EVA: An Encrypted Vector Arithmetic Language and Compiler for Efficient Homomorphic Computation

Roshan Dathathri
University of Texas at Austin, USA
roshan@cs.utexas.edu

Blagovesta Kostova
EPFL, Switzerland
blagovesta.pirelli@epfl.ch

Wei Dai
Microsoft Research, USA
wei.dai@microsoft.com

Kim Laine
Microsoft Research, USA
kilai@microsoft.com

Olli Saarikivi
Microsoft Research, USA
olsaarik@microsoft.com

Madan Musuvathi
Microsoft Research, USA
madanm@microsoft.com

Abstract

Fully-Homomorphic Encryption (FHE) offers powerful capabilities by enabling secure offloading of both storage and computation, and recent innovations in schemes and implementations have made it all the more attractive. At the same time, FHE is notoriously hard to use with a very constrained programming model, a very unusual performance profile, and many cryptographic constraints. Existing compilers for FHE either target simpler but less efficient FHE schemes or only support specific domains where they can rely on expert-provided high-level runtimes to hide complications.

This paper presents a new FHE language called Encrypted Vector Arithmetic (EVA), which includes an optimizing compiler that generates correct and secure FHE programs, while hiding all the complexities of the target FHE scheme. Bstered by our optimizing compiler, programmers can develop efficient general-purpose FHE applications directly in EVA. For example, we have developed image processing applications using EVA, with a very few lines of code.

EVA is designed to also work as an intermediate representation that can be a target for compiling higher-level domain-specific languages. To demonstrate this, we have re-targeted CHET, an existing domain-specific compiler for neural network inference, onto EVA. Due to the novel optimizations in EVA, its programs are on average 5.3× faster than those generated by CHET. We believe that EVA would enable a wider adoption of FHE by making it easier to develop FHE applications and domain-specific FHE compilers.

CCS Concepts: • Software and its engineering → Compilers; • Security and privacy → Software and application

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

PLDI ’20, June 15–20, 2020, London, UK
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7613-6/20/06...
$15.00
https://doi.org/10.1145/3385412.3386023

1 Introduction

Fully-Homomorphic Encryption (FHE) allows arbitrary computations on encrypted data without requiring the decryption key. Thus, FHE enables interesting privacy-preserving capabilities, such as offloading secure storage and secure computation to untrusted cloud providers. Recent advances in FHE theory [12, 13] along with improved implementations have pushed FHE into the realm of practicality. For instance, with appropriate optimization, we can perform encrypted fixed-point multiplications within a few microseconds, which matches the speed of 8086 processors that jumped the computing revolution. Future cryptographic innovations will further reduce the performance gap between encrypted and unencrypted computations.

Despite the availability of multiple open-source implementations [25, 28, 38, 41], programming FHE applications remains hard and requires cryptographic expertise, making it inaccessible to most programmers today. Furthermore, different FHE schemes provide subtly different functionalities and require manually setting encryption parameters that control correctness, performance, and security. We expect the programming complexity to only increase as future FHE schemes become more capable and performant. For instance, the recently invented CKKS scheme [13] supports fixed-point arithmetic operations by representing real numbers as integers with a fixed scaling factor, but requires the programmer to perform rescaling operations so that scaling factors and the cryptographic noise do not grow exponentially due...
to multiplications. Moreover, the so-called RNS-variant of the CKKS scheme [12] provides efficient implementations that can use machine-sized integer operations as opposed to multi-precision libraries, but imposes further restrictions on the circuits that can be evaluated on encrypted data.

To improve the developer friendliness of FHE, this paper proposes a new general-purpose language for FHE computation called Encrypted Vector Arithmetic (EVA). EVA is also designed to be an intermediate representation that is a backend for other domain-specific compilers. At its core, EVA supports arithmetic on fixed-width vectors and scalars. The vector instructions naturally match the encrypted SIMD — or batching — capabilities of FHE schemes today. EVA includes an optimizing compiler that hides all the complexities of the target FHE scheme, such as encryption parameters and noise. It ensures that the generated FHE program is correct, performant, and secure. In particular, it eliminates all common runtime exceptions that arise when using FHE libraries.

EVA implements FHE-specific optimizations, such as optimally inserting operations like rescaling and modulus switching. We have built a compiler incorporating all these optimizations to generate efficient programs that run using the Microsoft SEAL [41] FHE library which implements the RNS-variant of the CKKS scheme. We have built an EVA executor that transparently parallelizes the generated program efficiently, allowing programs to scale well. The executor also automatically reuses the memory used for encrypted messages, thereby reducing the memory consumed.

To demonstrate EVA’s usability, we have built a Python frontend for it. Using this frontend, we have implemented several applications in EVA with a very few lines of code and much less complexity than in SEAL directly. One application computes the length of a path in 3-dimensional space, which can be used in secure fitness mobile applications. We have implemented some statistical machine learning applications. We have also implemented two image processing applications, Sobel filter detection and Harris corner detection. We believe Harris corner detection is one of the most complex programs that have been evaluated using CKKS.

In addition, we have built a domain-specific compiler on top of EVA for deep neural network (DNN) inference. This compiler takes programs written in a higher-level language as input and generates EVA programs using a library of operations on higher-level constructs like tensors and images. In particular, our DNN compiler subsumes the recently proposed domain-specific compiler called CHET [18]. Our DNN compiler uses the same tensor kernels as CHET, except that it generates EVA programs instead of generating SEAL programs. Nevertheless, the optimizing compiler in EVA is able to outperform CHET in DNN inference by 5.3× on average.

In summary, EVA is a general-purpose language and an intermediate representation that improves the programmability of FHE applications by guaranteeing correctness and security, while outperforming current methods.

The rest of this paper is organized as follows. Section 2 gives background on FHE. Section 3 presents the EVA language. Section 4 gives an overview of the EVA compiler. We then describe transformations and analysis in the compiler in Sections 5 and 6 respectively. Section 7 briefly describes the domain-specific compilers we built on top of EVA. Our evaluation is presented in Section 8. Finally, related work and conclusions are presented in Sections 9 and 10 respectively.

2 Background and Motivation

In this section, we describe FHE (Section 2.1) and the challenges in using it (Section 2.2). We also describe an implementation of FHE (Section 2.3). Finally, we present the threat model assumed in this paper (Section 2.4).

2.1 Fully-Homomorphic Encryption (FHE)

An FHE scheme includes four stages: key generation, encryption, evaluation, and decryption. Most of the efficient FHE schemes, for example, BGV [6], BFV [19], and CKKS [13], are constructed on the Ring Learning with Errors (RLWE) problem [34]. At the time of key generation, a polynomial ring of degree N with integer coefficients modulo Q must be chosen to represent ciphertexts and public keys according to the security standard [1]. We call Q the ciphertext modulus. A message is encoded to a polynomial, and subsequently encrypted with a public key or a secret key to form a ciphertext consisting of two polynomials of degree up to N − 1. Encryption also adds to a ciphertext a small random error that is later removable in decryption.

FHE schemes are malleable by design. From the perspective of the user, they offer a way to encrypt integers (or fixed-point numbers in CKKS — see the next section) such that certain arithmetic operations can be evaluated on the resulting ciphertexts. Evaluation primarily includes four operations: addition of ciphertexts, addition of a ciphertext and a plaintext, multiplication of ciphertexts, and multiplication of a ciphertext and a plaintext. Decrypting (with a secret key) and decoding reveals the message, as if the computation was performed on unencrypted data.

Many modern FHE schemes also include a SIMD-like feature known as batching which allows a vector of values to be encrypted as a single ciphertext (N/2 values in CKKS). With batching, arithmetic operations happen in an element-wise fashion. Batching-compatible schemes can evaluate rotations which allow data movement inside a ciphertext. But evaluating each rotation step count needs a distinct public key.

2.2 Challenges in Using FHE

Programmers using FHE face significant challenges that must be overcome for correct, efficient, and secure computation. We discuss those challenges here to motivate our work.

Depth of Computation: Computations on ciphertexts increase the initially small error in them linearly on the
number of homomorphic additions and exponentially on the
multiplicative depth of the evaluation circuit. When the er-
ors get too large, ciphertexts become corrupted and cannot
be decrypted, even with the correct secret key. This bound is
in turn determined by the size of the encryption parameter
$Q$. Thus, to support efficient homomorphic evaluation of a
circuit, one must optimize the circuit for low depth.

**Relinearization**: Each ciphertext consists of 2 or more
polynomials (freshly encrypted ciphertexts consist of only 2
polynomials). Multiplication of two ciphertexts with $k$ and $l$
polynomials yields a ciphertext with $k + l + 1$ polynomials.
To prevent the number of polynomials from growing indef-
inently, an operation called relinearization is performed to
reduce it back to 2. Relinearization is costly and their optimal
placement is an NP-hard problem [10].

**CKKS and Approximate Fixed-Point**: The CKKS [13]
scheme introduced an additional challenge by only providing
approximate results (but much higher performance in return).
There are two main sources of error in CKKS: (i) error from
the encoding of values to polynomials being lossy, and (ii) the
noise added in every homomorphic operation being mixed
with the message. To counter this, CKKS adopts a fixed-
point representation by associating each ciphertext with
an unencrypted scaling factor. Using high enough scaling
factors allows the errors to be hidden.

CKKS further features an operation called *rescaling* that
scales down the fixed-point representation of a ciphertext.
Consider a ciphertext $x$ that contains the encoding of 0.25
multiplied by the scale $2^{10}$ (a relatively low scale). $x^2$ encodes
0.0625 multiplied by the scale $2^{20}$. Further powers would
rapidly overflow modest values of the modulus $Q$, requiring
impractically large encryption parameters to be selected.
Rescaling the second power by $2^{10}$ will truncate the fixed-
point representation to encode the value at a scale of $2^{20}$.

Rescaling has a secondary effect of also dividing the ci-
pertext’s modulus $Q$ by the same divisor as the ciphertext
itself. This means that there is a limited “budget” for rescal-
ing built into the initial value of $Q$. The combined effect for
CKKS is that $\log Q$ can grow linearly with the multiplicative
depth of the circuit. It is common to talk about the level of a
ciphertext as how much $Q$ is left for rescaling.

A further complication arises from the ciphertext after
rescaling being encrypted under fundamentally different
encryption parameters. To apply any binary homomorphic
operations, two ciphertexts must be at the same level, i.e.,
have the same $Q$. Furthermore, addition and subtraction
require ciphertexts to be encoded at the same scale due to
the properties of fixed-point arithmetic. CKKS also supports
a modulus switching operation to bring down the level of a
ciphertext without scaling the message. In our experience,
inserting the appropriate rescaling and modulus switching
operations to match levels and scales is a significantly difficult
process even for experts in homomorphic encryption.

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cipher</td>
<td>An encrypted vector of fixed-point values.</td>
</tr>
<tr>
<td>Vector</td>
<td>A vector of 64-bit floating point values.</td>
</tr>
<tr>
<td>Scalar</td>
<td>A 64-bit floating point value.</td>
</tr>
<tr>
<td>Integer</td>
<td>A 32-bit signed integer.</td>
</tr>
</tbody>
</table>

In the most efficient implementations of CKKS (so called
RNS-variants [11]), the truncation is actually performed by
dividing the encrypted values by prime factors of $Q$. Further-
more, there is a fixed order to these prime factors, which
means that from a given level (i.e., how many prime factors
are left in $Q$) there is only one valid divisor available for
rescaling. This complicates selecting points to rescale, as do-
ing so too early might make the fixed-point representation
so small that the approximation errors destroy the message.

**Encryption Parameters**: In CKKS, all of the concerns
about scaling factors, rescaling, and managing levels are in-
tricately linked with selecting encryption parameters. Thus,
a typical workflow when developing FHE applications in-
volves a lot of trial-and-error, and repeatedly tweaking the
parameters to achieve both correctness (accuracy) and per-
formance. While some FHE libraries warn the user if the
selected encryption parameters are secure, but not all of
them do, so a developer may need to keep in mind security-
related limitations, which typically means upper-bounding
$Q$ for a given $N$.

### 2.3 Microsoft SEAL

Microsoft SEAL [41] is a software library that implements
the RNS variant of the CKKS scheme. In SEAL, the modulus
$Q$ is a product of several prime factors of bit sizes up to 60
bits, and rescaling of ciphertexts is always done by dividing
away these prime factors. The developer must choose these
prime factors and order them correctly to achieve the desired
rescaling behavior. SEAL automatically validates encryption
parameters for correctness and security.

### 2.4 Threat Model

We assume a semi-honest threat model, as is typical for ho-
omorphic encryption. This means that the party perfor-
ing the computation (i.e., the server) is curious about the
encrypted data but is guaranteed to run the desired opera-
tions faithfully. This model matches for example the scenario
where the server is trusted, but a malicious party has read
access to the server’s internal state and/or communication
between the server and the client.

### 3 EVA Language

The EVA framework uses a single language as its input for-
mat, intermediate representation, and executable format. The

---

**Table 1. Types of values.**

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cipher</td>
<td>An encrypted vector of fixed-point values.</td>
</tr>
<tr>
<td>Vector</td>
<td>A vector of 64-bit floating point values.</td>
</tr>
<tr>
<td>Scalar</td>
<td>A 64-bit floating point value.</td>
</tr>
<tr>
<td>Integer</td>
<td>A 32-bit signed integer.</td>
</tr>
</tbody>
</table>
Table 2. Instruction opcodes and their semantics (see Section 2 for details on semantics of the restricted instructions).

<table>
<thead>
<tr>
<th>Opcode</th>
<th>Signature</th>
<th>Description</th>
<th>Restrictions</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEGATE</td>
<td>Cipher → Cipher</td>
<td>Negate each element of the argument.</td>
<td></td>
</tr>
<tr>
<td>ADD</td>
<td>Cipher × (Vector</td>
<td>Cipher) → Cipher</td>
<td>Add arguments element-wise.</td>
</tr>
<tr>
<td>SUB</td>
<td>Cipher × (Vector</td>
<td>Cipher) → Cipher</td>
<td>Subtract right argument from left one element-wise.</td>
</tr>
<tr>
<td>MULTIPLY</td>
<td>Cipher × (Vector</td>
<td>Cipher) → Cipher</td>
<td>Multiply arguments element-wise (and multiply scales).</td>
</tr>
<tr>
<td>ROTATELEFT</td>
<td>Cipher × Integer → Cipher</td>
<td>Rotate elements to the left by given number of indices.</td>
<td></td>
</tr>
<tr>
<td>ROTATERIGHT</td>
<td>Cipher × Integer → Cipher</td>
<td>Rotate elements to the right by given number of indices.</td>
<td></td>
</tr>
<tr>
<td>RELINEARIZE</td>
<td>Cipher → Cipher</td>
<td>Apply relinearization.</td>
<td>Not in input</td>
</tr>
<tr>
<td>MODSWITCH</td>
<td>Cipher → Cipher</td>
<td>Switch to the next modulus in the modulus chain.</td>
<td>Not in input</td>
</tr>
<tr>
<td>RESCALE</td>
<td>Cipher × Scalar → Cipher</td>
<td>Rescale the ciphertext (and divide scale) with the scalar.</td>
<td>Not in input</td>
</tr>
</tbody>
</table>

EVA language abstracts batching-compatible FHE schemes like BFV [19], BGV [6], and CKKS [12, 13], and can be compiled to target libraries implementing those schemes. Input programs use a subset of the language that omits details specific to FHE, such as when to rescale. In this section, we describe the input language and its semantics, while Section 4 presents an overview of the compilation to an executable EVA program.

Table 1 lists the types that values in EVA programs may have. The vector types Cipher and Vector have a fixed power-of-two size for each input program. The power-of-two requirement comes from the target encryption schemes.

We introduce some notation for talking about types and values in EVA. For Vector, a literal value with elements \( a_i \) is written \([a_1, a_2, \ldots, a_l] \) or as a comprehension \([a_i \mid i = 1 \ldots l] \). For the \( i \)th element of Vector \( a \), we write \( a_i \). For the product type (i.e., tuple) of two EVA types \( A \) and \( B \), we write \( A \times B \), and write tuple literals as \((a, b) \) where \( a \in A \) and \( b \in B \).

Programs in EVA are Directed Acyclic Graphs (DAGs), where each node represents a value available during execution. Example programs are shown in Figures 1(a) and 2(a). Nodes with one or more incoming edges are called instructions, which compute a new value as a function of its parameter nodes, i.e., the parent nodes connected to it. For the \( i \)th parameter of an instruction \( n \), we write \( n.parm_i \) and the whole list of parameter nodes is \( n.parms \). Each instruction \( n \) has an opcode \( n.op \), which specifies the operation to be performed at the node. Note that the incoming edges are ordered, as it corresponds to the list of arguments. Table 2 lists all the opcodes available in EVA. The first group are opcodes that frontends may generate, while the second group lists FHE-specific opcodes that are inserted by the compiler. The key to the expressiveness of the input language are the RotateLeft and RotateRight instructions, which abstract rotation (circular shift) in batching-compatible FHE schemes.

A node with no incoming edges is called a constant if its value is available at compile time and an input if its value is only available at run time. For a constant \( n \), we write \( n.value \) to denote the value. Inputs may be of any type, while constants can be any type except Cipher. This difference is due to the fact that the Cipher type is not fully defined before key generation time, and thus cannot have any values at compile time. The type is accessible as \( n.type \).

A program \( P \) is a tuple \((M, \text{Insts}, \text{Consts}, \text{Inputs}, \text{Outputs})\), where \( M \) is the length of all vector types in \( P \); Insts, Consts and Inputs are list of all instruction, constant, and input nodes, respectively; and Outputs identifies a list of instruction nodes as outputs of the program.

Next, we define execution semantics for EVA. Consider a dummy encryption scheme \( id \) that instead of encrypting Cipher values just stores them as Vector values. In other words, the encryption and decryption are the identity function. This scheme makes homomorphic computation very easy, as every plaintext operation is its own homomorphic counterpart. Given a map \( I : \text{Inputs} \to \text{Vector} \), let \( E_{id}(n) \) be the function that computes the value for node \( n \) recursively by using \( n.value \) or \( I(n) \) if \( n \) is a constant or input respectively and using \( n.op \) and \( E_{id}(\cdot) \) on \( n.parms \) otherwise. Now for a program \( P \), we further define its reference semantic as a function \( P_{id} \), which given a value for each input node maps each output node in \( P \) to its resulting value:

\[
P_{id} : \times_{n_i \in \text{Inputs}} n_i.type \to \times_{n_o \in \text{Outputs}} \text{Vector}
\]

\[
P_{id}(I(n_1^1), \ldots, I(n_j^{\text{Inputs}})) = (E_{id}(n_1^1), \ldots, E_{id}(n_o^{\text{Outputs}}))
\]

These execution semantics hold for any encryption scheme, except that output is also encrypted (i.e., Cipher type).

The EVA language has a serialized format defined using Protocol Buffers [23], a language and platform neutral data serialization format. Additionally, the EVA language has an in-memory graph representation that is designed for efficient analysis and transformation, which is discussed in Section 4.

4 Overview of EVA Compiler

In this section, we describe how to use the EVA compiler (Section 4.1). We then describe the constraints on the code
generated by EVA (Section 4.2). Finally, we give an overview of the execution flow of the EVA compiler (Section 4.3).

4.1 Using the Compiler

In this paper, we present the EVA compiler for the RNS variant of the CKKS scheme [12] and its implementation in the SEAL library [41]. Targeting EVA for other FHE libraries [25, 28, 38] implementing CKKS [12, 13] would be straightforward. The EVA compiler can also be adapted to support other batching-compatible FHE schemes like BFV [19] and BGV [6].

The EVA compiler takes a program in the EVA language as input. Along with the program, it needs the fixed-point scales or precisions for each input in the program and the desired fixed-point scales or precisions for each output in the program. The compiler then generates a program in the EVA language as output. In addition, it generates a vector of bit sizes that must be used to generate the encryption parameters as well as a set of rotation steps that must be used to generate the rotation keys. The encryption parameters and the rotations keys thus generated are required to execute the generated EVA program.

While the input and the output programs are in the EVA language, the set of instructions allowed in the input and the output are distinct, as listed in Table 2. The Relinearize, Rescale, and ModSwitch instructions require understanding the intricate details of the FHE scheme. Hence, they are omitted from the input program.

The input scales and the desired output scales affect the encryption parameters, and consequently, the performance and accuracy of the generated program. Choosing the right values for these is a trade-off between performance and accuracy (while providing the same security). Larger values lead to larger encryption parameters and more accurate but slower generated program, whereas smaller values lead to smaller encryption parameters and less accurate but faster generated program. Profiling techniques like those used in prior work [18] can be used to select the appropriate values.

4.2 Motivation and Constraints

There is a one-to-one mapping between instructions in the EVA language (Table 2) and instructions in the RNS-CKKS scheme. However, the input program cannot be directly executed. Firstly, encryption parameters are required to ensure that the program would be accurate. EVA can simply determine the bit sizes that is required to generate the parameters. However, this is insufficient to execute the program correctly because some instructions in the RNS-CKKS scheme have restrictions on their inputs. If these restrictions are not met, the instructions would just throw an exception at runtime.

Each ciphertext in the RNS-CKKS scheme has a coefficient modulus \( q \) \((Q = \prod_{i=1}^{r} q_i)\) and a fixed-point scale associated with it. All freshly encrypted ciphertexts have the same \( q \) but they may have different scale. The following constraints apply for the binary instructions involving two ciphertexts:

\[
\begin{align*}
    n.\text{parm}_1.\text{modulus} &= n.\text{parm}_2.\text{modulus} \\
    \text{if } n.\text{op} \in \{\text{Add, Sub, Multiply}\} \\
    n.\text{parm}_1.\text{scale} &= n.\text{parm}_2.\text{scale} \\
    \text{if } n.\text{op} \in \{\text{Add, Sub}\}
\end{align*}
\]

In the rest of this paper, whenever we mention Add regarding constraints, it includes both Add and Sub.

1In SEAL, if the coefficient modulus \( q \) is \( \{q_1, q_2, \ldots, q_r\} \), then \( q_i \) is a prime close to a power-of-2. EVA compiler (and the rest of this paper) assumes \( q_i \) is the corresponding power-of-2 instead. To resolve this discrepancy, when a Rescale instruction divides the scale by the prime, the scale is adjusted (by the EVA executor) as if it was divided by the power-of-2 instead.
Consider the example to compute $x^2y^3$ for ciphertexts $x$ and $y$ in Figure 1 (viewed as a dataflow graph). Constraint 2 is trivially satisfied because they are no Add instructions. Only Rescale and ModSwitch instructions modify $q$. Constraint 1 is also trivially satisfied due to the absence of these instructions. Nonetheless, without the use of Rescale instructions, the scales and the noise of the ciphertexts would grow exponentially with the multiplicative depth of the program (i.e., maximum number of Multiply nodes in any path) and consequently, the $\log_q$ of the coefficient modulus product $Q$ required for the input would grow exponentially. Instead, using Rescale instructions ensures that $\log_q Q$ would only grow linearly with the multiplicative depth of the program.

In Figure 1, the output has a scale of $2^{60} \cdot 2^{60} \cdot 2^{30} \cdot 2^{30} \cdot 2^{30}$ (with $x.scale = 2^{60}$ and $y.scale = 2^{30}$). This would require $Q$ to be at least $2^{210} \cdot s_r$, where $s_r$ is the user-provided desired output scale. To try to reduce this, one can insert Rescale after every Multiply, as shown in Figure 1(b). However, this yields an invalid program because it violates Constraint 1 for the last (bottom) Multiply (and there is no way to choose the same $q$ for both $x$ and $y$). To satisfy this constraint, ModSwitch instructions can be inserted, as shown in Figure 1(c). Both Rescale and ModSwitch drop the first element in their input $q$ (or consume the modulus), whereas Rescale also divides the scale by the given scalar (which is required to match the first element in $q$). The output now has a scale of $2^{60} \cdot 2^{30}$. This would require choosing $q = \{2^{30}, 2^{60}, 2^{60}, 2^{30}, 2^{30}\}$. Thus, although Figure 1(c) executes more instructions than Figure 1(a), it requires the same $Q$. A better way to insert Rescale instructions is shown in Figure 1(d). This satisfies Constraint 1 without ModSwitch instructions. The output now has a scale of $2^{60} \cdot 2^{30}$. We can choose $q = \{2^{60}, 2^{60}, 2^{30}, 2^{30}\}$, so $Q = 2^{150} \cdot s_r$. Hence, this program is more efficient than the input program.

If the computation was modified to $x^2 + y^3$ in Figure 1(d), then the last (bottom) Multiply would be replaced by Add and the program would violate Constraint 2 as Add would have operands with scales $2^{60}$ and $2^{30}$. Consider a similar but simpler example of $x^2 + x$ in Figure 2(a). One way to satisfy Constraints 1 and 2 is by adding Rescale and ModSwitch, as shown in Figure 2(b), which would require $q = \{2^{30}, 2^{30}, 2^{30}\}$. Another way is to introduce Multiply of $x$ and a constant 1 with $2^{30}$ scale to match the scale of Add operands, as shown in Figure 2(c), which would require $q = \{2^{60}, s_r\}$. Although the product $Q = 2^{60} \cdot s_r$ is same in both cases, the modulus length $r$ is different. Hence, the program in Figure 2(c) is more efficient due to a smaller $r$.

Multiply has another constraint. Each ciphertext consists of 2 or more polynomials. Multiply of two ciphertexts with $k$ and $l$ polynomials yields a ciphertext with $k + l + 1$ polynomials. Nevertheless, fewer the polynomials faster the Multiply, so we enforce it to be the minimum:

$$\forall i \text{ n.parm}_i \cdot \text{num}_\text{polynomials} = 2$$

if $n.op \in \{\text{Multiply}\}$

Relinearize reduces the number of polynomials in a ciphertext to 2. It can be inserted in the program in Figure 1(d) to satisfy this constraint, as shown in Figure 1(e).

Finally, we use $s_f$ to denote the maximum allowed rescale value in the rest of this paper ($\log_q s_f$ is also the maximum bit size that can be used for encryption parameters), i.e.,

$$n.parm_{2.value} \leq s_f$$

if $n.op \in \{\text{Rescale}\}$

In the SEAL library, $s_f = 2^{60}$ (which enables a performant implementation by limiting scales to machine-sized integers).

To summarize, FHE schemes (or libraries) are tedious for a programmer to reason about, due to all their cryptographic constraints. Programmers find it even more tricky to satisfy the constraints in a way that optimizes performance. The EVA compiler hides such cryptographic details from the programmer while optimizing the program.

4.3 Execution Flow of the Compiler

As mentioned in Section 3, the in-memory internal representation of the EVA compiler is an Abstract Semantic Graph, also known as a Term Graph, of the input program. In the rest of this paper, we will use the term graph to denote an Abstract Semantic Graph. In this in-memory representation, each node can access both its parents and its children, and for each output, a distinct leaf node is added as a child. It is straightforward to construct the graph from the EVA program and vice-versa, so we omit the details. We use the terms program and graph interchangeably in the rest of the paper.

Algorithm 1 presents the execution flow of the compiler. There are four main steps, namely transformation, validation, parameters selection, and rotations selection. The transformation step takes the input program and modifies it to satisfy the constraints of all instructions, while optimizing it. In the next step, the transformed program is validated to ensure that no constraints are violated. If any constraints are violated, then the compiler throws an exception. By doing this, EVA ensures that executing the output program will never lead to a runtime exception thrown by the FHE library. Finally, for
A graph transformation can be captured succinctly using (or data) flows from children to parents. In general, multiple forward pass from roots to leaves of the graph (a node rewrite rules until quiescence (no change). However, reducing 

The schedule in which these operations wherever needed. The schedule in which these local operations on a graph and the graph transformation itself is composed of applying these operations wherever needed. The schedule in which these local operations are applied may impact the correctness or efficiency of the transformation. Consider two schedules: (1) forward pass from roots to leaves of the graph (a node is scheduled for rewriting only after all its parents have already been rewritten), or (2) backward pass from leaves to roots of the graph (a node is scheduled for rewriting only after all its children have already been rewritten). Note that the rewriting operation may not do any modifications if its condition does not hold. In forward pass, state (or data) flows from parents to children. Similarly, in backward pass, state (or data) flows from children to parents. In general, multiple forward or backward passes may be needed to apply the rewrite rules until quiescence (no change).

5.1 Graph Rewriting Framework

A graph transformation can be captured succinctly using graph rewrite rules (or term rewrite rules). These rules specify the conditional transformation of a subgraph (or an expression) and the graph transformation consists of transforming all applicable subgraphs (or expressions) in the graph (or program). The graph nodes have read-only properties like the opcode and number of parents. In a graph transformation, some state or data may be stored on each node in the graph and the rewrite rules may read and update the state.

The rewrite rules specify local operations on a graph and the graph transformation itself is composed of applying these operations wherever needed. The schedule in which these local operations are applied may impact the correctness or efficiency of the transformation. Consider two schedules: (1) forward pass from roots to leaves of the graph (a node is scheduled for rewriting only after all its parents have already been rewritten), or (2) backward pass from leaves to roots of the graph (a node is scheduled for rewriting only after all its children have already been rewritten). Note that the rewriting operation may not do any modifications if its condition does not hold. In forward pass, state (or data) flows from parents to children. Similarly, in backward pass, state (or data) flows from children to parents. In general, multiple forward or backward passes may be needed to apply the rewrite rules until quiescence (no change).

5.2 Relinearize Insertion Pass

Each ciphertext is represented as 2 or more polynomials. Multiplying two ciphertexts each with 2 polynomials yields a ciphertext with 3 polynomials. The Relinearize instruction reduces a ciphertext to 2 polynomials. To satisfy Constraint 3, EVA must insert Relinearize after Multiply of two nodes with Cipher type and before another such Multiply.

The Relinearize rewrite rule (Figure 3) is applied for a node $n$ only if it is a Multiply operation and if both its parents (or parameters) have Cipher type. The transformation in the rule inserts a Relinearize node $n_i$ between the node $n$ and its children. In other words, the new children of $n$ will be only $n_i$ and the children of $n_i$ will be the old children of $n$. For the example graph in Figure 1(d), applying this rewrite rule transforms the graph into the one in Figure 1(e).

Optimal placement of relinearization is an NP-hard problem [10]. Our relinearization insertion pass is a simple way to enforce Constraint 3. More advanced relinearization insertion, with or without Constraint 3, is left for future work.

5.3 Rescale and ModSwitch Insertion Passes

Goal: The Rescale and ModSwitch nodes (or instructions) must be inserted such that they satisfy Constraint 1, so the goal of the Rescale and ModSwitch insertion passes is to insert them such that the coefficient moduli of the parents of any Add and Multiply node are equal.

While satisfying Constraint 1 is sufficient for correctness, performance depends on where Rescale and ModSwitch are inserted (as illustrated in Section 4.2). Different choices lead to different coefficient modulus $q = \{q_1, q_2, ..., q_n\}$, and consequently, different polynomial modulus $N$ for the roots (or inputs) to the graph (or program). Larger values of $N$ and $r$ increase the cost of every FHE operation and the memory of every ciphertext. $N$ is a non-decreasing function of $Q = \prod_{i=1}^{r} q_i$ (i.e., if $Q$ grows, $N$ either remains the same or grows as well). Minimizing both $Q$ and $r$ is a hard problem to solve. However, reducing $Q$ is only impactful if it reduces $N$, which is unlikely as the threshold of $Q$, for which $N$ increases, grows exponentially. Therefore, the goal of EVA is to yield the optimal $r$, which may or may not yield the optimal $N$. 

EVA includes a graph rewriting framework for arbitrary rewrite rules for a subgraph that consists of a node along with its parents or children. The rewrite rules for each graph transformation pass in EVA are defined in Figure 3. A single backward pass is sufficient for EAGER-MODSWITCH, while a single forward pass is sufficient for the rest. The rewrite rules assume the passes are applied in a specific order: WATERLINE-RESCALE, EAGER-MODSWITCH, MATCHSCALE, and RELINEARIZE (ALWAYS-RESCALE and LAZY-MODSWITCH are not used but defined only for clarity). For the sake of exposition, we will first describe RELINEARIZE pass before describing the other passes.

Algorithm 1: Execution of EVA compiler.

| Input: Program $P_i$ in EVA language |
| Input: Scales $S_i$ for inputs in $P_i$ |
| Input: Desired scales $S_d$ for outputs in $P_i$ |
| Output: Program $P_o$ in EVA language |
| Output: Vector $B_o$ of bit sizes |
| Output: Set $R_o$ of rotation steps |

1. $P_o = \text{Transform}(P_i, S_i)$
2. if $\text{Validate}(P_o) == \text{Failed}$ then
   1. Throw an exception
3. $B_o = \text{DetermineParameters}(P_o, S_i, S_d)$
4. $R_o = \text{DetermineRotationSteps}(P_i, S_i)$

the validated output program, the compiler selects the bit sizes and the rotation steps that must be used to determine the encryption parameters and the rotation keys respectively, before executing the output program. The transformation step involves rewriting the graph, which is described in detail in Section 5. The other steps only involve traversal of the graph (without changing it), which is described in Section 6.
**Constrained-Optimization Problem:** The only nodes that modify a ciphertext’s coefficient modulus are Rescale and ModSwitch nodes; that is, they are the only ones whose output ciphertext has a different coefficient modulus than that of their input ciphertext(s). Therefore, the coefficient modulus of the output of a node depends only on the Rescale and ModSwitch nodes in the path from the root to that node. To illustrate their relation, we define the term *rescale chain*.

**Definition 5.1.** Given a directed acyclic graph $G = (V, E)$:

For $n_1, n_2 \in V$, $n_1$ is a parent of $n_2$ if $\exists n_1, n_2 \in E$.

A node $r \in V$ is a root if $r.type = Cipher$ and $\exists n \in V$ s.t. $n$ is a root of $r$.

**Definition 5.2.** Given a directed acyclic graph $G = (V, E)$:

A path $p$ from a node $n_0 \in V$ to a node $n \in V$ is a sequence of nodes $p_0, ..., p_l \in V$ s.t. $n_0 = p_0, p_l = n$, and $\forall 0 \leq i < l, p_i \in V$ and $p_i$ is a parent of $p_{i+1}$. A path $p$ is said to be simple if $\forall 0 < i < l, p_i, op \neq Rescale$ and $p_i, op \neq ModSwitch$.

**Definition 5.3.** Given a directed acyclic graph $G = (V, E)$:

A rescale path $p$ to a node $n \in V$ is a sequence of nodes $p_0, ..., p_l$ s.t. $(\forall 0 \leq i \leq l, p_i, op \in \{Rescale, ModSwitch\})$, $\exists$ a simple path from a root to $n_0$, $p_0$, $\exists$ a simple path from $p_l$ to $n$, $(\forall 0 \leq i < l, \exists$ a simple path from $p_i$ to $p_{i+1}$), and $(n, op = Rescale$ or $n, op = ModSwitch) \implies (p_l = n)$.

A rescale chain of a node $n \in V$ is a vector $c$ s.t. $\exists$ a rescale path $p$ to $n$ and $(\forall 0 \leq i < |p|, (p_i, op = ModSwitch \implies c_i = \infty)$ and $(p_i, op = Rescale \implies c_i = (p_i, parm_2.value))$.

Note that $\infty$ is used here to distinguish ModSwitch from Rescale in the rescale chain.

A rescale chain $c$ of a node $n$ and $c'$ of a node $n'$ are *equal* if $(|c| = |c'|$ and $(\forall 0 \leq i < |c|, c_i = c'_i$ or $c_i = \infty$ or $c'_i = \infty))$.

A rescale chain $c$ of a node $n \in V$ is *conforming* if $\forall$ rescale chain $c'$ of $n$, $c$ is equal to $c'$.

Note that all the roots in the graph have the same coefficient modulus. Therefore, for nodes $n_1$ and $n_2$, the coefficient modulus of the output of $n_1$ is equal to that of $n_2$ if and only if there exists conforming rescale chains for $n_1$ and $n_2$, and the conforming rescale chain of $n_1$ is equal to that of $n_2$. Thus, we need to solve two problems simultaneously:

- **Constraints:** Ensure the conforming rescale chains of the parents of any Multiply or Add node are equal.
- **Optimization:** Minimize the length of the rescale chain of every node.

**Outline:** In general, the constraints problem can be solved in two steps:

- Insert Rescale in a pass (to reduce exponential growth of scale and noise).
- Insert ModSwitch in another pass so that the constraints are satisfied.

The challenge is in solving this problem in this way, while yielding the desired optimization.

**Always Rescale Insertion:** A naive approach of inserting Rescale is to insert it after every Multiply of Cipher.
nodes. We call this approach as always rescale and define it in the ALWAY-RESCALE rewrite rule in Figure 3. Consider the example in Figure 1(a). Applying this rewrite rule on this example yields the graph in Figure 1(b). For some MULTIPLY nodes (e.g., the bottom one), the conforming rescale chains of their parents do not exist or do not match. To satisfy these constraints, MonSwitch nodes can be inserted appropriately, as shown in Figure 1(c) (we omit defining this rule because it would require multiple forward passes). The conforming rescale chain length for the output is now more than the multiplicative depth of the graph. Thus, always rescale and its corresponding modswitch insertion may lead to a larger coefficient modulus (both in the number of elements and their product) than not inserting any of them.

**Insight:** Consider that all the roots in the graph have the same scale \( s \). For example in Figure 1(a), let \( x.\text{scale} = 2^{30} \) instead of \( 2^{60} \). Then, after always rescale (replace \( 2^{60} \) with \( 2^{30} \) in Figure 1(b)), the only difference between the rescale chains of a node \( n \) would be their length and not the values in it. This is the case even when roots may have different scales as long as all RESCALE nodes rescale by the same value \( s \). Then, a conforming rescale chain \( c_n \) for \( n \) can be obtained by adding MonSwitch nodes in the smaller chain(s). Thus, \( |c_n| \) would not be greater than the multiplicative depth of \( n \). The first key insight of EVA is that using the same rescale value for all RESCALE nodes ensures that \( |c_o| \) cannot be greater than the multiplicative depth of \( o \) (a tight upper bound). The multiplicative depth of a node \( n \) is not a tight lower bound for \( |c_n| \), as shown in Figure 1(d). The second key insight of EVA is that using the maximum rescale value \( s_f \) (satisfying Constraint 4) for all RESCALE nodes minimizes \( |c_o| \) because it minimizes the number of RESCALE nodes in any path.

**Waterline Rescale Insertion:** Based on our insights, the value to rescale is fixed to \( s_f = 2^{60} \) in SEAL. That does not address the question of when to insert RESCALE nodes. If the scale after RESCALE becomes too small, then the compiled message may lose accuracy irrevocably. We call the minimum required scale as waterline and use \( s_w \) to denote it. We choose \( s_w \) to be maximum of the scales of all roots. Consider a MULTIPLY node \( n \) whose scale after multiplication is \( s_n \). Then, a RESCALE in inserted between \( n \) and its children only if the scale after RESCALE is above the waterline, i.e., \( (s_n/s_f) \geq s_w \). We call this approach as waterline rescale and define the WATERLINE-RESCALE rewrite rule in Figure 3. This rule transforms the graph in Figure 1(a) to the one in Figure 1(d).

**ModSwitch Insertion:** For a node \( n \), let \( n.\text{level} \) denote its conforming rescale chain length. Let \( n.\text{relvel} \) denote the conforming rescale chain length of \( n \) in the transpose graph. A naive way to insert ModSwitch is to find a Add or Multiply node for which \( \text{level} \) of the parents do not match and insert the appropriate number of ModSwitch nodes between one of the parents and the node. We call this lazy insertion and define the LAZY-MODSWITCH rewrite rule in Figure 3. We call inserting it at the earliest feasible edge in the graph as eager insertion. The EAGER-MODSWITCH rewrite rule (Figure 3) finds a node for which \( \text{relvel} \) of the children do not match and inserts the appropriate number of ModSwitch nodes between some of the children and itself. If the \( \text{relvel} \) of the roots do not match, then there is another rule (omitted in Figure 3 for brevity) that inserts the appropriate ModSwitch nodes between some of the roots and their children.

Consider the \( x^2 + x + x \) example in Figure 4(a). Applying the LAZY-MODSWITCH and EAGER-MODSWITCH rewrite rules yields the graphs in Figures 4(b) and (c) respectively. The operands of Add after eager insertion use a smaller coefficient modulus than after lazy insertion, so Add would be faster if eager insertion is used. Thus, eager insertion leads to similar or more efficient code than lazy insertion.

**Matching Scales:** As illustrated in Section 4.2, it is easy to match scales of parents of Add by multiplying one of the parents and 1 with the appropriate scale. The MATCH-SCALE rewrite rule (Figure 3) takes this simple approach to satisfy Constraint 2 while avoiding introduction of any additional RESCALE or ModSwitch. For the example graph in Figure 2(a), applying this rewrite rule transforms the graph into the one in Figure 2(c).

**Optimality:** EVA selects encryption parameters (see Section 6.2) s.t. \( r = \max(\forall o \in \{\text{Outputs}\}, 1 + |c_o| + \lceil o.\text{scale}/s_w \rceil) \), where \( s_o \) is the desired scale for the output \( o \). WATERLINE-RESCALE is the only pass that determines \( |c_o| \) and \( o.\text{scale} \) for any output \( o \) (neither LAZY-MODSWITCH nor MATCH-SCALE modify that). If \( |c_o| \) is decreased by 1 (an element \( s_f \) from \( c_o \) is removed), then \( o.\text{scale} \) would increase by at least \( s_f \), so it would not decrease \( r \). Due to waterline rescale, \( o.\text{scale} < s_w \times s_f \), so RESCALE cannot be inserted to reduce \( o.\text{scale} \) by at least \( s_f \) (because the minimum required scale is \( s_w \)). Thus, EVA yields the minimal or optimal \( r \).

6 Analysis in EVA Compiler
In this section, we briefly describe our graph traversal framework (Section 6.1) and a few analysis passes (Section 6.2).
6.1 Graph Traversal Framework and Executor

EVA’s graph traversal framework allows either a forward traversal or a backward traversal of the graph. In the forward traversal pass, a node is visited only after all its parents are visited. Similarly, in the backward traversal pass, a node is visited only after all its children are visited. Graph traversals do not modify the structure of the graph (unlike graph rewriting) but a state on each node can be maintained.

**Parallel Implementation:** We implement an executor for the generated EVA program using the traversal framework. A node is said to be *ready or active* if all its parents (or children) in forward (or backward) pass have already been visited. These active nodes can be scheduled to execute in parallel as each active node only updates its own state (i.e., there are no conflicts). For example in Figure 1(e), the parents of the bottom *MULTIPLY* can execute in parallel. Each FHE instruction (node) can take a significant amount of time to execute, so it is useful to exploit parallelism among FHE instructions. The EVA executor automatically parallelizes the generated EVA program by implementing a parallel graph traversal using the Galois [20, 37] parallel library.

A node is said to *retire* if all its children (or parents) in forward (or backward) pass have already been visited. The state for the retired nodes will no longer be accessed, so it can be reused for other nodes. In Figure 1(e), after the *RELINEARIZE* is executed. The EVA executor automatically reuses the memory used for encrypted messages, thereby reducing the memory consumed.

6.2 Analysis Passes

**Validation Passes:** We implement a validation pass for each of the constraints in Section 4.2. All are forward passes. The first pass computes the rescale chains for each node and asserts that it is conforming. It also asserts that the conforming rescale chains of parents of *ADD* and *MULTIPLY* match, satisfying Constraint 1. The second and third passes compute a *scale* and *num_polynomials* for each node respectively and assert that Constraint 2 and 3 is satisfied respectively. If any assertion fails, an exception is thrown at compile-time. Thus, these passes elide runtime exceptions thrown by SEAL.

**Encryption Parameter Selection Pass:** Akin to encryption selection in CHET [18], the encryption parameter selection pass in EVA computes the conforming *rescale chain* and the scale for each node. For each leaf or output o after the pass, let c0 be the conforming rescale chains of o without ∞ in it and let s′ 0 = s0 * o.scale, where s0 is the desired output scale. s′ 0 is factorized into s0+s1+⋯+sk such that sk is a power-of-two, sk ≤ sf (= 260 in SEAL), and 0 ≤ i < k, si = sf. Let |s′0| denote the number of factors of s′ 0. Then EVA finds the output m with the maximum |cm|+|s′m|. The factors of sm are appended to cm and sf (the *special prime* that is consumed during encryption) in inserted at the beginning of cm. For each element s in cm, log2 s is applied to obtain a vector of bit sizes, which is then returned.

**Rotation Keys Selection Pass:** Similar to rotation keys selection in CHET [18], EVA’s rotation keys selection pass computes and returns the set of unique step counts used among all *ROTATELEFT* and *ROTATERIGHT* nodes in the graph.

7 Frontends of EVA

The various transformations described so far for compiling an input EVA program into an executable EVA program make up the *backend* in the EVA compiler framework. In this section, we describe two *frontends* for EVA, that make it easy to write programs for EVA.

7.1 PyEVA

We have built a general-purpose frontend for EVA as a DSL embedded into Python, called PyEVA. Consider the PyEVA program in Figure 5 for Sobel filtering, which is a form of edge detection in image processing. The *class* *Program* is a wrapper for the Protocol Buffer [23] format for EVA programs mentioned in Section 3. It includes a context manager, such that inside a *with* statement all operations are recorded in program. For example, the *inputEncrypted* function inserts an input node of type *Cipher* into the program currently in context and additionally returns an instance.
We make use of CHET’s data layout selection optimization, work as a graph of high-level tensor operations, and through this path length in 3-dimensional space. We also evaluate other applications, all results reported are an average over 20 different randomly generated inputs.

8 Experimental Evaluation

8.2 Deep Neural Network (DNN) Inference

Networks: We evaluate five deep neural network (DNN) architectures for image classification tasks that are summarized in Table 3:

- The three LeNet-5 networks are all for the MNIST [32] dataset, which vary in the number of neurons. The largest one matches the one used in the TensorFlow’s tutorials [42].
- Industrial is a network from an industry partner for privacy-sensitive binary classification of images.
- SqueezeNet-CIFAR is a network for the CIFAR-10 dataset [31] that uses 4 Fire-modules [15] and follows the SqueezeNet [27] architecture.

We obtain these networks (and the models) from the authors of CHET, so they match the networks evaluated in their paper [18]. Industrial is a FHE-compatible neural network that is proprietary, so the authors gave us only the network structure without the trained model (weights) or the test datasets. We evaluate this network using randomly generated numbers (between -1 and 1) for the model and the images. All the other networks were made FHE-compatible by CHET authors using average-pooling and polynomial activations instead of max-pooling and ReLU activations. Table 3 lists the accuracies we observed for these networks using unencrypted inference on the test datasets. We evaluate encrypted image inference with a batch size of 1 (latency).

<table>
<thead>
<tr>
<th>Network</th>
<th>No. of layers</th>
<th># FP operations</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeNet-5-small</td>
<td>2</td>
<td>2</td>
<td>159960</td>
</tr>
<tr>
<td>LeNet-5-medium</td>
<td>2</td>
<td>2</td>
<td>5791168</td>
</tr>
<tr>
<td>LeNet-5-large</td>
<td>2</td>
<td>2</td>
<td>21385674</td>
</tr>
<tr>
<td>Industrial</td>
<td>5</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>SqueezeNet-CIFAR</td>
<td>10</td>
<td>0</td>
<td>37759754</td>
</tr>
</tbody>
</table>

Table 3. Deep Neural Networks used in our evaluation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Input Scale (log P)</th>
<th>Output Scale</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cipher Vector Scalar</td>
<td>25 15 10</td>
<td>30</td>
<td>98.42</td>
</tr>
<tr>
<td>Vector Scalar</td>
<td>25 15 10</td>
<td>30</td>
<td>99.07</td>
</tr>
<tr>
<td>Scalar</td>
<td>25 20 10</td>
<td>25</td>
<td>99.34</td>
</tr>
<tr>
<td>Industrial</td>
<td>30 15 10</td>
<td>30</td>
<td>-</td>
</tr>
<tr>
<td>SqueezeNet-CIFAR</td>
<td>25 15 10</td>
<td>30</td>
<td>79.31</td>
</tr>
</tbody>
</table>

Table 4. Programmer-specified input and output scaling factors used for both CHET and EVA, and the accuracy of homomorphic inference in CHET and EVA (all test inputs).
Table 5. Average latency (s) of CHET and EVA on 56 threads.

<table>
<thead>
<tr>
<th>Model</th>
<th>CHET</th>
<th>EVA</th>
<th>Speedup from EVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeNet-5-small</td>
<td>3.7</td>
<td>0.6</td>
<td>6.2</td>
</tr>
<tr>
<td>LeNet-5-medium</td>
<td>5.8</td>
<td>1.2</td>
<td>4.8</td>
</tr>
<tr>
<td>LeNet-5-large</td>
<td>23.3</td>
<td>5.6</td>
<td>4.2</td>
</tr>
<tr>
<td>Median</td>
<td>70.4</td>
<td>9.6</td>
<td>7.3</td>
</tr>
<tr>
<td>SqueezeNet-CIFAR</td>
<td>344.7</td>
<td>72.7</td>
<td>4.7</td>
</tr>
</tbody>
</table>

Figure 6. Strong scaling of CHET and EVA (log-log scale).

Scaling Factors: The scaling factors, or scales in short, must be chosen by the user. For each network (and model), we use CHET’s profiling-guided optimization on the first 20 test images to choose the input scales as well as the desired output scale. There is only one output but there are many inputs. For the inputs, we choose one scale each for Cipher, Vector, and Scalar inputs. Both CHET and EVA use the same scales, as shown in Table 4. The scales impact both performance and accuracy. We evaluate CHET and EVA on all test images using these scales and report the accuracy achieved by fully-homomorphic inference in Table 4. There is negligible difference between their accuracy and the accuracy of unencrypted inference (Table 3). Higher values of scaling factors may improve the accuracy, but will also increase the latency of homomorphic inference.

Comparison with CHET Compiler: Table 5 shows that EVA is at least 4× faster than CHET on 56 threads for all networks. Note that the average latency of CHET is slower than that reported in their paper [18]. This could be due to differences in the experimental setup. The input and output scales they use are different, so is the SEAL version (3.1 vs.

Table 6. Encryption parameters selected by CHET and EVA (where $Q = \prod_{i=1}^{r} Q_i$).

<table>
<thead>
<tr>
<th>Model</th>
<th>CHET</th>
<th>EVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>log$_2 N$</td>
<td>r</td>
<td>log$_2 Q$</td>
</tr>
<tr>
<td>LeNet-5-small</td>
<td>15</td>
<td>480</td>
</tr>
<tr>
<td>LeNet-5-medium</td>
<td>15</td>
<td>480</td>
</tr>
<tr>
<td>LeNet-5-large</td>
<td>15</td>
<td>740</td>
</tr>
<tr>
<td>Industrial</td>
<td>16</td>
<td>1222</td>
</tr>
<tr>
<td>SqueezeNet-CIFAR</td>
<td>16</td>
<td>1740</td>
</tr>
</tbody>
</table>

3.3.1). We suspect the machine differences to be the primary reason for the slowdown because they use smaller number of heavier cores (16 3.2GHz cores vs. 56 2.2GHz cores). In any case, our comparison of CHET and EVA is fair because both use the same input and output scales, SEAL version, Channel-Height-Width (CHW) data layout, and hardware. Both CHET and EVA perform similar encryption parameters and rotation keys selection. The differences between CHET and EVA are solely due to the benefits that accrue from EVA’s low-level optimizations.

CHET relies on an expert-optimized library of homomorphic tensor kernels, where each kernel (1) includes FHE-specific instructions and (2) is explicitly parallelized. However, even experts cannot optimize or parallelize across different kernels as that information is not available to them. In contrast, EVA uses a library of vectorized tensor kernels and automatically (1) inserts FHE-specific instructions using global analysis and (2) parallelizes the execution of different instructions across kernels. Due to these optimizations, EVA is on average 5.3× faster than CHET. On a single thread (Figure 6), EVA is on average 2.3× faster than CHET and this is solely due to better placement of FHE-specific instructions. The rest of the improvement on 56 threads (2.3× on average) is due to better parallelization in EVA.

Both CHET and EVA have similar Relinearize placement. However, they differ in the placement of the other FHE-specific instructions — Rescale and ModSwitch. These instructions directly impact the encryption parameters (both CHET and EVA use a similar encryption parameter selection pass). We report the encryption parameters selected by CHET and EVA in Table 6. EVA selects much smaller coefficient modulus, both in terms of the number of elements $r$ in it and their product $Q$. Consequently, the polynomial modulus $N$ is one power-of-2 lower in all networks, except LeNet-5-large. Reducing $N$ and $r$ reduces the cost (and the memory) of each homomorphic operation (and ciphertext) significantly. In CHET, Rescale and ModSwitch used by the experts for a given tensor kernel may be sub-optimal for the program. On the other hand, EVA performs global (inter-procedural) analysis to minimize the length of the coefficient modulus, yielding much smaller encryption parameters.
Table 7. Compilation, encryption context (context), encryption, and decryption time for EVA.

<table>
<thead>
<tr>
<th>Model</th>
<th>Compilation</th>
<th>Context</th>
<th>Encrypt</th>
<th>Decrypt</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeNet-5-small</td>
<td>0.14</td>
<td>1.21</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>LeNet-5-medium</td>
<td>0.50</td>
<td>1.26</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>LeNet-5-large</td>
<td>1.13</td>
<td>7.24</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td>Industrial</td>
<td>0.59</td>
<td>15.70</td>
<td>0.12</td>
<td>0.03</td>
</tr>
<tr>
<td>SqueezeNet-CIFAR</td>
<td>4.06</td>
<td>160.82</td>
<td>0.42</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Table 8. Evaluation of EVA for fully-homomorphic arithmetic, statistical machine learning, and image processing applications on 1 thread (LoC: lines of code).

<table>
<thead>
<tr>
<th>Application</th>
<th>Vector Size</th>
<th>LoC</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-dimensional Path Length</td>
<td>4096</td>
<td>45</td>
<td>0.394</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>2048</td>
<td>10</td>
<td>0.094</td>
</tr>
<tr>
<td>Polynomial Regression</td>
<td>4096</td>
<td>15</td>
<td>0.104</td>
</tr>
<tr>
<td>Multivariate Regression</td>
<td>2048</td>
<td>15</td>
<td>0.094</td>
</tr>
<tr>
<td>Sobel Filter Detection</td>
<td>4096</td>
<td>35</td>
<td>0.511</td>
</tr>
<tr>
<td>Harris Corner Detection</td>
<td>4096</td>
<td>40</td>
<td>1.004</td>
</tr>
</tbody>
</table>

To understand the differences in parallelization, we evaluated CHET and EVA on 1, 7, 14, 28, and 56 threads. Figure 6 shows the strong scaling. We omit LeNet-5-small because it takes too little time, even on 1 thread. It is apparent that EVA scales much better than CHET. The parallelization in CHET is within a tensor operation or kernel using OpenMP. Such static, *bulk-synchronous* schedule limits the available parallelism. In contrast, EVA dynamically schedules the directed acyclic graph of EVA (or SEAL) operations asynchronously. Thus, it exploits the parallelism available across tensor kernels, resulting in much better scaling. The average speedup of EVA on 56 threads over EVA on 1 thread is 18.6× (excluding LeNet-5-small).

Comparison with Hand-Written LoLa: LoLa [7] implements hand-tuned homomorphic inference for neural networks, but the networks they implement are different than the ones we evaluated (and the ones in CHET). Nonetheless, they implement networks for the MNIST and CIFAR-10 datasets.

For the MNIST dataset, LoLa implements the highly-tuned CryptoNets [21] network (which is similar in size to LeNet-5-small). This implementation has an average latency of 2.2 seconds and has an accuracy of 98.95%. EVA takes only 1.2 seconds on a much larger network, LeNet-5-medium, with a better accuracy of 99.09%. For the CIFAR-10 dataset, LoLa implements a custom network which takes 730 seconds and has an accuracy of 74.1%. EVA takes only 72.7 seconds on a much larger network with a better accuracy of 79.34%.

LoLa uses SEAL 2.3 (which implements BFV [19]) which is less efficient than SEAL 3.3.1 (which implements RNS-CKKS [12]) but much more easier to use. EVA is faster because it exploits a more efficient FHE scheme which is much more difficult to manually write code for. Thus, EVA outperforms even highly tuned expert-written implementations like LoLa with very little programming effort.

Compilation Time: We present the compilation time, encryption context time, encryption time, and decryption time for all networks in Table 7. The encryption context time includes the time to generate the public key, the secret key, the rotation keys, and the relinearization keys. This can take a lot of time, especially for large N, like in SqueezeNet-CIFAR. Compilation time, encryption time, and decryption time are negligible for all networks.

8.3 Arithmetic, Statistical Machine Learning, and Image Processing

We implemented several applications using PyEVA. To illustrate a simple arithmetic application, we implemented an application that computes the length of a given encrypted 3-dimensional path. This computation can be used as a kernel in several applications like in secure fitness tracking on mobiles. For statistical machine learning, we implemented linear regression, polynomial regression, and multi-variate regression on encrypted vectors. For image processing, we implemented Sobel filter detection and Harris corner detection on encrypted images. All these implementations took very few lines of code (< 50), as shown in Table 8.

Table 8 shows the execution time of these applications on encrypted data using 1 thread. Sobel filter detection takes half a second and Harris corner detection takes only a second. The rest take negligible time. We believe *Harris corner detection is one of the most complex programs that have been evaluated using CKKS*. EVA enables writing advanced applications in various domains with little programming effort, while providing excellent performance.

9 Related Work

*Libraries for FHE*: SEAL [41] implements RNS variants of two FHE schemes: BFV [19] and CKKS [12, 13]. HElib [25] implements two FHE schemes: BGV [6] and CKKS. PALISADE [38] is a framework that provides a general API for multiple FHE schemes including BFV, BGV, and CKKS. For BFV and CKKS, PALSIDE is similar to SEAL as it only implements lower-level FHE primitives. On the other hand, EVA language abstracts batching-compatible FHE schemes like BFV, BGV, and CKKS while hiding cryptographic details from the programmer. Although EVA compiler currently generates code targeting only CKKS implementation in SEAL, it can be adapted to target other batching-compatible FHE scheme implementations or FHE libraries.
General-Purpose Compilers for FHE. To reduce the burden of writing FHE programs, general-purpose compilers have been proposed that target different FHE libraries. These compilers share many of the same goals as ours. Some of these compilers support general-purpose languages like Julia (cf. [2]), C++ (cf. [14]), and R (cf. [3]), whereas ALCHEMY [16] is the only one that provides its own general-purpose language. Unlike EVA, none of these languages are amenable to be a target for domain-specific compilers like CHET [18] because these languages do not support rotations on fixed power-of-two sized vectors. Nonetheless, techniques in these compilers (such as ALCHEMY’s static type safety and error rate analysis) are orthogonal to our contributions in this paper and can be incorporated in EVA.

All prior general-purpose compilers target (libraries implementing) either the BFV scheme [19] or the BGV scheme [6]. In contrast, EVA targets (libraries implementing) the recent CKKS scheme [12, 13], which is much more difficult to write or generate code for (compared to BFV or BGV). For example, ALCHEMY supports the BGV scheme and would require significant changes to capture the semantics (e.g., RESCALE) of CKKS. ALCHEMY always inserts MonSwitch after every ciphertext-ciphertext multiplication (using local analysis), which is not optimal for BGV (or BFV) and would not be correct for CKKS. EVA is the first general-purpose compiler for CKKS and it uses a graph rewriting framework to insert RESCALE and MonSwitch operations correctly (using global analysis) so that the modulus chain length is optimal. These compiler passes in EVA can be incorporated in other general-purpose compilers (to target CKKS).

Domain-Specific Compilers for FHE. Some prior compilers for DNN inferencing [4, 5, 18] target CKKS. CHET [18] is a compiler for tensor programs that automates the selection of data layouts for mapping tensors to vectors of vectors. The nGraph-HE [5] project introduced an extension to the Intel nGraph [17] deep learning compiler that allowed data scientists to make use of FHE with minimal code changes. The nGraph-HE compiler uses run-time optimization (e.g., detection of special plaintext values) and compile-time optimizations (e.g., use of ISA-level parallelism, graph-level optimizations). nGraph-HE2 [4] is an extension of nGraph-HE that uses a hybrid computational model – the server interacts with the client to perform non-HE compatible operations, which increases the communication overhead. Moreover, unlike CHET and EVA, neither nGraph-HE nor nGraph-HE2 automatically select encryption parameters.

To hide the complexities of FHE operations, all existing domain-specific compilers rely on a runtime of high-level kernels which can be optimized by experts. However, experts are limited to information within a single kernel (like convolution) to optimize insertion of FHE-specific operations and to parallelize execution. In contrast, EVA optimizes insertion of FHE-specific operations by using global analysis and parallelizes FHE operations across kernels transparently. Therefore, CHET, nGraph-HE, and nGraph-HE2 can target EVA instead of the FHE scheme directly to benefit from such optimizations and we demonstrated this for CHET.

Compilers for MPC. Multi-party computation (MPC) [22, 44] is another technique for privacy-preserving computation. The existing MPC compilers are mostly general-purpose [24] and even though it is possible to use them for deep learning applications, it is hard to program against a general-purpose interface. The EzPC compiler is a machine learning compiler that combines arithmetic sharing and garbled circuits and operates in a two-party setting [9]. EzPC uses ABY as a cryptographic backend [35].

Privacy-Preserving Deep Learning. CryptoNets, one of the first systems for neural network inference using FHE [21] and the consequent work on LoLa, a low-latency CryptoNets [7], show the ever more practical use of FHE for deep learning. CryptoNets and LoLa however use kernels for neural networks that directly translate the operations to the cryptographic primitives of the FHE schemes. There are also other algorithms and cryptosystems specifically for deep learning that rely on FHE (CryptoDL [26], [8], [29]), MPC (Chameleon [39], DeepSecure [40], SecureML [36]), oblivious protocols (MiniONN [33]), or on hybrid approaches (Gazelle [30], SecureNN [43]). None of these provide the flexibility and the optimizations of a compiler approach.

10 Conclusions
This paper introduces a new language and intermediate representation called Encrypted Vector Arithmetic (EVA) for general-purpose Fully-Homomorphic Encryption (FHE) computation. EVA includes a Python frontend that can be used to write advanced programs with little programming effort, and it hides all the cryptographic details from the programmer. EVA includes an optimizing compiler that generates correct, secure, and efficient code, targeting the state-of-the-art SEAL library. EVA is also designed for easy targeting of domain specific languages. The state-of-the-art neural network inference compiler CHET, when re-targeted onto EVA, outperforms its unmodified version by 5.3× on average. EVA provides a solid foundation for a richer variety of FHE applications as well as domain-specific compilers and auto-vectorizing compilers for computing on encrypted data.

Acknowledgments
This research was supported by the NSF grants 1406355, 1618425, 1705092, 1725322, and by the DARPA contracts FA8750-16-2-0004 and FA8650-15-C-7563. We thank Keshav Pingali for his support. We thank the anonymous reviewers and in particular our shepherd, Petar Tsankov, for their many suggestions in improving this paper.
References


