Université de Montréal

Programming tools for intelligent systems

with a case study in autonomous robotics

par

Breandan Considine

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Discipline

avril 2019

Sommaire

Summary

Table des matières

Liste des tableaux

Liste des figures

Chapitre 1

Introduction

Intelligent system: A computer system that uses techniques derived from artificial intelligence, particularly one in which such techniques are central to the operation or design of the system.

1.1. Stages in the software development lifecycle

The Waterfall Method comprises of five stages, but our contributions are limited to four.

1.1.1. Design

When designing intelligent systems, we need to iterate between requirements and design constraints. It is not sufficient for requirements engineers to hand an objective to the design team, or the design team to hand a design to the implementors. Rather, they must work in concert to produce a system that has the desired properties. Sometimes this means compromising, or redesigning the performance metrics to become more flexible.

1.1.2. Implementation

When implementing intelligent systems, we need to think carefully about languages and abstractions we use. If developers are implementing backpropogation, they will have little time to think about the high-level characteristics of these systems. This is no different from traditional software engineeering - we need to use the right abstractions for the job. With machine learning, the necessity of choosing appropriate implementatoins is even more pressing.

1.1.3. Verification

Because the space of many problems is intractably large, it is difficult to prove the correctness of a given solution. Instead, we need tools to verify the properties of a system under a given budget. In self-driving vehicles, human drivers average one fatality per hundred million miles (cite), this is incredibly difficult to test given our current methodology. Instead, we need better ways to probe the effectiveness of a candidate solution.

1.1.4. Maintenance

Maintenance of systems is radically different in machine learning. Often it is not possible to manually update the parameters of a model without starting from scratch. While there are some methods for transfer learning, if we are

Chapitre 2

Design: Programming tools for robotics systems

Programming tools are a bicycle for the mind. Complex systems need better tools for developers.

2.1. Software architecture of a robotics application

https://github.com/duckietown/hatchery

2.2. Structure of a ROS application

Defining the definition and structure of ROS services, messages, nodes, topics.

2.3. Foundations of a modern IDE

IDEs are more than just a souped-up text editor.

2.3.1. The parser

We can parse URDF, package and launch XML, and srv files.

2.3.2. Refactoring

Refactoring support is implemented.

2.3.3. Running and debugging

Assistance for running ROS applications.

2.4. More ROS Tools

Detecting and managing ROS installations.

Chapitre 3

Implementation: languages and compilers

3.1. Static and dynamic languages

Most machine learning and scientific computing today is done in dynamic languages, such as Python. In contrast, most of the industry uses statically typed languages for putting models into production (citation).

Dynamically typed languages are mostly for experimentation and prototyping. But are they scaleable to production systems?

Strong, static types are critical for reasoning about the behavior of complex programs. It provides an additional layer of safety against runtime errors.

3.2. Kotlin

Kotlin is a strong, statically typed language. It is well suited for building cross-platform systems.

3.3. Kotlin∇

Prior work has shown it is possible to encode a deterministic context-free grammar as a *fluent interface* [**4**] in Java. This result was strengthened to prove Java's type system is Turing complete [**5**]. As a practical consequence, we can use the same technique to perform shape-safe automatic differentiation (AD) in Java, using type-level programming. A similar technique is feasible in any language with generic types. We use *Kotlin*, whose type system is less expressive, but fully compatible with Java.

Differentiable programming has a rich history among dynamic languages like Python, Lua and JavaScript, with early implementations including projects like Theano, Torch, and TensorFlow. Similar ideas have been implemented in statically typed, functional languages, such as Haskell's Stalin∇ [**7**], DiffSharp in F# [**1**] and recently Swift [**9**]. However, the majority of existing automatic differentiation (AD) frameworks use a loosely-typed DSL, and few offer shape-safe tensor operations in a widely-used programming language.

Existing AD implementations for the JVM include Lantern [**8**], Nexus [**3**] and DeepLearning.scala [**2**], however these are Scala-based and do not interoperate with other JVM languages. Kotlin∇ is fully interoperable with vanilla Java, enabling broader adoption in neighboring languages. To our knowledge, Kotlin has no prior AD implementation. However, the language contains a number of desirable features for implementing a native AD framework. In addition to type-safety and interoperability, Kotlin∇ primarily relies on the following language features:

Operator overloading and infix functions allow a concise notation for defining arithmetic operations on tensor-algebraic structures, i.e. groups, rings and fields. **λ-functions and coroutines** support backpropogation with lambdas and shift-reset continuations, following [**6**] and [**8**]. **Extension functions** support extending classes with new fields and methods and can be exposed to external callers without requiring sub-classing or inheritance.

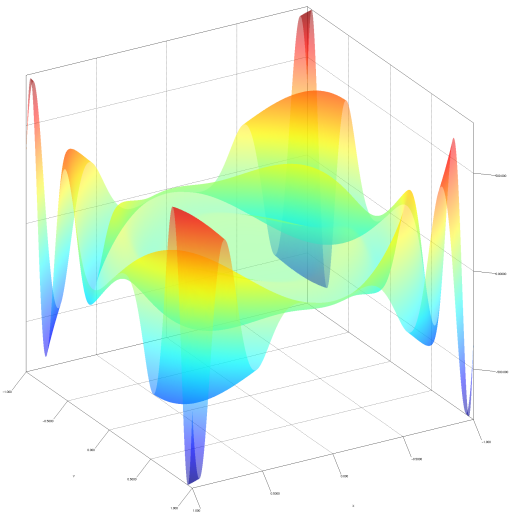
3.4. Usage

Kotlin∇ allows users to implement differentiable programs by composing operations on fields to form algebraic expressions. Expressions are lazily evaluated inside a numerical context, which may imported on a per-file basis or lexically scoped for finer-grained control over the runtime behavior.

Fig. 3.1. A basic Kotlin∇ program with two inputs and one output.

Above, we define a function with two variables and take a series of partial derivatives with respect to each variable. The function is numerically evaluated on the interval $(-1, 1)$ in each dimension and rendered in 3-space. We can also plot higher dimensional manifolds (e.g. the loss surface of a neural network), projected into four dimensions, and rendered

in three, where one axis is represented by time. To demonstrate, a display is required for animation purposes.



$z = \sin\big(10(x * x + y^2)\big)/10$, `plot`$\left(\frac{\partial^3 z}{\partial x^2 \partial y}\right)$

Fig. 3.2. Output generated by the program shown in Figure 3.1.

Kotlin∇ treats mathematical functions and programming functions with the same underlying abstraction. Expressions are composed recursively to form a data-flow graph (DFG). An expression is simply a `Function`, which is only evaluated once invoked with numerical values, e.g. `z(0, 0)`.

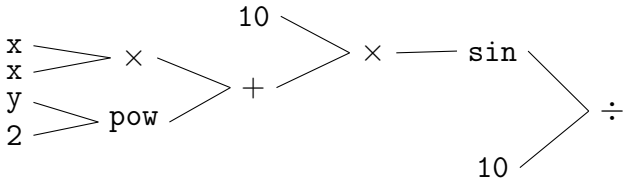


Fig. 3.3. Implicit DFG constructed by the original expression, `z`.

Kotlin∇ supports shape-shafe tensor operations by encoding tensor rank as a parameter of the operand's type signature. By enumerating type-level integer literals, we can define tensor operations just once using the highest literal, and rely on Liskov substitution to preserve shape safety for subtypes.

Fig. 3.4. Shape safe tensor addition for rank-1 tensors, $\forall L \leq 2$.

It is possible to enforce shape-safe vector construction as well as checked vector arithmetic up to a fixed `L`, but the full implementation is omitted for brevity. A similar pattern can be applied to matrices and higher rank tensors, where the type signature encodes the shape of the operand at runtime.

With these basic ingredients, we have almost all the features necessary to build an expressive shape-safe AD, but unlike prior implementations using Scala or Haskell, in a language that is fully interoperable with Java, while also capable of compiling to JVM bytecode, JavaScript, and native code.

In future work, we intend to implement a full grammar of differentiable primitives including matrix convolution, control flow and recursion. While Kotlin∇ currently implements arithmetic manually, we plan to wrap a BLAS such as cuBLAS or native linear algebra library for performance.

3.5. Type system

Describing the Kotlin∇ type system (formally).

Math	Infix	Prefix	Postfix	Type
$A+B$	`a + b` `a.plus(b)`	`plus(a, b)`		$(\mathtt{a} : \mathbb{R}^\tau \to \mathbb{R}^\pi, \mathtt{b} : \mathbb{R}^\lambda \to \mathbb{R}^\pi) \to (\mathbb{R}^? \to \mathbb{R}^\pi)$
$A-B$	`a - b` `a.minus(b)`	`minus(a, b)`		$(\mathtt{a} : \mathbb{R}^\tau \to \mathbb{R}^\pi, \mathtt{b} : \mathbb{R}^\lambda \to \mathbb{R}^\pi) \to (\mathbb{R}^? \to \mathbb{R}^\pi)$
AB	`a * b` `a.times(b)`	`times(a, b)`		$(\mathtt{a} : \mathbb{R}^\tau \to \mathbb{R}^{m*n}, \mathtt{b} : \mathbb{R}^\lambda \to \mathbb{R}^{n*p}) \to (\mathbb{R}^? \to \mathbb{R}^{m*p})$
$\frac{A}{B}$ AB^{-1}	`a / b` `a.div(b)`	`div(a, b)`		$(\mathtt{a} : \mathbb{R}^\tau \to \mathbb{R}^{m*n}, \mathtt{b} : \mathbb{R}^\lambda \to \mathbb{R}^{p*n}) \to (\mathbb{R}^? \to \mathbb{R}^{m*p})$
$-A$ $+A$		`-a` `+a`	`a.unaryMinus()` `a.unaryPlus()`	$(\mathtt{a} : \mathbb{R}^\tau \to \mathbb{R}^\pi) \to (\mathbb{R}^\tau \to \mathbb{R}^\pi)$
A+1 A-1	`a + 1` `a - 1`	`++a` `-a`	`a++, a.inc()` `a-, a.dec()`	$(\mathtt{a} : \mathbb{R}^\tau \to \mathbb{R}^{m*m}) \to (\mathbb{R}^\tau \to \mathbb{R}^{m*m})$

sin(a)
cos(a)
tan(a)
		`sin(a)`
		`cos(a)`
		`tan(a)`
			`a.sin()`
			`a.cos()`
			`a.tan()`
				$(\mathtt{a} : \mathbb{R} \to \mathbb{R}) \to (\mathbb{R} \to \mathbb{R})$

$\ln(A)$		`ln(a)` `log(a)`	`a.ln()` `a.log()`	$(\mathtt{a} : \mathbb{R}^\tau \to \mathbb{R}^{m*m}) \to (\mathbb{R}^\tau \to \mathbb{R}^{m*m})$
$\log_b A$	`a.log(b)`	`log(a, b)`		$(\mathtt{a} : \mathbb{R}^\tau \to \mathbb{R}^{m*m}, \mathtt{b} : \mathbb{R}^\lambda \to \mathbb{R}^{m*m}) \to (\mathbb{R}^? \to \mathbb{R})$
A^b	`a.pow(b)`	pow(a, b)		$(\mathtt{a} : \mathbb{R}^\tau \to \mathbb{R}^{m*m}, \mathtt{b} : \mathbb{R}^\lambda \to \mathbb{R}) \to (\mathbb{R}^? \to \mathbb{R}^{m*m})$
$\sqrt{a}$ $sqrt[3](a)$	`a.pow(1.0/2)` `a.root(3)`	`a.pow(1.0/2)` `a.root(3)`	`a.sqrt()` `a.cbrt()`	$(\mathtt{a} : \mathbb{R}^\tau \to \mathbb{R}^{m*m}) \to (\mathbb{R} \to \mathbb{R}^{m*m})$
$\frac{da}{db}$ $a'(b)$	`a.diff(b)`	`grad(a)[b]`	`d(a) / d(b)`	$(\mathtt{a} : C(\mathbb{R}^m)^*, \mathtt{b} : \mathbb{R} \to \mathbb{R}) \to (\mathbb{R}^m \to \mathbb{R})$
∇a		`grad(a)`	`a.grad()`	$(\mathtt{a} : C(\mathbb{R}^m)^*) \to (\mathbb{R}^m \to \mathbb{R}^m)$

3.6. Differentiable programming

The renaissance of modern deep learning is widely credited to progress in three research areas: algorithms, data and hardware. Among algorithms, most research has focused on deep learning architectures and representation learning. Equally important, arguably, is the role that automatic differentiation (AD) has played in facilitating the implementation of these ideas. Prior to the adoption of general-purpose AD libraries like Theano, PyTorch and TensorFlow, gradients needed to be programmed manually. The emergence of these and other libraries specifically tailored to deep learning simplified and accelerated the pace of machine learning, allowing researchers to focus on gradient-based optimization, network architectures and learning representations. Some of these ideas in turn, formed the basis for new methods in automatic differentiation, which continues to be an active area of research.

In deep learning, research primarily focuses on differentiable representations. In contrast, many problems we would like to solve are non-differentiable in their natural domain. For example, the structure of character-based representations are not easily differentiable as small changes to a word's symbolic representation can cause large differences to its semantic meaning. A key insight from deep learning research is that many problems require posing a good representation. For example, if we can represent words as a vector of real numbers, so that relations between words are spatially preserved by the vector representation, then we can perform gradient descent on such a space. Many classes of discrete problems can be relaxed to continuous surrogates by introducing a good representation.

As more domains found efficient representations, researchers observed neural networks were part of a broader class of differentiable architectures that could be built and interpreted in a manner not unlike computer programming [1], [2]. Hence the term *differentiable programming* was born.

Chapitre 4

Verification and validation

How do we test a machine that learns?

4.1. Adversarial test case generation

Neural networks and differentiable programming has provided a powerful new set of optimization tools for training learning algorithms. However these methods are often brittle to small variations in the input space, and have difficulty with generalization. In contrast, these same techniques used for probing the failure modes of neural networks can be applied to adversarial test case generation for traditional programs.

4.2. Background

Suppose we have a program $P : \mathbb{R} \to \mathbb{R}$ where:

$$P(x) = p_n \circ p_{n-1} \circ p_{n-2} \circ ... \circ p_1 \circ p_0 \tag{4.2.1}$$

From the chain rule of calculus, we know that:

$$\frac{dP}{dp_0} = \prod_{i=1}^{n} \frac{dp_i}{dp_{i-1}} \tag{4.2.2}$$

Imagine a single test $T : \mathbb{R} \to \mathbb{B}$. Consider the following example::

$$T : \forall x \in (0, 1), P(x) < C \tag{4.2.3}$$

How can we find a set of inputs that break the test under a fixed computational budget (i.e. constant number of program evaluations)? In other words:

$$D_T : \{x^i \sim \mathbb{R}(0, 1) \mid P(x^i) \implies \neg T\}, maximize|D_T| \tag{4.2.4}$$

If we have no information about the program implementation or its input distribution, D_P, we can do no better than random search (cf. no free lunch). However, if we know something about the input distribution, we could re-parameterize the distribution to incorporate our knowledge. Assuming the program already works on most common inputs (e.g. from a training set), we might want to sample $x \sim \frac{1}{D_P}$ for inputs that are infrequent, although this strategy is not guaranteed to yield any additional failures. If we knew how P were implemented, we might be able prioritize our search in input regions leading towards internal discontinuities (e.g. edge cases in software testing). However for functions that are continuous and differentiable, these heuristics are almost certainly insufficient. Finally, we could train a neural network to predict inputs that were likely to cause a program to fail a given specification. As input, the network would take the function and test cases, and as its output, produce values that were likely to violate T.

Another strategy, independent of how candidate inputs are selected, is to use some form of gradient based optimization in the search procedure. For example in (3) we could have a loss function:

$$\mathcal{L}(P, x) = C - P(x) \tag{4.2.5}$$

The gradient of the loss w.r.t. x (assuming P is fixed[1]) is:

$$\nabla\mathcal{L}(P, x) = -\frac{dP}{dx} \tag{4.2.6}$$

Where the vanilla gradient update step is defined as:

$$x^i_{n+1} = x^i_n - \alpha\frac{dP}{dx} \tag{4.2.7}$$

We hypothesize that if the implementation of P were flawed and a counterexample to (3) existed, as sample size increased, a subset of gradient descent trajectories would fail to converge, a portion would converge to local minima, and a subset of trajectories would discover inputs violating the program specification. How would such a search procedure look in practice? Consider the following:

[1]In contrast with backpropogation, where the parameters are updated.

Input : Program P, specification T, evaluation budget $Budget$

Output: D_T, the set of inputs which cause P to fail on T

$D_T = []$;

evalCount = 0;

while *evalCount* $\leq$*Budget* **do**

  sample candidate input x^i *according to selection strategy S;*

  if $P(x^i) \implies \neg T$ **then**

    append x^i *to* D_T

  else

    $n = 0$;

    $x_n^i = x^i$;

    while $n \leq C \wedge evalCount \leq Budget \wedge \neg converged$ **do**

      $n{+}{+}$;

      $x_n^i = x_{n-1}^i - \alpha \frac{dP}{dx}$;

      if $P(x_n^i) \implies T$ **then**

        append x_n^i *to* D_T

        break;

      end

      evalCount++;

    end

  end

  evalCount++; i++;

end

Algorithm 1: Algorithm for finding test failures. First select a candidate input x^i according to sampling strategy S (e.g. uniform random, or a neural network which takes P and T as input). If $P(x^i)$ violates T, we can append x^i to D_T and repeat. Otherwise, we follow the gradient of $\mathcal{L}(P, x)$ with respect to x and repeat until test failure, gradient descent convergence, or a fixed number of G.D. steps C are reached before resampling x^{i+1} from the initial sampling strategy S to ensure each G.D. trajectory will terminate before exhausting our budget.

4.3. Regression testing: a tool to prevent catastrophic forgetting

An endemic problem in modern deep learning is the problem of forgetting. In order to combat this issue, we turn to a classic software testing tool: regression testing.

Regular regression testing gives clear diagnostics about the behavior of these systems.

Chapitre 5

Case study: application for autonomous robotics

Here, we implement a mobile application using the above toolchain.

5.1. Design

Designed with Hatchery.

5.2. Implementation

Implementation includes Kotlin∇

5.3. Verification

Verified using property-based testing.

5.4. Maintenance

Deployed and CI-tested using Docker.

Chapitre 6

Software reproducibility

6.1. Operating systems and virtualization

Operating systems are a source of variability. Filesystems are complicated.

6.2. Dependency management

Packages are a source of variability. Dependency hell is NP Complete: https://research.swtch.com/version-sat

6.3. Containerization

One of the challenges of distributed software development across heterogeneous platforms is the problem of variability. With the increasing pace of software development comes the added burden of software maintenance. As hardware and software stacks evolve, so too must source code be updated to build and run correctly. Maintaining a stable and well documented codebase can be a considerable challenge, especially in a robotics setting where contributors are frequently joining and leaving the project. Together, these challenges present significant obstacles to experimental reproducibility and scientific collaboration.

In order to address the issue of software reproducibility, we developed a set of tools and development workflows that draw on best practices in software engineering. These tools are primarily built around containerization, a widely adopted virtualization technology in the software industry. In order to lower the barrier of entry for participants and minimize variability across platforms (e.g. simulators, robotariums, Duckiebots), we provide a state-of-the-art container infrastructure based on Docker, a popular container engine. Docker

allows us to construct versioned deployment artifacts that represent the entire filesystem and to manage resource constraints via a sandboxed runtime environment.

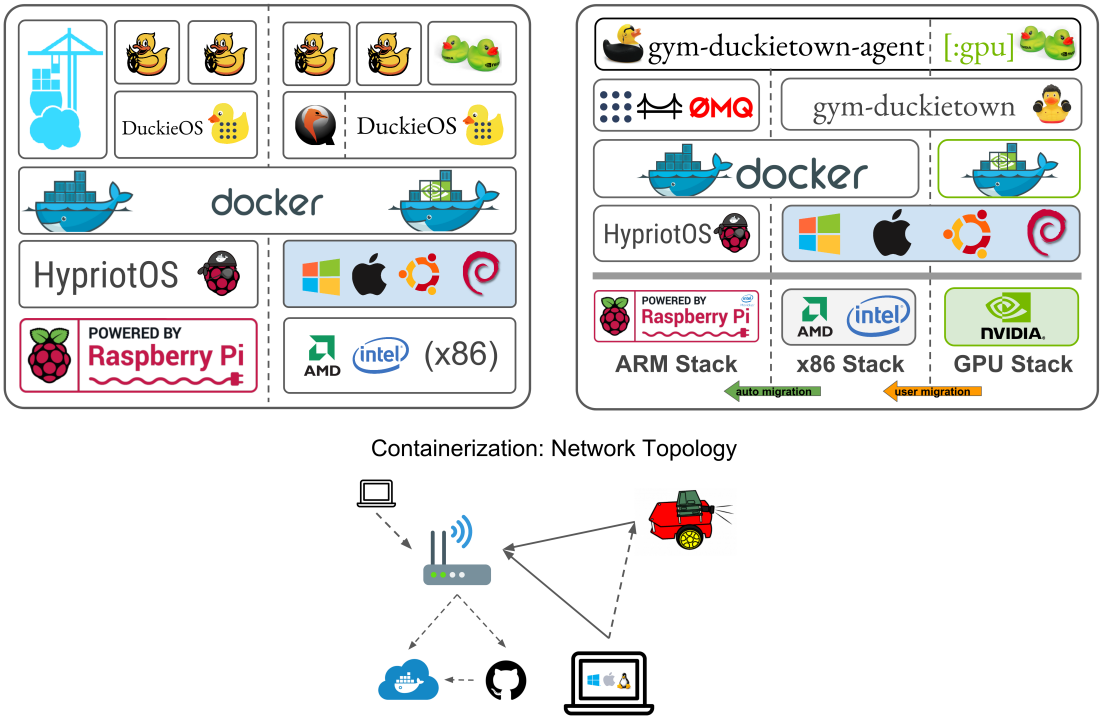


Fig. 6.1. AI-DO container infrastructure. Left: The ROS stack targets two primary architectures, x86 and ARM. To simplify the build process, we only build ARM artifacts, and emulate ARM on x86. Right: Reinforcement learning stack. Build artifacts are typically trained on a GPU, and transferred to CPU for evaluation. Deep learning models, depending on their specific architecture, may be run on an ARM device using an Intel NCS.

The Duckietown platform supports two primary instruction set architectures: x86 and ARM. To ensure the runtime compatibility of Duckietown packages, we cross-build using hardware virtualization to ensure build artifacts can be run on all target architectures. Runtime emulation of foreign artifacts is also possible, using a similar technique.[1] For performance and simplicity, we only build ARM artifacts and use emulation where necessary (e.g., on x86 devices). On ARM-native, the base operating system is HypriotOS, a lightweight Debian distribution with built-in support for Docker. For both x86 and ARM-native,

[1]For more information, this technique is described in further depth at the following URL: `https://www.balena.io/blog/building-arm-containers-on-any-x86-machine-even-dockerhub/`.

Docker is the underlying container platform upon which all user applications are run, inside a container.

Docker containers are sandboxed runtime environments that are portable, reproducible and version controlled. Each environment contains all the software dependencies necessary to run the packaged application(s), but remains isolated from the host OS and file system. Docker provides a mechanism to control the resources each container is permitted to access, and a separate Linux namespace for each container, isolating the network, users, and file system mounts. Unlike virtual machines, container-based virtualization like Docker only requires a lightweight kernel, and can support running many simultaneous containers with close to zero overhead. A single Raspberry Pi is capable of supporting hundreds of running containers.

While containerization considerably simplifies the process of building and deploying applications, it also introduces some additional complexity to the software development lifecycle. Docker, like most container platforms, uses a layered filesystem. This enables Docker to take an existing "image" and change it by installing new dependencies or modifying its functionality. Images may be based on a number of lower layers, which must periodically be updated. Care must be taken when designing the development pipeline to ensure that such updates do not silently break a subsequent layer as described earlier in Sec. **??**.

One issue encountered is the matter of whether to package source code directly inside the container, or to store it separately. If source code is stored separately, a developer can use a shared volume on the host OS for build purposes. In this case, while submissions may be reproducible, they are not easily modified or inspected. The second method is to ship code directly inside the container, where any changes to the source code will trigger a subsequent rebuild, effectively tying the sources and the build artifacts together. Including source code alongside build artifacts also has the benefit of reproducibility and diagnostics. If a competitor requires assistance, troubleshooting becomes much easier when source code is directly accessible. However doing so adds some friction during development, which has caused competitors to struggle with environment setup. One solution is to store all sources on the local development environment and rebuild the Docker environment periodically, copying sources into the image.

Chapitre 7

Conclusion

7.1. Future work

7.1.1. Requirements Engineering

Often it is not possible, or desirable to summarize the function of a complex system using a single variable. In multi-objective optimization, we have the notion of pareto-efficiency...

Tranditional software engineering has followed a rigorous process model and testing methodology. This model has guided the development of traditional software engineering, intelligent systems will require a reimagining of these ideas to build systems that adapts to its environment during operation. Intelligent systems are designed with objective functions, which are typically one- or low-dimensional metrics for evaluating the performance of the system. Most often, these take the form of a single criteria, such as an *error* or *loss* which can represent descriptive phenomena such as latency, safety, energy efficiency or any number of objective measures.

For example, in the design of a web based advertisement recommendation system, we can optimize for various objectives such as click rate, engagement, sales conversion. So long as we can measure these parameters, with today's powerful function approximators, we can optimize for any singly criterion or combination thereof. Much of the work involved in machine learning is to find representations which are amenable to learning, and preventing unintended consequences. For example, by optimizing for click rate, we create an artificial market for click bots. Similarly, in self driving cars, we often want to optimize for passenger safety. However by doing so naively, we create a vehicle that never moves, or always yields to nearby vehicles.

When building intelligent system developers must first ask, what are the requirements of the system? This process is often the most troublesome part, because the requirements must not be fuzzy specifications like traditional software engineering, but precise, programmable directives. "I would like the system to be fast," is not sufficiently precise. These kinds of requirements must be translated into statistical loss functions, so we must be very precise about how we specify our requirements. If we simply say, "The system must produce a valid response as quickly as possible, in less than 100ms," is better, but leaves open the possibility of returning an empty response.

In traditional software engineering, it is reasonable to assume the people who are implementing the system have some implicit knowledge and are generally well-intentioned human beings working towards the same goal. When building an intelligent system, a more reasonable assumption is that the entity implementing our requirements is a naive but powerful genie, and possibly an adversary. If we are to give it an optimization metric, it will take every conceivable shortcut to achieve that metric. If we are not careful about requirements engineering, this entity can produce a system that does not work, or has unintended consequences.

In the strictest sense, designing a good set of requirements is indistinguishable from implementing the system. With the right language abstractions (e.g. declarative programming), requirements and implementation can be the same thing. These ideas have been explored in recent decades with languages like SQL and Prolog. While these are toy systems, neural networks can express much larger classes of functions than traditional software engineering.

7.1.2. Continuous Delivery and Continual Learning

In CI/CD, software artifacts are "continously" updated. Similarly, intelligent systems must continously adapt to their environment and this trend will only accelerate.

Incremenental updates will grow increasingly smaller, until the program starts to alter itself after every input it processes.

We need tools to more effectively harness the stochasticity of these learning systems.

7.1.3. Developers, Operations, and the DevOps toolchain

Software engineers have begun to realize the value of bespoke tools that facilitate the process of shipping software, in addition to the software itself.

Teams building software are cybernetic systems, and require meta-programs for building code and organizational processes which enable them to ship code more efficiently.

Bibliography

[1] Atilim Gunes Baydin, Barak A. Pearlmutter, and Jeffrey Mark Siskind. Diffsharp: Automatic differentiation library. *CoRR*, abs/1511.07727, 2015.

[2] Yang Bo. DeepLearning.scala: A simple library for creating complex neural networks. 2018.

[3] Tongfei Chen. Typesafe abstractions for tensor operations (short paper). pages 45–50, 2017.

[4] Yossi Gil and Tomer Levy. Formal language recognition with the java type checker. 56, 2016.

[5] Radu Grigore. Java generics are Turing Complete. pages 73–85, 2017.

[6] Barak A Pearlmutter and Jeffrey Mark Siskind. Reverse-mode AD in a functional framework: Lambda the ultimate backpropagator. *ACM Transactions on Programming Languages and Systems (TOPLAS)*, 30(2):7, 2008.

[7] Barak A Pearlmutter and Jeffrey Mark Siskind. Using programming language theory to make automatic differentiation sound and efficient. pages 79–90, 2008.

[8] Fei Wang, Xilun Wu, Grégory M. Essertel, James M. Decker, and Tiark Rompf. Demystifying differentiable programming: Shift/reset the penultimate backpropagator. *CoRR*, abs/1803.10228, 2018.

[9] Richard Wei. First-class automatic differentiation in Swift: A manifesto. 2018.