ABSTRACT SEQUENTIAL MEMORY

HENRY STEINITZ

Contents

1. Introduction

Contemporary neural networks are complex systems composed of many submodules, each of which typically contains many learnable parameters which are trained in coordination with the entire network by gradient descent or other gradient-based methods. End-to-end training by gradient methods is possible because each module is required to be differentiable, which ensures error signals may flow through the network uninterrupted during an iterated application of the chain rule—backpropagation.

While end-to-end differentiability has led to an explosion of effective eural network architectures and unified learning algorithms, submodules are rarely evaluated outside the context of the complete network. Architectures are typically trained end-to-end and then evaluated on a standardized ML datasets. Without scientific evaluation, individual architecture choices, for example the choice of a recurrent memory architecture (e.g. GRU vs. LSTM), can have substantial-but-unnoticed effects on overall performance. This makes it difficult to target individual components responsible for the final model's strengths and weaknesses.

We suggest it's useful to build theoretical tools for classes of submodules common to many end-end differentiable architectures, and in this work, we build a simple theory of long-term, sequential, read/write memory.

2. Sequential memory

We begin with a simple model of long-term memory, where the goal is to remember a sequence of points $x_1, \dots x_n$ from an input space X by iteratively updating a running representation m_i in a memory space M. Then, we'd like reverse this process to recover previous x_i from m_n. Formally,

Definition 2.1. Let X be a Riemannian manifold called the input space. A sequence memory (M, π, γ) on X is a Riemannian manifold M with two smooth maps:

- $\pi : M \times X \to M$ called the save map.
- $\gamma : M \to X \times M$ called the recover map.

The save map induces the *iterated save map* $\pi^n : M \times X^n \to$ defined as

$$\pi^n(M, x_1, x_2, \dots, x_n) = \pi(((\dots \pi(\pi(M, x_1), x_2) \dots) x_{n-1}), x_n)$$

Note that we don't use superscripts to indicate dimension, ad if X has dimension n, then X^n has dimension $n \times N$. We can similarly define the *iterated recover map* $\gamma^n : M \to X^n \times M$. If we want to consider applications in deep learning architectures, we need to ensure that the save and recover maps are smooth.

3. Finite-dimensional memory is imperfect

Definition 3.1. A memory is perfect if for all $x \in X$, $m \in M$

$$\gamma(\pi(m, x)) = (m, x)$$

Theorem 3.2. *In a perfect memory the iterated save and iterated recovery maps are also inverses of each other.*

Proof. induction ∎

If we forget about th input and memory's manifold structure, any bijective save map $\pi : m \times X \to X$ and recover map $\gamma = \pi^{-1}$ will be perfect. Since we're interested in applications to machine learning and differentiable programming, we require that these maps are smooth. When this condition is imposed, no finite-dimensional memory can be perfect.

Theorem 3.3. *A perfect differential sequential memory must have infinite dimension.*

Proof. Note that $\pi : M \times X \to M$ is a differentable map with inverse γ^n. It follows that π is a diffeomorphism onto it's image$\pi^n(M \times X)$, which is a subset of M. So

$$\dim(\pi(M \times X)) = \dim M + \dim X <= \dim M$$

which only holds for a trivial input space $\dim X = 0$ when both $X$and M are finite dimensional. $\square$

4. Geometric recovery bounds

Theorem 4.1. *Suppose the recall map $\gamma \circ \pi : M \times X \to M \times X$ moves each point in $M, \times X$ at most K:*

$$\left|\gamma \circ \pi(m, x) - (m, x)\right| \leq K.$$

in other words, the distance between original inputs and their recall is bounded.

Proof. Geometric series $\square$

5. Perfect infinite-dimensional memory

Using an infinite-dimensional memory manifold, we may create a perfect and spatially-bounded memory for n-dimensional Euclidean space. Let $\mathbb{R}^n$, and M the manifold of smooth functions on $[-1, 1]^n$. Then M is infinite-dimensional. Define the save map as

$$\pi(m, x) = \int_0^t m(a) da + x$$

and the recover map as

$$\gamma(m) = \left(m(0), \frac{d}{dt} m\right).$$

The first coordinate of recall map evaluates to

$$\begin{aligned} \gamma(\pi(m, x))_1 &= \int_0^0 m(a) da + x \\ &= x \end{aligned}$$

By the fundamental theorem of calculus, the first coordinate of recall map evaluates to

$$\begin{aligned} \gamma(\pi(m, x))_2 &= \frac{d}{dt} \left(\int_0^t m(a) da + x\right) \\ &= \frac{d}{dt} \left(\int_0^t m(a) da\right) + \frac{d}{dt} x \\ &= m(t) + 0 \\ &= m(t) \end{aligned}$$