

Demystifying Differentiable Programming: Shift/Reset the Penultimate Backpropagator

FEI WANG, Purdue University, USA

DANIEL ZHENG, Purdue University, USA

JAMES DECKER, Purdue University, USA

XILUN WU, Purdue University, USA

GRÉGORIE M. ESSERTEL, 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 crucially 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 managing any auxiliary data structures. We further show how this formulation of AD can be fruitfully combined with multi-stage programming (staging), leading to an 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).

CCS Concepts: • **Software and its engineering** → **Domain specific languages**.

Additional Key Words and Phrases: Delimited Continuations, Multi-stage Programming, Differentiable Programming, Automated Differentiation

ACM Reference Format:

Fei Wang, Daniel Zheng, James Decker, Xilun Wu, Grégory M. Essertel, and Tiark Rompf. 2019. Demystifying Differentiable Programming: Shift/Reset the Penultimate Backpropagator. *Proc. ACM Program. Lang.* 3, ICFP, Article 96 (August 2019), 31 pages. <https://doi.org/10.1145/3341700>

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 now match or surpass 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 task. Hence, a crucial ability is that of backpropagating errors through the network to compute the gradient of a loss

Authors’ addresses: Fei Wang, Purdue University, USA; Daniel Zheng, Purdue University, USA; James Decker, Purdue University, USA; Xilun Wu, Purdue University, USA; Grégory M. Essertel, Purdue University, USA; Tiark Rompf, Purdue University, USA.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2019 Copyright held by the owner/author(s).

2475-1421/2019/8-ART96

<https://doi.org/10.1145/3341700>

function [Rumelhart et al. 1986]. Beyond this commonality, however, deep learning architectures vary widely. 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* [LeCun 2018; Olah 2015] — essentially, functional programming with first-class gradients — 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 and compiler writers, key players in this vision, are faced with the challenge of adding efficient and expressive program differentiation capabilities. Forms of automatic gradient 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 while the forward-mode flavor of AD is easy to grasp, descriptions of the reverse-mode flavor that generalizes backpropagation often appear mysterious to PL researchers. A notable exception is the seminal work of Pearlmutter and Siskind [2008], which cast AD in a functional programming framework and laid the groundwork for first-class, unrestricted, gradient operators in a functional language. Recent work by Elliott [2018] presented a unification of forward- and reverse-mode AD based on the “compiling to categories” approach [Elliott 2017], translating Haskell code to parameterized cartesian closed categories. However, the technique still needs primitive functor-level loop-style operations such as `map`, `sum`, and `zip`, and currently lacks support for general recursion or Turing-completeness.

The goal of the present work is to further demystify differentiable programming and reverse-mode AD for a PL audience, and to reconstruct the forward- and reverse-mode AD approaches based on well-understood program transformation techniques, without relying on category theory. We describe forward-mode AD as the 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 operators¹. 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].

Delimited control operators lead to an expressive implementation in the (define-by-run) style of PyTorch. We further show how to combine this approach with *multi-stage programming* to derive a framework in the (define-then-run) style of TensorFlow. The result is a highly-efficient and expressive DSL, dubbed Lantern², 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.

¹Our 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”.

²<https://github.com/feiwang3311/Lantern>

We first presented the idea of reverse-mode AD via delimited continuations and staging as a poster and accompanying abstract in the workshop track at ICLR [Wang and Rompf 2018], followed by a detailed tech-report on arXiv [Wang et al. 2018b]. We then presented this idea to the Machine Learning community at NeurIPS [Wang et al. 2018a], along with an evaluation of our prototypic implementation (which only supported a CPU backend) of the framework Lantern. The NeurIPS paper focused primarily on the intuitions and high-level ideas; no formal presentation was provided. The current paper presents a unified view of automatic differentiation from a PL perspective and extends earlier publications through the following contributions:

- We first bridge the conceptual distinction between automatic differentiation and symbolic differentiation by casting forward-mode AD as the application of standard high-school symbolic differentiation rules on ANF-transformed terms, with only constant expression size increase. Based on that insight, we define a formal transformation that implements forward-mode AD directly (Section 2).
- We then analyze reverse-mode AD, and relate its “there and back again” computation flow pattern to programs using nested continuations, as seen in CPS (continuation-passing style). By presenting detailed formal transformations (available as artifact online³) for reverse-mode AD based on CPS, with or without the use of control operators (shift/reset), in the target or meta-language, we reveal the formal relationship between reverse-mode AD and CPS transformation (Section 3).
- We demonstrate different ways to combine our forward- and reverse-mode AD for higher order gradients, and present a concrete OO-style class hierarchy for higher-order AD (Sections 2.4 2.5 3.7, available as artifact online³). We also discuss the question of mutability and describe one way to make reverse-mode AD purely functional via store-passing using an immutable map data-structure (Section 3.6).
- We illustrate the interplay between CPS transformation and staging, and relate the implementation of control flow operations (IF, WHILE, and TREE as a representative of recursive machine learning models) to formal rules for reverse-mode AD transformation. We also demonstrate examples showing intermediate code generation steps (Section 4).
- We demonstrate the performance of the complete Lantern framework on realistic benchmark models (TreeLSTM, SqueezeNet, ResNet, and DeepSpeech2) on GPU (Section 5).

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

2 DIFFERENTIABLE PROGRAMMING BASICS

Broadly speaking, a neural network is a specific kind of parameterized function approximator \hat{f}_w . The training process optimizes the parameters w to improve the approximation of an unknown *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 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 of the form \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 .

³<https://github.com/feiwang3311/demystifying-ad>

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

This problem is well recognized in the AD community and often cited as a 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:

$$\begin{array}{lll}
 d/dx \llbracket \text{let } y_1 = e_1 \text{ in} & = & \text{let } y_1 = e_1 & \text{in let } y'_1 = d/dx \llbracket e_1 \rrbracket \text{ in} \\
 \dots & & \dots & \\
 \text{let } y_n = e_n \text{ in} & & \text{let } y_n = e_n & \text{in let } y'_n = d/dx \llbracket e_n \rrbracket \text{ in} \\
 \text{let } z_1 = y_1 * y_2 \text{ in} & & \text{let } z_1 = y_1 * y_2 & \text{in let } z'_1 = y'_1 * y_2 + y_1 * y'_2 \text{ in} \\
 \text{let } z_2 = z_1 * y_3 \text{ in} & & \text{let } z_2 = z_1 * y_3 & \text{in let } z'_2 = z'_1 * y_3 + z_1 * y'_3 \text{ in} \\
 \dots & & \dots & \\
 \text{let } z_{n-1} = z_{n-2} * y_n \text{ in} & & \text{let } z_{n-1} = z_{n-2} * y_n & \text{in let } z'_{n-1} = z'_{n-2} * y_n + z_{n-2} * y'_n \\
 z_{n-1} \rrbracket & & \text{in } z'_{n-1} &
 \end{array}$$

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 binding. No expression is evaluated more often than in the original computation.

This example uses the standard symbolic differentiation rules for addition and multiplication, but also makes key use of the let rule in Figure 1, which splits a binding $\text{let } y = \dots$ into $\text{let } y = \dots$ and $\text{let } y' = \dots$. Using terminology from the AD community, we call y the *primal* and y' the *tangent*. The rules in Figure 1 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 bracket syntax to emphasize that symbolic differentiation is a syntactic transformation.

Symbolic differentiation of ANF-transformed terms maintains the asymptotic runtime-complexity. Let us consider a concrete example: $y = 2 * x + x * x * x$. We start from its ANF-transformed form.

$$\begin{array}{lll}
 d/dx \llbracket \text{let } y_1 = 2 * x \text{ in} & = & \text{let } y_1 = 2 * x \text{ in} & \text{let } y'_1 = 2 * x' \text{ in} \\
 \text{let } y_2 = x * x \text{ in} & & \text{let } y_2 = x * x \text{ in} & \text{let } y'_2 = x' * x + x * x' \text{ in} \\
 \text{let } y_3 = y_2 * x \text{ in} & & \text{let } y_3 = y_2 * x \text{ in} & \text{let } y'_3 = y'_2 * x + y_2 * x' \text{ in} \\
 \text{let } y = y_1 + y_3 \text{ in} & & \text{let } y = y_1 + y_3 \text{ in} & \text{let } y' = y'_1 + y'_3 \text{ in} \\
 y \rrbracket & & y' &
 \end{array}$$

Note that we differentiate with respect to x , and $x' = 1$. The computation of derivatives follows the rules in Figure 1. The final two let bindings compute the primal value y and the tangent y' of y with respect to x . The tangent is returned as the result of the program after transformation. We can confirm the correctness of the calculation easily, which reduces to $2 + 3 * x * x$.

For a generic straight-line program, we can see this pattern of computation of forward-mode AD (Figure 2). The abstract flow of the forward-mode AD is depicted in Figure 2 on the right. We use squares to denote value computations, and triangles to denote gradient computations. The transformed program interleaves value computations with gradient computations (Forward 1). We can further combine each value computation with its gradient computation (Forward 2), so that the transformation can be realized via operator overloading.

For straight-line programs, applying ANF conversion followed by symbolic differentiation achieves exactly the 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” is ultimately harmful and contributes to the mystical appearance of the field. We believe that

We denote p_{tk} as the k th parameter of t th computation,
 where $p_{tk} \in \{c\} \cup \{x\} \cup \{y_j | j < t\}$

let $y_1 = p_{11} \oplus p_{12}$ in
 let $y'_1 = {}^{d/dp_{11}} \llbracket p_{11} \oplus p_{12} \rrbracket * p'_{12} + {}^{d/dp_{12}} \llbracket p_{11} \oplus p_{12} \rrbracket * p'_{11}$ in
 let $y_2 = p_{21} \oplus p_{22}$ in
 let $y'_2 = {}^{d/dp_{21}} \llbracket p_{21} \oplus p_{22} \rrbracket * p'_{22} + {}^{d/dp_{22}} \llbracket p_{21} \oplus p_{22} \rrbracket * p'_{21}$ in
 ...
 let $y_n = p_{n1} \oplus p_{n2}$ in
 let $y'_n = {}^{d/dp_{n1}} \llbracket p_{n1} \oplus p_{n2} \rrbracket * p'_{n2} + {}^{d/dp_{n2}} \llbracket p_{n1} \oplus p_{n2} \rrbracket * p'_{n1}$ in
 y'_n

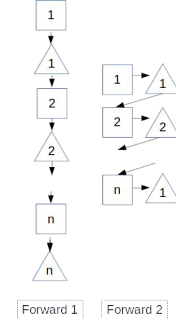


Fig. 2. Pattern of computation of forward-mode AD for generic straight-line program

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 Calculus

CORE LANGUAGE

Expressions:

EXP
 $e ::= c \mid x \mid e + e \mid e * e \mid \lambda x. e \mid @ e e \mid \text{let } x = e \text{ in } e$
 $\mid \text{fst } e \mid \text{snd } e \mid (e, e) \mid \text{ref } e \mid ! e \mid e := e$
 $\mid \text{inl } e \mid \text{inr } e \mid \text{case } e \text{ of } x \Rightarrow e \text{ or } x \Rightarrow e$
 $\mid \text{shift } x \text{ in } e \mid \langle e \rangle$

Values:

VAL
 structure left abstract

DERIVED CONSTRUCTS

Booleans and conditionals:

Value: True = inl ()
 Value: False = inr ()
 if b then t else e = case b of $y \Rightarrow t$ or $z \Rightarrow e$

Loops and recursion:

letrec $f = \lambda x. e_1$ in e_2 = let $f_0 = \lambda f_1. \lambda x. \text{let } f = @ f_1 f_1 \text{ in } e_1$ in
 let $f = @ f_0 f_0$ in e_2
 Loops: expressed as tail recursive functions

Tree data structures:

Example tree term: t = inr (inl 5, inl 6)

Syntactic sugar:

$y_1 += ! y_2$ = $y_1 := ! y_1 + ! y_2$
 let $(y, y') = e_1$ in e_2 = let $\tilde{y} = e_1$ in let $y = \text{fst } \tilde{y}$ in let $y' = \text{snd } \tilde{y}$ in e_2
 $e_1 ; e_2$ = let $_ = e_1$ in e_2

Fig. 3. Formal definition of the language we consider. It serves as both object- and meta-language (for transformation). We show the syntax of the core languages (untyped, but types can be added), as well as derived constructs that express branches, loops, recursion, and recursive data structures in a standard way. Syntactic sugar used in our presentation is also listed here.

We assume Barendregt's variable convention throughout, such that all bound variables are pairwise different and different from the free variables. This allows several rules to be simplified compared to other formulations (no need for variable substitutions in transformations).

For transformation, we assume that the target language is the same as the object language unless noted otherwise.

We now proceed beyond straight-line programs and formalize a variant of λ -calculus with let bindings, products, sum-type constructs (inl, inr, case), and mutable state (Figure 3). The language

For automatic differentiation (both forward-mode and reverse-mode in later sections), we use the following variable sugaring $\hat{\cdot}$ notation. This variable sugaring is not strictly necessary but we find it convenient for $+$ and $*$ rules. Also, note that this variable sugaring is always used at positions where we know for sure that the sugared variables bind with \mathbb{R} typed values, so that they must have gradients (denoted via variables with $'$).

$$\text{Variable Sugaring: } \hat{y} = (y, y')$$

Also note that for AD (both forward-mode and reverse-mode), we drop shift/reset terms from the source language, since the focus is to provide a semantics for AD in a standard language, and shift/reset will play a crucial role for the semantics of AD transformation in reverse mode. Our AD also supports mutable state in the source language.

$$\begin{aligned} \text{Transform}(f) &= \lambda x. \text{let } \hat{y} = @ \vec{\mathcal{D}}[f](x, 1) \text{ in } y' \\ \text{where } \vec{\mathcal{D}}[\cdot] : & \text{EXP} \rightarrow \text{EXP} \text{ is defined as below:} \end{aligned}$$

$$\begin{aligned} \vec{\mathcal{D}}[c] &= c \text{ if } c \notin \mathbb{R} \\ \vec{\mathcal{D}}[c] &= (c, 0) \text{ if } c \in \mathbb{R} \\ \vec{\mathcal{D}}[y] &= y \\ \vec{\mathcal{D}}[e_1 + e_2] &= \text{let } \hat{y}_1 = \vec{\mathcal{D}}[e_1] \text{ in} \\ & \quad \text{let } \hat{y}_2 = \vec{\mathcal{D}}[e_2] \text{ in} \\ & \quad (y_1 + y_2, y'_1 + y'_2) \\ \vec{\mathcal{D}}[e_1 * e_2] &= \text{let } \hat{y}_1 = \vec{\mathcal{D}}[e_1] \text{ in} \\ & \quad \text{let } \hat{y}_2 = \vec{\mathcal{D}}[e_2] \text{ in} \\ & \quad (y_1 * y_2, y_1 * y'_2 + y'_1 * y_2) \\ \vec{\mathcal{D}}[\lambda y. e] &= \lambda y. \vec{\mathcal{D}}[e] \\ \vec{\mathcal{D}}[@ e_1 e_2] &= @ \vec{\mathcal{D}}[e_1] \vec{\mathcal{D}}[e_2] \\ \vec{\mathcal{D}}[\text{let } y = e_1 \text{ in } e_2] &= \text{let } y = \vec{\mathcal{D}}[e_1] \text{ in } \vec{\mathcal{D}}[e_2] \\ \vec{\mathcal{D}}[\text{fst } e] &= \text{fst } \vec{\mathcal{D}}[e] \\ \vec{\mathcal{D}}[\text{snd } e] &= \text{snd } \vec{\mathcal{D}}[e] \\ \vec{\mathcal{D}}[\text{ref } e] &= \text{ref } \vec{\mathcal{D}}[e] \\ \vec{\mathcal{D}}[! e] &= ! \vec{\mathcal{D}}[e] \\ \vec{\mathcal{D}}[e_1 := e_2] &= \vec{\mathcal{D}}[e_1] := \vec{\mathcal{D}}[e_2] \\ \vec{\mathcal{D}}[(e_1, e_2)] &= (\vec{\mathcal{D}}[e_1], \vec{\mathcal{D}}[e_2]) \\ \vec{\mathcal{D}}[\text{inl } e] &= \text{inl } \vec{\mathcal{D}}[e] \\ \vec{\mathcal{D}}[\text{inr } e] &= \text{inr } \vec{\mathcal{D}}[e] \\ \vec{\mathcal{D}}[\text{case } e \text{ of } y_1 \Rightarrow e_1 \text{ or } y_2 \Rightarrow e_2] &= \text{case } \vec{\mathcal{D}}[e] \text{ of } y_1 \Rightarrow \vec{\mathcal{D}}[e_1] \text{ or } y_2 \Rightarrow \vec{\mathcal{D}}[e_2] \end{aligned}$$

Fig. 4. Transformation rules for forward-mode AD. Note that there is no metalanguage redex generated in the transformation, so by default, all constructs on the right-hand-sides are dynamic/target language constructs. Rules that are different from the standard are highlighted in blue.

also contains delimited control operators shift and reset (denoted via $\langle \cdot \rangle$), which will be used in later sections. Note that the language is untyped, though types can be added in a standard way. Control operators (shift/reset) and mutable state are orthogonal features, so their interaction does not pose any difficulties.

We define a new differentiation operator $\vec{\mathcal{D}}[e]$, where the arrow indicates forward-mode, and provide the forward-mode AD transformation rules in Figure 4. Note that differentiation is still with respect to a fixed x . However, we always transform abstractions (for any non-abstraction term e , we add an η -redex, and perform $@ \vec{\mathcal{D}}[\lambda x. e](x, 1)$). By Barendregt's variable convention, $\vec{\mathcal{D}}[\cdot]$ never applies to the special variable x , thus the $\vec{\mathcal{D}}[x]$ rule is elided in the formal presentation.

Compared to Section 2.1, 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 due

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 handles computations in different scopes easily because, in function applications, the `let` insertions require both the primal and tangent of the parameters to perform the tangent computation. Since the transformation is purely local, working with pairs of numeric expressions makes it immediately clear that this strategy can be implemented easily in standard programming languages by operator overloading. This is standard practice, which we illustrate through our implementation in Scala (Figure 5).

```
// 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 and test.
val df = grad(x => 2*x + x*x*x)
forall { x => df(x) == 2 + 3*x*x }
```

Fig. 5. Forward-mode AD in Scala (operator overloading)

The `NumF` class encapsulates the primal as `x` and the tangent as `d`, with arithmetic 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 `NumF => NumF` (Figure 5 upper right). Internally, `grad` invokes its argument function with a tangent value of 1 and returns the tangent field of the function result. In line with the previous sections, we only handle scalar functions, but the approach generalizes to multidimensional functions as well. An example using the `grad` operator is shown in Figure 5 lower right. Note that the constant 2 is implicitly converted to `new NumF(2.0, 0.0)` (tangents of constants are 0.0 because constants do not change). The use of `Double` instead of a generic number type is simply for clarity of presentation. Note how the implementations in Figure 5 correspond directly to the formal rules in Figure 4.

2.4 Nested Gradient Invocation and Perturbation Confusion

In the current implementation, we can compute the gradient of any function of type `NumF => NumF` with respect to any given value using forward-mode AD. However, our `grad` function is not truly first-class, since we cannot apply it in a nested fashion, as in `grad(grad(f))`. This prevents us from computing higher order derivatives, and from solving nested min/max problems in the form of:

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

Yet, even this somewhat restricted operator has a few subtleties. There is a common issue 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! Unexpected.
  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 the 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 invocations differentiate with respect to different variables (outer grad wrt. x , inner grad wrt. y), and that their gradient updates should not be mixed. We do not provide any new solutions for perturbation confusion, but our implementation can be easily extended to support known solutions, either based on *dynamic tagging* or based on types as realized in Haskell⁵, which lifts tags into the type system using rank-2 polymorphism, just like the ST monad [Launchbury and Peyton Jones 1994].

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, while preventing perturbation confusion. 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 tags into account to avoid confusing 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 = {...}
```

This class hierarchy provides a flexible way to compose higher-order gradient computation (implementation available online³). Alternative implementations that use parametric types and type classes instead of OO-style inheritance are also possible.

This concludes the core ideas of forward-mode AD. Implementations based on operator overloading are simple and direct, and exist in many languages. As noted earlier, we propose that forward-mode AD 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 value-derivative pairs (i.e. primals and tangents).

3 DIFFERENTIABLE PROGRAMMING WITH REVERSE-MODE AD

Forward-mode AD is straightforward to implement and generalizes to functions with multiple inputs and outputs. However, it is inefficient for functions with many inputs, and neural networks generally have many inputs and few outputs. 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 this leads to $O(n)$ more operations than the original function. Is there a better approach?

We consider again $f: \mathbb{R}^n \rightarrow \mathbb{R}$ represented as a straight-line program in ANF, i.e., as a sequence of let $y_j = e_j$ expressions, with inputs x_i and output y_m . The basic intuition is: 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 $d/dy_j y_m$ and $d/dx_i y_m$. For this, we need a way to compute derivatives starting with $d/dy_m y_m = 1$, and accumulate derivatives backwards through the program until we reach the inputs x_i . This form of AD is called reverse-mode AD, and is the basis for backpropagation for 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 $n \gg m$.

But how do the gradients propagate backward? The basic idea is rooted in the *chain rule of differentiation*, which states that:

$$d/du f(g(u)) = d/dv f(v) * d/du g(u) \text{ where } v = g(u)$$

To interpret the chain rule in English, it says that the “sensitivity” of $f(g(u))$ to changes in u is the “sensitivity” of $f(v)$ to changes in v , where $v = g(u)$, amplified by the “sensitivity” of $g(u)$ to changes in u .

⁵<http://conway.rutgers.edu/~cshah/wiki/blog/posts/Differentiation>

For the $e_1 * e_2$ term in the grammars in Figure 1, we may be tempted to write:

$$\begin{aligned} dy/d\llbracket e_1 \rrbracket &= dy/d\llbracket e_1 * e_2 \rrbracket * d\llbracket u * e_2 \rrbracket/du \\ dy/d\llbracket e_2 \rrbracket &= dy/d\llbracket e_1 * e_2 \rrbracket * d\llbracket e_1 * u \rrbracket/du, \text{ where } u \text{ is fresh variable} \end{aligned}$$

The rules can be read as: the “sensitivity” of y to $\llbracket e_1 \rrbracket$ is the “sensitivity” of y to $\llbracket e_1 * e_2 \rrbracket$ amplified by the “sensitivity” of $\llbracket e_1 \rrbracket$ ’s context to $\llbracket e_1 \rrbracket$, and the “sensitivity” of y to $\llbracket e_2 \rrbracket$ is the “sensitivity” of y to $\llbracket e_1 * e_2 \rrbracket$ amplified by the “sensitivity” of $\llbracket e_2 \rrbracket$ ’s context to $\llbracket e_2 \rrbracket$. For direct correlation between the above grammars and the chain rule, just do the following substitutions in the first transformation rule: the first y to $f(g(u))$, $\llbracket e_1 \rrbracket$ to u , the second y to $f(v)$, $\llbracket e_1 * e_2 \rrbracket$ to v , and $\llbracket u * e_2 \rrbracket$ to $g(u)$.

However, the above transformation rules are not exactly correct. If both e_1 and e_2 contains x , then the sensitivity of y to x should be the sum of $dy/d\llbracket x \rrbracket$ ’s that occurred in multiple places (in both $dy/d\llbracket e_1 \rrbracket$ and $dy/d\llbracket e_2 \rrbracket$). This accumulation of gradients is often modeled by mutable references and += operations on the mutable references. We call this *destination-passing style*, where the reference cells accumulating the gradients are passed to the operations in the backward pass. (An alternative pure functional implementation is discussed in Section 3.6.) Let us again try our running example $y = 2 * x + x * x * x$ (after ANF-transformation), and map out the procedures for reverse-mode AD (Figure 6 top left). Note that we have to run a *forward-pass* first to compute and remember intermediate values, and then a *backward-pass* to accumulate the gradients. This is simply due to the fact that the derivative of the “contexts” may depend on the intermediate values computed in the forward pass.

For the running example, we deliberately reverse the statements in the backward-pass (computation flow should follow the double arrows in the figure), so that the forward-pass and the backward-pass of the same computation are on the same row. A more general presentation of reverse-mode AD for straight-line programs is given in Figure 6 bottom.

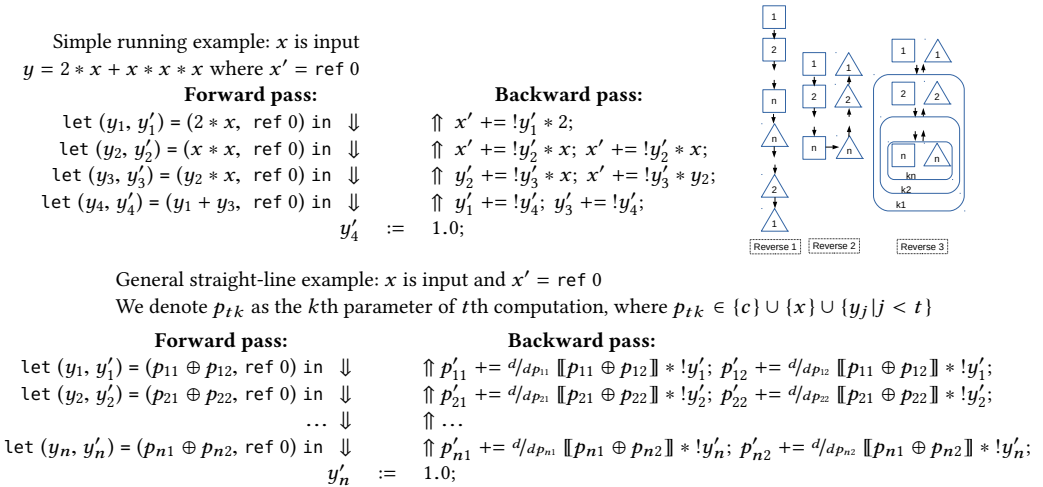


Fig. 6. Running examples of reverse-mode AD, the transformation of general straight-line programs, and abstract computation flow that motivated continuation-passing style.

Now if we look at the abstract computation flow shown in Figure 6 top right, in comparison with forward-mode AD, the computation flow of reverse-mode AD processes all value computations in the forward order, then processes all gradient computations in the reverse order (Reverse 1). We can “fold” the gradient calculations up in parallel with the value calculations (Reverse 2), like in our examples. We can further nest the computations into continuations (k_1, k_2, \dots, k_n in Reverse 3),

following the inspiration from “There and Back again” [Danvy and Goldberg 2005], and look for ways to model the computation as a sequence of function calls, where the call path implements the forward pass and the return path implements the backward pass.

With this intuition, it is not hard to see that a transformation to continuation-passing style (CPS) provides exactly the right structure, i.e., for each computation step, the subsequent forward-backward combinations are contained in a set of nested continuations (as k_1, k_2, \dots, k_n in Reverse 3, Figure 6 upper-right). In contrast to regular CPS, our continuations do return and can be followed by other computation/program statements. This kind of continuations, which behaves more like regular callback functions, is called *delimited continuations* [Felleisen 1988].

3.1 Implementation Using Operator Overloading

We first express the idea via an implementation in Scala that directly follows the intuitions in Figure 6, where each overloaded operator is provided with a (delimited) continuation k . The code is shown in Figure 7. Just like in forward-mode AD, we associate values and their gradients as two fields of a class, here `NumR0`. Every operator takes a delimited continuation k , which is expected to take the intermediate variable y , handle the rest of the forward pass after this computation step, as well as the leading part of the backward pass before this step. Once the continuation returns, the gradients ($y.d$ and possibly other gradients in the closure) should have been correctly updated, and the operator then updates the gradients of the dependent variables using `+=` operations.

```
// Differentiable real number type.
class NumR0(val x: Double, var d: Double) {
  def +(that: NumR0) = { (k: NumR0=>Unit) =>
    val y = new NumR0(this.x + that.x, 0.0); k(y)
    this.d += y.d; that.d += y.d
  }
  def *(that: NumR0) = { (k: NumR0=>Unit) =>
    val y = new NumR0(this.x * that.x, 0.0); k(y)
    this.d += that.x * y.d
    that.d += this.x * y.d
  } ...
}

// Differentiation operator.
def grad(f: NumR0 => (NumR0=>Unit)=>Unit )(x:Double)={
  val z = new NumR0(x, 0.0)
  f(z)((r: NumR0) => r.d = 1.0)
  z.d
}
// Example: 2*x + x*x*x.
val df = grad { x => k =>
  (2*x) (y1=>( x*x ) (y2=>(y2 * x ) (y3=>(y1 + y3)(k))))
}
forAll { x =>
  df(x) == 2 + 3*x*x
}
```

Fig. 7. Automatic Differentiation in Scala: reverse-mode AD in continuation-passing style (left), grad function definition and use case (right). Handling of continuations is highlighted. Note that `val` and `var` mean immutable and mutable variables respectively in Scala. Constants are implicitly lifted to `NumR0`s. Code first appeared in [Wang and Rompf 2018].

However, this implementation is not yet taking care of the generation of delimited continuations. As a consequence, it is cumbersome to use. Even for our simple running example $y = 2 * x + x * x * x$, we have to explicitly construct delimited continuations for each step (shaded box in Figure 7 lower-right). Fortunately, there exist *delimited control operators* [Felleisen 1988] that enable programming with delimited continuations in a direct style, without making continuations explicit. As a next step, we are going to use the shift/reset pair of operators [Danvy and Filinski 1990] to simplify our implementation.

3.2 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. They are readily available in Scala as a compiler plug-in [Rompf et al. 2009], thus we can simply use them in our `NumR` implementation. In Figure 8, 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 intermediate code into a continuation, and passes it to the shift construct as the parameter k [Rompf et al. 2009]. As a result, the implementation of NumR with shift/reset operators is almost identical to the CPS NumR0 implementation in Figure 7 (modulo the added shift). The implementation also corresponds to formal translation rules we provide in Section 3.3 and especially Figure 9.

The shift/reset operators in Scala are tracked by types annotated in the form of $A @cps[B]$. Semantically, this means that the shift construct can be used anywhere A -typed values are needed, but it must be within a reset context of type B . For reverse-mode AD, we expect the continuation k to be of type $NumR \Rightarrow Unit$, and the body of shift to be of type $Unit$.

```
// Differentiable number type.
class NumR(val x: Double, var d: Double) {
  def +(that: NumR) = shift {(k: NumR=>Unit) =>
    val y = new NumR(this.x + that.x, 0.0); k(y)
    this.d += y.d; that.d += y.d
  }
  def *(that: NumR) = shift {(k: NumR=>Unit) =>
    val y = new NumR(this.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: 2*x + x*x*x.
val df = grad { x =>
  2*x + x*x*x
}
forall { x =>
  df(x) == 2 + 3*x*x
}
```

Fig. 8. Automatic Differentiation in Scala: reverse-mode using delimited continuations, with shift/reset operators (left), grad function definition, and use case (right). Handling of continuations (shaded boxes) is confined to implementation logic and does not leak into user code. Constants are implicitly lifted to NumRs. Code first appeared in [Wang and Rompf 2018].

3.3 Reverse-Mode AD for Lambda Calculus

We now formalize reverse-mode AD as a transformation based on the same lambda calculus as used for the forward mode (Figure 3). The straightforward first step is to make use of shift/reset control operators in the target language to capture continuations delimited at the end of AD computation. We provide formal rules for this transformation in Figure 9, matching the Scala implementation in Figure 8. Note that the arrow of the new differentiation operator $\overleftarrow{\mathcal{D}}[\![\cdot]\!]$ indicates reverse-mode. Similar to the forward differentiation operator, the differentiation is still with respect to a fixed x , but the $\overleftarrow{\mathcal{D}}[\![\cdot]\!]$ operator never encounters the special x , since we only transform the abstraction $\lambda x.e$, assuming Barendregt’s variable convention.

Making use of shift/reset control operators in the target language, the formal rules in Figure 9 precisely capture the idea of the abstract nested computation flow in Figure 6 and the Scala implementation in Figure 8. However, what if we want to use a target language that does not provide shift/reset operators? This can be achieved by moving the uses of shift/reset into the meta-language [Danvy and Filinski 1992] (so that they are used at the time of translation), and generating target terms in explicit CPS (without shift/reset). We provide formal rules for this transformation in Figure 10. The *result* of this translation matches the Scala implementation in Figure 7. Note that in this and following figures, we use overline/underline notations (adapted from Danvy and Filinski [1992]) to mark static/meta-language constructs (overline), and dynamic/target language constructs (underline). We also introduce a wavy underline notation for handling proper tail calls, with special reduction-upon-construction logic (Figure 10 lower). Note that the wavy underline notation for “let” means that let bindings should be removed if and only if the right-hand side of the let binding is just a variable/symbol, so this wavy underline normalization performs only renaming, not full substitution. This rule is not strictly necessary for properly tail-recursive calls,

$$\begin{aligned}
\text{Transform}(f) &= \lambda x. \text{let } \hat{x} = (x, \text{ref } 0) \text{ in} \\
&\quad \langle \text{let } \hat{z} = @ \overleftarrow{\mathcal{D}}[f] \hat{x} \text{ in } z' := 1.0 \rangle; \\
&\quad ! x' \\
\text{where } \overleftarrow{\mathcal{D}}[\cdot] : &\quad \text{EXP} \rightarrow \text{EXP} \text{ is defined as below:} \\
\overleftarrow{\mathcal{D}}[c] &= c \text{ if } c \notin \mathbb{R} \\
\overleftarrow{\mathcal{D}}[c] &= (c, \text{ref } 0) \text{ if } c \in \mathbb{R} \\
\overleftarrow{\mathcal{D}}[y] &= y \\
\overleftarrow{\mathcal{D}}[e_1 + e_2] &= \text{let } \hat{y}_1 = \overleftarrow{\mathcal{D}}[e_1] \text{ in} \\
&\quad \text{let } \hat{y}_2 = \overleftarrow{\mathcal{D}}[e_2] \text{ in} \\
&\quad \text{shift } k \text{ in let } \hat{y} = (y_1 + y_2, \text{ref } 0) \text{ in} \\
&\quad \quad @ k \hat{y}; \\
&\quad \quad y'_1 += ! y'; \\
&\quad \quad y'_2 += ! y' \\
\overleftarrow{\mathcal{D}}[e_1 * e_2] &= \text{let } \hat{y}_1 = \overleftarrow{\mathcal{D}}[e_1] \text{ in} \\
&\quad \text{let } \hat{y}_2 = \overleftarrow{\mathcal{D}}[e_2] \text{ in} \\
&\quad \text{shift } k \text{ in let } \hat{y} = (y_1 * y_2, \text{ref } 0) \text{ in} \\
&\quad \quad @ k \hat{y}; \\
&\quad \quad y'_1 += ! y' * y_2; \\
&\quad \quad y'_2 += ! y' * y_1 \\
\overleftarrow{\mathcal{D}}[\lambda y. e] &= \lambda y. \overleftarrow{\mathcal{D}}[e] \\
\overleftarrow{\mathcal{D}}[@ e_1 e_2] &= @ \overleftarrow{\mathcal{D}}[e_1] \overleftarrow{\mathcal{D}}[e_2] \\
\overleftarrow{\mathcal{D}}[\text{let } y = e_1 \text{ in } e_2] &= \text{let } y = \overleftarrow{\mathcal{D}}[e_1] \text{ in } \overleftarrow{\mathcal{D}}[e_2] \\
\overleftarrow{\mathcal{D}}[\text{fst } e] &= \text{fst } \overleftarrow{\mathcal{D}}[e] \\
\overleftarrow{\mathcal{D}}[\text{snd } e] &= \text{snd } \overleftarrow{\mathcal{D}}[e] \\
\overleftarrow{\mathcal{D}}[\text{ref } e] &= \text{ref } \overleftarrow{\mathcal{D}}[e] \\
\overleftarrow{\mathcal{D}}[! e] &= ! \overleftarrow{\mathcal{D}}[e] \\
\overleftarrow{\mathcal{D}}[e_1 := e_2] &= \overleftarrow{\mathcal{D}}[e_1] := \overleftarrow{\mathcal{D}}[e_2] \\
\overleftarrow{\mathcal{D}}[(e_1, e_2)] &= (\overleftarrow{\mathcal{D}}[e_1], \overleftarrow{\mathcal{D}}[e_2]) \\
\overleftarrow{\mathcal{D}}[\text{inl } e] &= \text{inl } \overleftarrow{\mathcal{D}}[e] \\
\overleftarrow{\mathcal{D}}[\text{inr } e] &= \text{inr } \overleftarrow{\mathcal{D}}[e] \\
\overleftarrow{\mathcal{D}}[\text{case } e \text{ of } y_1 \Rightarrow e_1 \text{ or } y_2 \Rightarrow e_2] &= \text{case } \overleftarrow{\mathcal{D}}[e] \text{ of } y_1 \Rightarrow \overleftarrow{\mathcal{D}}[e_1] \text{ or } y_2 \Rightarrow \overleftarrow{\mathcal{D}}[e_2]
\end{aligned}$$

Fig. 9. Transformation of reverse-mode AD with shift/reset and mutable state in the target language (identical to interpretation except for the handling of environments). Rules that are different from standard transformation are highlighted in blue. Note that in arithmetic rules (+ and *), the computations for both forward-pass and backward-pass are defined in the same rule, with the captured continuation k executed in between. This programming pattern directly fits the abstract computation flow in Figure 6 upper-right, where continuations are triggered in-between forward computations and backward computations. The transformation is also local.

but it removes unnecessary symbol bindings for the case expression in abstraction (supporting implementation of this transformation in Scala with examples is available online³).

It is of course also possible to express the CPS transformation without shift/reset entirely by switching the meta-language code to CPS. This can be achieved formally by applying the same transformation as above to the meta-language translation code. The result is that occurrences of shift/reset are fully erased from the right-hand sides of the translation (Figure 11). A complete version of the formal presentation with standard interpretations/transformations and examples of loops and recursions is available online³.

$$\begin{aligned}
\text{Transform}(f) &= \underline{\lambda}x. \underline{\text{let}} \hat{x} = (x, \underline{\text{ref}} 0) \underline{\text{in}} \\
&\quad \underline{\text{@}} \overleftarrow{\mathcal{D}}[f] (\underline{\lambda}m. \underline{\text{@}} (\underline{\text{@}} m \hat{x}) (\underline{\lambda}z. \underline{\text{let}} \hat{z} = z \underline{\text{in}} z' := 1.0)); \\
&\quad \underline{!} x' \\
\text{where } \overleftarrow{\mathcal{D}}[\cdot] : &\quad \text{EXP} \rightarrow \text{EXP} \text{ is defined as below:} \\
\overleftarrow{\mathcal{D}}[c] &= \underline{\lambda}\kappa. \underline{\text{@}} \kappa c \text{ if } c \notin \mathbb{R} \\
\overleftarrow{\mathcal{D}}[c] &= \underline{\lambda}\kappa. \underline{\text{@}} \kappa (c, \underline{\text{ref}} 0) \text{ if } c \in \mathbb{R} \\
\overleftarrow{\mathcal{D}}[y] &= \underline{\lambda}\kappa. \underline{\text{@}} \kappa y \\
\overleftarrow{\mathcal{D}}[e_1 + e_2] &= \underline{\lambda}\kappa. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_1] (\underline{\lambda}p_1. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_2] (\underline{\lambda}p_2. \text{\#} \\
&\quad \underline{\text{let}} \hat{y}_1 = p_1 \underline{\text{in}} \underline{\text{let}} \hat{y}_2 = p_2 \underline{\text{in}} \\
&\quad \underline{\text{let}} \hat{y} = (y_1 + y_2, \underline{\text{ref}} 0) \underline{\text{in}} \\
&\quad \underline{\text{@}} \kappa \hat{y}; \\
&\quad y'_1 \pm = \underline{!} y'; \\
&\quad y'_2 \pm = \underline{!} y'') \\
\overleftarrow{\mathcal{D}}[e_1 * e_2] &= \underline{\lambda}\kappa. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_1] (\underline{\lambda}p_1. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_2] (\underline{\lambda}p_2. \text{\#} \\
&\quad \underline{\text{let}} \hat{y}_1 = p_1 \underline{\text{in}} \underline{\text{let}} \hat{y}_2 = p_2 \underline{\text{in}} \\
&\quad \underline{\text{let}} \hat{y} = (y_1 * y_2, \underline{\text{ref}} 0) \underline{\text{in}} \\
&\quad \underline{\text{@}} \kappa \hat{y}; \\
&\quad y'_1 \pm = \underline{!} y' * y_2; \\
&\quad y'_2 \pm = \underline{!} y' * y_1)) \\
\overleftarrow{\mathcal{D}}[\underline{\lambda}y. e] &= \underline{\lambda}\kappa. \underline{\text{@}} \kappa (\underline{\lambda}y. \underline{\lambda}k. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e] (\underline{\lambda}m. \underline{\text{@}} k m)) \\
\overleftarrow{\mathcal{D}}[\underline{\text{@}} e_1 e_2] &= \underline{\lambda}\kappa. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_1] (\underline{\lambda}m. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_2] (\underline{\lambda}n. \underline{\text{@}} (\underline{\text{@}} m n) (\underline{\lambda}a. \underline{\text{@}} \kappa a))) \\
\text{\# } \overleftarrow{\mathcal{D}}[\underline{\text{let}} y = e_1 \underline{\text{in}} e_2] &= \underline{\lambda}\kappa. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_1] (\underline{\lambda}y_1. \underline{\text{let}} y = y_1 \underline{\text{in}} \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_2] \kappa) \\
\overleftarrow{\mathcal{D}}[\underline{\text{fst}} e] &= \underline{\lambda}\kappa. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e] (\underline{\lambda}y. \underline{\text{@}} \kappa (\underline{\text{fst}} y)) \\
\overleftarrow{\mathcal{D}}[\underline{\text{snd}} e] &= \underline{\lambda}\kappa. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e] (\underline{\lambda}y. \underline{\text{@}} \kappa (\underline{\text{snd}} y)) \\
\overleftarrow{\mathcal{D}}[\underline{\text{ref}} e] &= \underline{\lambda}\kappa. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e] (\underline{\lambda}y. \underline{\text{@}} \kappa (\underline{\text{ref}} y)) \\
\overleftarrow{\mathcal{D}}[\underline{!} e] &= \underline{\lambda}\kappa. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e] (\underline{\lambda}y. \underline{\text{@}} \kappa (\underline{!} y)) \\
\overleftarrow{\mathcal{D}}[e_1 := e_2] &= \underline{\lambda}\kappa. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_1] (\underline{\lambda}y_1. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_2] (\underline{\lambda}y_2. \underline{\text{@}} \kappa (y_1 := y_2))) \\
\overleftarrow{\mathcal{D}}[(e_1, e_2)] &= \underline{\lambda}\kappa. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_1] (\underline{\lambda}y_1. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_2] (\underline{\lambda}y_2. \underline{\text{@}} \kappa ((y_1, y_2)))) \\
\overleftarrow{\mathcal{D}}[\underline{\text{inl}} e] &= \underline{\lambda}\kappa. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e] (\underline{\lambda}y. \underline{\text{@}} \kappa (\underline{\text{inl}} y)) \\
\overleftarrow{\mathcal{D}}[\underline{\text{inr}} e] &= \underline{\lambda}\kappa. \underline{\text{@}} \overleftarrow{\mathcal{D}}[e] (\underline{\lambda}y. \underline{\text{@}} \kappa (\underline{\text{inr}} y)) \\
\overleftarrow{\mathcal{D}}[\underline{\text{case}} e \text{ of } y_1 \Rightarrow e_1 \text{ or } y_2 \Rightarrow e_2] &= \underline{\lambda}\kappa. \underline{\text{let}} k = \underline{\lambda}a. \underline{\text{@}} \kappa a \underline{\text{in}} \underline{\text{@}} \overleftarrow{\mathcal{D}}[e] (\underline{\lambda}v. \underline{\text{case}} v \\
&\quad \underline{\text{of}} y_1 \Rightarrow \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_1] (\underline{\lambda}m. \underline{\text{@}} k m) \\
&\quad \underline{\text{or}} y_2 \Rightarrow \underline{\text{@}} \overleftarrow{\mathcal{D}}[e_2] (\underline{\lambda}n. \underline{\text{@}} k n))
\end{aligned}$$

Fig. 11. Transformation of source language for reverse-mode AD in CPS (meta-language does not contain shift/reset). Rules that are different from standard CPS transformation are highlighted by color. Note that in the plus rule and the multiplication rule (labeled by #), we avoided using variable sugaring in $\bar{\lambda}p_1$ and $\bar{\lambda}p_2$ so that we can introduce a dynamic let binding for them. The dynamic let bindings are necessary to preserve sharing, evaluation order, and asymptotic complexity, since the right-hand sides of them are accessed multiple times via $y_1, y'_1, y_2,$ and y'_2 .

and shared goals. The implementation proposed by Pearlmutter and Siskind returns a pair of a value and a backpropagator: $x \mapsto (v, dy/dv \mapsto dy/dx)$ for backward propagation. Doing this correctly requires a non-local program transformation, as noted in that paper. Further tweaks are required if a lambda uses variables from an outer scope: there must be some mechanism that allows backpropagation for captured variables, not just the function inputs.

In contrast to Pearlmutter and Siskind [2008], using delimited continuations with shift/reset operators enables reverse-mode AD with only local transformations. Any underlying non-local transformations are implicitly resolved 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. The proposed implementation is also extremely concise, to the point that it can serve as a *specification* of reverse-mode AD and can be used to teach AD to students.

3.5 Relation to Tape-Based Approaches

From our implementation using delimited continuations, we can derive a classic tape-based formulation of reverse-mode AD. We first realize that conceptually, our use of delimited continuations builds an implicit representation of a tape-like structure on the call stack instead of representing it as an explicit data structure on the heap. We can map this implicit structure back to the heap, by accumulating the gradient-update code that follows the invocations of `k` into closures, and storing their composition in a global mutable variable, which is used to explicitly invoke the backward pass. After this change, all invocations of continuations become tail calls, and hence delimited continuations or control operators are no longer necessary. The downside of this approach is the potentially costly management of heap-allocated closures, and, crucially, a less straightforward mapping to staged or *define-then-run* AD implementations that reify computation graphs, which falls out very naturally for CPS-based formulations (see Section 4).

We show a Scala implementation in Figure 12, noting that a similar implementation has been proposed by Roesch et al. [2018] in their framework Relay. It is easy to see that this implementation can be defunctionalized [Danvy and Nielsen 2001; Reynolds 1998] to obtain a classic tape-based AD formulation, and thus can be seen as a refunctionalized version [Danvy and Millikin 2009] of such a classic tape datastructure.

```
// Refunctionalized tape.
var tape = (u: Unit) => ()
// Differentiable number type.
class NumB(val x: Double, var d: Double) {
  def +(that: NumB) = {
    val y = new NumB(this.x + that.x, 0.0)
    tape = ((x:Unit)=> this.d += y.d) andThen tape
    tape = ((x:Unit)=> that.d += y.d) andThen tape
    y
  }
  def *(that: NumB) = {
    val y = new NumB(this.x * that.x, 0.0)
    tape = ((x:Unit)=> this.d += that.x*y.d) andThen tape
    tape = ((x:Unit)=> that.d += this.x*y.d) andThen tape
    y
  } ...
}

// Differentiation operator.
def grad(f: NumB => NumB)(x: Double) = {
  val z = new NumB(x, 0.0)
  f(z).d = 1.0
  tape()
  z.d
}

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

Fig. 12. Reverse-mode Automatic Differentiation with a global refunctionalized tape. The `andThen` infix operator is for function composition in Scala. Since new additions are composed before the old tape, calling `tape()` will play the tape in reverse order of insertion.

3.6 Purely Functional Implementation

Since our presentation makes central use of the mutable state, an interesting question is whether a purely functional formulation is also possible. For example, since the continuation `k` takes a new `NumR`, updates its gradient, and returns `Unit`, why not simply let `k` return the new gradient and avoid mutation? The type of `k` would change to `Double => Double` accordingly. Unfortunately, this simple change is not enough, because the continuation `k` may update the gradients of *more than one* `NumRs`. If earlier `NumRs` are also involved in the computations in `k`, then `k` needs to update their gradients too, but returning just a `Double` without side-effects cannot achieve that. Thus, a pure functional implementation is easy to achieve for straight-line programs [Elliott 2018], but not for ones with complex control flow and especially nested lambdas.

Based on this observation, we can build a purely functional implementation by adding a layer of indirection. Each NumR is assigned a unique id, and we change the type of continuations to $\text{NumR} \Rightarrow \text{Map}[\text{Id}, \text{Double}]$, returning an immutable map from NumR ids to their calculated gradient updates. In essence, this model uses a reified functional store for gradient updates instead of storing the gradients directly in the Scala heap. Since there is no conceptual simplification, we prefer the model based on direct mutation for our presentation.

3.7 Nested Invocations For Higher-Order Gradients

Just like with forward-mode AD in Section 2.5, we are interested in extending the reverse-mode AD implementation to support nested invocations of the grad operator. The way to achieve this nesting of reverse-mode AD within reverse-mode AD (i.e., reverse-of-reverse) is to use multiple levels of continuations, and their corresponding higher-order control operators such as `shift2` [Danvy and Filinski 1990]. Unfortunately, we cannot directly implement this in Scala, since the Scala compiler only provides a single CPS transform layer. However, we can manually embed a shift/reset layer within another shift/reset via explicit CPS to create a similar functionality as `shift2` (Figure 13).

```
// Type of contexts (explicit or implicit CPS).
type Ctx = ((Unit => Unit) => Unit)
type Ctxi = Unit @cps[Unit]
// Definition of (restricted) shift2.
def shift2(body: (NumRR=>Ctxi)=>Ctxi): NumRR @cps[Ctx] =
  shift { k: (NumRR => Ctx) => k2: (Unit => Unit) =>
    def kk(y: NumRR) = shift((k3: Unit=>Unit) => k(y)(k3))
    reset { body(kk); k2() }
  }
// Differentiable number class.
class NumRR(val x: NumR, var d: NumR) {
  def * (that: NumRR) = shift2 {k: (NumRR => Ctxi) =>
    val y = new NumRR(x * that.x, new NumR(0.0,0.0))
    k(y)
    this.d = this.d + y.d * that.x
    that.d = that.d + y.d * this.x
  }
  ...
}

// Differentiation operator.
def gradRR(f: NumRR => NumRR @cps[Ctx])(x: NumR) =
  shift { (k: NumR => Unit) =>
    val z = new NumRR(x, new NumR(0.0,0.0))
    val ff = reset {
      f(z).d = new NumR(1.0,0.0)
      (k2: Unit => Unit) => k2()
    }
    ff((u: Unit) => k(z.d))
  }
}

// Example: 2*x + x*x*x.
val df = grad { gradRR { x =>
  2*x + x*x*x
}}
forall { x =>
  df(x) == 6*x
}
```

Fig. 13. Second order reverse-of-reverse AD via explicit CPS in shift/reset.

Another way to realize higher-order gradients is to nest forward-mode AD in reverse-mode AD (i.e. compute the first-order gradient via reverse-mode AD, and higher order gradient via forward-mode AD). This approach is practically efficient if the higher-order gradient is for functions $\mathbb{R}^m \rightarrow \mathbb{R}^n$, where n is relatively large compared to m . This "forward-of-reverse" combination can efficiently compute Hessians as the Jacobian of gradients [Baydin et al. 2018], and Hessian-vector products in a single forward-of-reverse pass [Christianson 1992].

4 REIFYING COMPUTATION GRAPHS VIA MULTI-STAGE PROGRAMMING

CPS conversion puts reverse-mode AD on a firm basis, rooted in programming language concepts. Extending the Num type to tensors and relaying tensor operations to high performant libraries provides all the necessary machinery for a deep learning framework in the expressive PyTorch-style that performs gradient computation as part of the normal program execution ("define-by-run").

From PyTorch-style to TensorFlow-style. However, TensorFlow-style frameworks have traditionally been more performant than define-by-run ones, by constructing a restricted dataflow model before executing gradient computation, which offers a larger optimization surface on the tensor IR level ("define-then-run"). Can we also realize a TensorFlow-style framework, but with a richer and more standard IR language, better supporting native control flow and recursion?

TensorFlow-style via Multi-Stage Programming. This question can be naturally addressed by leveraging the idea found in the formal model of moving the use of shift/reset into the metalanguage to generate code in CPS (Section 3). We use *multi-stage programming* (staging) as a practical way to realize the overline/underline distinction found in the formal model. Modern tools such as LMS (Lightweight Modular Staging) [Rompf and Odersky 2010] blend normal program execution with IR construction. In LMS, a type constructor $\text{Rep}[T]$ is used to mark staged expressions. That is to say, all $\text{Rep}[T]$ -typed variables (whether directly labeled or type-inferred) will trigger LMS-based IR construction. Through type inference and advanced operator overloading, normal syntax can be used to stage built-in control-flow constructs such as `if`, `for`, and `while`. We can relate staging via LMS to the formal rules in Figure 10, though the different stages are determined by types in LMS [Rompf 2016].

To show the flavor of LMS as well as how to make use of LMS in our reverse-mode AD to reify computation graphs (LMS-based IR), let us walk through our running example again: $y = 2 * x + x * x * x$, where we simply focus on first-order reverse-mode AD such that x is of type `NumR`. To stage our running example, the most important change is the type signature of the `NumR` class:

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

Here, the $\text{Rep}[T]$ type of x and d states that all handling of x and d will construct nodes in LMS-IR. The $\text{Rep}[\text{Var}[\text{Double}]]$ maps to staged mutable reference (such as type `double&` in C++), which allows us to accumulate gradients d by reference. Note that our presentation is isomorphic to staging `NumR` as in $\text{Rep}[\text{NumR}]$, since both fields of `NumR` are already staged. However, staging only the fields of `NumR` gives us a more concise generated code.

There are no fundamental challenges with staging our reverse-mode AD in CPS using LMS, as it is a well-known insight that multi-stage programs that use continuations at generation time can generate code in CPS [Bondorf 1992; Danvy and Filinski 1992] (relating to formal rules in Figure 10). LMS can also be set up to generate low-level, efficient code in C++ and CUDA. This enables a TensorFlow-style framework with rich analysis and optimization opportunities, much like an aggressive whole-program compiler.

The apparent downsides of TensorFlow-style systems, however, are the rather clunky user programming model 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 current static frameworks thanks to staging (in particular, the LMS framework). Of course, TensorFlow can also be viewed as a staged programming model, but the staged language is a restricted dataflow language. On the other hand, LMS provides a rich staged language that includes subroutines, recursion, and more.

We show below how CPS code generation is supported in a natural form, in straight-line code, branches, loops, and recursion. Note that our setup is mostly similar to Figure 10, where only the metalanguage has `shift/reset`, but not identical, since stages are controlled by types (more redexes can be simplified). Also, we will refer to generic types (A, B, and C) for control flow constructs in the following part of this section to illuminate the abstraction of branches, loops, and recursion.

4.1 Staging Reverse-Mode AD: Straight-Line Code

We begin by investigating how to stage and perform AD on straight-line programs (i.e., those without loops, branches, or recursion). Let us start with a very simple straight-line program.

```
def snippet(in: Rep[Double]): Rep[Double] = grad(x => x * x)(in)
```

We show the code after reducing arithmetic operations, `grad` function, and `shift/reset` control operators (left), and the generated pseudo-LMS-IR (middle). Note that in this example, since the `NumR` class itself is not $\text{Rep}[T]$ -typed (both fields of `NumR` are), fields of `NumR` will trigger IR-construction for code generation, but all `NumR` object construction and field accesses will be staged away. The IR

can be used to generate C++ code (with optimizations including dead code elimination, constant folding, etc., shown on the right):

```
def snippet(in: Rep[Double]): Rep[Double] = {
  val z = new NumR(in, 0.0)
  val y = new NumR(z.x * z.x, 0.0)
  y.d = 1.0
  z.d += y.d * z.x
  z.d += y.d * z.x
  z.d
}

def snippet(in) = {
  d0 = ref 0
  v1 = in * in; d1 = ref 0
  d1 := 1.0
  d0 += ! d1 * in
  d0 += ! d1 * in
  ! d0
}

double snippet(in: double) =
  return 2 * in;
```

4.2 Staging Reverse-Mode AD: Conditionals

The conditionals are closely related to the case rule in Figure 10 (e.g., reset in both branches). We define a syntactically different IF operator that takes a $\text{Rep}[\text{Boolean}]$ condition and two $(\Rightarrow \text{Rep}[A] \text{ @cps}[\text{Rep}[B]])$ typed parameters for the then- and else-branches. In Scala, $\Rightarrow T$ typed parameters are *passed by name*, so that the parameters are evaluated each time they are used. Following the case rule, the IF function accesses the delimited continuation k via `shift`, lifts k to a dynamic function $k1$ (to avoid code duplication), and invokes $k1$ with both the then-branch and the else-branch argument. In LMS, the `fun` function lifts static functions to dynamic ones [Rompf 2016]:

```
def fun(f: Rep[A] => Rep[B]): Rep[A => B]
```

We use overline/underline to mark function names in function applications and definitions, and control flow constructs to indicate their stages (meta-language vs target language). This is similar to the formal rules in Section 3, but added here purely for reasons of readability.

```
def IF(c: Rep[Boolean])(a: => Rep[A] @cps[Rep[B]])(b: => Rep[A] @cps[Rep[B]]): Rep[A] @cps[Rep[B]] =
  shift { k: (Rep[A] => Rep[B]) =>
    // Emit k1 as a dynamic function to avoid code duplication.
    val k1 = fun (k)
    // Emit conditional, with each branch enclosed by a reset.
    if (c) reset(k1(a)) else reset(k1(b))
  }
```

Below is an example using the IF construct. For readability, we only selectively label *some* names with overline/underline to highlight stages of *some* constructs.

```
def snippet(in: Rep[Double]): Rep[Double] = grad(x => IF(x.x > 0.0){ -1.0*x*x }{ x*x })(in)
```

We show the code after we reduce `grad`, `fun`, and some `shift/reset` (upper left), the pseudo-IR (right), and the generated C++ code with optimizations including inlining and hoisting (lower left):

```
def snippet(in: Rep[Double]): Rep[Double] = {
  val z = new NumR(in, 0.0)
  val k1: Rep[NumR => Unit] = (t => t.d = 1.0)
  // elide process of reset block (similar to straight-line program)
  if (z.x > 0.0) reset { k1 (-1.0 * z * z) }
  // elide process of reset block (similar to straight-line program)
  else reset { k1 (z * z) }
  z.d
}

double Snippet(double in) {
  auto k = [&](double x, double& d) { d = 1.0; };
  double d = 0.0;
  if (in > 0.0) { k(-in * in, d); return -2.0 * in * d; }
  else { k(in * in, d); return 2.0 * in * d; }
}

def snippet(in) = {
  d0 = ref 0
  k = (x, d) => d := 1.0
  if (in > 0.0) {
    v1 = - in * in; d1 = ref 0
    k(v1, d1)
    d0 += ! d1 * (- in)
    d0 += ! d1 * (- in)
    ! d0
  } else {
    v1 = in * in; d1 = ref 0
    k(v1, d1)
    d0 += ! d1 * in
    d0 += ! d1 * in
    ! d0
  }
}
```

4.3 Staging Reverse-Mode AD: Loops

Differentiable loop constructs are important for deep learning, for example in recurrent neural networks. By the rules of CPS transformation, loops need to be transformed into tail-recursive

functions. A loop construct consists of an initial value $\text{Rep}[A]$, a loop guard, and a loop body of type $\text{Rep}[A] \Rightarrow \text{Rep}[A] @\text{cps}[\text{Rep}[B]]$ as parameters. The loop guard can be either $\text{Rep}[A] \Rightarrow \text{Rep}[\text{Boolean}]$, like a while construct, or simply a $\text{Rep}[\text{Int}]$, like a for construct. The actual loop logic can be described as follows: if the loop guard is true, recursively call the loop after invoking 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] => Rep[Boolean])(b: Rep[A] => Rep[A] @cps[Rep[B]]): Rep[A] @cps[Rep[B]] =
  shift { k: (Rep[A] => Rep[B]) =>
    // tail recursive function implementing loop semantics.
    def loop: Rep[A => B] = fun { (x: Rep[A]) =>
      if (c(x)) reset(loop(b(x))) else reset(k(x))
    }
    loop(init)
  }
```

Below is an example using the WHILE construct:

```
def snippet(in: Rep[Double]): Rep[Double] = grad(x => WHILE(x)(t => t.x > 1.0)(t => t * 0.5))(in)
```

We show the code after we reduce grad, fun, some shift/reset, and application of b, c, and k (left), the pseudo-IR (middle), and the generated C++ code (right):

<pre>def snippet(in:Rep[Double]):Rep[Double]= { val z = new NumR(in, 0.0) def loop: Rep[A => B] = (t => // elide process of reset block // similar to straight-line program if (t.x > 1.0) reset(loop(t * 0.5)) else t.d = 1.0) loop(z) z.d }</pre>	<pre>def snippet(in) = { d0 = ref 0 def loop = (x, d) => { if (x > 1.0) { v1 = x * 0.5; d1 = ref 0 loop(v1, d1) d += 0.5 * !d1 } else d := 1.0 } loop(in, d0) !d0 }</pre>	<pre>double Snippet(double in) { double d = 0.0; auto loop = [&](double x, double& d) { if (x > 1.0) { double d1 = 0.0; loop(0.5 * x, d1); d += 0.5 * d1; } else d = 1.0; }; loop(in, d); return d; }</pre>
---	---	--

We can also relate our WHILE definition to the formal rules. Although we did not include an explicit letrec rule in Figure 10, Figure 3 explains how letrec can be derived from let, λ , $@$, and case constructs, the rules of which are identical to the standard transformations in Figure 10. Thus, the letrec rule is also the same as in the standard transformation [Danvy and Filinski 1992] (modulo wavy underline notation), which we recap here:

$$\# \overline{\mathcal{D}}[\text{letrec } f = \lambda x. e_1 \text{ in } e_2] = \overline{\text{shift}} \ k \ \overline{\text{in}} \ \text{letrec } f = \lambda x. \lambda k_1. \overline{\lambda} @ \ k_1 \ \overline{\mathcal{D}}[e_1] \ \overline{\text{in}} \ \overline{\lambda} k \ \overline{\mathcal{D}}[e_2]$$

Note that the rule is simplified due to Barendregt's variable convention (no variable substitution needed). However, specific constraints of loops guarantee that the recursive calls always appear in tail positions, and the recursive function should only be applied once, in the original location of the loop. That means the abstraction of continuation (k_1 in the rule) can be optimized away, just like the contification optimization for closures. The formal rule should now be adapted (note that the transformation rule of apply for f should also adapt accordingly) to:

$$\# \overline{\mathcal{D}}[\text{letrec } f = \lambda x. e \text{ in } @ f y] = \overline{\text{shift}} \ k \ \overline{\text{in}} \ \text{letrec } f = \lambda x. \overline{\lambda} @ \ k \ \overline{\mathcal{D}}[e] \ \overline{\text{in}} \ @ f y$$

The abstract body e in the above rule is most likely a conditional construct that needs to be translated to IF. Our WHILE definition can be derived after normalization of IF (containing shift) with its surrounding reset context.

4.4 Staging Reverse-Mode AD: Functions & Recursion

As a true differentiable programming framework, we aim to handle general forms of recursion. This is useful in deep learning: one application is processing tree-structured data, such as sentence parse trees (see Section 5). We have already seen in Section 4.2 how we can use fun to generate

staged functions in LMS, but how do we make those differentiable? The answer is simply to provide an equivalent of `fun` that generates a staged function in CPS (`FUN` below):

```
def FUN(f: Rep[A] => Rep[B] @cps[Rep[C]]) = (y: Rep[A]) => shift {k: (Rep[B] => Rep[C]) =>
  val f1 = Fun((t: Rep[A], k1: Rep[B => C]) => reset(k1(f(t))))
  f1(y, fun(k))
}
```

With this `FUN` subroutine, implementing a differentiable tree traversal is straightforward. We can define a `TREE` abstraction to recursively traverse a `Rep[Tree]` data structure. For empty trees, the `init` value is returned directly. For non-empty trees, the function `b` composes the recursive results from the subtrees.

```
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 is not empty, recurse on subtrees and compose results, otherwise return the initial values.
    F (tree.notEmpty) { b(f(tree.left), f(tree.right)) } init
  }
  f(t)
}
```

How do the above implementations relate to the formal `let rec` rule? The `FUN` definition is similar to the representation of $\lambda y. \overline{\mathcal{D}}[\text{letrec } f = \lambda t. e \text{ in } @ f y]$, which can be transformed as below:

$$\begin{aligned} & \lambda y. \overline{\mathcal{D}}[\text{letrec } f = \lambda t. e \text{ in } @ f y] \\ &= \lambda y. \overline{\text{shift } k \text{ in } \text{letrec } f = \lambda t. \lambda k_1. \overline{\text{@ } k_1 \overline{\mathcal{D}}[e] \text{ in } @} (@ f y) (\lambda a. \overline{\text{@ } k a})} \\ &= \lambda y. \overline{\text{shift } k \text{ in } \text{letrec } f = \lambda t. \lambda k_1. \overline{\text{@ } k_1 (\overline{\text{@ } t. \overline{\mathcal{D}}[e]} t) \text{ in } @} (@ f y) (\lambda a. \overline{\text{@ } k a})} \end{aligned}$$

Note that in this term, we use **blue color** to highlight the sub-term that corresponds to the parameter of `FUN` (also marked in blue). We can also see that the body of `TREE` (i.e. `f(t)` in the code) evaluates to $\overline{\mathcal{D}}[\text{letrec } f = \lambda t. e \text{ in } @ f t]$. Below is an example using the `TREE` construct:

```
def snippet(tree:Rep[Tree], in:Rep[Double]):Rep[Double] = grad(x => TREE(x)(tree){(l, r) => l * r * tree.value})(in)
```

The code after reducing `grad`, `FUN`, `TREE`, and some `shift/reset` is shown below, The generated pseudo LMS-IR and the generated C++ code are in Figure 14.

```
def snippet(tree: Rep[Tree], in: Rep[Double]): Rep[Double] = {
  val z = new NumR(in, 0.0)
  val k: Rep[NumR => Unit] = (x => x.d = 1.0)
  def f1(t: Rep[Tree], k0: NumR => Unit) = if (t.notEmpty) {
    val k_l: Rep[NumR => Unit] = (l =>
      // elide process of reset block (similar to straight-line program)
      val k_r: Rep[NumR => Unit] = (r => reset{ k0(l * r * t.value) })
      f1(t.right, k_r)
    )
    f1(t.left, k_l)
  } else k0(z)
  f1(tree, k)
  z.d
}
```

With the above implementations, we have established a staged reverse-mode AD framework that supports branches, loops, and recursion. Though implementing these control-flow operators requires some engineering, they simply combine CPS transformation with staging in the standard way (as shown in Figure 10). The resulting framework provides a programming interface that is similar in style and expressiveness to PyTorch. It also generates an intermediate representation with inlined AD logic (pure manipulation of `Doubles` or `Tensors`) which allows extensive optimizations similar in style to 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 for accessibility and simplicity. A good, *selective*, CPS transform (that transforms only the minimum necessary code to CPS) is nontrivial to implement [Rompf et al. 2009].

```

def snippet(tree, in) = {
  d0 = ref 0
  k = (x, d) => d := 1.0
  f1 = (t, k0) => if (t.isEmpty) {
    k_l = (x_l, d_l) => {
      k_r = (x_r, d_r) => {
        v0 = t.value
        v1 = x_l * x_r * v0; d1 = ref 0
        k0(v1, d1)
        d_l += x_r * v0 * (! d1)
        d_r += x_l * v0 * (! d1)
      }
      f1(t.right, k_r)
    }
    f1(t.left, k_l)
  } else k0(in, d0)
  f1(tree, k)
  ! d0
}

double Snippet(Tree tree, double in) {
  double d = 0.0;
  auto k = [&](double x, double d) { d = 1.0; };
  auto rec = [&](Tree tree, function<void(double, double)> k) {
    if (tree.isEmpty) {
      auto k_l = [&](double x_l, double d_l) {
        auto k_r = [&](double x_r, double d_r) {
          double x_t = tree.value; double dt = 0.0;
          k(x_l * x_r * x_t, dt);
          d_l += x_r * x_t * dt;
          d_r += x_l * x_t * dt;
        };
        rec(tree.right, k_r);
      };
      rec(tree.left, k_l);
    } else k(in, d);
  };
  rec(tree, k);
  return d;
}

```

Fig. 14. Generated pseudo LMS-IR (left) and the generated C++ code (right) of the tree example.

5 EVALUATION

So far, we have shown both how to implement plain PyTorch-style reverse-mode AD using delimited continuations, and how to mix in multi-stage programming for TensorFlow-style graph reification. Now, we extend our implementation to tensor operations, and present a system, Lantern², that scales our described approach to real-world deep learning workloads.

The snippet below shows the basic structure of the staged tensor API:

```

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

```

The type of field `dimension`, `Array[Rep[Int]]`, indicates that the tensor rank (number of dimensions) is always known at staging time. The `TensorR` class takes two `Tensors`, one as the value, and the other as the gradient. Operators in `TensorR` are overloaded with `shift` constructs, providing access to delimited continuations. In analogy with the CPS-style implementation in Section 3, class `Tensor` takes the role of `Double`, and `TensorR` that of `NumR`.

The `Tensor` class provides all tensor-level operations including element-wise operations with broadcasting, matrix multiplication, convolution, and so on. Lantern provides abstractions to run each of those operations either on CPU or GPU. Implementations make use of BLAS library functions (for CPU) and cuBLAS/cuDNN library functions (for GPU), but also include nested for loops (for CPU) and custom CUDA kernels (for GPU).

We would like to stress that the efficiency of Lantern can be further improved by more sophisticated backend engineering, which is not the focus of this paper. One direction is tensor IR level optimization similar to TVM [Chen et al. 2018] and Glow [Rotem et al. 2018], including operator fusion, and systematic operation scheduling. LMS provides fusion and array facilities [Rompf et al. 2013] that have been used in OptiML [Sujeeth et al. 2011] and other DSLs based on the Delite compiler framework [Brown et al. 2016, 2011]. These could be leveraged for Lantern as well.

Tensor IR level optimization is naturally supported by define-then-run systems (e.g. Lantern and TensorFlow), but not by define-by-run systems (e.g. PyTorch, though recently PyTorch 1.0 moves towards this direction by extracting computation graphs using Torch Script [PyTorch 2019]). Another important direction is advanced batching support, either in the form of *autobatching* à la Dynet [Neubig et al. 2017b] or *dynamic batching* à la TensorFlow Fold [Looks et al. 2017]. Advanced batching support is particularly useful in dynamic models where manual batching is challenging. Another use is suggesting optimal batch sizes based on model and hardware details (e.g. GPU memory size).

Even with the current level of backend engineering, our evaluation shows that Lantern is competitive on contemporary machine learning models, thus pushing the boundaries of existing frameworks in various dimensions (expressivity and efficiency).

5.1 Recursive Neural Network: TreeLSTM

In this and the following sections, we evaluate Lantern on several commonly used machine learning models. Our evaluation focuses on expressivity (the ability to express various kinds of machine learning models) and efficiency (the runtime for training those machine learning models on a single GeForce GTX 1080 Ti with CUDA 10, using PyTorch version 1.0rc, TensorFlow version 1.12.0-rc0). Runtime results are reported as the median runtime of 5 epochs with 3 repeats. The correctness of the computed gradients is implicit (we have extensive unit tests, and we check our gradients with PyTorch). We elide loss curves and other dimensions of evaluation such as hyperparameter-tuning/cross-validation/testing.

We start the evaluation with TreeLSTM, which is a state-of-the-art *recursive* machine learning model that heavily depends on dynamic control flow guided by structural training data. Models like this are useful for handling natural language parse trees and abstract syntax trees of programming languages. At the same time, such dynamic models pose interesting challenges to machine learning frameworks.

We showcase TreeLSTM on the Sentiment Classification task [Tai et al. 2015] using the Stanford Sentiment Treebank dataset [Chuang 2013]. The dataset contains sentences of movie reviews, which are parsed into binary trees based on language semantics. Each leaf node contains a word that can be mapped to a known numeric vector (embedding) based on word semantics. TreeLSTM should model a function (Bi-LSTM) that recursively computes hidden states for all nodes, and a mapping from hidden states to sentiment scores, both of which are then trained end-to-end by minimizing *softmax-cross-entropy loss* with regard to true sentiment labels of each node.

The recursive function that computes hidden states of node i can be written as below:

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

where h_i represents hidden state of node i , `Embedding` represents the known mapping from words to their embeddings, $i.\text{word}$, $i.\text{left}$, $i.\text{right}$ represent the optional word, left-child, and right-child of node i , and Bi-LSTM is a variant of LSTM that can handle two hidden states as inputs.

It is easy to express this model in Lantern since Lantern supports unrestricted control flow including branches, loops, and recursion. We show the core of the TreeLSTM model in Lantern (function `lossFun`) below. The TREE abstraction (Section 4.4) makes the code very concise. Users merely need to supply the (anonymous) function that computes the state of the current node from the states of the left and right children (`lState/rState`):

```
def lossFun(node: Rep[Tree]) = {
  val initialState = State(loss = 0, hidden = 0, cell = 0)
  val resultState = TREE(initialState)(node) { (lState, rState) =>
    val embedding      = IF (node.isLeaf) { Embedding(node.word) } { 0 }
    val (hidden, cell) = BiLSTM(embedding, lState, rState)
    val loss           = softmaxCrossEntropyLoss(Linear(hidden), node.score)
    State(loss, hidden, cell)
  }
  resultState.loss
}
```

This model can also be expressed easily in define-by-run frameworks like PyTorch where computation graphs are constructed on-the-fly (via Python recursive functions in this case). However, as we can see from the runtime (Figure 15), training TreeLSTM in PyTorch is very slow (more than 4 times slower than Lantern), mainly due to the overhead incurred for each individual computation step. In comparison, another define-by-run framework called DyNet [Neubig et al. 2017b] has a

more lightweight internal graph representation and optimized C++ backend [Neubig et al. 2017a], which makes it run faster than PyTorch (see DyNetNB, short for DyNet-No-Batching), though still about 1.6 times slower than Lantern.

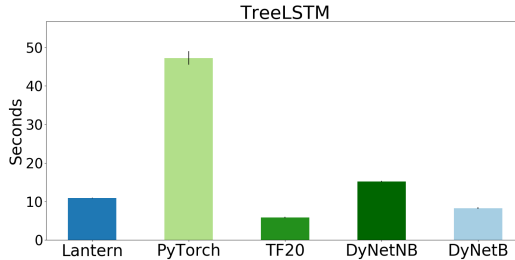


Fig. 15. Running time of TreeLSTM for different frameworks.

On the other hand, TensorFlow has trouble expressing TreeLSTM, or any other recursive neural network models. This is mainly due to limitations of TensorFlow’s static graph construction interface, which does not support recursion. As a consequence, TensorFlow can neither define static computation graphs that are recursive and covers structural data of different shapes (like Lantern), nor define computation graphs dynamically based on each structural data (like PyTorch). Other ways to flatten structural data into sequences and model them with recurrent neural networks often incur high memory overhead. Interestingly, TensorFlow Fold [Looks et al. 2017], a library on top of TensorFlow, manages to push this limited static computation graph even further for training recursive neural networks such as TreeLSTM. The main idea is that, given a set of static computation graphs of different shapes, TensorFlow Fold rewrites them into one static computation graph that handles all given graphs, by extensively using extra concat and gather operations to move data around. The added benefit is that instances of the same operations at the same depth can be batched together (in machine learning, batching refers to processing multiple pieces of data simultaneously, often in a mini-batch), which makes training more efficient. Indeed TensorFlow Fold (TF20, short for TensorFlow Fold at batch size 20) is the most efficient framework in our evaluation of TreeLSTM.

The dynamic batching approach of TensorFlow Fold is not the only way to batch training data for recursive neural networks. DyNet provides another strategy called autobatching. Being a dynamic framework, DyNet has the freedom to construct and manipulate computation graphs on-the-fly, including automatically batching nodes in computation graphs based on node types, data dimensions, and node dependencies. However, it should be noted that using autobatching efficiently requires good batching heuristics from the framework, and some input from the user, who controls the partitioning of computation graphs considered for autobatching. In our case, DyNetB (short for DyNet-Batching) shows about 50% improvement on GPU runtime when we allow autobatching within each input structure, but not across multiple input structures. Although both batched frameworks outperform Lantern, it should be noted that dynamic batching (as in TensorFlow Fold) could also be added to Lantern with additional engineering effort.

5.2 Convolutional Neural Networks: SqueezeNet and ResNet50

For non-recursive models, batching simply means adding an extra dimension to the input data, which is supported in Lantern. We now evaluate representative convolutional neural networks.

SqueezeNet [Iandola et al. 2016] and ResNet50 [He et al. 2016] are contemporary convolutional neural network models for image classification. SqueezeNet uses a carefully designed CNN architecture so that it contains fewer parameters, but shows a similar level of accuracy as larger models. ResNet50 belongs to the ResNet family of CNN architectures, which makes use of batch-normalization, residual connections, and other techniques for fast/stabilized training.

PyTorch and TensorFlow implementations of these models exist on GitHub. We hand-wrote identical Lantern models and also imported existing ONNX (Open Neural Network Exchange) models into Lantern. We evaluate these implementations on the CIFAR-10 dataset [Krizhevsky 2012]. As shown in Figure 16, Lantern and TensorFlow perform slightly better than PyTorch on SqueezeNet, and all three models have similar runtime performance on ResNet50. This result is expected since SqueezeNet and ResNet50 are mostly composed of convolution layers, which dominate the runtime cost. Convolution layers heavily rely on hardware-specific library functions such as cuDNN API functions, rendering other graph-level optimizations (of TensorFlow and potentially Lantern) insignificant. However, getting Lantern on par with PyTorch and TensorFlow took non-trivial effort: memory management techniques were crucial when using cuDNN API functions.

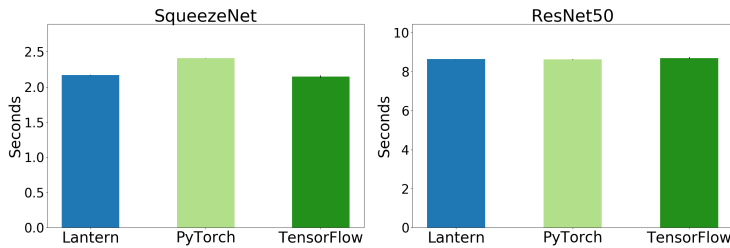


Fig. 16. Running time of SqueezeNet and ResNet50 for different frameworks.

5.3 DeepSpeech2

DeepSpeech2 [Amodei et al. 2015] is a representative deep neural network for automatic speech recognition (ASR), which reaches state-of-the-art performance on real-world datasets. DeepSpeech2 is the most complex model in our evaluation: it is a real production model with convolutional, batch norm, and RNN layers, and is trained with the CTC (Connectionist Temporal Classification) loss function. A variant of this model is included in the MLPerf benchmark suite [mlperf.org 2018].

We evaluated DeepSpeech2 models on the Librispeech [Panayotov et al. 2015] dataset, but skipped TensorFlow because it uses a custom CPU implementation of CTCLoss, making a fair comparison impossible. Lantern and PyTorch models both use bidirectional RNNs with ReLU activation and SGD with momentum.

At the time of writing, Lantern is ~10% faster than PyTorch (Figure 17) on this model, most likely because Lantern spent extra time to select better CuDNN kernel functions for CNN in the first mini-batch. Additional engineering can probably further improve Lantern’s efficiency, via better hand-written kernel functions (such as softmax, activation, and so on). However, surpassing the performance of existing frameworks on these well-tuned model implementations has not been our primary goal; the aim of this paper was merely to demonstrate the scope of Lantern, and to validate our fundamental design.

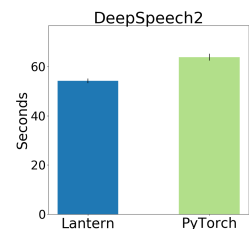


Fig. 17. Running time of DeepSpeech2 for different frameworks.

6 RELATED WORK

Automatic Differentiation: A History. Gradient-based optimization lies at the heart of machine learning, with backpropagation [Rumelhart et al. 1986], an application of differentiation, as a key ingredient for training neural networks. The fundamental idea of automatic differentiation (AD) emerged in the 1950s as programs that calculate derivatives alongside the normal computation [Beda et al. 1959; Nolan 1953]. A formal introduction to forward-mode AD appeared in the 1960s [Wengert 1964]. The application of gradient descent to large-scale optimization first arose in control theory

[Bryson and Ho 1975; Bryson and Denham 1962], although the underlying ideas are of course much older. In the 1970s, Linnainmaa [1976] introduced the concept of reverse-mode AD and the related idea of computation graphs, which are now widely used by modern machine learning frameworks. Speelpenning [1980] implemented reverse-mode AD in a general-purpose language, 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 Hecht-Nielsen [1988] brought together the work from both communities.

Automatic Differentiation: A PL View. AD has also received attention from the programming language community, with recent proposals to generalize neural network models to differentiable functional programs [Fong et al. 2017; Olah 2015]. This development is also fueled by modern deep learning frameworks, which define neural networks “very much like a regular program” [Abadi et al. 2017; LeCun 2018]. Some recent research demonstrates this direct correspondence between the two fields by implementing differentiable analogs of traditional data structures [Grefenstette et al. 2015] and machine models [Graves et al. 2014]. Another line of work has aimed to formalize AD, both forward-mode [Siskind and Pearlmutter 2008] and reverse-mode [Pearlmutter and Siskind 2008]. There exist high-level languages with first-class AD operators [Siskind and Pearlmutter 2016], as well as flexible AD library implementations, e.g., DiffSharp [Baydin et al. 2016]. A Haskell implementation of forward-mode AD was proposed by Elliott [2009]. Swift for TensorFlow [TensorFlow 2019] integrates AD as a first-class feature in a general purpose language. Shaikhha et al. [2018] demonstrated that forward-mode AD can sometimes outperform reverse-mode AD when combined with aggressive fusion and code motion techniques in a functional array programming language. This could also be achieved in Lantern using existing fusion and array facilities in LMS [Rompf et al. 2013], as used in OptiML [Sujeeth et al. 2011] and other DSLs based on the Delite compiler framework [Brown et al. 2016, 2011; Rompf et al. 2011]. Baydin et al. [2018] provided a thorough review of AD and deep learning from a functional programming perspective.

A Tale of Two Styles. Most modern deep learning frameworks compute gradients of training loss with respect to neural network parameters in one of two ways [Baydin et al. 2018]. The first is to let users define computation graphs using a domain-specific language (DSL) and to interpret graph operations at runtime. Computation graphs represent entire programs and are more amenable to global analysis and optimizations like operator fusion. However, graph-building DSLs are limited in expressivity, contain unintuitive control structures, and are difficult to debug. Frameworks such as Theano [Al-Rfou et al. 2016] and TensorFlow [Abadi et al. 2016] belong to this category. The other way is to integrate general-purpose programming languages with reverse-mode AD as a library, of which Torch [Collobert et al. 2011], PyTorch [Paszke et al. 2017a,b], Autograd [Maclaurin 2016], and Chainer [Tokui et al. 2015] are well-known representatives. Caffe [Jia et al. 2014], MXNet [Chen et al. 2015], and CNTK [Seide and Agarwal 2016] are somewhere in the middle. The tight integration between host languages and AD frameworks of the pure-library category has certain usability benefits, such as natural control flow and easy debugging, but comes at the expense of efficiency. Neural network exchange formats such as ONNX [ONNX working groups 2017] aim to bridge this gap by enabling an easy conversion between frameworks.

Previous attempts at building source-to-source deep learning compilers mostly focus on either the define-by-run or define-then-run approach, as noted by Baydin et al. [2018]. Tangent [van Merriënboer et al. 2017; Wiltschko 2017] implements a source-to-source compiler in Python which supports automatic differentiation, but this framework constrains the host language to a limited subset of Python. DLVM [Wei et al. 2017a,b] compiles deep learning programs written in Swift into a domain-specific SSA IR, performs analyses and transformations (including source code

transformation AD), and generates code via LLVM. Swift for TensorFlow [TensorFlow 2019] mixes the two approaches: it enables imperative-style programs but uses a “graph program extraction” compiler transform to automatically extract tensor code and build computation graphs.

Staging: A Unification of the Two Styles. The present work aims to reap the benefits of both styles by using a computation graph DSL that really is a general-purpose programming language. Our transformation of high-level neural networks to low-level code is fueled by the idea of multi-stage programming (staging). More than 30 years ago, Jørring and Scherlis [1986] observed that many computations can be naturally separated into stages distinguished 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 blend normal program execution with the delayed construction of an *intermediate program representation* (IR), which may be a computation graph, or in more traditional systems, an abstract syntax tree (AST). We use 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].

Delimited Continuations: A Simpler Essence. Lantern relies on delimited continuations [Danvy and Filinski 1990, 1992; Danvy and Nielsen 2003], as implemented in Scala [Rompf et al. 2009]. In parallel to our work, which first appeared as tech report on arXiv [Wang et al. 2018b], Elliott [2018] proposed a generalized view of AD based on the paradigm of “compiling to categories” [Elliott 2017]. The paper echoes our view of AD as a specific form of symbolic differentiation and also mentions continuations for reverse AD, but overall it approaches the problem from a very different categorical perspective. In comparison, our work proposes what we think is an “even simpler essence” of automatic differentiation. In particular, we show that continuations are central to reverse-mode AD, but that category theory is optional. Focusing on continuations as the key enabler makes reverse-mode AD (and hence gradient-descent optimization) immediately applicable to basically any program, including in-graph recursion and higher-order functions.

7 CONCLUSIONS

With this paper, we set out to demystify automatic differentiation by examining it through the lens of program transformation. We established 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 using operator overloading that is no more complex than forward-mode AD.

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 expressivity of pure library approaches (e.g., PyTorch).

Based on these two ideas, we have built a deep learning framework named Lantern. With native C++/CUDA backends, Lantern attains competitive performance for a variety of state-of-the-art deep learning models, such as SqueezeNet, ResNet, DeepSpeech2, and TreeLSTM.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers and especially our anonymous shepherd for numerous thorough and thoughtful comments and suggestions, especially the suggestion of encoding multi-level CPS via nesting explicit CPS transformation in `shift/reset`. This work was supported in part by NSF awards 1553471 and 1564207, DOE award DE-SC0018050, as well as gifts from Google, Facebook, and VMware.

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 <http://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, Ziyi 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 Shabani, Étienne Simon, Sigurd Spieckermann, S. Ramana Subramanyam, Jakob 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 <http://arxiv.org/abs/1605.02688>
- Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Y. Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Y. Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqiang Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. 2015. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *CoRR* abs/1512.02595 (2015).
- 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).
- Anders Bondorf. 1992. Improving Binding Times Without Explicit CPS-Conversion. In *Proceedings of the Conference on Lisp and Functional Programming*. ACM Press, 1–10. <https://doi.org/10.1145/141471.141483>
- Kevin J. Brown, HyoukJoong Lee, Tiark Rompf, Arvind K. Sujeeth, Christopher De Sa, Christopher R. Aberger, and Kunle Olukotun. 2016. Have abstraction and eat performance, too: optimized heterogeneous computing with parallel patterns. In *CGO*. ACM, 194–205.
- Kevin J. Brown, Arvind K. Sujeeth, HyoukJoong Lee, Tiark Rompf, Hassan Chafi, Martin Odersky, and Kunle Olukotun. 2011. A Heterogeneous Parallel Framework for Domain-Specific Languages. In *PACT*. IEEE Computer Society, 89–100.
- 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).
- Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Q. Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. 2018. TVM: An Automated End-to-End Optimizing

- Compiler for Deep Learning. In *OSDI*. USENIX Association, 578–594.
- Bruce Christianson. 1992. Automatic Hessians by reverse accumulation. *IMA J. Numer. Anal.* 12, 2 (1992), 135–150.
- Jason Chuang. 2013. Stanford Sentiment Treebank. <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*.
- Olivier Danvy and Andrzej Filinski. 1990. Abstracting Control. In *LISP and Functional Programming*. 151–160.
- Olivier Danvy and Andrzej Filinski. 1992. Representing Control: A Study of the CPS Transformation. *Mathematical Structures in Computer Science* 2, 4 (1992), 361–391.
- Olivier Danvy and Mayer Goldberg. 2005. There and back again. *Fundamenta Informaticae* 66, 4 (2005), 397–413.
- Olivier Danvy and Kevin Millikin. 2009. Refunctionalization at work. *Sci. Comput. Program.* 74, 8 (2009), 534–549.
- Olivier Danvy and Lasse R. Nielsen. 2001. Defunctionalization at Work. In *PPDP*. ACM, 162–174.
- Olivier Danvy and Lasse R. Nielsen. 2003. A first-order one-pass CPS transformation. *Theor. Comput. Sci.* 308, 1-3 (2003), 239–257.
- 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>
- Conal Elliott. 2017. Compiling to categories. *PACMPL* 1, ICFP (2017), 27:1–27:27.
- Conal Elliott. 2018. The simple essence of automatic differentiation. *PACMPL* 2, ICFP (2018), 70:1–70:29.
- Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science* 14, 2 (1990), 179–211.
- Matthias Felleisen. 1988. The Theory and Practice of First-Class Prompts. In *POPL*. ACM Press, 180–190.
- 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>
- Edward Grefenstette, Karl Moritz Hermann, Mustafa Suleyman, and Phil Blunsom. 2015. Learning to Transduce with Unbounded Memory. In *Advances in Neural Information Processing Systems* 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 1828–1836. <http://papers.nips.cc/paper/5648-learning-to-transduce-with-unbounded-memory.pdf>
- 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)
- Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size. *CoRR* abs/1602.07360 (2016).
- 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>
- Alex Krizhevsky. 2012. Learning Multiple Layers of Features from Tiny Images. *University of Toronto* (05 2012).
- John Launchbury and Simon L. Peyton Jones. 1994. Lazy Functional State Threads. In *PLDI*. ACM, 24–35.
- Yann LeCun. 2018. Deep Learning est mort. Vive Differentiable Programming! <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. *ICLR* (2017).
- Dougal Maclaurin. 2016. *Modeling, Inference and Optimization with Composable Differentiable Procedures*. Ph.D. Dissertation. mlperf.org.
2018. A broad ML benchmark suite for measuring performance of ML software frameworks, ML hardware accelerators, and ML cloud platforms. <https://mlperf.org/>.

- Graham Neubig, Chris Dyer, Yoav Goldberg, Austin Matthews, Waleed Ammar, Antonios Anastasopoulos, Miguel Ballesteros, David Chiang, Daniel Clothiaux, Trevor Cohn, Kevin Duh, Manaal Faruqui, Cynthia Gan, Dan Garrette, Yangfeng Ji, Lingpeng Kong, Adhiguna Kuncoro, Gaurav Kumar, Chaitanya Malaviya, Paul Michel, Yusuke Oda, Matthew Richardson, Naomi Saphra, Swabha Swayamdipta, and Pengcheng Yin. 2017a. DyNet: The Dynamic Neural Network Toolkit. *CoRR* abs/1701.03980 (2017).
- Graham Neubig, Yoav Goldberg, and Chris Dyer. 2017b. On-the-fly Operation Batching in Dynamic Computation Graphs. In *NIPS*. 3974–3984.
- John F Nolan. 1953. Analytical differentiation on a digital computer.
- Christopher Olah. 2015. Neural Networks, Types, and Functional Programming. <http://colah.github.io/posts/2015-09-NN-Types-FP/>.
- ONNX working groups. 2017. ONNX: Open Neural Network Exchange format. <https://onnx.ai/>
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2015), 5206–5210.
- 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. 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.
- PyTorch. 2019. Torch Script. <https://pytorch.org/docs/master/jit.html> [Online; accessed 1-June-2019].
- 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)
- John C. Reynolds. 1998. Definitional Interpreters for Higher-Order Programming Languages. *Higher-Order and Symbolic Computation* 11, 4 (1998), 363–397.
- Jared Roesch, Steven Lyubomirsky, Logan Weber, Josh Pollock, Marisa Kirisame, Tianqi Chen, and Zachary Tatlock. 2018. Relay: A New IR for Machine Learning Frameworks. *CoRR* abs/1810.00952 (2018).
- Tiark Rompf. 2016. The Essence of Multi-stage Evaluation in LMS. In *A List of Successes That Can Change the World (Lecture Notes in Computer Science)*, Vol. 9600. Springer, 318–335.
- 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 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.
- Tiark Rompf, Arvind K. Sujeeth, Nada Amin, Kevin J. Brown, Vojin Jovanovic, HyoukJoong Lee, Manohar Jonnalagedda, Kunle Olukotun, and Martin Odersky. 2013. Optimizing data structures in high-level programs: new directions for extensible compilers based on staging. In *POPL*. ACM, 497–510.
- Tiark Rompf, Arvind K. Sujeeth, HyoukJoong Lee, Kevin J. Brown, Hassan Chafi, Martin Odersky, and Kunle Olukotun. 2011. Building-Blocks for Performance Oriented DSLs. In *DSL (EPTCS)*, Vol. 66. 93–117.
- Nadav Rotem, Jordan Fix, Saleem Abdulrasool, Summer Deng, Roman Dzhabarov, James Hegeman, Roman Levenstein, Bert Maher, Nadathur Satish, Jakob Olesen, Jongsoo Park, Artem Rakhov, and Misha Smelyanskiy. 2018. Glow: Graph Lowering Compiler Techniques for Neural Networks. *CoRR* abs/1805.00907 (2018).
- 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.
- Amir Shaikhha, Andrew Fitzgibbon, Dimitrios Vytiniotis, Simon Peyton Jones, and Christoph Koch. 2018. Efficient Differentiable Programming in a Functional Array-Processing Language. *CoRR* abs/1806.02136 (2018).
- 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, 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.
- 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. 2019. Swift For TensorFlow. <https://www.tensorflow.org/swift> [Online; accessed 1-June-2019].
- 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
- Fei Wang, James Decker, Xilun Wu, Gregory Essertel, and Tiark Rompf. 2018a. Backpropagation with Callbacks: Foundations for Efficient and Expressive Differentiable Programming. In *NeurIPS*.
- Fei Wang and Tiark Rompf. 2018. A Language and Compiler View on Differentiable Programming. *ICLR Workshop Track* (2018). <https://openreview.net/forum?id=SJxJtYkPG>
- Fei Wang, Xilun Wu, Grégory M. Essertel, James M. Decker, and Tiark Rompf. 2018b. Demystifying Differentiable Programming: Shift/Reset the Penultimate Backpropagator. *CoRR* abs/1803.10228 (2018).
- 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. <https://research.googleblog.com/2017/11/tangent-source-to-source-debuggable.html>