Universal Approximation of Input-Output Maps by Temporal Convolutional Nets

Joshua Hanson and Maxim Raginsky
University of Illinois at Urbana-Champaign

Background:
- Convolutional nets are replacing recurrent nets in a growing number of sequence-to-sequence modeling applications (machine translation, audio generation, language modeling). Why?
- Both are inherently suited to modeling systems with limited long-term dependencies.
- Recurrent nets have theoretically infinite memory, but often fail to learn long sequences. Infinite memory is also usually unnecessary in practice.
- The lack of feedback elements in convolutional nets provides a computational advantage (copies of the input sequence can be processed in parallel).

Questions:
- When do convolutional architectures provide better approximation than recurrent architectures?
- How do we quantify “limited long-term dependencies”?

Definitions:
- i/o map — nonlinear operator $F: S \rightarrow \mathcal{S}$ where $S := \{ u = (u_h)_{h \in \mathbb{Z}} \}$
- Right shift — $(F u)_h := u_{h+1}$, $h \in \mathbb{Z}$
- Window — $(W_{w,\gamma} u)_h := u_{\gamma h + (|\gamma| w - w_0)}$ for $w_0 \leq \gamma h < w_0 + w$
- Causal — $w_0 = 0$
- Time-invariant — $\forall k \in \mathbb{Z}$, $(F^k u)_h := (F^k (F^{k-1} \cdots (F u)_h))_{\gamma h}$ for $\gamma h \geq 0$
- Approximately finite memory on $M \subset \mathcal{S}$ — exists $\epsilon > 0$ such that $\sup_{w_0,M} \sup_{\gamma \leq 0} \sup_{h \in \mathbb{Z}} |(F u)_h - (F^{\gamma h + w_0} u)_h| \leq \epsilon$

Main Result:
Assumption 1: The i/o map $F$ has approximately finite memory on $M(R)$. 
Assumption 2: For any $t \in \mathbb{Z}$, the functional $\hat{F}_t : R^{t+1} \rightarrow R$ is uniformly continuous on $[-R, R]^{t+1}$ with modulus and inverse modulus of continuity
- $\omega_F(\delta) := \sup \{ |\hat{F}_t(x) - \hat{F}_t(x')| : x, x' \in [-R, R]^{t+1}, |x - x'|_\infty \leq \delta \}$
- $\omega_F^T(\delta) := \sup \{ \delta > 0 : \omega_F(\delta) \leq \epsilon \}$

Theorem: Let $F$ be a causal, time-invariant i/o map satisfying Assumptions 1 and 2. Then for any $\epsilon > 0, \gamma \in (0,1)$ there exists a ReLU TCF $F$ with context length $m = m_F(\gamma, \epsilon)$, width $m + 2$, depth $\frac{\log(1 + R^2)}{\log(1 + \gamma)}$, such that
- $\sup_{w_0,M} \sup_{\gamma \leq 0} \sup_{h \in \mathbb{Z}} |(F u)_h - (F^{\gamma h + w_0} u)_h| \leq \epsilon$

Remark: The role of $\gamma$ is to trade off context length and depth. Tuning $\gamma$ can reduce the number of computation units needed to achieve $\epsilon$-accuracy.

Fading Memory:
- Set of weighting sequences — $W := \{ w \in S : w_h \in (0,1], w_h \downarrow 0 \text{ as } t \rightarrow \infty \}$
- $w$-fading memory on $M \subset \mathcal{S}$ — $\forall w > 0 \exists \delta > 0 \forall \bar{w}, \bar{v} \in M \forall \gamma \in \mathbb{Z}$
- $\max_{\gamma \geq 0} \inf_{w_0, \gamma} \sup_{h \in \mathbb{Z}} |(F u)_h - (F^{\gamma h + w_0} u)_h| \leq \delta 

Proposition: An i/o map $F$ satisfies Assumptions 1 and 2 if and only if it has $w$-fading memory on $M(R)$ for arbitrary $w \in W$.

Figure 1 (left): Definition of approximately finite memory

Figure 2 (right): Sliding context window of ReLU TCF

$$f(x_{t+1}) = \int f(x_t, u_t) \, g(y_t)$$

where $f : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$, $g : \mathbb{R}^n \rightarrow \mathbb{R}$, and $\xi$. We consider systems that are

Uniformly asymptotically incrementally stable on $M \subset \mathcal{S}$ — there exists a function $\beta : \mathbb{R}_+ \times \mathbb{R} \rightarrow (0,\infty)$ of class $\mathcal{K}$ such that

$$\sup_{x \in M} \psi_{\beta} \left( z \right) \leq \beta \left( z, \varphi_{\beta} \left( z \right) \right)$$

where $\varphi_{\beta} \left( z \right)$ is the solution to the system above for $x_0 = z$ and input $u$. 

Theorem: Assume $f$ and $g$ are Lipschitz, $\psi_{\beta} \left( z \right)$ remains in a compact set for all $u \in M$ and all $t \in \mathbb{Z}_+$, and $\beta$ is summable over its second argument. Then the i/o map for the above system satisfies Assumptions 1 and 2.

Examples: Contractivity, Lur’e systems + circle criterion, Denylovich criterion