Instruction level automatic differentiation on a reversible Turing machine

Jin-Guo Liu[1,*] and Taine Zhao[2]

[1]*Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China*
[2]*Department of Computer Science, University of Tsukuba*

This paper considers instruction level differential programming, i.e. knowing only the backward rule of basic instructions like +, -, * and /, differentiate a program with proper performance. We will review briefly why instruction level automatic differentiation is hard for current machine learning package even for a source to source automatic differentiation package. Then we propose a reversible Turing machine implementation to achieve instruction level automatic differentiation.

I. INTRODUCTION

There are two modes of automatic differentiation (AD) [1], the tangent mode [2] and the adjoint mode. Consider a multi-in multi-out function $\vec{y} = f(\vec{x})$, the tangent mode computes a column of its Jacobian $\frac{\partial \vec{y}}{\partial x_i}$ efficiently, where x_i is one of the input variables. Whereas the adjoint mode computes a row of Jacobian $\frac{\partial y_i}{\partial \vec{x}}$ efficiently.

Most popular automatic differentiation package implements the adjoint mode differentation, since the adjoint mode is computational more efficient in variational applications, where the output loss is always a scalar. However, implementing adjoint mode AD requires a program's intermediate state for back propagation, including

1. computation graph,
2. and input variables of nodes in computation graph.

A computational graph is a directed acyclic graph (DAG) that records the relation between data (edges) and functions (nodes). In Pytorch [3] and Flux [4], every variable has a tracker field that stores its parent information, i.e. the input data and function that generate this variable. TensorFlow [5] implements a static computational graph before actual computation happens. Source to source automatic differentiation package Zygote [4, 6] use a intermediate representation of a program as the computation grpah, which known as the static single assignment (SSA) form. It can back propagate over a native julia code. Its intermediate caching is handled by a global storage.

Due to the computation graph recording and intermediate state caching, some problems are observed in most machine learning packages. First of all, these package requires a lot primitive functions with programmer defined backward rules. This is not nessesary. At the lowest level, these primitive functions are compiled to instructions, and these instructions are from a finite set of '+', '-', '*', '/', conditional jump statements et. al. Define backward rules for basic instructions and the automatic differentiation should just work. With instruction level computational graph, it should be possible to obtain gradients like `exp`, or even linear algebras functions like singular value decomposition (SVD) and eigenvalue decomposition (ED) [7–9] automatically. The recent development of backward rules of linear algebra functions greatly change the research code in physics [10**?**]. However, these rules are generated manually. Secondly, inplace functions are not handled properly in the language of computational graph, even in source to source AD engine Zygote, it is not trivil to implement. Most functions in BLAS and LAPACK are implemented as inplace functions, the lack of automatic differentiation support to inplace functions cause bad memory performance and make the memory wall problem in machine learning even more sevire. [11]. Thridly, obtaining higher order gradients are not efficient in these packages. In most machine learning package, people back propagate the whole program of obtaining lower order gradients to obtain higher order gradients. The repeated use of back propagation cause exponential overhead with respect to the order of gradients. A better approach to obtain higher order gradients is through Taylor propagation like in JAX [12]. However Taylor propagation requires writing rules for all primitives. Besides the exponential overhead, the sorce to source AD engine Zygote suffers from the significant overhead of just in time compiling in Julia language [13].

These machine learning packages can not use instructions as the computational graph for practical reasons. The cost of memorizing the computational graph and intermediate caching kills the performance for more than two orders in simple loops (as we will show latter). A even more serious problem is the memory consumption for caching intermediate results increases linearly as time. Even in many traditional deep network like recurrent neural network and residual neural networks, where the depth is only several thousand, the memory wall can be big problem. Making use of reversibility is a promising approach to save memory. People use information buffer [14] and reversible activation functions to reduce the memory allocations in recurrent neural network [15] and residual neural networks [16]. However, the use of reversiblity in these cases are not general purposed.

Hence we develop a embeded domain specific language (eDSL) in Julia language that implements reversible Turing machine. [17, 18]. The gradient of any program writting in this eDSL can be obtained in comparable time with the forward function. The implementation of AD is similar to `ForwardDiff` [2] but runs backward. There has been

* cacate0129@iphy.ac.cn

some prototypes of reversible languages like Janus [19], R (not the popular one) and object oriented ROOPL, reversible instruction sets like Pendulum ISA [20] and BobISA. These languages have reversible control flow that allows user input additional information, the postcondition in control flows to help programs run backward. The main motivation to make a program reversible is to support reversible devices. Reversible devices do not have a lower bound of energy consumption as indicated by Landauer principle [21]. However CMOS devices still has two orders [18] space to optimize regarding this lower bound. The main contribtion of our work is breaking the information barrier between machine learning community and reversible programming community, providing yet another strong motivating to develop a reversible programming. Our eDSL borrows the design of reversible control flow in the Janus, meanwhile provides multiple dispatch based abstraction. With these additional features, differentiating over a general program requires less than 100 lines. Our AD engine generates native julia code, is compatible with current Julia language. Potensial applications includes

1. AD rule generation,
2. control problem in robotics [22],
3. differentiating over reversible integrators [23],

In this paper, we first introduce the design of this eDSL in Sec. II and the back propagation of Jacobians and Hessians on this DSL in Sec. IV. In Sec. V, we introduce several examples. In Sec. **??**, we discuss on several important issues, like how time space tradeoff works, reversible instructions and hardwares and finally an outlook to some open problems to be solved.

II. REVERSIBLE DOMAIN SPECIFIC LANGUAGE

In a mordern programming language, functions are pushed to a global stack for scheduling. The memory layout of a function is consisted of input arguments, a function frame with informations like return address and saved memory segments, local variables and working stack. After each call, the function clears the input arguments, function frame, local variables and working stack and only stores the return value. In the invertible programming style, this kind of design pattern is nolonger the best practise, input variables can not be easily discarded after a function call, since discarding information may ruin reversibility. Hence, a reversible instruction/function call in NiLang is a mapping between a same set of symbols.

NiLang is a reversible eDSL in Julia that simulates reversible Turing machine without actual hardware or even instruction level support. The grammar is shown in Appendix A. A function definition is composed of function head and statements, the interprecation of a reversible function to native Julia language consists three stages. The first stage preprocess human input to a reversible IR. It checks human input and removes redundancy in grammar. To be specific

- adds missing `@deanc` to make sure `@anc` and `@deanc` statements appear in pairs,
- expand `@routine` macro,
- expand the symbol in postcondition field as precondition.

Here, the macro `@anc <a> = <expr>` binds `<a>` to an initial value specified by `<expr>`, while `@deanc <a> = <expr>` deallocates the variable `<a>`, before doing that, it checks the variable is restored to its initial value. This underlines the difference between irreversible assign statements and reversible ancilla statements. `@anc` and `@deanc` always appear in pairs inside a function call, a while statement or a for statement. `@deanc` can be added automatically if not given. Similar designs in Janus and R are `local/delocal` statement and `let` statement.

`@routine <name> Stmt` is a statements recorder, it records a statement to variable `<name>`. When `@routine <name>` or `@routine <name>` is called, the statement or inverse statement is inserted to that position. A condition expression contains two parts, the precondition and postcondition, if the precondition is not changed, we can use the precondition as postcondition by inserting ~ in the postcondition field.

The second stage transforms this reversible IR to its reversed version.

statement	**reverse**
`<f>(<args>)`	`(~<f>)(<args>)`
`<out!> += <f>(<args>)`	`<out!> -= <f>(<args>)`
`<out!> .+= <f>.(<args>)`	`<out!> .-= <f>.(<args>)`
`<out!> ⊻= <f>(<args>)`	`<out!> ⊻= <f>(<args>)`
`<out!> .⊻= <f>(<args>)`	`<out!> .⊻= <f>(<args>)`
`@anc <a> = <expr>`	`@deanc <a> = <expr>`

`begin`
`  <Stmts>`
`end`
	`begin`
	`  ~(<Stmts>)`
	`end`

`if (<pre>, <post>)`
`  <Stmts>`
`else`
`  <Stmts>`
`end`
	`if (<post>, <pre>)`
	`  ~(<Stmts>)`
	`else`
	`  ~(<Stmts>)`
	`end`

`while (<pre>, <post>)`
`  <Stmts>`
`end`
	`while (<post>, <pre>)`
	`  ~(<Stmts>)`
	`end`

`for <i>=<m>:<s>:<n>`
`  <Stmts>`
`end`
	`for <i>=<m>:-<s>:<n>`
	`  ~(<Stmts>)`
	`end`

`@safe <expr>`	`@safe <expr>`

Table I. A collection of reversible statements.

A condition expression is a two-element tuple, it also allows user putting additional postcondition in control flows to help reverse the program. A postcondition is a boolean expression that evaluated after the controlled body being executed. The `@safe` macro can be followed by an arbituary statement, it

allows user to use external statements that does not break reversibility. `@safe @show var` is often used for debugging.

The third stage is translating this reversible IR to native Julia code. It

- adds `@instr` before each instruction and function call statement,
- attach a return statement after the function definition which returns the modified input variables as the output,
- adds statements to check the consistency between preconditions and postconditions to ensure reversibility,
- compile the inverse function at the same time.

The macro `@instr` assign the output of a function to the argument list of a function. Hence, the values are changed while the symbol table is not changed. Here, the statement `out! += x * y` calls instruction `⊕(*)(out!, x, y)`, which means accumulate a product of two variables to target symbol. Combine it with `@instr` that assigns each output to each input, we simulate a mutable operations.

A. Types and dataviews

A constructor is also a reversible function, it changes the behavior of data or derive a new field from a data. Its reverse function is a "destructor", a destructor does not deallocate memory directly but unpacks data.

Before introducing dataviews, let's first consider the following example

```
grad(arr[3].value) += x * y
```

In Julia, this statement will raise a syntax error, since a function call can not be assigned. Meanwhile `arr[3]` might be a immutable type. In our eDSL, assigning a single argument function call or a immutable type is allowed.

In our interpreter, `grad(arr[3].value) += x * y` is translated to `@instr grad(arr[3].value) += x * y` at the thrid stage. To execute the instruction, `@instr` translate the statement to

```
res = (*)(grad(arr[3].value), x, y)
arr[3] = chfield(arr[1], Val(:value),
    chfield(arr[3].value, grad, res[1]))
x = res[2]
y = res[3]
```

`⊕(*)(grad(arr[3].value), x, y)` computes the output, which is a tuple of length 3. `chfield` is used to infer the new data by modifying a field or a bijective mapping of a field, it returns a new object. The second and third arguments can be assigned back directly. A dataview of a data can be data itself, a field of its view, an array element of its view, or a bijective mapping of its view. If the default constructor of a type is not overwritten by user, NiLang can modify a field of that type automatically. For a bijective mapping of a field, user need to specify the behavior of a dataview by overloading `chfield` function.

III. TRAINING BY CONSISTENCY

We compute the output as

$$y = f(x), \tag{1}$$

and we have an expected output $\hat{y}$. In traditional machine learning, we define a loss and minimize it. However, this is not how human brain works. [**?**] Since the $\hat{y}$ is different with y, the network start to "think", if the output is $\hat{y}$, what should the input (including network parameters) be? So, we feed $\hat{y}$ back to the output end of network, and inverse the tape, with the runtime information generated by the input image. The network will compute a different value to network parameters, then the network feel "confused", and chose a value between old and new values.

Let's look at the following example

```
@i function f1(out!, x, y)
    y += identity(x)
    out! -= exp(x)
    out! += exp(y)
end

@i function f2(out!, x, y)
    y += identity(x)
    out! -= exp(x)
    x -= log(-out!)
    out! += exp(y)
end

function train(f)
    loss = Float64[]
    y = 1.6
    for i=1:100
        out!, x = 0.0, 0.3
        @instr f(out!, x, y)
        push!(loss, out!)
        out! = 1.0
        @instr (~f)(out!, x, y)
    end
    loss
end
```

Loss function `f1` and `f2` computes $f(x,y) = e^{(y+x)} - e^x$ and stores the output in a new memory `out!`, the target is to find a y that make the output x equal to the target value 1. After 200 steps training, y runs into the fixed point and x is equal to 1 upto machine precision.

This training is similar the recursive convergence method that widely used in mathematics and physics. However, in

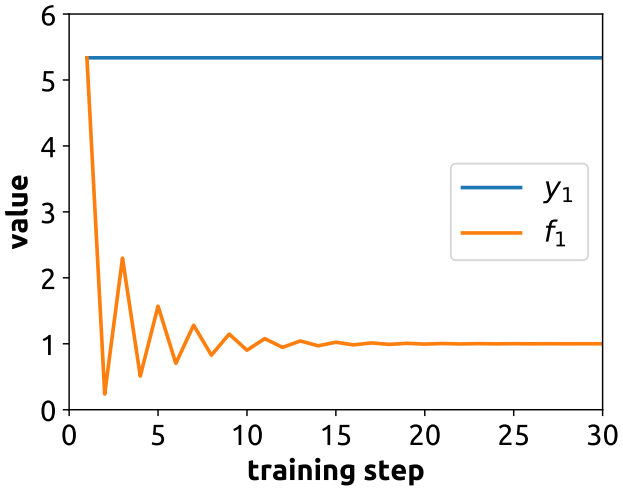

Figure 1. Target function value as a function of self-consistent training step.

programming language but it has a loophole, it is vulnarable to loss of injectivity. For example, if the result is accumulated to another variable $z+ = x$, and take z as the loss, then the loss can not accumulated to the target value correctly since the inverse of above operation is $z- = x$ and x is unchanged in the inverse run. This is because the redundancy introduced in operation $z = z + e^x$, which keeps both x and z, the change of z shifting z directly is apparently a lazier approach to fit z with target value 1, to figure out the correct causality, corresponding change in output side to x and y are required. This redundancy can be detected by computing the entanglement entropy (or mutual information) between two output datas. We call this type of invertibility inference, which should be avoided in this type of training. Then how shall we calculate $f(x) = e^x$ like functions, which are not partly invertible from the numeric perspective? Here we introduce the notion of "weak invertibility"

$$@asserty == 0 \tag{2}$$
$$\texttt{fl}(y)+ = exp(\texttt{fl}(x)) \tag{3}$$
$$\texttt{fl}(x)- = log(y) \tag{4}$$

We can keep x to ensure numerical invertibility, meanwhile erase most informations inside it. Since x is effectively 0, thus no redundancy (or entanglement entropy) in output, in this way, the above training is still valid.

A. Compactness of data

Measure of compactness by entanglement entropy. In invertible programming, we define the entanglement between output space garbage space as the compactness of information.

If the output information is compact, then we have $(x, x_a) \to (y, y_a)$ mutual information is

$$I(x, y) = S(x, y) - S(x) - S(y) \tag{5}$$
$$S(x, x_a) = S(y, y_a) \tag{6}$$
$$I(x, x_a) = 0 \tag{7}$$
$$I(y, y_a) = S(y, y_a) - S(y_a) \tag{8}$$

B. Traditional Invertible Computing Devices

It is possible to design and fabricate a fully reversible processor using resources which are comparable to a conventional microprocessor. Existing CAD tools and silicon technology are sufficient for reversible computer design.

Like quantum computer, it should be able to reset to qubits 0. The reset operation, sometimes can be expensive. In a quantum device, this reset operation can be done by dissipation and dicoherence (i.e. interact with environment and get thermalized). Alternatively, a immediate feedback loop can be introduced to a quantum device, where a classical computer is regarded as the dissipation source.

IV. TAYLOR PROPAGATION ON A REVERSIBLE TURING MACHINE

Taylor propagation is exponentially (as the order) more efficient in obtaining higher order gradients than differentiating lower order gradients recursively. The later requires traversing the computational graph repeatedly. In JAX, in order to support Taylor propagation, the propagation rules for part of primitives manually defined. The exhaused support requires much more effort than the first order gradient propagation. Instruction level automatic differentiation is more flexible in obtaining higher order gradients like Hessian.

A. First order gradient

Given a node $\vec{y} = f(\vec{x})$ in a computational graph, we can propagate the Jacobians in tangent mode like

$$J^O_{x_i} = J^O_{y_j} J^{y_j}_{x_i} \tag{9}$$

and the adjoint mode

$$J^{y_j}_I = J^{y_j}_{x_i} J^{x_i}_I \tag{10}$$

Here, I is the inputs and O is the outputs. Einstein's notation is used so that duplicated indices are summed over. Tagent mode instruction level automatic differentiation can be implemented easily in a irreversible language with dual numbers, here we focus on the adjoint mode. The backward rule can be described in tensor network language as shown in Fig. 2.

In reversible programming, we have the following implementation

1. The program runs forward and computes outputs,

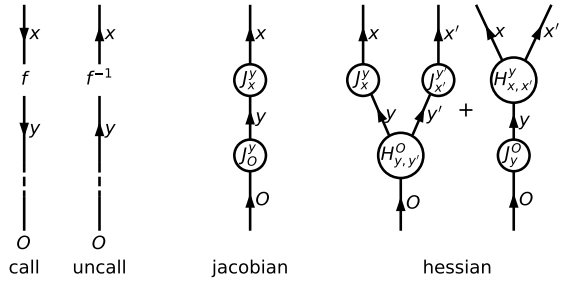

Figure 2. Adjoint rules for Jacobians and Hessians in tensor network language.

2. call constructor `GVar` that transfer a floating point number type to a `GVar` type.
3. Add 1 to the gradient field of the loss variable,
4. The program uncomputes outputs and recover all inputs, gradients are the `grad` fields of input variables.

Here, `GVar` is a "reversible type". `GVar(x)` wraps a variable into a GVar, which attaches a zero gradient field to a variable just like the dual number in tangent mode automatic differentiation. Its inverse *GVar* deallocate the gradient field safely and returns its value field. Here, "safely" means it will check the gradient field to make sure it is in 0 state. When a `instruct` meets a `GVar`, besides computing its value field $value(\vec{x}) \leftarrow f^{-1}(value(\vec{y}))$, it also updates the gradient field $grad(\vec{x}) = f^{-1}(value(\vec{x}), grad(\vec{y}))$. Since f^{-1} is bijective, $J^{\vec{y}}_{\vec{x}}$ can be easily obtained by inspecting its inverse function f. The final output is stored in the gradient field, when then gradient is not used anymore, the faithful reversible programming style to compute gradients would be uncomputing the whole process to obtain gradient, which increases the hyrachy by 1. Whenever the hyrachy increase by 1, the computational overhead doubles comparing with its irreversible counter part.

B. Second order gradient

The second order gradient can also be back propagated

$$\begin{aligned} H^{f}_{y_L,y'_L} &= 0 \\ H^{f}_{y_{i-1},y'_{i-1}} &= J^{y_i}_{y_{i-1}} H^{f}_{y_i,y'_i} J^{y'_i}_{y'_{i-1}} + J^{f}_{y_i} H^{y_i}_{y_{i-1},y'_{i-1}} \end{aligned} \tag{11}$$

In tensor network language, it can be represented as in Fig. 2. This approach can be easily extended to higher orders, or taylor propagation. However, this is not the widely adopted approach to compute higher order gradients. Althought backpropagating higher order gradients directly is exponentially faster than back propagating the computational graph for computing lower order gradients for computing higher order gradients, one has to extending the backward rules for each primitive rather than reusing existing ones. Here, we emphasis that with instruction level AD, rewritting backward rules for primitives turns out to be not so difficult.

C. Gradient on ancilla problem

Ancilla can also carry gradients during computation, sometimes these gradients can not be uncomputed even if their parents can be uncomputed regoriously. In these case, we simply "drop" the gradient field instead of raising an error. In this subsection, we prove doing this is safe, i.e. does not have effect on rest parts of program.

Ancilla is the fixed point of a function, which means

$$\begin{aligned} &b, y \leftarrow f(x, a) \text{ , where } b == a \\ &\frac{\partial b}{\partial x} = 0 \end{aligned} \tag{12}$$

During the computation, the gradient field does not have any effect to the `value` field of variables. The key question is will the loss of gradient part in ancilla affect the reversibility of the gradient part of argument variables. The gradient of argument variable is defined as $\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y}\frac{\partial y}{\partial x} + \frac{\partial L}{\partial b}\frac{\partial b}{\partial x}$, where the second term vanish naturally.

D. Implementation

The automatic differentiation engine is so short that we present the function defintion as follows

```
@i function (g::Grad)(args...; kwargs...)
    @safe @assert count(x -> x isa Loss, args) == 1
    @anc iloss = 0
    @routine getiloss begin
        for i=1:length(args)
            if (tget(args,i) isa Loss, iloss==i)
                iloss += identity(i)
                (~Loss)(tget(args,i))
            end
        end
    end

    g.f(args...; kwargs...)
    GVar.(args)
    grad(tget(args,iloss)) += identity(1.0)
    (~g.f)(args...; kwargs...)

    ~@routine getiloss
end
```

The program first checks the input parameters and locate the loss variable. Then `Loss` unwraps the loss variable, the location of loss variable is transferred to the ancilla `iloss` of integer type.

`GVar`

```
julia> using NiLang, NiLang.AD

julia> x, y = GVar(0.5), GVar(0.6)
(GVar(0.5, 0.0), GVar(0.6, 0.0))

julia> @instr grad(x) += identity(1.0)

julia> @instr x += identity(y)

julia> y
GVar(0.6, -1.0)

julia> @instr grad(x) -= identity(1.0)

julia> @instr (~GVar)(x)

julia> x
1.1
```

Broadcasting is supported. To avoid possible confusing, tuple indexing is forbidden delebrately, one can use `tget(tuple, 2)` to get the second element of a tuple.

V. EXAMPLES

A. Computing Fibonacci Numbers

An example that everyone likes

```
@i function rfib(out, n::T) where T
    @anc n1 = zero(T)
    @anc n2 = zero(T)
    @routine init begin
        n1 += identity(n)
        n1 -= identity(1.0)
        n2 += identity(n)
        n2 -= identity(2.0)
    end
    if (value(n) <= 2, ~)
        out += identity(1.0)
    else
        rfib(out, n1)
        rfib(out, n2)
    end
    ~@routine init
end
```

To compute the first Fibonacci number that is greater or equal to 100

```
@i function rfib100(n)
    @safe @assert n == 0
    while (fib(n) < 100, n != 0)
        n += identity(1.0)
    end
end
```

Here, `fib` and `rfib` are defined in Appendix V A. The reversible `while` statement contains two statements, the precondition and postcondition. Before entering the while statement, the program check the postcondition to make sure it is false. After each iteration, postcondition returns true. The inverse function exchanges the precondition and postcondition so that the repeatition of loop body is not changed.

B. exp function

An exp function can be computed using Taylor expansion

$$\text{out!+} = \sum_n \frac{x^n}{n!} \tag{13}$$

At a first glance, this is a recursive algorithm that mimics pebble game. Define the term for accumulation $s_n \equiv \frac{x^n}{n!}$, the recursion relation is written as $s_n = \frac{x s_{n-1}}{n}$. There is no known constant memory algorihm to pebble game. Notice that by viewing $*=$ and $/=$ a pair of reversible operations, we can uncompute $s_{n-1} = \frac{n s_n}{x}$ to deallocate memory. By allowing loss of several digit precision of result, implementing the constant memory reversible exp function is possible

```
using NiLang, NiLang.AD

@i function iexp(out!, x::T; atol::Float64=1e-14)
        where T
    @anc anc1 = zero(T)
    @anc anc2 = zero(T)
    @anc anc3 = zero(T)
    @anc iplus = 0
    @anc expout = zero(T)

    out! += identity(1.0)
    @routine r1 begin
        anc1 += identity(1.0)
        while (value(anc1) > atol, iplus != 0)
            iplus += identity(1)
            anc2 += anc1 * x
            anc3 += anc2 / iplus
            expout += identity(anc3)
            # speudo inverse
            anc1 -= anc2 / x
            anc2 -= anc3 * iplus
            SWAP(anc1, anc3)
        end
    end

    out! += identity(expout)

    ~@routine r1
end
```

The definition of SWAP instruction can be found in Appendix B. The two lines bellow the comment `# pseudo inverse` "uncompute" variables `anc1` and `anc2` to avalue very close to zero. Although the final output is not exact due to the rounding error, the reversibility is not harmed. Rouding error in output is not as toxical as that in ancilla, in the latter case, error accumulates in the whole program. In the second for loop inside the inverse notation `~`, we uncompute all ancilla bits rigorously. The `while` statement takes two conditions, the precondition and postconditoin. Precondition `val(anc1) > atol` indicates when to break the forward pass and post condition `!isapprox(iplus, 0.0)` indicates when to break the backward pass.

The appendix gives another example of QR decomposition.

To obtain the gradient, one can wrap the loss with `Loss` type.

```
julia> out!, x = 0.0, 1.6
(0.0, 1.6)

julia> @instr iexp'(Loss(out!), x)

julia> grad(x)
4.9530324244260555

julia> out!, x = 0.0, 1.6
(0.0, 1.6)

julia> @instr iexp''(Loss(out!), x)

julia> collect_hessian()
2×2 Array{Float64,2}:
 0.0  0.0
 0.0  4.95303
```

`iexp'` returns an object of type `Grad{typeof(exp)}`, it is a reversible function. Due to the non-locality of Hessians, we use a global tape for Hessian propagation. Whenever a new variable is created, the tape allocates a larger ring. It is global to ease the memory allocations of ancillas, the Hessian in ancilla is important to reach correct result.

```
julia> out!, x = 0.0, 1.6
(0.0, 1.6)

julia> @instr iexp''(Loss(out!), x)

julia> collect_hessian()
2×2 Array{Float64,2}:
 0.0  0.0
 0.0  4.95303
```

The final result can be obtained by calling a global function `collect_hessian`.

C. QR decomposition

Not only simple functions, linear algebra functions

```
@i function iqr(Q, R, A::AbstractMatrix{T}) where T
    @anc anc_norm = zero(T)
    @anc anc_dot = zeros(T, size(A,2))
    @anc ri = zeros(T, size(A,1))
    for col = 1:size(A, 1)
        ri .+= identity.(A[:,col])
        for precol = 1:col-1
            idot(anc_dot[precol], Q[:,precol], ri)

     R[precol,col] += identity(anc_dot[precol])
            for row = 1:size(Q,1)

     ri[row] -= anc_dot[precol] * Q[row, precol]
            end
        end
        inorm2(anc_norm, ri)

        R[col, col] += anc_norm^0.5
        for row = 1:size(Q,1)
            Q[row,col] += ri[row] / R[col, col]
        end

        ~(ri .+= identity.(A[:,col]);
        for precol = 1:col-1
            idot(anc_dot[precol], Q[:,precol], ri)
            for row = 1:size(Q,1)

     ri[row] -= anc_dot[precol] * Q[row, precol]
            end
        end;
        inorm2(anc_norm, ri)
        )
    end
end
```

where `idot` and `inorm2` are implemented as

```
@i function idot(out, v1::AbstractVector{T}, v2)
        where T
    @anc anc1 = zero(T)
    for i = 1:length(v1)
        anc1 += identity(v1[i])
        CONJ(anc1)
        out += v1[i]*v2[i]
        CONJ(anc1)
        anc1 -= identity(v1[i])
    end
end

@i function inorm2(out, vec::AbstractVector{T})
        where T
    @anc anc1 = zero(T)
    for i = 1:length(vec)
        anc1 += identity(vec[i])
        CONJ(anc1)
        out += anc1*vec[i]
        CONJ(anc1)
        anc1 -= identity(vec[i])
    end
end
```

One can easily check the gradient of this naive implementation of QR decomposition is correct

```
using Test
A = randn(4,4)
q = zero(A)
r = zero(A)

@i function test1(out, q, r, A)
    iqr(q, r, A)
    out += identity(q[1,2])
end

@i function test2(out, q, r, A)
    iqr(q, r, A)
    out += identity(r[1,2])
end

@test check_grad(test1, (Loss(0.0), q, r, A);
        atol=0.05, verbose=true)
@test check_grad(test2, (Loss(0.0), q, r, A);
        atol=0.05, verbose=true)
```

Here, the `check_grad` function is a gradient checker function defined in `NiLangCore.ADCore`.

D. Unitary Matrices

Recurrent networks with a unitary parametrized network ease the gradient exploding and vanishing problem [24–26]. Among different parametrization schemes, the most elegant one is [26], which parametrized the unitary matrix with two-level unitary operations, any unitary matrix of size $N \times N$ can be parametrized by $k = N(N-1)/2$ two level unitary matrices [27]. All these two-level unitary matrices can be applied in $O(1)$ time as a two register instruction. Hence a real unitary matrix can be parametrized compactly by k rotation angles, each represents a rotation operations between datas in two target parameters.

```
@i function umm!(x, , Nin::Int, Nout::Int)
    @anc k = 0
    for j=1:Nout
        for i=Nin-1:-1:j
            k += identity(1)
            ROT(x[i], x[i+1], [k])
        end
    end

    # uncompute k
    for j=1:Nout
        for i=Nin-1:-1:j
            k -= identity(1)
        end
    end
end
```

`ROT(a!, b!, θ)` is the rotation instruction, θ is the rotation angle, which represents rotating data in `a!` and `b!` by an

angle of θ.

$$R(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{14}$$

Its backward rule of a `ROT` instruction is

$$\begin{aligned} \bar{\theta} &= \sum \frac{\partial R(\theta)}{\partial \theta} \odot (\bar{y} x^T) \\ &= \mathrm{Tr}\left[\frac{\partial R(\theta)}{\partial \theta}^T \bar{y} x^T \right] \\ &= \mathrm{Tr}\left[R\left(\frac{\pi}{2} - \theta\right) \bar{y} x^T \right] \end{aligned} \tag{15}$$

We bind the adjoint function of `ROT` to its reverse `IROT`, and define a new function that dispatch to `GVar` variables

```
@i function IROT(a!::GVar, b!::GVar, ::GVar)
    IROT(value(a!), value(b!), value())
    NEG(value())
    value() -= identity(/2)
    ROT(grad(a!), grad(b!), value())
    grad() += value(a!) * grad(a!)
    grad() += value(b!) * grad(b!)
    value() += identity(/2)
    NEG(value())
    ROT(grad(a!), grad(b!), /2)
end
```

VI. DISCUSSION AND OUTLOOK

This could be used to reduce the momory cost in normalizing flow, time reversible integrator, recurrent neural network and residual neural network.

A. Time Space Tradeoff

So far, we have introduced the eDSL. There are many other designs of reversible language and instruction set. The reversible Turing machine may have either a space overhead propotional to computing time T or a computational overhead that sometimes can be even exponential given limited space. In the simplest g-segment trade off scheme [28, 29], it takes $Time(T) = T^{\log_g(4g-2)}$ and $Space(T) = (g-1)S \log_g T$. This section, we try to convince the reader that the overhead of reversible computing is not as terrible as people thought.

First, one should notice that even in the worst case, the overhead of a reversible program is not more than a traditional machine learning package. In pytorch, a tensor memorize every input of a primitive. The program is apparently reversible since it does not discard any information. For deep neural networks, people used checkpointing trick to trade time with space [30], which is also a widely used trick in reversible programming [17]. Reversible programming AD is sometimes more memory efficient. Comparing with logging computational graph, reversible programming has the advantage of supporting inplace functions, which is difficult is both tranditional and source to source AD framework. The example of parametrizing unitary matrice costs zero memory allocation.

Second, some computational overhead of running recursive algorithms with limited space resources can be mitigated by "pseudo uncomputing" without sacrifising reversibility like in the `iexp` example.

Third, making reversible programming an eDSL rather than a independant language allows flexible choices between reversibility and computational overhead. For example, in order to deallocate the gradient memory in a reversible language one has to uncompute the whole process of obtaining this gradient. In this eDSL, we can just deallocate the memory irreversibly, i.e. trade energy with time. This underlines the fact that a reversible program is more suited in a program with small granularity. We can quantify it by introducing

***Definition* 1** (program granularity)**.** The logarithm of the ratio between the execution time of a reversible program and its irreversible counter part.

$$\log_2 \frac{Time(T)}{T} \tag{16}$$

In the lowest granuality, instruction design, we need ancilla bits. Defining primitive functions like `iexp` requires uncomputing ancillas. Deallocating the gradient further increase the granularity by ~ 1. After the training, the inverse training of whole programs should be done to deallocate all the memory used for training. As a result, the program complexity increase exponentially as the granuality increase. The granularity can be decreased by flattening the functions, since the uncomputing of ancillas can be executed at any level of granularity.

One should notice the memory advantage of reversible programming to machine learning does comes from reversibility itself, but from a better data tracking strategy inspired from invertible programming. Normally, a reversible program is not as memory efficient as its irreversible couterpart due to the additional requirement of no information loss. A naive approach that keeping track of all informations will cost an additional space $O(T)$, where T stands for the excution time in a irreversible TM, the longer the program runs, the larger the memory usage is. This is exactly the approach to keeping reversibility in most machine learning packages in the market. The point it, an reversible Turing Machine is able to trade space with time. In some cases, it may cause polynomial overhead than its irreversible counterpart.

B. Instructions and Hardwares

So far, our eDSL is not really compiled to instructions, instead it runs on a irreversible host Julia. In the future, it can be compiled to low level instructions and is executable on a reversible devices. For example, the control flow defined

in this NiLang can be compiled to reversible instructions like conditioned `goto` statement. It is designed in such a way that the target instruction is a `comefrom` statement which specifies the postcondition. [20]

Arithmetic instructions should be redesigned to support better reversible programs. The major obstacle to exact reversibility programming is current floating point adders used in our computing devices are not exactly reversible. There are proposals of reversible floating point adders [31, 32] that introduces garbage bits to ensure reversibility. In other words, to represent a 64 bit floating point number requires more than 64 bits in storage. Reversible multiplier is also possible in similar approach. [33] With these infrastructure, a reversibile program can be executed without suffereing from the irreversibility from rounding error. In machine learning field, people using information buffer in multiplication operations [14] in an approach to enforce invertibility in a memory efficient way.

Reversible programming is not nessesarily related to reversible hardwares, reversible programs is a subset of irreversible programs hence can be simulated efficiently with CMOS technologies [20]. However, only using reversible hardwares [] can break the energy efficiency barrier by Landauer principle. Reversible hardwares are not nessesarily related to reversible gates such as Toffoli gate and Fredkin gate. Devices with the ability of recovering signal energy is able able to save energy, which is known as generalized reversible computing. [34, 35] In the following, we comment briefly on a special type of reversible device Quantum computer.

C. Quantum Computers

One of the fundamental difficulty of building a quantum computer is, unlike a classical state, an unknown quantum state can not be copied. Quantum random access memory [] is very hard to design and implement, it is known to have many caveats [**?**]. A quantum state in a environment will decoherence and can not be recovered, this underlines the simulation nature of quantum devices. Reversible computing does not enjoy the quantum advantage from entanglement, nor the quantum disadvantages from non-cloning and decoherence. Only the limitation of reversibility is retained, reversibility comes from the fact that micro scopic processes are all unitary. In microscopic world, irreversibility is rare, it can come from the interaction with classical devices, like environment induces decaying, qubit state resetting, measurements and classical feedbacks to quantum devices. These are typically harder to implement on a quantum device.

Given the fundamental limitations of quantum decoherence and non-cloning and the microscopic reversible nature. It is reasonable to have a reversible computing device to bridge the gap between classical and universal quantum computing. By introducing entanglement little by little, we can accelerate some basic components like a reversible adder. The quantum fourier transformation provides a shortcut to the adder bt introducing only one additional gate, the CPHASE gate even though it is a classical in classical out algorithm. [] Interestingly, by introducing rotation gates $Ry(\theta)$ and $Rz(\theta)$ in the reversible programming, we make NiLang a path integral based universal quantum simulator as shown in Appendix **??**. The compiling theory developed for reversible programming will have profounding effect to quantum computers.

D. Outlook

So far NiLang is not full ready for productivility. It can be improved from multiple perspectives, compiling support to merge the uncomputing can decrease granularity and hence reduce overhead. Additional instructions like stack operations and logical operations are under consideration. It is also interesting to see how it can be combined with a high performance quantum simulator like `Yao`. It is able to provide control flow to `Yao`'s QBIR. By porting a quantum simulator. it is interesting to see how quantum simulator can improve the instruction design. Notice a quantum fourier transformation (QFT) based quantum adder and multiplier is sometimes more efficient than a classical adder [36] [**JG: Is this true?**]. Reversible programming is known to have the advantage in parallel computing [37] and debugging [38], it is interesting to see how it combines with other parts of Julia packages like CUDAnative [39] and Debugger, it would be interesting to see our eDSL running on a GPU with little synchronization overhead [**JG: Is this even possible?**] and using an interative debugging with bidirectional move.

VII. ACKNOWLEDGMENTS

Jin-Guo Liu thank Lei Wang for motivating the project with possible applications reversible integrator, normalizing flow and neural ODE. Xiu-Zhe Luo for discussion on the implementation details of source to source automatic differetiation, Shuo-Hui Li for helpful discussion on differential geometry. Damian Steiger for telling me the `comefrom` joke. Tong Liu and An-Qi Chen for helpful discussion on quantum adders and multipliers. The authors are supported by the National Natural Science Foundation of China under the Grant No. 11774398, the Strategic Priority Research Program of Chinese Academy of Sciences Grant No. XDB28000000 and the research funding from Huawei Technologies under the Grant No. YBN2018095185.

[1] L. Hascoet and V. Pascual, ACM Transactions on Mathematical Software (TOMS) **39**, 20 (2013).

[2] J. Revels, M. Lubin, and T. Papamarkou, "Forward-mode automatic differentiation in julia," (2016), arXiv:1607.07892

[cs.MS].
[3] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, in *NIPS Autodiff Workshop* (2017).
[4] M. Innes, E. Saba, K. Fischer, D. Gandhi, M. C. Rudilosso, N. M. Joy, T. Karmali, A. Pal, and V. Shah, "Fashionable modelling with flux," (2018), arXiv:1811.01457 [cs.PL].
[5] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," (2015), software available from tensorflow.org.
[6] M. Innes, A. Edelman, K. Fischer, C. Rackauckas, E. Saba, V. B. Shah, and W. Tebbutt, CoRR **abs/1907.07587** (2019), arXiv:1907.07587.
[7] M. Seeger, A. Hetzel, Z. Dai, E. Meissner, and N. D. Lawrence, "Auto-differentiating linear algebra," (2017), arXiv:1710.08717 [cs.MS].
[8] Z.-Q. Wan and S.-X. Zhang, "Automatic differentiation for complex valued svd," (2019), arXiv:1909.02659 [math.NA].
[9] C. Hubig, "Use and implementation of autodifferentiation in tensor network methods with complex scalars," (2019), arXiv:1907.13422 [cond-mat.str-el].
[10] H.-J. Liao, J.-G. Liu, L. Wang, and T. Xiang, Physical Review X **9** (2019), 10.1103/physrevx.9.031041.
[11] "Breaking the Memory Wall: The AI Bottleneck," https://blog.semi.org/semi-news/breaking-the-memory-wall-the-ai-bottleneck.
[12] J. Bettencourt, M. J. Johnson, and D. Duvenaud, (2019).
[13] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, SIAM review **59**, 65 (2017).
[14] D. Maclaurin, D. Duvenaud, and R. Adams, in *Proceedings of the 32nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 37, edited by F. Bach and D. Blei (PMLR, Lille, France, 2015) pp. 2113–2122.
[15] M. MacKay, P. Vicol, J. Ba, and R. B. Grosse, in *Advances in Neural Information Processing Systems 31*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Curran Associates, Inc., 2018) pp. 9029–9040.
[16] J. Behrmann, D. Duvenaud, and J. Jacobsen, CoRR **abs/1811.00995** (2018), arXiv:1811.00995.
[17] K. S. Perumalla, *Introduction to reversible computing* (Chapman and Hall/CRC, 2013).
[18] M. P. Frank, IEEE Spectrum **54**, 3237 (2017).
[19] C. Lutz, "Janus: a time-reversible language," (1986), *Letter to R. Landauer*.
[20] C. J. Vieri, *Reversible Computer Engineering and Architecture*, Ph.D. thesis, Cambridge, MA, USA (1999), aAI0800892.
[21] R. Landauer, IBM journal of research and development **5**, 183 (1961).
[22] M. Giftthaler, M. Neunert, M. Stäuble, M. Frigerio, C. Semini, and J. Buchli, Advanced Robotics **31**, 12251237 (2017).
[23] D. N. Laikov, Theoretical Chemistry Accounts **137** (2018), 10.1007/s00214-018-2344-7.
[24] M. Arjovsky, A. Shah, and Y. Bengio, CoRR **abs/1511.06464** (2015), arXiv:1511.06464.
[25] S. Wisdom, T. Powers, J. R. Hershey, J. L. Roux, and L. Atlas, "Full-capacity unitary recurrent neural networks," (2016), arXiv:1611.00035 [stat.ML].
[26] L. Jing, Y. Shen, T. Dubcek, J. Peurifoy, S. A. Skirlo, M. Tegmark, and M. Soljacic, CoRR **abs/1612.05231** (2016), arXiv:1612.05231.
[27] C.-K. LI, R. ROBERTS, and X. YIN, International Journal of Quantum Information **11**, 1350015 (2013).
[28] C. H. Bennett, SIAM Journal on Computing **18**, 766 (1989), https://doi.org/10.1137/0218053.
[29] R. Y. Levine and A. T. Sherman, SIAM Journal on Computing **19**, 673 (1990).
[30] T. Chen, B. Xu, C. Zhang, and C. Guestrin, CoRR **abs/1604.06174** (2016), arXiv:1604.06174.
[31] M. Nachtigal, H. Thapliyal, and N. Ranganathan, in *2011 11th IEEE International Conference on Nanotechnology* (2011) pp. 451–456.
[32] T. D. Nguyen and R. V. Meter, "A space-efficient design for reversible floating point adder in quantum computing," (2013), arXiv:1306.3760 [quant-ph].
[33] M. Nachtigal, H. Thapliyal, and N. Ranganathan, in *10th IEEE International Conference on Nanotechnology* (2010) pp. 233–237.
[34] M. P. Frank, in *35th International Symposium on Multiple-Valued Logic (ISMVL'05)* (2005) pp. 168–185.
[35] M. P. Frank, in *Reversible Computation*, edited by I. Phillips and H. Rahaman (Springer International Publishing, Cham, 2017) pp. 19–34.
[36] T. Haener, M. Soeken, M. Roetteler, and K. M. Svore, Lecture Notes in Computer Science , 162174 (2018).
[37] D. R. Jefferson, ACM Transactions on Programming Languages and Systems (TOPLAS) **7**, 404 (1985).
[38] B. Boothe, ACM SIGPLAN Notices **35**, 299 (2000).
[39] T. Besard, C. Foket, and B. D. Sutter, CoRR **abs/1712.03112** (2017), arXiv:1712.03112.

Appendix A: NiLang Grammar

Terminologies

- *ident*, symbols
- *num*, numbers
- ϵ, empty statement
- *JuliaExpr*, native julia expression
- [ ], zero or one repetitions.

⟨Stmts⟩ ::= ϵ
| ⟨Stmt⟩
| ⟨Stmts⟩ ⟨Stmt⟩
⟨Stmt⟩ ::= ⟨BlockStmt⟩
| ⟨IfStmt⟩
| ⟨WhileStmt⟩
| ⟨ForStmt⟩
| ⟨InstrStmt⟩
| ⟨RevStmt⟩
| ⟨@anc⟩ ⟨Stmt⟩
| ⟨@routine⟩ ⟨Stmt⟩
| ⟨@safe⟩ *JuliaExpr*
| ⟨CallStmt⟩
⟨BlockStmt⟩ ::= *begin* ⟨Stmts⟩ *end*
⟨RevCond⟩ ::= *(JuliaExpr , JuliaExpr)*
⟨IfStmt⟩ ::= *if* ⟨RevCond⟩ ⟨Stmts⟩ [*else* ⟨Stmts⟩] *end*
⟨WhileStmt⟩ ::= *while* ⟨RevCond⟩ ⟨Stmts⟩ *end*
⟨Range⟩ ::= *JuliaExpr : JuliaExpr* [*: JuliaExpr*]
⟨ForStmt⟩ ::= *for ident* = ⟨Range⟩ ⟨Stmts⟩ *end*
⟨CallStmt⟩ ::= *JuliaExpr (* [⟨DataViews⟩] *)*
⟨Constant⟩ ::= *num* | π
⟨InstrBinOp⟩ ::= += | -= | ⊻=
⟨InstrTrailer⟩ ::= [.] *(* [⟨DataViews⟩] *)*
⟨InstrStmt⟩ ::= ⟨DataView⟩ ⟨InstrBinOp⟩ *ident* [⟨InstrTrailer⟩]
⟨RevStmt⟩ ::= ~ ⟨Stmt⟩
⟨@routine⟩ ::= @*routine ident* ⟨Stmt⟩
⟨AncArg⟩ ::= *ident = JuliaExpr*
⟨@anc⟩ ::= @*anc* ⟨AncArg⟩
| @*deanc* ⟨AncArg⟩
⟨@safe⟩ ::= @*safe JuliaExpr*
⟨DataViews⟩ ::= ϵ
| ⟨DataView⟩
| ⟨DataViews⟩ , ⟨DataView⟩
⟨DataView⟩ ::= ⟨DataView⟩ *[JuliaExpr]*
| ⟨DataView⟩ . *ident*
| *JuliaExpr (* ⟨DataView⟩ *)*
| ⟨Constant⟩
| *ident*

Dataview is a special bijective mapping of an object or a field (or item) of an object. The dataview can feedback to parent data with the `chfield` method, so that the modified object can generate desired dataview.

Appendix B: Instruction Table

Even though $\oplus(identity)$, $\oplus(*)$, $\oplus(/)$ and their reverse together with control flows are sufficient to write an arbituary differentiable program. For convinience we provide more,

$\text{SWAP}(a, b) \to b, a$
$\text{ROT}(a, b, \theta) \to a\cos\theta - b\sin\theta, b\cos\theta + a\sin\theta, \theta$
$\text{IROT}(a, b, \theta) \to a\cos\theta + b\sin\theta, b\cos\theta - a\sin\theta, \theta$
$y \mathrel{+}= a\hat{}b \to y + a^b, a, b$
$y \mathrel{+}= \exp(x) \to y + e^x, x$
$y \mathrel{+}= \log(x) \to y + \log x, x$
$y \mathrel{+}= \sin(x) \to y + \sin x, x$
$y \mathrel{+}= \cos(x) \to y + \cos x, x$
$y \mathrel{+}= \text{abs}(x) \to y + \lvert x\rvert, x$
$\text{NEG}(y) \to -y$
$\text{CONJ}(y) \to y'$

Table II. A collection of reversible instructions.

Appendix C: Julia based eDSL implementation details

Macro 'invfunc' defines a invertible function, ancillas are binded to a function, since ancillas are umcomputed to 0 at the end of call, so that it can be used repeatedly in a function, it is like a class variable in a class, with no side effects. When the *JuliaExpr* is a function all, it must be pure.

Some variables can be uncomputed to 0, but we choose not to for performance reason. For example `infer!(argmax, i, x)` which computes the location of maximum value in x and store the output to i, if we uncompute it, it doubles additional computational time. Here, we trade off the memory with computation time. As a result, we must feed `imax` to the function, so that this variable can be manipulated in outer scopes.