

Demystifying Differentiable Programming: Shift/Reset the Penultimate Backpropagator

FEI WANG, Purdue University, USA

XILUN WU, Purdue University, USA

GREGORY ESSERTEL, Purdue University, USA

JAMES DECKER, Purdue University, USA

TIARK ROMPF, Purdue University, USA

Deep learning has seen tremendous success over the past decade in computer vision, machine translation, and gameplay. This success rests in crucial ways on *gradient-descent optimization* and the ability to “learn” parameters of a neural network by backpropagating observed errors. However, neural network architectures are growing increasingly sophisticated and diverse, which motivates an emerging quest for even more general forms of *differentiable programming*, where arbitrary parameterized computations can be trained by gradient descent. In this paper, we take a fresh look at automatic differentiation (AD) techniques, and especially aim to demystify the *reverse-mode* form of AD that generalizes backpropagation in neural networks.

We uncover a tight connection between reverse-mode AD and delimited continuations, which permits implementing reverse-mode AD purely via operator overloading and without any auxiliary data structures. We further show how this formulation of AD can be fruitfully combined with multi-stage programming (staging), leading to a highly efficient implementation that combines the performance benefits of deep learning frameworks based on explicit reified computation graphs (e.g., TensorFlow) with the expressiveness of pure library approaches (e.g., PyTorch).

1 INTRODUCTION

Under the label *deep learning*, artificial neural networks have seen a remarkable renaissance over the last decade. After a series of rapid advances, they are now matching or surpassing human performance in computer vision, machine translation, and gameplay. Common to all these breakthroughs is the underlying dependency on optimization by gradient descent: a neural network “learns” by adjusting its parameters in a direction that minimizes the observed error on a training set. Hence, a crucial ability is that of backpropagating errors through the network to compute the gradient of its loss function [Rumelhart et al. 1986]. Beyond this commonality, however, deep learning architectures vary widely (see Figure 1). In fact, many of the practical successes are fueled by increasingly sophisticated and diverse network architectures that in many cases depart from the traditional organization into layers of artificial neurons. For this reason, prominent deep learning researchers have called for a paradigm shift from deep learning towards *differentiable programming*¹²—essentially, functional programming with a first-class gradient operator—based on the expectation that further advances in artificial intelligence will be enabled by the ability to “train” arbitrary parameterized computations by gradient descent.

Programming language designers, key players in this vision, are faced with the challenge of adding efficient and expressive program differentiation capabilities. Forms of automatic gradient

¹<http://colah.github.io/posts/2015-09-NN-Types-FP/>

²<https://www.facebook.com/yann.lecun/posts/10155003011462143>

computation that generalize the classic backpropagation algorithm are provided by all contemporary deep learning frameworks, including TensorFlow and PyTorch. These implementations, however, are ad hoc, and each framework comes with its own set of trade-offs and restrictions. In the academic world, automatic differentiation (AD) [Speelpenning 1980; Wengert 1964] is the subject of study of an entire community. Unfortunately, results disseminate only slowly between communities, and differences in terminology make typical descriptions of AD appear mysterious to PL researchers, especially those concerning the reverse-mode flavor of AD that generalizes backpropagation. A notable exception is the seminal work of Pearlmutter and Siskind [2008], which has cast AD in a functional programming framework and laid the groundwork for first-class, unrestricted, gradient operators in a functional language.

The goal of the present paper is to further demystify differentiable programming and AD for a PL audience, and to reconstruct the forward- and reverse-mode AD approaches based on well-understood program transformation techniques. We describe forward-mode AD as symbolic differentiation of ANF-transformed programs, and reverse-mode AD as a specific form of symbolic differentiation of CPS-transformed programs. In doing so, we uncover a deep connection between reverse-mode AD and delimited continuations.

In contrast to previous descriptions, this formulation suggests a novel view of reverse-mode AD as a purely local program transformation which can be realized entirely using operator overloading in a language that supports shift/reset [Danvy and Filinski 1990] or equivalent delimited control operators³. By contrast, previous descriptions require non-local program transformations to carefully manage auxiliary data structures (often called a *tape*, *trace*, or *Wengert-list* [Wengert 1964]), either represented explicitly, or in a refunctionalized form as in Pearlmutter and Siskind [2008].

We further show how this implementation can be combined with staging, using the LMS (Light-weight Modular Staging) framework [Rompf and Odersky 2010]. The result is a highly efficient and expressive DSL, dubbed Lantern, that reifies computation graphs at runtime in the style of TensorFlow [Abadi et al. 2016], but also supports unrestricted control flow in the style of PyTorch [Paszke et al. 2017a]. Thus, our approach combines the strengths of these systems without their respective weaknesses, and explains the essence of deep learning frameworks as the combination of two well-understood and orthogonal ideas: staging and delimited continuations.

The rest of this paper is organized around our contributions as follows:

- We derive forward-mode AD from high-school symbolic differentiation rules in an effort to provide accessibility to PL researchers at large (Section 2).
- We then present our reverse-mode AD transformation based on delimited continuations and contrast it with existing methods (Section 3).
- We combine our reverse-mode AD implementation in an orthogonal way with staging, removing interpretive overhead from differentiation (Section 4).
- We present Lantern, a deep learning DSL implemented using these techniques, and evaluate it on several case studies, including recurrent and convolutional neural networks, tree-recursive networks, and memory cells (Section 5).

Finally, Section 6 discusses related work, and Section 7 offers concluding thoughts.

³Our description reinforces the functional “Lambda, the ultimate backpropagator” view of Pearlmutter and Siskind [2008] with an alternative encoding based on delimited continuations, where control operators like *shift/reset* act as a powerful front-end over λ -terms in CPS — hence, as the “penultimate backpropagator”.

Demystifying Differentiable Programming: Shift/Reset the Penultimate Backpropagator

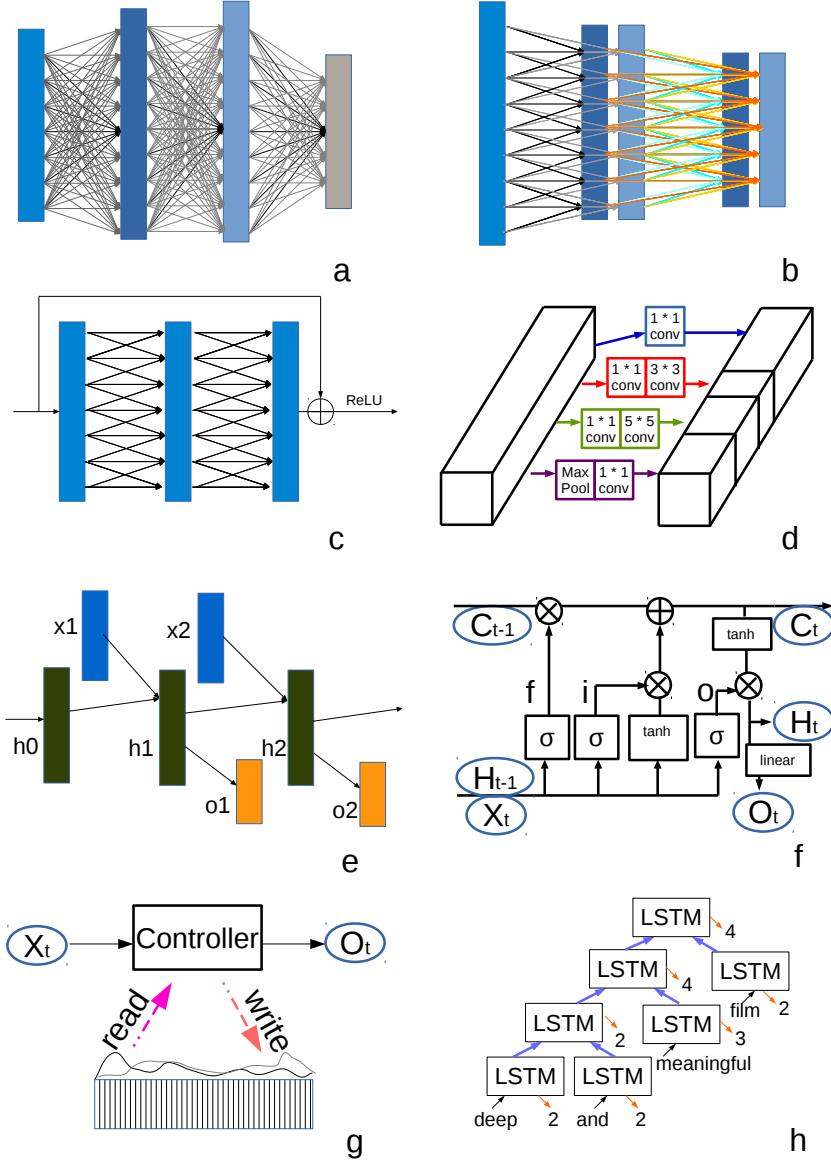


Fig. 1. Different deep learning models and components: (a) fully connected neural network (b) convolutional neural network (c) ResNet with identity mapping (d) Inception module (e) simple recurrent neural network (RNN) (f) Long short-term memory (LSTM) (g) neural Turing machine (h) tree-structured LSTM. See Section 5.1 for details.

2 DIFFERENTIABLE PROGRAMMING BASICS

Figure 1 shows various neural network architectures and modules that are in common use today. We postpone a detailed discussion of these architectures until Section 5.1, and instead first focus on the use of gradient-based optimization. Broadly speaking, a neural network is a special kind of parameterized function approximator \hat{f}_w . The training process optimizes the parameters w to improve the approximation of an uncomputable *ground truth* function f based on training data.

$$f : A \rightarrow B \quad \hat{f}_w : A \rightarrow B \quad w \in P$$

For training, we take observed input/output samples $(a, f(a)) \in A \times B$ and update w according to a *learning rule*. In typical cases where the functions f and \hat{f}_w are maps $\mathbb{R}^n \rightarrow \mathbb{R}^m$ and w is in the form of \mathbb{R}^k , we want to find the weights w that achieve the smallest error or loss $L(w) = \|f(a) - \hat{f}_w(a)\|$ on a given training set, in the hope that the training set is representative enough that the quality of the approximation of \hat{f}_w will generalize to other inputs of f .

While there exist many ways to update w , the method which has gained the most ground is gradient descent. This is largely due to the fact that gradients can be efficiently computed even for extremely large numbers of parameters. We briefly describe gradient descent, as follows:

Given a training sample $(a, f(a)) \in A \times B$ and some initialization of w at w^i , both the loss $L(w^i)$ and the gradient $\nabla L(w^i)$ ⁴ can be computed: The gradient marks the direction which increases the loss $L(w^i)$ most rapidly, and the gradient descent algorithm dictates that w should be updated in the direction of the negative gradient by a small step defined by the *learning rate* r .

$$w^{i+1} = w^i - r * \nabla L(w^i)$$

This update step is iterated many times. In reality, however, gradient descent is almost never used in this pure form. Most commonly used are *stochastic gradient descent* (SGD) flavors that operate on batches of training samples at a time. Popular variants include momentum [Qian 1999], adagrad [Duchi et al. 2011], and Adam [Kingma and Ba 2014].

An important property is that differentiability is compositional. For traditional neural networks (i.e., those organized into layers), we have a simple function composition $\hat{f}_w = \hat{f}_{n, w_n} \circ \dots \circ \hat{f}_{1, w_1}$ where each \hat{f}_{i, w_i} represents a layer. Other architectures similarly compose and enable end-to-end training. A popular example demonstrating this is that of image captioning, which composes convolutional neural networks (CNN) [LeCun et al. 1990] and recurrent neural networks (RNN) [Elman 1990].

Imagine, however, that \hat{f}_w and by extension $L(w)$ is not just a simple sequence of function composition, but is instead defined by a *program*, e.g., a λ -term with complex control flow. How, then, should $\nabla L(w)$ be computed?

2.1 From Symbolic Differentiation to Forward-Mode AD

Symbolic differentiation techniques to obtain the derivative of an expression are taught in high schools around the world. Some of the well-known rules are shown in Figure 2 (we will get to the one dealing with let expressions shortly). As such, this seems a candidate for a first approach. However, we hit a problem: some of the rules may cause the code to explode, not only in size, but also in terms of computation cost.

Consider the following example:

⁴The gradient ∇f of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined as the vector of partial derivatives of f with respect to each of its parameters: $\nabla f(u) = (\frac{\partial f(u)}{\partial u_1}, \frac{\partial f(u)}{\partial u_2}, \dots, \frac{\partial f(u)}{\partial u_n})$

Demystifying Differentiable Programming: Shift/Reset the Penultimate Backpropagator

Syntax:	Symbolic differentiation rules:
$e ::= c$	$d/dx \llbracket c \rrbracket = 0$
$ x$	$d/dx \llbracket x \rrbracket = 1$
$ e + e$	$d/dx \llbracket e_1 + e_2 \rrbracket = d/dx \llbracket e_1 \rrbracket + d/dx \llbracket e_2 \rrbracket$
$ e * e$	$d/dx \llbracket e_1 * e_2 \rrbracket = d/dx \llbracket e_1 \rrbracket * e_2 + e_1 * d/dx \llbracket e_2 \rrbracket$
$ \text{let } x = e \text{ in } e$	$d/dx \llbracket \text{let } y = e_1 \text{ in } e_2 \rrbracket = \text{let } y = e_1 \text{ in}$ $\text{let } y' = d/dx \llbracket e_1 \rrbracket \text{ in}$ $d/dx \llbracket e_2 \rrbracket$ $d/dx \llbracket y \rrbracket = y' \quad (y \neq x)$

Fig. 2. Symbolic differentiation for a simple expression language, extended with let expressions.

$$\begin{aligned}
 d/dx \llbracket e_1 * e_2 * \dots * e_n \rrbracket &= d/dx \llbracket e_1 \rrbracket * e_2 * \dots * e_n + \\
 &e_1 * d/dx \llbracket e_2 \rrbracket * \dots * e_n + \\
 &\dots + \\
 &e_1 * e_2 * \dots * d/dx \llbracket e_n \rrbracket
 \end{aligned}$$

The size- n term on the left-hand side is transformed into n size- n terms, which is a non-linear increase. Worse, each e_i is now evaluated n times.

This problem is well recognized in the AD space and often cited as major motivation for more efficient approaches. In fact, many AD papers go to great lengths to explain that “AD is not symbolic differentiation” [Baydin et al. 2018; Pearlmutter and Siskind 2008]. However, let us consider what happens if we convert the program to administrative normal form (ANF) [Flanagan et al. 1993] first, binding each intermediate result in a let expression:

$$\begin{array}{lll}
 d/dx \llbracket \text{let } y_1 = \dots \text{ in} & = & \text{let } y_1 = \dots \quad \text{in let } y'_1 = \dots \text{ in} \\
 \dots & & \dots \\
 \text{let } y_n = \dots \text{ in} & & \text{let } y_n = \dots \quad \text{in let } y'_n = \dots \text{ in} \\
 \text{let } z_1 = y_1 * y_2 \text{ in} & & \text{let } z_1 = y_1 * y_2 \quad \text{in let } z'_1 = y'_1 * y_2 + y_1 * y'_2 \text{ in} \\
 \text{let } z_2 = z_1 * y_3 \text{ in} & & \text{let } z_2 = z_1 * y_3 \quad \text{in let } z'_2 = z'_1 * y_3 + z_1 * y'_3 \text{ in} \\
 \dots & & \dots \\
 \text{let } z_{n-1} = z_{n-2} * y_n \text{ in} & & \text{let } z_{n-1} = z_{n-2} * y_n \quad \text{in let } z'_{n-1} = z'_{n-2} * y_n + z_{n-2} * y'_n \text{ in} \\
 z_{n-1} \rrbracket & & z'_{n-1}
 \end{array}$$

After ANF-conversion, the expression size increases only by a constant factor. The program structure remains intact, and just acquires an additional let binding for each existing one. No expression is evaluated more often than before.

This example uses the standard symbolic differentiation rules for addition and multiplication but also makes key use of the let rule in Figure 2, which splits a binding $\text{let } y = \dots$ into $\text{let } y = \dots$ and $\text{let } y' = \dots$. Following terminology from the AD community, we call y the *primal* and y' the *tangent*. The rules in Figure 2 work with respect to a fixed x , which we assume by convention does not occur bound in any $\text{let } x = \dots$ expression. All expressions are of type \mathbb{R} , so a derivative can be computed for any expression. We write $d/dx \llbracket e \rrbracket$ using syntax brackets to emphasize that symbolic differentiation is a syntactic transformation.

For straight-line programs, applying ANF conversion and then symbolic differentiation achieves exactly the same result as standard presentations of forward-mode AD. Hence, it seems to us that the AD community has taken a too narrow view of symbolic differentiation, excluding the possibility of let-bindings, and we believe that repeating the mantra “AD is not symbolic differentiation”

$$\begin{array}{lll}
 e ::= \dots \mid \lambda x. e \mid e e & \overrightarrow{\mathcal{D}}_x \llbracket e_1 + e_2 \rrbracket = & \overrightarrow{\mathcal{D}}_x \llbracket \lambda y. e \rrbracket = \lambda \overrightarrow{\mathcal{D}}_x \llbracket y \rrbracket. \overrightarrow{\mathcal{D}}_x \llbracket e \rrbracket \\
 \tau ::= \mathbb{R} \mid \tau \rightarrow \tau & \text{let } (a, a') = \overrightarrow{\mathcal{D}}_x \llbracket e_1 \rrbracket \text{ in} & \overrightarrow{\mathcal{D}}_x \llbracket e_1 e_2 \rrbracket = \overrightarrow{\mathcal{D}}_x \llbracket e_1 \rrbracket \overrightarrow{\mathcal{D}}_x \llbracket e_2 \rrbracket \\
 & \text{let } (b, b') = \overrightarrow{\mathcal{D}}_x \llbracket e_2 \rrbracket \text{ in} & \overrightarrow{\mathcal{D}}_x \llbracket \text{let } y = e_1 \text{ in } e_2 \rrbracket = \\
 & (a + b, a' + b') & \text{let } \overrightarrow{\mathcal{D}}_x \llbracket y \rrbracket = \overrightarrow{\mathcal{D}}_x \llbracket e_1 \rrbracket \text{ in } \overrightarrow{\mathcal{D}}_x \llbracket e_2 \rrbracket \\
 \overrightarrow{\mathcal{D}}_x \llbracket c^{\mathbb{R}} \rrbracket = (c, 0) & \overrightarrow{\mathcal{D}}_x \llbracket e_1 * e_2 \rrbracket = & \overrightarrow{\mathcal{D}}_x \llbracket \mathbb{R} \rrbracket = \mathbb{R} \times \mathbb{R} \\
 \overrightarrow{\mathcal{D}}_x \llbracket x^{\mathbb{R}} \rrbracket = (x, 1) & \text{let } (a, a') = \overrightarrow{\mathcal{D}}_x \llbracket e_1 \rrbracket \text{ in} & \overrightarrow{\mathcal{D}}_x \llbracket \tau_1 \rightarrow \tau_2 \rrbracket = \overrightarrow{\mathcal{D}}_x \llbracket \tau_1 \rrbracket \rightarrow \overrightarrow{\mathcal{D}}_x \llbracket \tau_2 \rrbracket \\
 \overrightarrow{\mathcal{D}}_x \llbracket y^{\mathbb{R}} \rrbracket = (y, y') & \text{let } (b, b') = \overrightarrow{\mathcal{D}}_x \llbracket e_2 \rrbracket \text{ in} & \\
 \overrightarrow{\mathcal{D}}_x \llbracket y^{\tau \neq \mathbb{R}} \rrbracket = y & (a * b, a * b' + a' * b) &
 \end{array}$$

Fig. 3. (a) formal grammar of enriched expression language, (b) gradients of the expressions in the enriched formal grammar

is ultimately harmful and contributes to the mystical appearance of the field. We believe that understanding sophisticated AD algorithms as *specific forms* of symbolic differentiation will overall lead to a better understanding of these techniques.

2.2 Forward-Mode AD for Lambda Terms

We now proceed beyond straight-line programs and enrich our grammar with lambdas and applications in Figure 3. A consequence of this is having to distinguish between number-typed expressions and function-typed expressions, where only numeric expressions are differentiable. We define a new differentiation operator $\overrightarrow{\mathcal{D}}_x \llbracket e^\tau \rrbracket$, where the arrow indicates forward mode and where τ is the type associated with the expression e . We omit the τ for a cleaner presentation if τ is not explicitly used. We use the same notation to transform variables in argument positions, and to explain how types are transformed. The key strategy for numeric values is to always pair the primal value with its tangent ($\overrightarrow{\mathcal{D}}_x \llbracket \mathbb{R} \rrbracket = \mathbb{R} \times \mathbb{R}$), including through function arguments and results. This generalizes the paired let-bindings from Figure 2. Note that differentiation is still with respect to a fixed x . Compared to the previous section, we no longer rely on an ANF-pre-transform pass. Instead, the rules for addition and multiplication insert let bindings directly. It is important to note that the resulting program may not be in ANF due to nested let-bindings, but code duplication is still eliminated thanks to the strict pairing of primals and tangents.

Readers acquainted with forward-mode AD will note that this methodology is standard [Baydin et al. 2018], though the presentation is not.

2.3 Implementation using Operator Overloading

Pairing the primal and tangent values for numeric expressions is quite convenient, because when dealing with function application, the let-insertion needs both the primal and tangent of the parameter for the tangent computation. Since the transformation is purely local, working with pairs for numeric expressions makes it immediately clear that this strategy can be implemented easily in standard programming languages by overloading operators. This is standard practice, which we illustrate through our implementation in Scala (Figure 4).

The NumF class encloses both the primal as x and tangent as d , with operators overloaded to compute primal and tangent values at the same time. To use the forward-mode AD implementation, we still need to define an operator `grad` to compute the derivative of any function $\text{NumF} \Rightarrow \text{NumF}$ (Figure 4, top right). Internally, `grad` invokes its argument function with a tangent value of 1 at the given position and returns the tangent field of the function result. In line with the previous sections,

```
// differentiable number type
class NumF(val x: Double, val d: Double) {
  def +(that: NumF) =
    new NumF(this.x + that.x, this.d + that.d)
  def *(that: NumF) =
    new NumF(this.x * that.x,
              this.d * that.x + that.d * this.x)
  ...
}

// differentiation operator
def grad(f: NumF => NumF)(x: Double) = {
  val y = f(new NumF(x, 1.0))
  y.d
}

// example
val df = grad(x => 2*x + x*x*x)
forAll { x =>
  df(x) == 2 + 3*x*x }
}
```

Fig. 4. Forward-mode AD in Scala

we only handle scalar functions, but the approach generalizes to multidimensional functions as well. An example using the `grad` operator is shown in the bottom right of Figure 4.

2.4 Nested Gradient Invocation and Perturbation Confusion

In the current implementation, we can compute the gradient of any functions in the type `NumF => NumF` at any given values using forward mode AD. However, our `grad` function is not yet truly first class, since we cannot apply it in a nested fashion as in `grad(grad(f))`, which prevents us from computing higher order derivatives, or solve nested min/max problems in the form of:

$$\min_x \max_y f(x, y)$$

Yet, even this somewhat restricted operator gives rise to a few subtleties. There is a common issue associated with functional implementations of AD that, like ours, expose a gradient operator within the language. In the simple example shown below, the inner call to `grad` should return 1, meaning that the outer `grad` should also return 1.

```
grad { x: NumF =>
  val shouldBeOne = grad(y => x + y)(1) // evaluates to 2 instead of 1 !
  val z = NumF(shouldBeOne, 0)
  x * z
} (1)
```

However, this is not what happens. The inner `grad` function will also collect the tangent from `x`, thus returning 2 as the gradient of `y`. The outer `grad` will then give a result of 2 as gradient of `x`. This issue is called *perturbation confusion* because the `grad` function is confusing the perturbation (i.e., derivative) of a free variable used within the closure with the perturbation of its own parameter.

The root of this problem is that the two `grad` functions are associated with different final results; their gradient updates should not be mixed. We do not provide any new solutions to perturbation confusion, and hence consider this issue orthogonal to our work. Our implementation can be easily extended to support known ways to prevent confusion of gradients, either based on *dynamic tagging*, or through a type-based solution as realized in Haskell⁵, which lifts the tags into the type system using rank-2 polymorphism, in the same way as the ST monad [Launchbury and Peyton Jones 1994].

⁵<http://conway.rutgers.edu/%7Eeccshan/wiki/blog/posts/Differentiation/>

2.5 First-Class Gradient Operator

While not the main focus of our work, we outline one way in which our NumF definition can be changed to support first-class gradient computation and to prevent perturbation confusion at the same time. Inspired by DiffSharp [Baydin et al. 2016], we change the class signatures as shown below. We unify NumF and Double in the same abstract class Num, and add a dynamic tag value tag. The grad operator needs to assign a new tag for each invocation, and overloaded operators need to take tag values into account so as not to confuse different ongoing invocations of grad.

```
abstract class Num
class NumV(val x: Double) extends Num
class NumF(val x: Num, val d: Num, val tag: Int) extends Num {...}
def grad(f: Num => Num)(x: Num): Num = {...}
```

Alternative implementations using parameterized types and type classes instead of OO-style inheritance are also possible.

This concludes the core ideas of forward-mode automated differentiation (AD). Forward-mode AD implementations using operator overloading exist in many languages, as it is a simple and direct choice. As proposed previously, forward-mode AD can be viewed as a specific kind of symbolic differentiation, either using standard differentiation rules after ANF-conversion, or using transformation rules that insert let-bindings on the fly, operating on a pair composed of a value and a gradient (or a primal and a tangent).

3 DIFFERENTIABLE PROGRAMMING IN REVERSE MODE

Forward-mode AD is straightforward to implement, preserves the temporal and spatial complexity of a computation, and also generalizes to functions with multiple inputs and multiple outputs. However, it is inefficient when the number of inputs is large, which is the case for loss functions of neural networks. To compute the gradient of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, we have to compute n forward derivatives either sequentially or simultaneously, but in either case leading to an $O(n)$ slowdown compared to the original program. Is there a better way?

We consider again $f : \mathbb{R}^n \rightarrow \mathbb{R}$ represented as a straight-line program in ANF, i.e., as a sequence of $\text{let } y_j = e_j$ expressions, with inputs x_i and output y_m . The basic intuition is that instead of computing all $n * m$ internal derivatives $d/dx_i y_j$ as in forward mode, we would rather only compute the $m + n$ derivatives $d/dy_j y_n$ and $d/dx_i y_n$. For this, we need to find a way to compute these derivatives starting with $d/dy_n y_n = 1$, and working our way backwards through the program until we reach the inputs x_i . Hence, this form of AD is called reverse mode, or backpropagation when referring to neural networks. The approach generalizes to functions $\mathbb{R}^n \rightarrow \mathbb{R}^m$ with multiple outputs, and is generally more efficient than forward-mode AD when $n \gg m$.

The distinct feature of reverse-mode AD is the two phases of computations: the forward propagation that computes all the primal values, and the backward propagation that computes all the gradients (also called *adjoints* or *sensitivities* in the AD research community). This idea is illustrated by the simple example of computing x^4 via $y = x * x, z = y * y$ below.

Forward pass:	Backward pass:
let $y = x * x$	let $z' = d/dz z = 1.0$
let $z = y * y$	let $y' = d/dy z = 2 * y$
	let $x' = d/dx z = d/dx y * d/dy z = 2 * x * y'$
	$= 4 * x * x * x$

The backward propagation depends in a crucial way on the *chain rule*:

Forward pass:	Backward pass:
let $x'_0 = \text{ref } 0$	$x'_n += 1$
let $x_1 = x_{11} \oplus x_{12}$	$x'_{n2} += \frac{d}{dx_{n2}} \llbracket x_{n1} \oplus x_{n2} \rrbracket * !x'_n$
let $x'_1 = \text{ref } 0$	$x'_{n1} += \frac{d}{dx_{n1}} \llbracket x_{n1} \oplus x_{n2} \rrbracket * !x'_n$
let $x_2 = x_{21} \oplus x_{22}$...
let $x'_2 = \text{ref } 0$	$x'_{22} += \frac{d}{dx_{22}} \llbracket x_{21} \oplus x_{22} \rrbracket * !x'_2$
...	$x'_{21} += \frac{d}{dx_{21}} \llbracket x_{21} \oplus x_{12} \rrbracket * !x'_2$
let $x_n = x_{n1} \oplus x_{n2}$	$x'_{12} += \frac{d}{dx_{12}} \llbracket x_{11} \oplus x_{12} \rrbracket * !x'_1$
let $x'_n = \text{ref } 0$	$x'_{11} += \frac{d}{dx_{11}} \llbracket x_{11} \oplus x_{12} \rrbracket * !x'_1$

Fig. 5. Reverse-mode AD: general pattern for a straight-line program in ANF. The forward pass is the original program, extended with allocating mutable variables to hold the adjoints. The adjoints are successively computed by the backward pass.

$$\frac{d}{dx} f(g(x)) = \text{let } y = g(x) \text{ in} \\ \text{let } y' = \frac{d}{dx} g(x) \text{ in} \\ \frac{d}{dx} y * \frac{d}{dy} f(y)$$

This rule is used in the last line to compose $\frac{d}{dx} y$ with $\frac{d}{dy} z$ for $\frac{d}{dx} z$.

A more general presentation of reverse-mode AD is given in Figure 5. Note that there are several differences when compared with forward-mode AD. For one, because an intermediate result may be used in multiple calculation steps, and one reverse propagation step can only capture the partial gradient from the associated calculation, the gradients are typically stored in mutable variables and modified by relevant reverse propagation steps. In addition to this, reverse-mode also must remember values from the forward propagation – whether by additional data structure such as graphs or stacks or other means – to support partial derivative computation in the reverse propagations.

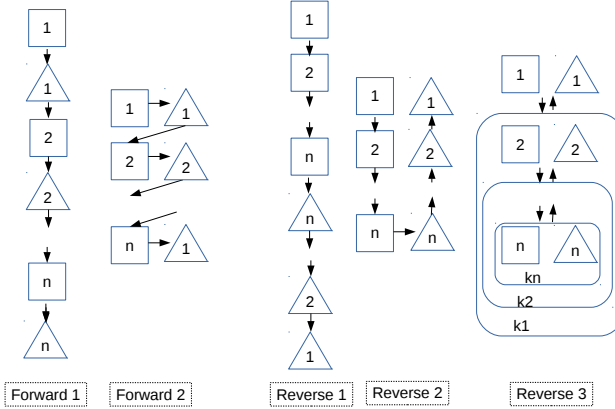


Fig. 6. Computation flow of forward-mode and reverse-mode AD

3.1 Implementing Reverse-Mode AD with Continuations

The computation flows of forward mode and reverse mode are illustrated in Figure 6. In forward mode, the computation is interleaving value calculations in squares and gradient calculations in

triangles (Forward 1). It is easy to pair the neighboring value calculations and gradient calculations (Forward 2) so that the forward mode can be achieved by operator overloading. In reverse mode, the computation has to finish value calculations first, then carry out gradient calculations in reverse order. It is tempting to simply “fold” the gradient calculations up in parallel with the value calculations, to achieve a similar looking presentation (Reverse 2). Although we cannot directly implement the Reverse 2 mode since the computation flows both down and up, we take inspiration from “There and Back again” [Danvy and Goldberg 2005], and look for ways to model the computation in a sequence of function calls, where the call path implements the forward pass and the return path implements the backward path. With this intuition, it is not hard to see that a transformation to continuation-passing style (CPS) provides exactly the right structure. We show the central rules for a continuation-based transformation in Figure 7. In contrast to the standard CPS rules for addition and multiplication, the continuation parameter k in the reverse AD rule $\overleftarrow{\mathcal{D}}_x \llbracket e \rrbracket k$ is a *delimited continuation*, since the adjoints are updated *after* the call to the continuation returns. The Reverse 3 mode in Figure 6 depicts this scoped model.

$$\begin{aligned}
 &\text{Standard CPS transformation:} \\
 &\llbracket x \rrbracket k = k(x) \\
 &\llbracket e_1 * e_2 \rrbracket k = \llbracket e_1 \rrbracket (\lambda y_1. \llbracket e_2 \rrbracket (\lambda y_2. k(y_1 * y_2))) \\
 &\text{Reverse AD with CPS transformation:} \\
 &\overleftarrow{\mathcal{D}}_x \llbracket x \rrbracket k = k(x) \\
 &\overleftarrow{\mathcal{D}}_x \llbracket e_1 * e_2 \rrbracket k = \overleftarrow{\mathcal{D}}_x \llbracket e_1 \rrbracket (\lambda(y_1, y'_1). \overleftarrow{\mathcal{D}}_x \llbracket e_2 \rrbracket (\lambda(y_2, y'_2). \\
 &\quad \text{let } y = y_1 * y_2 \text{ in} \\
 &\quad \text{let } y' = \text{ref } 0 \text{ in} \\
 &\quad k(y, y') \\
 &\quad y'_1 += y_2 * !y' \\
 &\quad y'_2 += y_1 * !y'))
 \end{aligned}$$

Fig. 7. Reverse-mode AD: CPS transformation rules to implement the transform in Figure 5

3.2 Implementation Using Operator Overloading

Our first implementation in Scala is mechanical, following directly the rules in Figure 7. It is shown in Figure 8. Just like in forward-mode AD, we associate the value and the gradient closely as two fields of a class, here `NumR`. Each operator takes a delimited continuation which updates the gradient using the value of the intermediate result. As shown in the computation flow Figure 6, the continuation is then expected to handle the rest of the forward propagation after this computation step, as well as the beginning of the backward propagation before this step. Once the continuation returns, the operator updates the gradients of the dependent values as side effects.

However, it is still cumbersome to use this implementation. For a simple model such as $y = 2 * x + x * x * x$, we have to explicitly construct delimited continuations for each step (last line in Figure 8). Fortunately, *delimited control operators* exist that enable programming with delimited continuations in a direct style, without making the continuations explicit.

3.3 Implementation using Control Operators

The shift and reset operators [Danvy and Filinski 1990] work together to capture a partial return path up to a programmer-defined bound: in our case the remainder of the forward pass. In Figure 9,

Demystifying Differentiable Programming: Shift/Reset the Penultimate Backpropagator

```
// differentiable number type
class NumR(val x: Double, var d: Double) {
  def +(that: NumR) = { (k: NumR=>Unit) =>
    val y = new NumR(x + that.x, 0.0)
    k(y)
    this.d += y.d
    that.d += y.d
  }
  def *(that: NumR) = { (k: NumR=>Unit) =>
    val y = new NumR(x * that.x, 0.0)
    k(y)
    this.d += that.x * y.d
    that.d += this.x * y.d
  }
  ...
}

// differentiation operator
def gradR(f: NumR => (NumR=>Unit)=>Unit)(x: Double)={
  val z = new NumR(x, 0.0)
  f(z)(r => r.d = 1.0)
  z.d
}

// example: 2*x + x*x*x
val df = gradR { x =>
  (2*x) (y1=> ( x*x ) (y2=> (y2 *x ) (y3=> y1 + y2)))
}
forAll { x =>
  df(x) = 2 + 3*x*x
}
```

Fig. 8. Automatic Differentiation in Scala: (a) reverse-mode AD in continuation-passing style (b) grad function definition and use case. Handling of continuations is highlighted.

the keyword `shift` provides access to a delimited continuation that reaches up the call chain to the nearest enclosing `reset`. The Scala compiler transforms all the code in between accordingly [Rompf et al. 2009]. The implementation of `NumR` with `shift/reset` operators is almost identical to `NumR` (modulo added `shift`). Note that the `shift` operator returns a CPS-annotated type `NumR @diff`, where type `diff` is defined as `cps[Unit]`.

It is important to note at this point that this method may appear similar to [Pearlmutter and Siskind \[2008\]](#). However, there are substantial differences despite the similarities and shared goals. Pearlmutter and Siskind proposed an implementation which returns a pair of a value and a backpropagator: $x \mapsto (v, dv/dy \mapsto dx/dy)$. Doing this correctly requires a non-local program transformation, and requires further tweaks if a lambda uses variables from an outer scope, in which case a channel needs to be established that allows backpropagation for closed-over variables, not just the function parameter.

Using delimited continuations with `shift/reset` operators, by contrast, enables reverse-mode AD with only local transformations. Any underlying non-local transformations are taken care of implicitly by `shift` and `reset`. Beyond this, it is also worth noting that our method can allocate all closures and mutable variables on the stack, i.e., we never need to return closures that escape their allocation scope. Indeed, our computation graph is never reified, but instead remains implicit in the function call stack. A consequence of this is that tail calls become proper function calls. The proposed implementation is also extremely concise, to the point of being able to serve as a *specification* of reverse-mode AD and be used to teach AD to students.

3.4 Nested Invocations For High-Order Gradient

In a similar situation as with forward-mode AD in Section , we are interested in extending the current implementation to support nested invocations of the `grad` operator. The only restriction is that we cannot invoke reverse-mode AD within reverse-mode AD (i.e., reverse-reverse) due to the fact that `shift2`, a theoretically well-understood construct [Danvy and Filinski 1990], is

```

// differentiable number type
class NumR(val x: Double, var d: Double) {
  def +(that: NumR) = shift {(k: NumR=>Unit)=>
    val y = new NumR(x + that.x, 0.0)
    k(y)
    this.d += that.x * y.d
    that.d += this.x * y.d
  }
  def *(that: NumR) = shift {(k: NumR=>Unit)=>
    val y = new NumR(x * that.x, 0.0)
    k(y)
    this.d += that.x * y.d
    that.d += this.x * y.d
  }
  ...
}

// differentiation operator
def grad(f: NumR => NumR @cps[Unit] )(x: Double) = {
  val z = new NumR(x, 0.0)
  reset { f(z).d = 1.0 }
  z.d
}

// example
val df = grad(x => 2*x + x*x*x)
forAll { x =>
  df(x) = 2 + 3*x*x
}
    
```

Fig. 9. Automatic Differentiation in Scala: (a) reverse mode using delimited continuations, with shift/reset operators (b) grad function definition and use case. Handling of continuations is confined to implementation logic and does not leak into user code.

unavailable in Scala. However, it is still interesting to nest forward-mode AD with reverse-mode AD for higher-order derivatives. For a typical $\mathbb{R}^n \rightarrow \mathbb{R}$ function, forward-reverse is particular interesting for computing first-order gradient with reverse-mode, and second-order gradient with forward mode. In particular, we can compute Hessians as the Jacobian of gradients [Baydin et al. 2018], and Hessian vector products in a single reverse pass [Christianson 1992].

This is done by unifying NumR under the same abstract Num from Section , and using the dynamic tagging method previously described to address perturbation confusion bugs.

```
class NumR(val x: Num, var d: Num, tag: Int) extends Num {...}
```

4 IMPLEMENTATION IN LMS

CPS conversion puts our reverse-mode AD on a firm basis, rooted in programming language concepts. Extending the Num type to tensors will provide all the necessary machinery for a deep learning framework. However, running Scala code on a JVM will not be efficient for practical deep learning tasks. For practical use, we obviously need to make use of the low-level implementations that run on deep-learning-specific hardware such as GPUs, TPUs, etc.

One way to achieve this goal is to wrap the low-level implementations as Scala function calls in a library. Doing so allows us to shift the burden of each tensor operation to a library function call. This is very similar to frameworks such as Torch and PyTorch, which construct *dynamic* computation graphs at runtime, and exploit efficient low-level implementations for tensor operations. This paradigm has become quite popular due to its ease of use and flexibility in working with control flow constructs such as branches, loops, and recursion.

A better way is to transform high level implementation in Scala to low-level code, via multi-stage programming (staging). Modern staging tools such as Lightweight Modular Staging (LMS) [Rompf and Odersky 2010] blend normal program execution with IR construction which is similar to but

more general than computation graphs in TensorFlow. The flavor of LMS is shown in the following example.

```
def totalScore(names: Rep[Vector[String]]) = {
  val scores = for ((a,i) <- names.zipWithIndex) yield (i * nameScore(a))
  scores.sum
}
def nameScore(name: Rep[String]) = {
  name.map(c => c - 64).sum
}
```

Here, a simple score is computed for a vector of strings. The types `Rep[Vector[String]]` and `Rep[String]` are the only giveaways that an IR is constructed. The implementation crucially relies on type inference and advanced operator overloading capabilities, which extend to built-in control flow constructs like `if`, `for`, and `while` so that normal syntax can be used. As shown in the example, an expression for `((a,i) <- names) yield ...` becomes a series of method calls with closure arguments `names.map((a,i) => ...)`.

Staging our reverse-mode AD with LMS will not face fundamental challenges, because it is a well-known insight that a multi-stage program that uses continuations can generate code in CPS [Biernacka et al. 2005]. LMS can also be set up to generate low-level code such as C++, allowing efficient back-ends. This makes our framework similar to frameworks such as TensorFlow or Theano. The main benefit of these approaches is that they offer a larger surface for analysis and optimization, much like in an aggressive whole-program compiler. High-level optimization among tensor operations can be applied to make training and inference more efficient.

The apparent downsides of TensorFlow and Theano, however, are the rather clunky programming interface offered by current frameworks, the absence of sophisticated control flow constructs, and the inability to use standard debugging facilities. However, our system largely avoids the downsides of the current static frameworks by using the idea of staging (in particular, the LMS framework). Of course, TensorFlow can be viewed as utilizing staging as well, but the staged language is a restricted dataflow model. On the other hand, LMS provides a rich staged language that includes subroutines, recursion, and so on.

We show below how LMS is added to our CPS-based autodiff system, and demonstrate cases of employing branches, loops, and recursion in a natural form. One added intellectual benefit of this is the achievement of actual code transformations from an unmodified API.

4.1 Staging Straight-Line Programs

We begin by looking at how we can stage and perform AD on straight-line programs (i.e., those without loops, branches, or recursion). We already have the basic `Num` class definition with operators in CPS which can support straight-line programs. Therefore, all that is required is the labeling of some types `T` as `Rep[T]` for staging in LMS. One choice that can be ruled out quickly is `Rep[Num]`, because it means that the generated code has to maintain `Num` class and objects. An obvious suggestion may be as follows:

```
class Num(val x: Rep[Double], var d: Rep[Double]) {...}
```

This works for straight-line programs, but poses challenges for programs with nested scopes. As has already been discussed in Section 3, a destination-passing style is needed in which we can pass a reference of the gradient field to the nested scope. Ergo, the proper solution is to wrap the gradient `d` as a staged variable:

```
class Num(val x: Rep[Double], val d: Rep[Var[Double]]) {...}
```

Here, the `Rep[Var[Double]]` is generating a pointer to double in C++, which allows us to pass the gradient `d` by reference, following destination-passing style. Given this basic setting of `Rep` types, we will then refer to general types (`A`, `B`, `C`) in the following part of this section, to illustrate the abstraction of branches, loops, and recursion.

4.2 Staging Programs with Branches

We cannot simply use the overridden `if` operator in LMS for branches because we need to manage continuations using shift operators. Thus, we define a standalone `IF` function, taking a `Rep[Boolean]`, and two `(\Rightarrow Rep[A] @cps[Rep[B]])` typed parameters for the then-branch and else-branch respectively. In Scala, `\Rightarrow T` typed parameters are *passed by name*, so that the parameters are evaluated lazily in the function body. The `IF` function, just like the operators in the `Num` class, is a shift construct taking a delimited continuation `k`. The function needs to invoke the continuation either with the then-branch parameter or the else-branch parameter, based on the value of the condition. Note that we have to encase both branches with `reset`:

```
def fun(f: Rep[A]  $\Rightarrow$  Rep[B]): Rep[A  $\Rightarrow$  B] // LMS support for staging a function
def IF(c: Rep[Boolean])(a:  $\Rightarrow$ Rep[A] @cps[Rep[B]])(b:  $\Rightarrow$ Rep[A] @cps[Rep[B]]): Rep[A] @cps[Rep[B]] =
  shift { k:(Rep[A]  $\Rightarrow$  Rep[B])  $\Rightarrow$ 
    val k1 = fun(k) // generate lambda for k
    if (c) reset(k1(a)) else reset(k1(b))
  }
```

If we simply pass `k` to the `if` statement without generating `k1` as a lambda, we will have code duplication in both branches. The code increase can be exponential, given multiple consecutive `IF` invocations.

4.3 Staging with Loops or Traversing Sequences

Sometimes, deep learning deals with sequential data of different lengths using recurrent neural networks (RNN), such as words viewed as an array of characters or audio streams viewed as a sequence of floats. As a differentiable programming framework, it is only natural to support loop construct. How do we handle loop generation with CPS? For sure, we cannot directly use the `while` or `for` operators in Scala.

It is clear that the a loop needs to be transformed into a recursive function in CPS. In LMS, recursive functions can be constructed by:

```
def f: Rep[A  $\Rightarrow$  B] = fun { x  $\Rightarrow$  ... f(...) ... }
```

A loop construct takes at least an initial value `Rep[A]`, a loop guard and a loop body of type `Rep[A] \Rightarrow Rep[A] @cps[Rep[B]]` as parameters. The loop guard can be either `Rep[A] \Rightarrow Rep[Boolean]`, like a `while` construct, or simply a `Rep[Int]`, like a `for` construct. The actual loop logic can be described as follows: if the loop guard is true, call the loop body; else call the continuation. The `WHILE` construct is defined below, mimicking the standard `while` loop.

```
def WHILE(init: Rep[A])(c: Rep[A]  $\Rightarrow$  Rep[Boolean])(b: Rep[A]  $\Rightarrow$  Rep[A] @cps[Rep[B]]):
Rep[A] @cps[Rep[B]] = shift {
  k:(Rep[A]  $\Rightarrow$  Rep[B])  $\Rightarrow$ 
  lazy val loop: Rep[A]  $\Rightarrow$  Rep[B] = fun { (x: Rep[A])  $\Rightarrow$ 
    if (c(x)) RST(loop(b(x))) else RST(k(x))
  }
  loop(init)
}
```

4.4 Staging with Recursion or Traversing Structural Data

As a differentiable programming framework, we would also like to handle more general forms of recursion. This is also very useful in deep learning, for processing structural data such as language parse trees. The standard TensorFlow framework cannot process parse trees with different sizes and shapes due to a lack of expressivity of its computation graph construction.

Before rushing into our recursion abstraction, it may be beneficial to build an intuition about how to manage recursion in CPS. We begin by examining a simple recursive program on a list.

```
def compute(a: Int, b: Int): Int = ...
def traverseList(l: List) = {
  if (l.isEmpty) 0
  else compute(traverseList(l.tail), l.head)
}
```

In this example, the `traverseList` function is not tail-recursive. In continuation-passing style, we must form a lambda as $(x \Rightarrow \text{compute}(x, \text{l.head}))$, and put this lambda before the continuation that comes after. This is the support we want to build first (as shown in `FUN`), which takes a function of type $\text{Rep}[A] \Rightarrow \text{Rep}[B] \text{ @cps}[\text{Rep}[C]]$, and returns a function of the same type. Inside `FUN`, the parameter `f` is squeezed between the continuation `k` and the initial value `x`.

With that support, implementing a tree traversal in `TREE` is straightforward. Here, the $\text{Rep}[A]$ type in `FUN` is $\text{Rep}[\text{Tree}]$. For empty trees, the `init` is directly passed along. For non-empty trees, the result of the left child and right child are composed by the `b` function supplied by the user.

```
def FUN(f: Rep[A] => Rep[B] @cps[Rep[C]]): {
  val f1 = fun((x, k) => reset(k(f(x)))) // put f in between k and x
  { (x: Rep[A]) => shift { k: (Rep[B] => Rep[C]) => f1((x, fun(k))) } }
}
def TREE(init: Rep[B])(t: Rep[Tree])(b: (Rep[B], Rep[B]) => Rep[B] @cps[Rep[C]]):
Rep[B] @cps[Rep[C]] = {
  def f = FUN { tree: Rep[Tree] =>
    if (tree.isEmpty) init
    else b(f(tree.left), f(tree.right))
  }
  f(t)
}
```

With all of the above implementations in place, we have established a framework capable of supporting branches, loops, and recursion. Although building these control flow operators takes some engineering, this is simply providing abstraction through which we may generate CPS code.

This framework provides a highly expressive programming interface in the style of PyTorch, as well as generating an intermediate representation free from AD logic. This allows pure manipulation of Doubles or Tensors, which in turn allows extensive optimization in the style of TensorFlow.

We note in passing that while it is, naturally, an option to implement CPS at the LMS IR level, we choose to forgo this route in favor of the presented implementation simply for clarity and accessibility.

5 IMPLEMENTATION AND CASE STUDIES

Up to this point, we have shown both how to do reverse-mode AD using delimited continuations and how to do this efficiently using staging. In this section, we delve into “real” deep learning. We present the implementation and evaluation of a prototype system called Lantern which extends

the material presented in preceding sections by providing a staged tensor API with functionality similar to that of standard deep learning frameworks. As we demonstrate in this section, Lantern performs competitively on cutting edge deep learning applications which push the boundaries of existing frameworks in various dimensions, i.e., expressivity or performance.

A snippet is shown here for the basic structure of the staged tensor API. Our `Tensor` class holds a `Rep[Array[Double]]` and any number of dimensions in `Array[Int]`, and handles all tensor-level computations including element-wise operations, matrix multiplications and convolution. The `TensorR` class takes two `Tensors`, one as the value, the other as the gradient. Operators in `TensorR` are overloaded with shift constructs, providing entry to delimited continuations. By comparing with the CPS-style implementation in Section 3, the `Tensor` class is equivalent to `Double`, and `TensorR` is equivalent to `NumR`.

```
class Tensor(val data: Rep[Array[Double]], val shape: Array[Int]) {...}
class TensorR(val x: Tensor, val d: Tensor){...}
```

Note that Lantern’s staged tensor API implements tensor operations in naive for loops, and is not yet optimized on the Tensor IR. A more complete way is to interface Lantern’s tensor IR with standard tensor compiler pipelines (e.g. XLA, NNVM) [Distributed (Deep) Machine Learning Community 2018; TensorFlow team 2018] or to purpose-built compiler frameworks that directly extend LMS (e.g. Delite and OptiML [Sujeeth et al. 2014, 2011]). However, even with its relatively simple tensor API, Lantern already achieves performance competitive with modern deep learning frameworks such as PyTorch and TensorFlow.

5.1 Deep Learning Basics

For completeness, we present a small overview of deep learning basics for an audience unfamiliar with the subject. Experts are invited to move directly to Section 5.2.

5.1.1 Simple Layered Neural Networks Using Fully Connected Layers. Most deep learning models are built on different kinds of neural networks. Inspired by neural cells in human brains, which individually have the capacity to compose incoming signals and fire outgoing signals in a non-linear manner, neural networks are composed of neural nodes that each takes some number of inputs and provides one output by a non-linear transformation. Traditional neural networks are arranged by computation layers (Figure 1a), with one input layer, one output layer, and several hidden layers in between. Nodes in each layer linearly compose the information from all nodes in the previous layer, then feed the output to the next layer after a nonlinear transformation (e.g. tanh, sigmoid, or ReLU). Among the deep learning community, these are often called fully-connected (FC) layers.

Actual implementations use vectors to represent FC layers, and matrices as linear weights connecting two layers, as in $Out = \tanh(Weight * In + Bias)$.

Note that the nonlinear transformation after each layer is critical, otherwise the whole neural network can be collapsed into one linear transformation (this is easy to see given a linear algebra formula). Layered neural networks are essentially composed nonlinear functions, each having similar functional structure, but different and adjustable parameters. Together, multiple composed functions can emulate a wide range of complex nonlinear functions.

Neural networks use gradient descent to search for proper parameters in a vast parameter space. Given a set of training data with inputs and targets, the difference between function outputs and targets can be quantified as one numerical value, called “loss”. Partial derivatives are computed for all parameters with regard to the loss, and parameters are modified towards the direction of negative partial derivatives by a small step (also called learning rate in deep learning terminology). With many iterations, a proper set of parameters can be found. The partial derivatives are normally

produced by reverse-mode automated differentiation (AD), which is presented in detail in previous sections.

By stacking many hidden layers (hence, *deep* learning), the neural networks are able to emulate generic functions to a very high level of complexity. However, naively stacking hidden layers creates difficulties in training and learning, and too many parameters used inefficiently leads to overly large models and overfitting to training data. The deep learning community has come up with various neural network architectures to address this problem and also to adapt to specific tasks. For instance, convolutional layers (CONV) [LeCun et al. 1990] are used mostly in image processing (Figure 1b).

5.1.2 Variants of Convolutional Neural Networks. Instead of connecting with every node in the previous layer, nodes in convolutional layers only connect with a small block of the previous layer using small parameter blocks called “kernels”. Each kernel scans the previous layer to generate the next layer, and multiple kernels generate multiple layers (or “channels”, in deep learning terms). Compared to FC layers, CONV layers use parameters more efficiently by utilizing the concept of “weight tying”. Since kernels can pick up local patterns in the previous layer, their performance in image processing has historically been very good.

Although better than the FC layers, very deep CONV layers also showed deteriorating performance even on training losses. The understanding is that although a deeper neural network is (in theory) more powerful than a shallower one, it is actually very hard for a deeper neural network to emulate a shallower neural network, since identity mapping is hard to learn for nonlinear transformations in the neural network layers. To overcome this problem, ResNet was introduced (Figure 1c). In ResNet, an identity mapping was added to send unchanged data to two layers downstream just before the nonlinear transformation (such as ReLU). By passing data unchanged, a very deep ResNet can be trained with good performance [He et al. 2016].

Another variety of CONV layer is an inception layer (Figure 1d). In inception neural networks [Szegedy et al. 2017], a previous layer can be mapped to the next layer by different CONV layers, thus avoiding having to choose hyperparameters for convolution operations. Inception layers also reduce the number of parameters using 1×1 CONV layers that shrink the previous layers in the dimension of channels, thus allowing parameters to be used more efficiently.

5.1.3 Variants of Recurrent and Recursive Neural Networks. Although CONV layers and variants achieve good performance in image processing, a big missing ability is sequential context. A convolutional neural network (CNN) may analyze each frame of a movie well, but can never form connections between frames. To handle inputs where sequential context is important – such as speech recognition or language translation – we need another class of neural networks called recurrent neural networks (RNN) [Kombrink et al. 2011]. The simplest class of RNN, also called a “vanilla RNN” (Figure 1e), has a defining feature of persistent internal memory, called a “hidden layer” (h_0 , h_1 , and h_2 in the figure). Initialized to zero, the hidden layer is used together with the inputs (x_1 and x_2) to produce the next hidden layer, which in turn is used to generate output (o_1 and o_2) and maintain a persistent memory about all previous inputs. Even the simplest RNN can achieve academically interesting results such as language modeling by characters (details in Section 5.2).

It is theoretically possible for a vanilla RNN to learn patterns from sequences of any length. However, training a vanilla RNN for long sequences in practice faces issues such as exploding and vanishing gradients. A simple intuition is that when the same linear transformation is reused multiple times, the values will either get very big or very small, just like multiplying with a number repeatedly. This issue is addressed by Long Short Term Memory (LSTM) (Figure 1f) in a similar fashion as ResNet’s identity mapping.

LSTM [Hochreiter and Schmidhuber 1997] uses two types of persistent memory: the hidden layer, and the cell state. While the hidden layer (H_{t-1}) is often concatenated with the input layer (X_t) for computation (just like in a vanilla RNN), the cell state is only linearly modified by a forget gate f and an input gate i . The updated cell state is then used to generate new hidden layer and output layer, through an output gate o . Detailed equations can be found in Section 5.3, but the key take-away point is that learning long sequences becomes easier by using cell states.

In practice, LSTM is widely used as the basic RNN building block, while vanilla RNNs are generally only used in small examples like tutorials. One avenue for exploiting LSTM for more complicated tasks is to supply an external memory for LSTM to read from and write to, as in the case of Neural Turing Machine (NTM) (Figure 1g). In order to make the whole system differentiable, the reading and writing operations are applied to the entire memory space, just to varying extents (this is called the “attention mechanism” in deep learning terminology). Given training data, NTM learns basic programming abilities such as copy, sort, and associative recall. Note that the NTM also uses other types of neural networks as the controller [Graves et al. 2014].

Another use case for LSTM is to handle structural data such as language parse tree (Figure 1h). Unlike a sequence of data where the LSTM processes each element one by one, in TreeLSTM the LSTM block needs to traverse the tree in an order such that states of the children are fed into the parent as previous states [Tai et al. 2015]. Thus, given a tree-structured data, the LSTM usually must traverse it in a recursive way, thereby gaining the name *recursive* neural networks. Recursive neural networks performs better than recurrent neural networks for tree-structured input, but pose interesting challenges for the expressivity of the deep learning framework. We evaluated TreeLSTM in Section 5.4.

5.2 Vanilla RNN Implementation and Evaluation

We begin our evaluation with a vanilla RNN implementation, `min-char-rnn`.⁶ This vanilla RNN model analyzes a paragraph character by character, and builds a language model predicting the frequency distribution of the next character given the sequence of passed characters. The characters are simply embedded as one-hot vectors, and the hidden vector and loss are updated by simple rules as shown below. Note that $*$ represents matrix vector multiplication, and x_i and y_i are the one-hot embeddings of the input character and the target character.

$$\begin{aligned} h_{i+1} &= \tanh(W_1 * x_i + W_2 * h_i + b_1) \\ e_{i+1} &= \exp(W_3 * h_{i+1} + b_2) \\ p_{i+1} &= e_{i+1} / \text{sum}(e_{i+1}) \\ \text{loss} & -= \log(p_{i+1} \text{ dot } y_i) \end{aligned}$$

Implementation of `min-char-rnn` in Lantern (Figure 10) is very similar to the Numpy implementation provided at the url given in the footnote below. The recurrent nature of the neural network was realized by the `LOOP` construct, which is simply build on the `WHILE` using a `Rep[Int]` index counter as loop guard. We do not have to explicitly provide code for gradient calculations, unlike the Numpy implementation. The training loop is almost identical, with the only difference being that Lantern (similar to PyTorch) must clear the gradient after each training step.

Implementations in PyTorch and TensorFlow are very straightforward, with simple tutorials widely available online. PyTorch’s implementation follows the general rule of putting all parameters and forward propagation logic in the `torch.nn.Module` class, and then allowing a `torch.nn.optim` object to handle gradient descent. TensorFlow, meanwhile, has more encapsulated `BasicRNNCell`

⁶<https://gist.github.com/karpathy/d4dee566867f8291f086>

Demystifying Differentiable Programming: Shift/Reset the Penultimate Backpropagator

```
val pars = ... // all trainable parameters
def lossFun(inputs: Rep[Array[Int]], targets: Rep[Array[Int]]) = {
  val in = (init_loss, init_hidden_vector)
  val out = LOOP(in)(inputs.length){i => in =>
    val xi, yi = ... // one-hot encoding of inputs(i) and targets(i)
    val new_hidden = ((pars(0) dot xi) + (pars(1) dot in._2) + pars(2)).tanh()
    val unnormalized_prob = ((pars(3) dot new_hidden) + pars(4)).exp()
    val normalized_prob = unnormalized_prob / unnormalized_prob.sum()
    val new_loss = in._1 - (normalized_prob dot yi).log()
    (new_loss, new_hidden)
  }
  out(0)
}
for (n <- (0 until maxIter): Rep[Range]) {
  val inputs, targets = next_training_data()
  // grad_loss returns the final result of the forward propagation
  val loss = grad_loss(lossFun(inputs, targets))
  for (par <- pars) {
    par.x -= par.d * learning_rate
    par.clear_grad()
  }
}
```

Fig. 10. Code snippet of vanilla RNN implementation in Lantern

and `AdagradOptimizer` interfaces. Both systems offer insights about how Lantern can be optimized for ease of use.

When it comes to runtime performance, Lantern’s generated and compiled C++ code outstrips the competition somewhat handily, running 5000 iterations in approximately 2.6 seconds. with the closest existing implementation being Numpy, at around 7 seconds. TensorFlow’s implementation, on the other hand, takes around 18 seconds, with PyTorch running for more than 40 seconds (Figure 11 b).⁷ We note here that we distinguish runtime for model training and preparation (everything other than the training loop, including initialization of library and model, data loading, and model compilation as in TensorFold). The runtime comparison in general is unsurprising; both Lantern and Numpy use transformed programs which flatten out the AD logic as simple calculations and function calls. In the meantime, full-fledged systems easily have more overhead since they support more features, and are expected to appear inefficient when competing on very small models. TensorFlow particularly has a longer preparation time in general, due to initialization of Curl library.

We also plot the training loss by training steps (Figure 11 a), to show that all systems reduce training loss at a similar pace. It indicates that Lantern’s implementation is correctly doing gradient descent.

We note that for all evaluations presented, we are exclusively concerned with expressivity, training loss reduction, and training time. Whether the model generalizes well to testing data is the

⁷All experiments were run using a single thread on a laptop with a dual-core AMD A9-9410 RADEON CPU @1.70GHz and 8GB of SODIMM Synchronous 2400 MHz RAM.

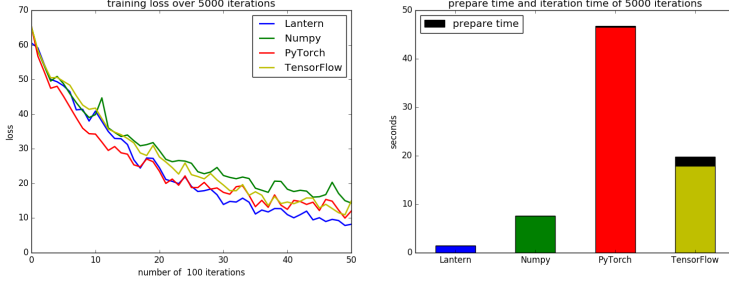


Fig. 11. (a) Training loss of Vanilla RNN model in different frameworks, (b) Training time of Vanilla RNN model in different frameworks

problem of the model or the algorithm, not the concern of the framework, and thus out of scope for this evaluation.

5.3 LSTM Implementation and Evaluation

By simple adaptation, we can implement LSTM model based on the Vanilla RNN model. As briefly mentioned in section 5.1, LSTM address the exploding and vanishing gradient problem of Vanilla RNN, with a more sophisticated gate-based model and a stable cell state for learning long-distance dependency in sequences. The detailed equations are shown below.

$$\begin{aligned}
 f_t &= \sigma(W_f * [h_{t-1}, x_t] + b_f) \\
 i_t &= \sigma(W_i * [h_{t-1}, x_t] + b_i) \\
 o_t &= \sigma(W_o * [h_{t-1}, x_t] + b_o) \\
 \tilde{c}_t &= \tanh(W_c * [h_{t-1}, x_t] + b_c) \\
 c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
 h_t &= o_t \odot \tanh(c_t)
 \end{aligned}$$

In the above equations, σ represents the sigmoid operation, $[h_{t-1}, x_t]$ means simple concatenation of the two vectors, and \odot means element-wise multiplication. As shown in the equations, the cell state c_t is only modified by “forgetting” some information controlled by the forget gate f_t , and adding some new information controlled by the input gate i_t . The output gate o_t controls which part of the cell state is used for generating hidden state h_t . Both cell state c_t and hidden state h_t need to be passed to the next recurrence.

Adding these gates to Lantern is easy: we simply add more operations in the LOOP body (Figure 12). The extra complexity does not pose any control flow challenges. For consistency, we evaluate using the same training data as the vanilla RNN. Running the generated and compiled C++ code from Lantern for 5000 iteration takes approximately 5 seconds, while the PyTorch and TensorFlow implementations runs for approximately 60 seconds and 25 seconds for the same workload (Figure 13 b). We elect not to implement the Numpy version for this or the following experiments, as performing manual differentiation is overwhelming for larger models, and thus infeasible in any practical setting.

When examining training loss, all three frameworks reduced training loss in a similar pace, which is reasonably faster than their vanilla RNN counterparts (Figure 13 a).

Demystifying Differentiable Programming: Shift/Reset the Penultimate Backpropagator

```

val pars = ... // all trainable parameters
def lossFun(inputs: Rep[Array[Int]], targets: Rep[Array[Int]]) = {
  val in = (init_loss, init_hidden, init_cell)
  val out = LOOP(in)(inputs.length){i => in =>
    val xi, yi = ... // one-hot encoding of inputs(i) and targets(i)
    val f_i = ((pars(0) dot in._2) + (pars(1) dot xi) + pars(2)).sigmoid() // forget gate
    val i_i = ((pars(3) dot in._2) + (pars(4) dot xi) + pars(5)).sigmoid() // input gate
    val o_i = ((pars(6) dot in._2) + (pars(7) dot xi) + pars(8)).sigmoid() // output gate
    val C_i = ((pars(9) dot in._2) + (pars(10) dot xi) + pars(11)).tanh() // cell update
    val c_i = f_i * in._3 + i_i * C_i // new cell state
    val h_i = o_i * c_i.tanh() // new hidden state
    val e_i = ((pars(12) dot h_i) + pars(13)).exp() // unnormalized_prob
    val p_i = e_i / e_i.sum() // normalized_prob
    val loss = in._1 - (p_i dot yi).log() // new loss
    (loss, h_i, c_i)
  }
  out(0) // return the final loss
}

```

Fig. 12. Code snippet of LSTM implementation in Lantern

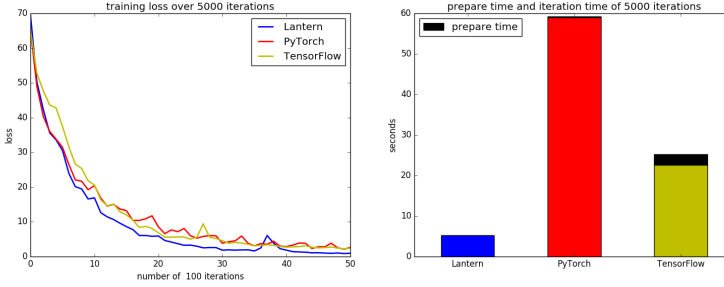


Fig. 13. (a) Training loss of LSTM in different frameworks (b) Training time of LSTM in different frameworks

5.4 Tree-Structured LSTM Implementation and Evaluation

For handling structural data that is more complicated than sequences, tree-structured LSTM (or TreeLSTM) is required. As briefly introduced in Section 5.1, TreeLSTM adapts its data flow to structural data of different sizes and shapes and captures the structural information that is otherwise missed by sequential NNs. As analyzed in detail in Section 4, a recursive control flow can be implemented using the described TREE abstraction (Figure 14).

We implemented Sentiment Classification using the dataset from the Stanford Sentiment Treebank [Chuang 2013] following the work of [Tai et al. 2015]. The dataset contains sentences for movie reviews, with each sentence parsed into a tree. Each leaf node contains a word, and each non-leaf node contains two children nodes but no word. All nodes have a label (0 to 4 in fine grained subtasks) reflecting how positive the node is. The goal is to train a LSTM which analyzes the parse tree in this manner:

$$h_i = \text{TreeLSTM}(\text{Embedding}(\text{word}), h_{i.\text{left}}, h_{i.\text{right}})$$

```

val pars = ... // all trainable parameters
def lossFun(node: Rep[Tree]) = {
  val in = (init_loss, init_hidden, init_cell)
  val out = TREE(in)(node) { (in_l, in_r) =>
    val target = one_hot(node.score)
    val embedding_tensor = IF (node.isLeaf) {Embedding(node.word)}{0}
    val i_gata = IF (node.isLeaf)
      {(pars(0) dot embedding_tensor) + pars(1)).sigmoid()}
      {(pars(2) dot in_l._2 + pars(3) dot in_r._2 + pars(1)).sigmoid()}
    val fl_gate    = ... // IF branch
    val fr_gate    = ... // IF branch
    val o_gate     = ... // IF branch
    val cell_update = ... // IF branch
    val new_cell    = i_gate * cell_update + fl_gate * in_l._3 + fr_gate * in_r._3
    val new_hidden  = o_gate * new_cell.tanh()
    val prob        = softmax(new_hidden)
    val new_loss    = in_l._1 + in_r._1 - (prob dot target).log()
    (new_loss, new_hidden, new_cell)
  }
  out(0)
}

```

Fig. 14. Code snippet of TreeLSTM implementation in Lantern

Here h_i is the hidden vector and the cell state (default when describing LSTM) associated with node i , and the Embedding is a large lookup table which maps each word to a 300-sized array, reflecting the semantic distances between all words in the vocabulary. TreeLSTM differs from simple LSTM by taking two previous states, from both the left and right children. For leaf nodes, the previous states are zero, as is the embedding for non-leaf nodes. The hidden vector from each node can be used to compute a softmax of labels, thus generating a loss by comparing with the true label for each node. By training, the total loss (or average loss per sentence) should be reduced; thus the TreeLSTM learns to evaluate reviews in a parse-tree format.

Implementations of this task in PyTorch⁸ and TensorFlow⁹ are available online, and we perform only minor adaptations for our experiments. Compared with these, Lantern’s implementation (shown above) looks simpler by use of the TREE abstraction. In fact, it looks much like the LSTM implementation, just with slightly more logic handling for whether the node is leaf or non-leaf.

It is interesting to note that implementing this task in PyTorch is not a big challenge, as PyTorch does not construct true computation graphs. For each training data, whether it is structurally identical or different, PyTorch always constructs a new computation trace by linking `torch.autograd.Variables` with operators. This is sometimes referred to as a *dynamic* computation graph.

On the other hand, standard TensorFlow machinery cannot handle structural data of various shapes, since it constructs *static* computation graphs that cannot adapt to different trees. Lantern’s methodology is closer to this TensorFlow style, but thanks to a much richer staging language

⁸<https://github.com/tpro1995/TreeLSTMSentiment>

⁹https://github.com/tensorflow/tensorflow/blob/master/tensorflow_fold/g3doc/sentiment.ipynb

provided by LMS, Lantern computation graphs can be expressed as recursive functions and function closures, which handle structural data easily.

The TensorFlow implementations in evaluation actually depend on TensorFlow Fold [Looks et al. 2017], a library that compiles static TensorFlow models based on a given set of structural data. As a somewhat ad hoc solution, this implementation may seem even more mysterious than the already clunky standard TensorFlow. However, by extensively remodeling computation graphs based on data, TensorFlow Fold can run TreeLSTM in batches, which is not supported by Lantern or PyTorch.

For simplicity (and to see much quicker convergence), we use a smaller set of training data (the dev-set, containing only 1101 sentences) in the experiment (Figure 15). Here we measure runtime by epoch (one complete traversal of the training data), because TensorFlow Fold is using different batch sizes. Lantern spends about 31 seconds per epoch, whereas PyTorch requires 75 seconds. TensorFlow Fold running on batch-size 20 is very efficient, using only 21 seconds per epoch. However, forcing TensorFlow Fold to run at batch-size 1 is very slow, at 125 seconds per epoch. It is worth noting that TensorFlow Fold uses a visible amount of preprocessing time due to the extensive graph reconstruction by Fold.

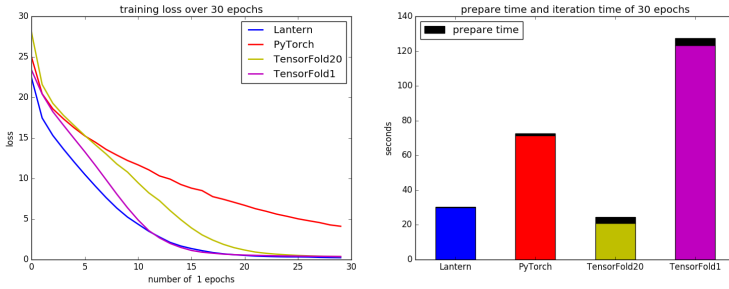


Fig. 15. (a) Training loss of TreeLSTM in different frameworks (b) Running time of TreeLSTM in different frameworks

The plot for training loss reduction diverges slightly among the frameworks. Both Lantern and TensorFlow Fold at batch-size 1 achieve fast convergence, while PyTorch and TensorFlow Fold at batch-size 20 lag behind. We are not fully clear why the training loss reduction is different between Lantern and PyTorch, even after efforts to carefully unify the hyperparameters and training conditions. TensorFlow Fold with larger batch-sizes may have a justification for learning slower, as the parameters are not updated as frequently as the other frameworks/settings, though this is simply intuition.

5.5 A Simple CNN

In order to evaluate Lantern on a simple convolutional neural network (CNN), we elect to use the widely-used MNIST dataset. MNIST is a relatively simple computer vision dataset containing handwritten digits (0-9). Neural networks are trained to classify these digits and correctly determine which digit is shown.

Lantern implementation is shown in Figure 16. Both TensorFlow and PyTorch have publicly available tutorials which operate over the MNIST dataset, though PyTorch’s implementation targets a smaller model. For uniformity, we choose to use this smaller example, and have adapted the TensorFlow tutorial accordingly. The training set is composed of 60,000 28×28 pixel grayscale images, with the testing set containing an additional 10,000 images.

```

val pars = ... // all trainable parameters
def trainFun(input: TensorR, target: Rep[Int]) = { (dummy: TensorR) =>
    val resL1 = input.conv(pars(0)).maxPool(stride).relu()
    val resL2 = resL1.conv(pars(1)).maxPool(stride).relu()
    val resL3 = ((pars(2) dot resL2.resize(in3)) + pars(3)).relu().dropout(0.5f)
    val resL4 = (pars(4) dot resL3) + pars(5)
    val res = resL4.logSoftmax()
    res.nllLoss(target)
}
    
```

Fig. 16. Code snippet of CNN implemented in Lantern

In order to evaluate Lantern using a CNN on the MNIST dataset, we must first build the model used in PyTorch’s MNIST tutorial implementation. This model is composed of two convolutional layers: the first with kernels of 5×5 and a 10-channel output, the second with kernels of 5×5 and a 20-channel output. Both layers have a max pooling of stride 2, and use ReLU as an activation function. These layers are followed by two linear layers (i.e., fully connected), between which we have a dropout of 50%. The first of these linear layers receives 320 inputs and produces 50 outputs which the second layer receives as inputs and ultimately produces 10 outputs of its own. Finally, we elect to use the `logSoftmax` function in order to compute the prediction of our CNN. We present the implementation of this as follows:

Once the gradient has been computed by our backpropagation, we use a stochastic gradient descent (SGD) algorithm with a learning rate of 5×10^{-4} when given a batch of size 1, or 5×10^{-2} when given a batch of size 100.

With our model trained, we run the MNIST benchmark using Lantern, PyTorch, and TensorFlow (Figure 17). Lantern does not currently support batches beyond size 1, but we include larger batch sizes for PyTorch and TensorFlow for completeness. Lantern completes the benchmark with an average time of 40 seconds per epoch (batch size 1). PyTorch, meanwhile, has an average of 140 seconds per epoch for batch size 1, and an average of 30 seconds for batch size 100. TensorFlow, on the other hand, has an average of 200 seconds for batch size 1, and an average of 70 seconds for batch size 100. The training loss reductions were similar in all frameworks/settings.

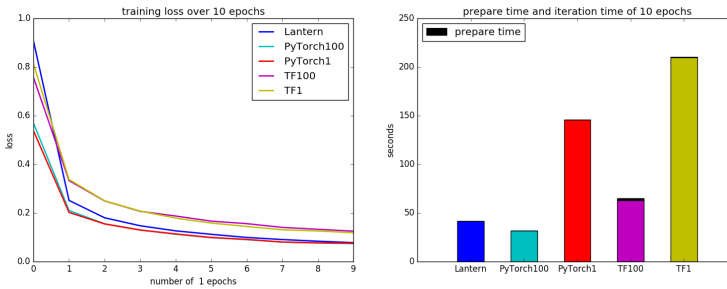


Fig. 17. (a) Training loss of CNN in different frameworks (b) Training time of CNN in different frameworks

6 RELATED WORK

Gradient-based optimization lies at the heart of machine learning. Backpropagation [Rumelhart et al. 1986] can be viewed as a special case of reverse-mode AD, and is a key ingredient for gradient descent, the primary family of algorithms used to train machine learning models. The fundamental idea of automatic differentiation (AD) emerged in the 1950s, around the time when programs emerged that had to perform calculations of derivatives [Beda et al. 1959; Nolan 1953]. A formal introduction to forward-mode AD appeared in 1960s [Wengert 1964].

The application of gradient descent first arose in control theory [Bryson and Ho 1975; Bryson and Denham 1962]. In the 1970s, Linnainmaa [1976] introduced the concept of reverse-mode AD and the concept of computational graphs which are now widely used by modern machine learning frameworks. Speelpenning [1980] implemented reverse-mode AD in a general-purpose programming language in 1980, which is considered the first implementation of reverse-mode AD that performed gradient computations automatically. At the same time, backpropagation was invented and reinvented within the machine learning community [Parker et al. 1985; Rumelhart et al. 1986; Werbos 1974]. This divergence continued until 1989 when Hecht-Nielsen [1988] cited the work from both communities.

Most modern deep learning frameworks are required to compute gradients of training loss in order to update weights in the neural network (backpropagation). Baydin et al. [2018] describe how this can be done in two ways. The first is to have users define a computational graph using some domain-specific language (DSL) and interpret operators along the graph at runtime. This computational graph structure can help DSL compilers optimize the operator-interpreting process for better performance while at the cost of the expressiveness of the DSL. This limit could make developing neural network models challenging in terms of lack of code reuse, unintuitive control structures, and difficulty in debugging. Many mainstream frameworks including Torch [Collobert et al. 2011], Theano [Al-Rfou et al. 2016], Caffe [Jia et al. 2014], TensorFlow [Abadi et al. 2016], and CNTK [Seide and Agarwal 2016] belong to this category. The other method proposed is to integrate general-purpose programming languages and truly reverse-mode automatic differentiation, of which PyTorch [Paszke et al. 2017a,b], MXNet [Chen et al. 2015], autograd [Maclaurin 2016], and Chainer [Tokui et al. 2015] are well-known representatives. The tight integration between host language and automatic differentiation of this category brings many benefits to users, including flexible control statements and easy debugging. In fact, the gap between these two styles is being bridged, with ONNX [ONNX working groups 2017] as some of the earliest known work. ONNX allows users to convert their deep learning models from one framework to another and many popular frameworks from both paradigms are developing ONNX support.

Differentiable programming is of joint interest to the machine learning and programming language communities. As deep learning models becomes more and more sophisticated, researchers have noticed that building blocks into a large neural network model is similar to using functions, and that some powerful neural network patterns are analogous to higher-order functions in functional programming [Fong et al. 2017; Olah 2015]. This is also thanks to the development of modern deep learning frameworks which make defining neural networks “very much like a regular program” [Abadi et al. 2017; LeCun 2018]. Some recent research demonstrates this direct mapping between these two areas by implementing differentiable analogues of traditional data structures [Cortes et al. 2015], and with differentiable programming, neural networks could do more than expected [Graves et al. 2014]. In the functional programming community, a similar effort is being undertaken. Siskind and Pearlmutter implemented forward-mode AD as an operator [Siskind and Pearlmutter 2008] and reverse-mode AD as lambda [Pearlmutter and Siskind 2008], all within a functional framework. After this, they augmented a high-level language with first-class AD using operator overloading

[Siskind and Pearlmutter 2016] and implemented a differentiable functional programming library called DiffSharp [Baydin et al. 2016]. A Haskell implementation of forward-mode AD was proposed by Elliott [2009]. For a thorough view of AD and deep learning from functional programming perspective, we advise readers refer to this survey [Baydin et al. 2018].

The present work tries to find a balance between those two proposed methods from two orthogonal perspectives: introducing automatic differentiation using operator overloading, and implementing a neural network model compiler without reducing the expressiveness of the host language. Previous works have attempted implementing a source-to-source deep learning compiler, but have always focused on only one of the two methods proposed by Baydin et al. [2018]. For example, Tangent [van Merriënboer et al. 2017; Wiltchko 2017] implements a source-to-source compiler in Python which supports automatic differentiation, but this framework constrains the expressiveness of the host language to a limited subset of Python. DLVM [Wei et al. 2017a,b] focuses more on the compiler side. It compiles deep learning code written in Swift into a domain-specific IR, performs some specific code analysis and transformations on this IR and generate code via LLVM targeting GPU.

Our transformation of high-level implementations of neural network models to low-level code is fueled by the idea of multi-stage programming (staging). Already more than 30 years ago, Jørring and Scherlis [1986] observed that many computations can be naturally separated into stages delineated by frequency of execution or availability of data. The idea to treat staging as an explicit *programming model* was popularized, among others, by Taha and Sheard [2000]. Since then, modern staging approaches have been proposed which blend normal program execution with delayed construction of an *intermediate program representation* (IR), which may be a computation graph or a more customary abstract syntax tree (AST). An example is the Lightweight Modular Staging (LMS) framework [Rompf and Odersky 2010], which provides a rather seamless implementation of staging in the Scala language and has been utilized in a range of existing applications [Rompf and Amin 2015; Rompf et al. 2015; Sujeeth et al. 2011]. Lantern, our deep learning framework which is built on top of LMS, achieves a balance between source code expressiveness and target code performance via this compiler building technique.

Lantern also relies on delimited continuations, as implemented in Scala [Rompf et al. 2009]. Our success here depends greatly upon delimited continuations in order to embed reverse-mode automatic differentiation in an elegant way.

7 CONCLUSIONS

With this paper, we set out to demystify automatic differentiation by examining it through the lense of program transformation. In doing so, we uncovered a tight connection between reverse-mode AD and delimited continuations. With the help of delimited continuation control operators, we provided an implementation of reverse-mode AD by pure local transformations via operator overloading and without any auxiliary data structures. Our work follows on the functional “Lambda, the ultimate backpropagator” view of Pearlmutter and Siskind [2008] with a powerful frontend over lambda terms in CPS – hence building the “penultimate backpropagator.”

We further combined this formulation of AD with multi-stage programming (staging), which leads to a highly efficient implementation that combines the performance benefits of deep learning frameworks based on explicit reified computation graphs (e.g., TensorFlow) with the expressiveness of pure library approaches (e.g., PyTorch).

Based on these two ideas, we built our deep learning framework prototype, which we named Lantern, as another step towards practical *differentiable programming*. With a simple backend that generates C++ code natively, we already show competitive performance based on a few deep learning benchmarks such as vanilla RNN, LSTM, TreeLSTM, and CNN. With the perspective of

Demystifying Differentiable Programming: Shift/Reset the Penultimate Backpropagator

plugging our front-end with known tensor IR processing back-end such as XLA, NNVM or Delite, we provide the necessary core for deep learning frameworks emphasizing both efficiency and expressivity.

REFERENCES

- Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *CoRR* abs/1603.04467 (2016). [arXiv:1603.04467](https://arxiv.org/abs/1603.04467) [http://arxiv.org/abs/1603.04467](https://arxiv.org/abs/1603.04467)
- Martin Abadi, Michael Isard, and Derek G. Murray. 2017. A Computational Model for TensorFlow (An Introduction). <http://dl.acm.org/citation.cfm?doid=3088525.3088527>
- Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermüller, Dzmitry Bahdanau, Nicolas Ballas, Frédéric Bastien, Justin Bayer, Anatoly Belikov, Alexander Belopolsky, Yoshua Bengio, Arnaud Bergeron, James Bergstra, Valentin Bisson, Josh Blecher Snyder, Nicolas Bouchard, Nicolas Boulanger-Lewandowski, Xavier Bouthillier, Alexandre de Brébisson, Olivier Breuleux, Pierre Luc Carrier, Kyunghyun Cho, Jan Chorowski, Paul F. Christiano, Tim Cooijmans, Marc-Alexandre Côté, Myriam Côté, Aaron C. Courville, Yann N. Dauphin, Olivier Delalleau, Julien Demouth, Guillaume Desjardins, Sander Dieleman, Laurent Dinh, Melanie Ducoffe, Vincent Dumoulin, Samira Ebrahimi Kahou, Dumitru Erhan, Ziye Fan, Orhan Firat, Mathieu Germain, Xavier Glorot, Ian J. Goodfellow, Matthew Graham, Çağlar Gülçehre, Philippe Hamel, Iban Harlouchet, Jean-Philippe Heng, Balázs Hidasi, Sina Honari, Arjun Jain, Sébastien Jean, Kai Jia, Mikhail Korobov, Vivek Kulkarni, Alex Lamb, Pascal Lamblin, Eric Larsen, César Laurent, Sean Lee, Simon Lefrançois, Simon Lemieux, Nicholas Léonard, Zhouhan Lin, Jesse A. Livezey, Cory Lorenz, Jeremiah Lowin, Qianli Ma, Pierre-Antoine Manzagol, Olivier Mastropietro, Robert McGibbon, Roland Memisevic, Bart van Merriënboer, Vincent Michalski, Mehdi Mirza, Alberto Orlandi, Christopher Joseph Pal, Razvan Pascanu, Mohammad Pezeshki, Colin Raffel, Daniel Renshaw, Matthew Rocklin, Adriana Romero, Markus Roth, Peter Sadowski, John Salvatier, François Savard, Jan Schlüter, John Schulman, Gabriel Schwartz, Iulian Vlad Serban, Dmitriy Serdyuk, Samira Shabanian, Étienne Simon, Sigurd Spieckermann, S. Ramana Subramanyam, Jakub Sygnowski, Jérémie Tanguay, Gijs van Tulder, Joseph P. Turian, Sebastian Urban, Pascal Vincent, Francesco Visin, Harm de Vries, David Warde-Farley, Dustin J. Webb, Matthew Willson, Kelvin Xu, Lijun Xue, Li Yao, Saizheng Zhang, and Ying Zhang. 2016. Theano: A Python framework for fast computation of mathematical expressions. *CoRR* abs/1605.02688 (2016). [arXiv:1605.02688](https://arxiv.org/abs/1605.02688) [http://arxiv.org/abs/1605.02688](https://arxiv.org/abs/1605.02688)
- Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. 2018. Automatic differentiation in machine learning: a survey. *CoRR* abs/1502.05767 (2018).
- Atilim Günes Baydin, Barak A. Pearlmutter, and Jeffrey Mark Siskind. 2016. DiffSharp: An AD Library for .NET Languages. *CoRR* abs/1611.03423 (2016).
- L. M. Beda, L. N. Korolev, N. V. Sukkikh, and T. S. Frolova. 1959. *Programs for automatic differentiation for the machine BESM*. Technical Report. Institute for Precise Mechanics and Computation Techniques, Academy of Science, Moscow, USSR. (In Russian).
- Malgorzata Biernacka, Dariusz Biernacki, and Olivier Danvy. 2005. An Operational Foundation for Delimited Continuations in the CPS Hierarchy. *Logical Methods in Computer Science* 1, 2 (2005). [https://doi.org/10.2168/LMCS-1\(2:5\)2005](https://doi.org/10.2168/LMCS-1(2:5)2005)
- A Bryson and Yu-Chi Ho. 1975. Applied optimal control: Optimization, estimation, and control (revised edition). *Levittown, Pennsylvania: Taylor & Francis* (1975).
- Arthur E Bryson and Walter F Denham. 1962. A steepest-ascent method for solving optimum programming problems. *Journal of Applied Mechanics* 29, 2 (1962), 247–257.
- Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. 2015. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274* (2015).
- Bruce Christianson. 1992. Automatic Hessians by reverse accumulation. *IMA J. Numer. Anal.* 12, 2 (1992), 135–150.
- Jason Chuang. 2013. Stanford Sentiment Treebank. (2013). <https://nlp.stanford.edu/sentiment/treebank.html>
- Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. 2011. Torch7: A Matlab-like Environment for Machine Learning. In *BigLearn, NIPS Workshop*.
- Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 2015. *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*. <http://papers.nips.cc/book/advances-in-neural-information-processing-systems-28-2015>
- Olivier Danvy and Andrzej Filinski. 1990. Abstracting Control. In *LISP and Functional Programming*. 151–160.

- Olivier Danvy and Mayer Goldberg. 2005. There and back again. *Fundamenta Informaticae* 66, 4 (2005), 397–413.
- Distributed (Deep) Machine Learning Community. 2018. NNVM: Open Compiler for AI Frameworks. (2018). <https://github.com/dmlc/nnvm>
- John C. Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research* 12 (2011), 2121–2159. <http://dl.acm.org/citation.cfm?id=2021068>
- Conal Elliott. 2009. Beautiful differentiation. In *International Conference on Functional Programming (ICFP)*. <http://conal.net/papers/beautiful-differentiation>
- Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science* 14, 2 (1990), 179–211.
- Cormac Flanagan, Amr Sabry, Bruce F. Duba, and Matthias Felleisen. 1993. The Essence of Compiling with Continuations. In *PLDI*. ACM, 237–247.
- Brendan Fong, David I Spivak, and Rémy Tuyéras. 2017. Backprop as Functor: A compositional perspective on supervised learning. *arXiv preprint arXiv:1711.10455* (2017).
- Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural Turing Machines. *CoRR* abs/1410.5401 (2014). arXiv:1410.5401 <http://arxiv.org/abs/1410.5401>
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. IEEE Computer Society, 770–778.
- Robert Hecht-Nielsen. 1988. Theory of the backpropagation neural network. *Neural Networks* 1, Supplement-1 (1988), 445–448. [https://doi.org/10.1016/0893-6080\(88\)90469-8](https://doi.org/10.1016/0893-6080(88)90469-8)
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.
- Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *CoRR* abs/1408.5093 (2014). arXiv:1408.5093 <http://arxiv.org/abs/1408.5093>
- Ulrik Jørring and William L. Scherlis. 1986. Compilers and Staging Transformations. In *POPL*. ACM Press, 86–96.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). arXiv:1412.6980 <http://arxiv.org/abs/1412.6980>
- Stefan Kombrink, Tomas Mikolov, Martin Karafiát, and Lukás Burget. 2011. Recurrent Neural Network Based Language Modeling in Meeting Recognition. In *INTERSPEECH*. ISCA, 2877–2880.
- John Launchbury and Simon L. Peyton Jones. 1994. Lazy Functional State Threads. In *PLDI*. ACM, 24–35.
- Yann LeCun. 2018. LcCun’s facebook pose on Differentiable programming. (2018). <https://www.facebook.com/yann.lecun/posts/10155003011462143>
- Yann LeCun, Bernhard E Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne E Hubbard, and Lawrence D Jackel. 1990. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*. 396–404.
- Seppo Linnainmaa. 1976. Taylor expansion of the accumulated rounding error. *BIT Numerical Mathematics* 16, 2 (1976), 146–160.
- Moshe Looks, Marcello Herreshoff, DeLesley Hutchins, and Peter Norvig. 2017. Deep learning with dynamic computation graphs. *arXiv preprint arXiv:1702.02181* (2017).
- Dougal Maclaurin. 2016. *Modeling, Inference and Optimization with Composable Differentiable Procedures*. Ph.D. Dissertation.
- John F Nolan. 1953. Analytical differentiation on a digital computer. (1953).
- Christopher Olah. 2015. Neural Networks, Types, and Functional Programming. (2015). <http://colah.github.io/posts/2015-09-NN-Types-FP/>
- ONNX working groups. 2017. ONNX: Open Neural Network Exchange format. (2017). <https://onnx.ai/>
- D.B. Parker, Massachusetts Institute of Technology, and Sloan School of Management. 1985. *Learning Logic: Casting the Cortex of the Human Brain in Silicon*. Massachusetts Institute of Technology, Center for Computational Research in Economics and Management Science. <https://books.google.com/books?id=2kS9GwAACAAJ>
- Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. 2017a. PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration. (2017). www.pytorch.org
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017b. Automatic differentiation in PyTorch. (2017).
- Barak A. Pearlmutter and Jeffrey Mark Siskind. 2008. Reverse-mode AD in a functional framework: Lambda the ultimate backpropagator. *ACM Trans. Program. Lang. Syst.* 30, 2 (2008), 7:1–7:36.
- Ning Qian. 1999. On the momentum term in gradient descent learning algorithms. *Neural Networks* 12, 1 (1999), 145–151. [https://doi.org/10.1016/S0893-6080\(98\)00116-6](https://doi.org/10.1016/S0893-6080(98)00116-6)
- Tiark Rompf and Nada Amin. 2015. Functional pearl: a SQL to C compiler in 500 lines of code. In *ICFP*. ACM, 2–9.
- Tiark Rompf, Kevin J. Brown, HyoukJoong Lee, Arvind K. Sujeeth, Manohar Jonnalagedda, Nada Amin, Georg Ofenbeck, Alen Stojanov, Yannis Klonatos, Mohammad Dashti, Christoph Koch, Markus Püschel, and Kunle Olukotun. 2015. Go Meta! A Case for Generative Programming and DSLs in Performance Critical Systems. In *SNAPL (LIPICs)*, Vol. 32. Schloss

Demystifying Differentiable Programming: Shift/Reset the Penultimate Backpropagator

- Dagstuhl - Leibniz-Zentrum fuer Informatik, 238–261.
- Tiark Rompf, Ingo Maier, and Martin Odersky. 2009. Implementing first-class polymorphic delimited continuations by a type-directed selective CPS-transform. In *ICFP*. ACM, 317–328.
- Tiark Rompf and Martin Odersky. 2010. Lightweight modular staging: a pragmatic approach to runtime code generation and compiled DSLs. In *GPCE*. ACM, 127–136.
- David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature* 323, 6088 (1986), 533.
- Frank Seide and Amit Agarwal. 2016. Cntk: Microsoft’s open-source deep-learning toolkit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2135–2135.
- Jeffrey Mark Siskind and Barak A. Pearlmutter. 2008. Nesting forward-mode AD in a functional framework. *Higher-Order and Symbolic Computation* 21, 4 (2008), 361–376.
- Jeffrey Mark Siskind and Barak A. Pearlmutter. 2016. Efficient Implementation of a Higher-Order Language with Built-In AD. *CoRR* abs/1611.03416 (2016).
- Bert Speelpenning. 1980. *Compiling fast partial derivatives of functions given by algorithms*. Ph.D. Dissertation.
- Arvind K. Sujeeth, Kevin J. Brown, HyoukJoong Lee, Tiark Rompf, Hassan Chafi, Martin Odersky, and Kunle Olukotun. 2014. Delite: A Compiler Architecture for Performance-Oriented Embedded Domain-Specific Languages. *ACM Trans. Embedded Comput. Syst.* 13, 4s (2014), 134:1–134:25.
- Arvind K. Sujeeth, HyoukJoong Lee, Kevin J. Brown, Tiark Rompf, Hassan Chafi, Michael Wu, Anand R. Atreya, Martin Odersky, and Kunle Olukotun. 2011. OptiML: An Implicitly Parallel Domain-Specific Language for Machine Learning. In *ICML*. Omnipress, 609–616.
- Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In *AAAI*. AAAI Press, 4278–4284.
- Walid Taha and Tim Sheard. 2000. MetaML and multi-stage programming with explicit annotations. *Theor. Comput. Sci.* 248, 1-2 (2000), 211–242.
- Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *CoRR* abs/1503.00075 (2015). arXiv:1503.00075 <http://arxiv.org/abs/1503.00075>
- TensorFlow team. 2018. XLA Overview. (2018). <https://www.tensorflow.org/performance/xla/>
- Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: a next-generation open source framework for deep learning. In *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, Vol. 5.
- B. van Merriënboer, A. B. Wiltschko, and D. Moldovan. 2017. Tangent: Automatic Differentiation Using Source Code Transformation in Python. *ArXiv e-prints* (Nov. 2017). arXiv:cs.MS/1711.02712
- Richard Wei, Vikram S. Adve, and Lane Schwartz. 2017a. DLVM: A modern compiler infrastructure for deep learning systems. *CoRR* abs/1711.03016 (2017).
- Richard Wei, Lane Schwartz, and Vikram Adve. 2017b. A modern compiler infrastructure for deep learning systems with adjoint code generation in a domain-specific IR. In *NIPS AutoDiff Workshop*.
- R. E. Wengert. 1964. A simple automatic derivative evaluation program. *Commun. ACM* 7, 8 (1964), 463–464. <https://doi.org/10.1145/355586.364791>
- Paul Werbos. 1974. Beyond regression: New tools for prediction and analysis in the behavior science. *Unpublished Doctoral Dissertation, Harvard University* (1974).
- Alex Wiltschko. 2017. Tangent: Source-to-Source Debuggable Derivatives. (2017). <https://research.googleblog.com/2017/11/tangent-source-to-source-debuggable.html>