Lecture 10: Architectural benefits

Announcements:
* Next 5 lectures on zoom.
* HW2 probably out Oct 2.

Original material, now "open questions"

1. Architectural benefits of convolution & attention layers.
2. Benefits of activation choice
3. Modeling distributions
4. Low norm approximation
5. Effect of depth (non-algorithmic)
   (i) polynomial networks (sum-product networks)
   (ii) many layers to fewer layers
   (iii) depth 3 vs depth 2
All based on "flatten mapping $\Delta : \mathbb{R} \to \mathbb{R}^\star$"

$$\begin{align*}
\Delta(x) &= \begin{cases} 
2x & \text{if } x \in [0, 1] \\
-x & \text{otherwise}
\end{cases} \\
\Delta(\Delta(x)) &= \begin{cases} 
2(2x - 1) & \text{if } 2x - 1 \in [0, 1] \\
-2(2x - 1) & \text{otherwise}
\end{cases}
\end{align*}$$

$$\Delta^\infty(x) = \begin{cases} 
2^k x & \text{if } 2^k x \in [0, 1] \\
-2^k x & \text{otherwise}
\end{cases}$$

**Remarks:**
Approximating this with shallow network needs exponentially many nodes.

One attempt to capture fractal structure:

For $x \in [0, 1], \Delta^k(x) = \Delta(\Delta^k(x)) = \Delta(2^k x)$

Proof: Induction on layers: $\forall k \in \mathbb{N}, \Delta^k(x) = \Delta(2^k x)$

Base case: $\Delta^0(x) = \Delta(2^0 x) = x$.

IH: Assume $\Delta^k(x) = \Delta(2^k x)$.

Then $\Delta^{k+1}(x) = \Delta(\Delta^k(x)) = \Delta(2^k x) = \Delta(2^{k+1} x)$.

Application of $\Delta^k$:

1. Approximate $x^2$ and $x^y$ up to accuracy $\epsilon$ with linear maps $L_\epsilon$ and $L_y$.
2. Polynomial $\Rightarrow$ smooth function.
3. If $x(\epsilon) = x(\epsilon) \forall x \in [0, 1]$, then $f(\Delta^k(x)) = f(\Delta(2^k x))$.
4. In particular, can use $\Delta$ to obtain 2D representations.
Consider progressive affine approximations of $u^3$:

$$S_i = \frac{\Delta}{2} \frac{1}{2}, \frac{3}{2}, \ldots, \frac{i}{2}$$

$h_i$ = affine interpolant of $u^3$ along $S_i$.

Note: $h_i(x) = h_0(x) + \sum_{k=i}^{\Delta} \left( h_{i+k}(x) - h_i(x) \right)$

$$= x + \sum_{k=1}^{\Delta} \left( h_{i+k}(x) - h_i(x) \right)$$

Consider $h_{i+k}(x) - h_{i}(x)$

Case 1: $x \in S_i$:
$S_{ii} \supseteq S_i$, $h_{i+k}(x) - h_i(x) = x - x^3 = 0$.

Case 2: $x \in S_{ii} \setminus S_i$:
Pick $k \leq i$ s.t.

$$h_{i+k}(x) - h_i(x) = -\frac{\Delta^3}{24} x^4 + \text{higher order terms}$$

Case 3: $x \not\in S_{ii} \setminus S_i$:

$$h_{i+k}(x) - h_i(x) = -\frac{\Delta^3}{24} x^4$$

Theorem: Let $h_i$ be the progressive affine interpolation of $u^3$ along $S_i$.

1. $h_i$ can be written as Roll model with $S$: case $i \leq \Delta$.

Proof: 0 What we had before, but rename $\Delta$.

2. Essentially same proof as $h_i(x) = h_{i+k}(x)$, see notes.

**Remark:** 0 Need $O(n^2)$ steps locally to get $E$ accuracy.

2. Multiplication via polynomial identity: $x^4 = \frac{1}{3} (x^2 + y^2)^2$.