DEPARTAMENTO DE
MATEMÁTICA

O poder da Aprendizagem Profunda

Felipe Kaminsky Riffel
Universidade Federal de Santa Catarina

31 de março de 2025

Artigo:

LIN, Henry W.; TEGMARK, Max; ROLNICK, David. Why does deep and cheap learning work so well? Journal of Statistical Physics, v. 168, n. 6, p. 1223–1247, 2017.

Três problemas principais da teoria de redes neurais:

- Expressabilidade: que funções podemos expressar?

[1] Traduzido de "Learnability"

Três problemas principais da teoria de redes neurais:

- Expressabilidade: que funções podemos expressar?
- Eficiência: quão complexa a rede tem que ser?

[1] Traduzido de "Learnability"

Três problemas principais da teoria de redes neurais:

- Expressabilidade: que funções podemos expressar?
- Eficiência: quão complexa a rede tem que ser?
- "Aprendibilidade": quão rápido a rede consegue aprender a ajustar os bons parâmetros? [1]

[1] Traduzido de "Learnability"

Três problemas principais da teoria de redes neurais:

- Expressabilidade: que funções podemos expressar?
- Eficiência: quão complexa a rede tem que ser?
- "Aprendibilidade": quão rápido a rede consegue aprender a ajustar os bons parâmetros? [1]

Aqui, focamos nos dois primeiros: **Expressabilidade** e **Eficiência**.

[1] Traduzido de "Learnability"

Problema: "como redes neurais funcionam bem na prática, se o número de funções possíveis é exponencialmente maior que o número de redes possíveis?"

Exemplo: imagem preta/branca de 1MP vetor de 1000000 entradas com 256 valores possíveis (valor em cada pixel)

Exemplo: imagem preta/branca de 1MP   vetor de 1000000 entradas com 256 valores possíveis (valor em cada pixel)

$$\Longleftrightarrow \qquad \begin{pmatrix} x_1 \\ \vdots \\ x_{1000000} \end{pmatrix}$$

$$x_i \in I_{256} := \{1, 2, 3, \ldots, 256\}$$

Exemplo: imagem preta/branca de 1MP vetor de 1000000 entradas com 256 valores possíveis (valor em cada pixel)

$$\Longleftrightarrow \qquad \begin{pmatrix} x_1 \\ \vdots \\ x_{1000000} \end{pmatrix}$$

$$x_i \in I_{256} := \{1, 2, 3, \ldots, 256\}$$

Nº total de imagens possíveis: $256^{1000000}$.

Exemplo: imagem preta/branca de 1MP   vetor de 1000000 entradas com 256 valores possíveis (valor em cada pixel)

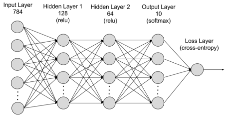
$$\Longleftrightarrow \quad \begin{pmatrix} x_1 \\ \vdots \\ x_{1000000} \end{pmatrix}$$

$$x_i \in I_{256} := \{1, 2, 3, \ldots, 256\}$$

Nº total de imagens possíveis: $256^{1000000}$.

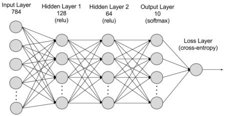
Se existe $p : I_{256} \to (0, 1)$ que associa cada imagem a uma probabilidade, p deve ter uma lista $256^{1000000}$ valores (!!!)

Porém, redes neurais relativamente simples conseguem calcular bem a tarefa.



$$p(\text{Gato}|\mathbf{x}) = \mathbf{83}\%$$

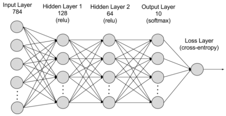
Porém, redes neurais relativamente simples conseguem calcular bem a tarefa.



$$p(\text{Gato}|\mathbf{x}) = \mathbf{83}\%$$

A **matemática** ajuda a explicar: as redes neurais conseguem diminuir drasticamente a explosão combinatória de número de parâmetros em relação ao número de valores;

Porém, redes neurais relativamente simples conseguem calcular bem a tarefa.



$$p(\text{Gato}|\mathbf{x}) = \mathbf{83}\%$$

A **matemática** ajuda a explicar: as redes neurais conseguem diminuir drasticamente a explosão combinatória de número de parâmetros em relação ao número de valores;

A razão também é **física**: as leis sugerem que os datasets de interesse são, em sua maioria, advindos de distribuições simples.

Considere $\mathbf{x} \in \mathbb{R}^d$. Sejam $A_i : \mathbb{R}^d \to \mathbb{R}^d$ operadores afim, i.e.,

$$A_i = W_i - b_i$$

com $W_i \in \mathbb{R}^{m_i \times n_i}$ e $b_i \in \mathbb{R}^{n_i}$.

Considere $\mathbf{x} \in \mathbb{R}^d$. Sejam $A_i : \mathbb{R}^d \to \mathbb{R}^d$ operadores afim, i.e.,

$$A_i = W_i - b_i$$

com $W_i \in \mathbb{R}^{m_i \times n_i}$ e $b_i \in \mathbb{R}^{n_i}$.

Dadas $\sigma_i : \mathbb{R}^{n_i} \to \mathbb{R}^{m_i}$ não linear, chamamos de rede neural feedforward uma função $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ da forma:

$$\mathbf{f}(\mathbf{x}) = \sigma_n A_n \ldots \sigma_2 A_2 \sigma_1 A_1 \mathbf{x}. \tag{1}$$

σ_i pode ser qualquer operador não linear. Escolhas comuns são, dado $\mathbf{x} = (x_1, \ldots, x_n)$:

- Função local: escolha $\sigma : \mathbb{R} \to \mathbb{R}$ não linear e aplique ponto a ponto $\sigma_i(\mathbf{x}) = (\sigma(x_1), \ldots, \sigma(x_n))$;

σ_i pode ser qualquer operador não linear. Escolhas comuns são, dado $\mathbf{x} = (x_1, \ldots, x_n)$:

- Função local: escolha $\sigma : \mathbb{R} \to \mathbb{R}$ não linear e aplique ponto a ponto $\sigma_i(\mathbf{x}) = (\sigma(x_1), \ldots, \sigma(x_n))$;
- Max-pooling: $\sigma_i(\mathbf{x}) = \max_{j=1,\ldots,n}(x_j)$;

σ_i pode ser qualquer operador não linear. Escolhas comuns são, dado $\mathbf{x} = (x_1, \ldots, x_n)$:

- Função local: escolha $\sigma : \mathbb{R} \to \mathbb{R}$ não linear e aplique ponto a ponto $\sigma_i(\mathbf{x}) = (\sigma(x_1), \ldots, \sigma(x_n))$;
- Max-pooling: $\sigma_i(\mathbf{x}) = \max_{j=1,\ldots,n}(x_j)$;
- Softmax:

$$\sigma_i(\mathbf{x}) = \frac{1}{\sum_{j=1}^{n} e^{x_j}}(e^{x_1}, \ldots, e^{x_n}).$$

Teorema 1 (Aproximação da multiplicação)

Seja $\mathbf{f}$ rede neural da forma $\mathbf{f}(\mathbf{x}) = A_2 \sigma A_1 \mathbf{x}$, onde σ é aplicação não linear ponto a ponto qualquer. Considere as camadas de entrada, escondida e de saída com tamanhos 2, 4 e 1 respectivamente. Então, $\mathbf{f}$ pode aproximar uma porta de multiplicação arbitrariamente bem.
Ou seja, dado $\varepsilon > 0$, para qualquer σ não linear (aplicada ponto a ponto), existem $A_1 : \mathbb{R}^2 \to \mathbb{R}^4, A_2 : \mathbb{R}^4 \to \mathbb{R}$ tais que a rede $f(x) = A_2 \sigma A_1 \mathbf{x}$ é tal que, dado $x = (u\ v)^T$ qualquer

$$|f(x) - uv| < \varepsilon$$

para u, v em um compacto qualquer.

Seja $\sigma : \mathbb{R} \to \mathbb{R}$ não linear qualquer suficientemente suave. Na expansão de Taylor em torno de $x = 0$:

$$\sigma(u) = \sigma(0) + \sigma'(0)u + \frac{u^2}{2}\sigma''(0) + \mathcal{O}(u^3).$$

Sem perda de generalidade, considere $\sigma''(0) \neq 0$ (ou então, ajuste b_1 para que $\sigma''(A_{1,1}x - b_{1,1}), \sigma''(A_{1,2}x - b_{1,2}) \neq 0$, que deve existir dado que é não linear).

Seja $\sigma : \mathbb{R} \to \mathbb{R}$ não linear qualquer suficientemente suave. Na expansão de Taylor em torno de $x = 0$:

$$\sigma(u) = \sigma(0) + \sigma'(0)u + \frac{u^2}{2}\sigma''(0) + \mathcal{O}(u^3).$$

Sem perda de generalidade, considere $\sigma''(0) \neq 0$ (ou então, ajuste b_1 para que $\sigma''(A_{1,1}x - b_{1,1}), \sigma''(A_{1,2}x - b_{1,2}) \neq 0$, que deve existir dado que é não linear).

Então,

$$\begin{aligned} m(u, v) &:= \frac{\sigma(u+v) + \sigma(-u-v) - \sigma(u-v) - \sigma(v-u)}{4\sigma''(0)} \\ &= \sigma''(0)\frac{(u+v)^2 + (-u-v)^2 - (u-v)^2 - (v-u)^2 + \mathcal{O}((u+v)^3)}{4\sigma''(0)} \\ &= uv + \mathcal{O}((u+v)^3) \end{aligned}$$

Ou seja, $m(u, v) = uv + \mathcal{O}((u + v)^3)$, de modo que $\lim_{u^2+v^2 \to 0} \frac{m(u,v)-uv}{u^2+v^2} = 0$.

Ou seja, $m(u,v) = uv + \mathcal{O}((u+v)^3)$, de modo que $\lim_{u^2+v^2\to 0} \frac{m(u,v)-uv}{u^2+v^2} = 0$.
Veja que, $m(u,v) = A_2 \sigma A_1 (u\ v)^T$, onde:

$$A_1 = \begin{pmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{pmatrix}, A_2 = (4\sigma''(0))^{-1} \begin{pmatrix} 1 & 1 & -1 & -1 \end{pmatrix}.$$

Ou seja, $m(u,v) = uv + \mathcal{O}((u+v)^3)$, de modo que $\lim_{u^2+v^2\to 0} \frac{m(u,v)-uv}{u^2+v^2} = 0$.
Veja que, $m(u,v) = A_2 \sigma A_1 (u\ v)^T$, onde:

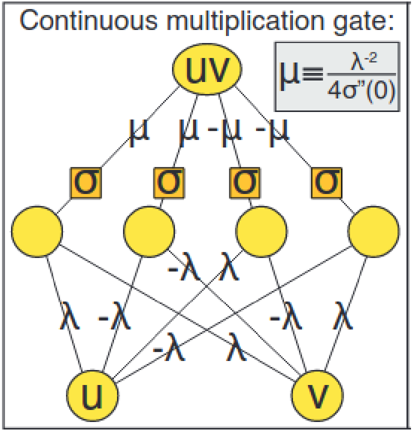
$$A_1 = \begin{pmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ -1 & 1 \end{pmatrix}, A_2 = (4\sigma''(0))^{-1} \begin{pmatrix} 1 & 1 & -1 & -1 \end{pmatrix}.$$

Taylor fornece uma estimativa local, sendo boa para $u, v \approx 0$. Para u, v num compacto de raio qualquer, tome λA_1 e $\lambda^{-2} A_2$ na definição de $\mathbf{f}$, de modo a obter

$$f(x) = (\lambda^{-2} A_2)\sigma(\lambda A_1)x = \lambda^{-2}(\lambda u \lambda v) = uv,$$

tornando a estimativa tão boa quanto se queira. $\square$

Figura: Ilustração da arquitetura da rede no teorema anterior



Fonte: Lin, et.al. (2017)

MTM

Obrigado!

Contato: riffel.felipe@grad.ufsc.br

Repositório com os experimentos desenvolvidos:
https://github.com/felipekriffel/TCC-Regularizacao-EIT