shapes. Likewise, we use broadcasting to multiply the input vector x by a larger diagonal tensor D. We then perform a matrix multiplication with the broadcasted vector and the matrix M. Thus, our program constructs a single $k \times k$ matrix M, and it does so only once rather than actually tiling it as shown in equation 4. We further found that a more aggressive gradient clipping was needed when training Seq2seq-LSTM models that use VVMAs; otherwise, the gradient grew extremely large and as a result eventually overflowed. We believe that this is because gradients accumulate as we propagate them back to a single small matrix M.

4 Results

In this section, we present empirical results showing that VVMAs can substitute different types of weight matrices in neural networks (NNs). Specifically, we use our VVMAs in Seq2seq-LSTM and Transformer NMT. We report some theoretical speedups that VVMAs provide when using a TPU-style architecture. We then present a small ablation study where we modify our VVMAs by removing the diagonal terms $D_{i,j}$ or by varying the value of k. We also compare VVMA to standard low-rank approximations. Finally, we show that our technique extends to language modelling with Transformer-XL, and beyond NLP tasks.

Unless otherwise noted, all results in this section use VVMAs with $k = 32$. That is, the matrix W in the neural network is replaced with a VVMA that uses a 32×32 matrix M along with 32×32 diagonal matrices $D_{i,j}$ as shown in equation 4.

4.1 Neural Machine Translation

We tested our VVMAs on NMT: we integrated them as part of Seq2seq-LSTM and Transformer models, as they are most commonly used today.

4.1.1 Sequence-to-Sequence Models

For the Seq2seq-LSTM models (Cho et al., 2014; Sutskever et al., 2014), we slightly modified the code by Luong et al. (2017), and we ran it on the two benchmarks provided in the repository. In particular, we used WMT datasets to train German-English and English-German models. We further used IWSLT datasets to train Vietnamese-English and English-Vietnamese models. We prepared the datasets according to the instructions found in the repository. For the German-English and English-German models, we used newstest2015 for testing.

Both models are Seq2seq models with LSTM layers and attention mechanism. We used four VVMAs for the LSTM cells: for the forget gate, for the input gate, for the output gate, and for the cell state vector. We also used VVMAs for the matrices in the attention mechanism.

For the Seq2seq-LSTM models, we decreased the gradient clipping value from 5 to 1 in order to prevent the gradient from overflowing. We also decreased the batch size to 32, to fit the models on a single GPU. We trained for 340,000 iterations for German-English and English-German, and 48,000 for Vietnamese-English and English-Vietnamese.

The results comparing the original models with models that use VVMAs are shown in Table 1. We can see that the BLEU scores decrease when using VVMAs, which should be expected given that the overall number of parameters in the model decreased noticeably. Overall, when taking into account the number of parameters, the observed decrease in the BLEU scores is very reasonable.

4.1.2 Transformer Models

For the Transformer models (Vaswani et al., 2017), we replaced the matrices in the feed-forward layers with VVMAs.[1] We trained these models on WMT datasets for German-English and English-German translation. We prepared the datasets according to the instructions found in the repository that we modified. We used the base Transformer models with a hidden size of 512 (rather than the big models, which have a hidden size of 1024). We trained these models with a batch size of 2048 for 6 epochs.

In Table 1, we present our results on the Transformer models with VVMAs. We achieved reasonable BLEU scores compared to the original Transformer. For German-English, the original model had 61.4M parameters and an uncased test BLEU score of 29.66. The VVMA model had 37M parameters and a BLEU score of 28.5. For English-German, the original model had 61.4M parameters and a BLEU score of 24.57. The VVMA model had 37M parameters and a BLEU score of 23.13. To recap, each matrix in these models was replaced by VVMAs except for the embedding and the projection matrices. We found that restricting these with the VVMA constraints had a sizable negative impact on performance.

4.2 Theoretical Speedups

We also calculated two measures for the inference time of the models described in Section 4.1: (*i*) the estimated number of clock cycles, and (*ii*) the number of floating point operations (FLOPs). Both roughly correspond to the real time needed to perform the inference at run time. We computed the former for a TPU-style architecture with one matrix multiply unit of size $k \times k$, and we estimated the latter for the original and the VVMA models using Equations 2 and 3 with $k = 32$, $t = 1$, and sequence lengths of 25. Note that the vector-vector operation before M takes zero extra clock cycles, as illustrated in Figures 1 and 2.

[1]We modified code from `github.com/tensorflow/models/tree/master/official/transformer`

This happens because we pipeline these vector-vector processes as we feed the data into the matrix multiply unit. Moreover, we initialize these operations while loading weights into the matrix multiply unit. We used a TensorFlow profiling tool in order to measure the number of FLOPs in our models. Looking at Table 1, we can see that the original Seq2seq-LSTM models require three to four times more clock cycles and roughly twice as many FLOPs compared to the VVMA models.

For the Transformer models with VVMAs, we saw less noticeable speed-ups. For similar accuracy, the estimated number of clock cycles and FLOPs were roughly 1.7 and 1.5 times more in the original Transformer models compared to models with VVMAs. This is expected since we use VVMAs only in the feed-forward layers. We tried to use VVMAs for the attention layers as well, but this led to larger decrease in accuracy, due to the significant reduction in the number of parameters.

As the Transformer is already getting noticeable impact in industrial settings, e.g., for machine translation and Web search, there is active research in developing more efficient Transformer architectures (Sanh et al., 2019; Kitaev et al., 2020; Beltagy et al., 2020; Zaheer et al., 2020). Thus, with each new version of a Transformer architecture, new VVMA experiments would be needed in order to measure the potential improvements in efficiency that VVMA would yield.

4.3 Ablation Study

Next, we performed an ablation study for the Seq2seq-LSTM models described in Section 4.1 for the English-Vietnamese machine translation task. In particular, we slightly modified the VVMAs in the Seq2seq-LSTM models by removing the diagonal terms $D_{i,j}$ or by changing the value of k.

Here, we trained with a batch size of 32 for 48,000 steps. In order to prevent the gradient from overflowing, we needed to multiply the shared matrix M by a scaling factor of 0.1 when removing the diagonal terms $D_{i,j}$. The results are shown in Table 2. We can see that removing the diagonal terms significantly decreases the BLEU scores for our models, while changing the value of k has no significant impact. Additionally, Figure 3 presents BLEU scores as the number of clock increases. We can see that compared to their original counterparts, VVMA models do not yield degradation in performance when then number of clocks gets large.

Architecture	k	Diags	# Params	BLEU	# Clocks	FLOPs
Original	N/A	N/A	27.5M	25.34	34.8M	54.9M
VVMA	32	T	17.1M	24.42	10.3M	34.1M
VVMA	32	F	16.7M	15.62	10.3M	33.8M
VVMA	16	T	17.4M	24.76	22.7M	34.8M
VVMA	64	T	16.9M	23.96	5.0M	33.9M

Table 2: Ablation study for English-Vietnamese NMT with Seq2seq-LSTM models. Here, k is the size of M in VVMAs, Diags shows whether diagonal terms are present (T=true, F=false), then follow the number of parameters, BLEU score, number of clocks and FLOPs. Original's clock is on a TPU with a block size of 32.

Architecture	k	# Params	# Clocks	PPL
Original	N/A	151.1M	99.4M	24.05
VVMA	32	138.2M	67.0M	30.70
VVMA	64	138.1M	35.9M	30.55
QRNN	N/A	151.0M	N/A	33.0

Table 3: Language modeling on WikiText-103 using Transformer-XL with and without VVMA, as well as using QRNN. (Original: TPU with a block size of 32.)

n / Fit ($\times 10^3$)	VVMA	Low-rank	Optimal
1,024	3.0	2.9	2.9
2,048	6.1	5.9	5.8
4,096	12.2	11.9	11.7

Table 4: VVMA's closeness of fit to a target matrix is comparable to that of (*i*) standard low-rank approximation and (*ii*) optimal approximation, but it is orders of magnitude faster at inference time.

4.4 Comparison to Standard Low-Rank Approximation

First, note that the rank of VVMA is maximum k for a $k \times k$ sharing matrix. To prove that, we can represent the matrix in equation 4 as a product of matrices of maximal rank k. Then, we can use the property that $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$.

Second, we compare to low-rank approximation. We fix $k = 128$ and we choose n =1,024; 2,048; 4,096. We sample a random matrix and we fit VVMA parametrization to it using Adam (Kingma and Ba, 2015) with a learning rate of 0.0001 ran for $30,000$ steps, and using the Frobenius norm as a loss. We do the same experiment with $UV^\top$ low-rank rank-p approximation, where p is chosen to match the number of parameters in VVMA. Additionally, we use Eckart–Young–Mirsky's theorem to get the Optimal low-rank fit. Table 4 shows some Frobenius norm losses from these experiments. We can see that VVMA's expressiveness is comparable to standard low-rank approximation; note, however, that standard low-rank approximation does not yield the inference speedups of VVMA.

4.5 Extension to Language Modelling

Even though the main focus of this paper is the contribution of VVMA to neural machine translation, we also demonstrate that VVMA is compatible to state-of-the-art language modelling architectures. For that purpose, we perform an experiment on WikiText-103 (Merity et al., 2017) using the Transformer-XL model (Dai et al., 2019).

In this experiment, we directly integrate VVMA into the Transformer-XL architecture, keeping all hyper-parameter values as in the original Transformer-XL paper (Dai et al., 2019), except for reducing the batch size to 30, in order to fit the optimization on two GPUs. We chose to replace the weights of the attention mechanism with VVMA. Replacing the weights of the positional feed-forward layers drastically decreases the number of parameters, which yields poor performance, as Transformer-XL's perplexity is sensitive to the number of parameters. We present our results in Table 3, where we can see that VVMA with a block size of 64 yields reasonable performance, and the perplexity decreases noticeably with the reduction of parameters.

4.6 Extension to Other Areas

We further extended our VVMAs beyond NLP, to image classification. We modified the convolutional filters in ResNet (He et al., 2016) to use VVMAs and we trained on CIFAR-10.[2] We prepared the CIFAR-10 dataset following the instructions in the repository we modified. We trained all ResNet models with a batch size of 128 for 250 epochs. Figure 3 bottom shows the accuracy of the ResNet models as a function of the number of parameters. We can see that the ResNet models with VVMAs outperform the original ResNet models when keeping the number of parameters fixed.

[2]We modified code from `github.com/tensorflow/models/tree/master/official/resnet`

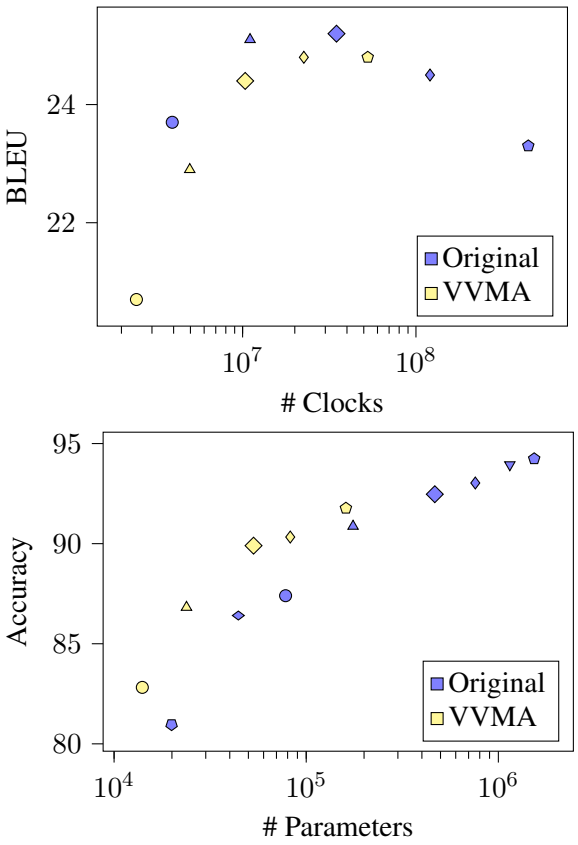

Figure 3: BLEU scores and validation accuracy as a function of the number of trainable parameters in the original and in the VVMA Seq2seq-LSTM models for English-Vietnamese (top) and ResNet (He et al., 2016) models on CIFAR-10 (bottom). The number of parameters is varied by changing the depth and the size of the hidden state. Unique shapes with different colors refer to the same Seq2seq-LSTM model, with the original model in blue and the VVMA model in yellow.

5 Discussion

Below, we discuss new AI hardware that could optimize inference for neural networks via VVMAs. This hardware would decrease latency at inference time rather than decreasing the training time.

Tensor Processing Unit. As mentioned above, Google's Tensor Processing Units (TPU) has a dedicated matrix multiply unit (Jouppi et al., 2017). We believe that a modified version of the TPU could take advantage of VVMAs. The necessary modifications would be relatively simple. As illustrated in Figure 2, we would add a dedicated vector-vector unit before the matrix multiply unit, and we would pipeline it and initialize it at the same time as the matrix multiply unit. As seen in Section 4.2, this would noticeably decrease the number of inference clock cycles in Seq2seq-LSTM models.

Tensor Cores. NVIDIA's newest GPUs have dedicated matrix multiply units called *Tensor Cores*, which can perform 4×4 matrix multiplications in a single clock cycle (Markidis et al., 2018). Adding vector-vector units before each Tensor Core would make them more efficient for VVMAs. The largest speedup would come from the time spent loading matrices from the memory into the Tensor Cores. For instance, if multiple Tensor Cores share the same matrix elements, this would decrease the latency when performing inference.

Optical Processing Unit. A newer, more experimental architecture, is to use VVMAs with optical computing. Shen et al. (2017) proposed to use an Optical Processing Unit (OPU) to perform matrix multiplications at the speed of light. Likewise, it is possible to use an OPU in order to accelerate inference on a neural network. Note, however, that the OPU would run into some of the same problems that the TPU has. That is, there will be a large delay when loading the matrix weights from the memory onto the OPU. Thus, we propose to add an electronic vector-vector unit before the OPU, which would be pipelined and initialized as weights are loaded onto the OPU. This extra unit will not increase the overall latency of a system that uses an OPU because the input vectors will still need to be fetched from the digital electronic memory. Likewise, performing vector-vector operations with the input data will not significantly increase the latency of the entire system.

6 Conclusion and Future Work

We have proposed a novel vector-vector-matrix architecture for low-latency inference, and we have demonstrated theoretical and empirical speed-ups for Seq2seq-LSTM and Transformer models, with application to neural machine translation, language modeling, and image classification. We hope that this work would bring the novel concept of AI *co-design* (between software and hardware) to the domain of NLP applications.

In future work, we plan to optimize the low-level code and to develop new hardware to deploy VVMAs in real-world applications. Distilling models to their VVMA counterparts would be an interesting experiment, and potentially an orthogonal enhancement to pre-existing frameworks (Sanh et al., 2019). VVMAs could also be an orthogonal contribution to other factorizations of NLP models, such as in (Lan et al., 2020).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15, San Diego, CA, US.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2007.14062*.

Wenlin Chen, James T. Wilson, Stephen Tyree, Kilian Q. Weinberger, and Yixin Chen. 2015. Compressing neural networks with the hashing trick. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, ICML '15, pages 2285–2294, Lille, France.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Kyunghyun Cho, Bart van Merriënboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP '14, pages 1724–1734, Doha, Qatar.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, EMNLP '19, Florence, Italy.

Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Deming Chen, Marianne Winslett, Hassan Sajjad, and Preslav Nakov. 2020. Compressing large-scale transformer-based models: A case study on BERT. *arXiv preprint arXiv:2002.11985*.

Fei Gao, Lijun Wu, Lu Zhao, Tao Qin, Xueqi Cheng, and Tie-Yan Liu. 2018. Efficient sequence learning with group recurrent networks. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 799–808, New Orleans, LA, US.

Yunhui Guo. 2018. A survey on methods and theories of quantized neural networks. *arXiv preprint arXiv:1808.04752*.

Song Han, Junlong Kang, Huizi Mao, Yiming Hu, Xin Li, Yubin Li, Dongliang Xie, Hong Luo, Song Yao, Yu Wang, Huazhong Yang, and William J. Dally. 2016a. ESE: efficient speech recognition engine with compressed LSTM on FPGA. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, FPGA '17, pages 75–84, Monterey, CA, US.

Song Han, Huizi Mao, and William J. Dally. 2016b. Deep compression: Compressing deep neural network with pruning, trained quantization and Huffman coding. In *Proceedings of the 4th International Conference on Learning Representations*, ICLR '16, San Juan, Puerto Rico.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems 28*, NIPS '15, pages 1135–1143, Montréal, Canada.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, Las Vegas, NV, US.

Norman P. Jouppi, Cliff Young, Nishant Patil, David A. Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, Richard C. Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon. 2017. In-datacenter performance analysis of a tensor processing unit. In *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture*, ISCA '17, pages 1–12, Toronto, ON, Canada.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, ICLR '15, San Diego, CA, US.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *Proceedings of the 8th International Conference on Learning Representations*, ICLR '20, Addis Ababa, Ethiopia.

Guillaume Klein, Yoon Kim, Yuntian Deng, Josep Maria Crego, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for*

Computational Linguistics-System Demonstrations, ACL '17, pages 67–72, Vancouver, Canada.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the 8th International Conference on Learning Representations*, ICLR '20, Addis Ababa, Ethiopia.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. *Nature*, 521:436–444.

Yann LeCun, John S. Denker, and Sara A. Solla. 1990. Optimal brain damage. In *Advances in Neural Information Processing Systems 2*, NIPS '89, pages 598–605, Denver, CO, US.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Łukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by summarizing long sequences. In *Proceedings of the 6th International Conference on Learning Representations*, ICLR '18, Vancouver, Canada.

Minh-Thang Luong, Eugene Brevdo, and Rui Zhao. 2017. Neural machine translation (seq2seq) tutorial. *https://github.com/tensorflow/nmt*.

Stefano Markidis, Steven Wei Der Chien, Erwin Laure, Ivy Bo Peng, and Jeffrey S. Vetter. 2018. NVIDIA tensor core programmability, performance & precision. In *Proceedinngs of the 2018 IEEE International Parallel and Distributed Processing Symposium Workshops*, IPDPSW '18, pages 522–531.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, Toulon, France.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, CoNLL '16, pages 280–290.

Jerry Quinn and Miguel Ballesteros. 2018. Pieces of eight: 8-bit neural machine translation. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '18, pages 114–120, New Orleans, LA, US.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, EMNLP '15, pages 379–389, Lisbon, Portugal.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Yichen Shen, Nicholas C. Harris, Scott Skirlo, Mihika Prabhu, Tom Baehr-Jones, Michael Hochberg, Xin Sun, Shijie Zhao, Hugo Larochelle, Dirk Englund, and Marin Soljačić. 2017. Deep learning with coherent nanophotonic circuits. *Nature Photonics*, 11:441–446.

Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. 2019. Adaptive attention span in transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, ACL '19, pages 331–335, Florence, Italy.

Ilya Sutskever, Oriol Vinayals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems 27*, NIPS '14, pages 3104–3112, Montréal, Canada.

Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. 2017. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of IEEE*, 105(12):2295–2329.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems: Advances in Neural Information Processing Systems*, NIPS '17, pages 5998–6008, Long Beach, CA, US.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, C. Alberti, S. Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and A. Ahmed. 2020. Big Bird: Transformers for longer sequences. *arXiv preprint arXiv:2007.14062*.

Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '18, pages 6848–6856, Salt Lake City, UT, US.