La propagation avant en tant que matrice

Samuel Leblanc

Université de Sherbrooke

14 novembre 2024

Références

i. Marco Armenta et Pierre-Marc Jodoin. "The Representation Theory of Neural Networks". In : *Mathematics* 9.24 (2021)

ii. Marco Armenta, Thomas Brüstle, Souheila Hassoun et Markus Reineke. "Double Framed Moduli Spaces of Quiver Representations". In : *Linear Algebra and its Applications* 650 (2022), p. 98-131

iii. Marco Armenta, Thierry Judge, Nathan Painchaud, Youssef Skandarani, Carl Lemaire, Gabriel Gibeau Sanchez, Philippe Spino et Pierre-Marc Jodoin. "Neural Teleportation". In : *Mathematics* 11.2 (2023)

iv. L., Aiky Rasolomanana et Marco Armenta. *Hidden Activations Are Not Enough : A General Approach to Neural Network Predictions*. 2024. arXiv : 2409.13163 [cs.LG]

$$\mathcal{D} = \left\{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \right\}$$

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

But : Trouver ψ telle que $\psi(x_i) = y_i$.

$$\Psi(W, f) = \begin{pmatrix} \alpha & \gamma \\ \beta & \delta \end{pmatrix} \circ \begin{pmatrix} f \\ f \end{pmatrix} \circ \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

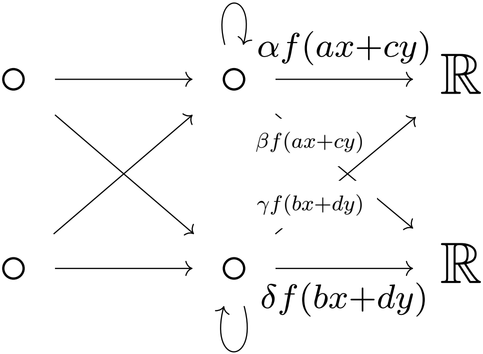
$$
\begin{tikzcd}
\mathbb{R} \arrow[r, "a"] \arrow[dr, "b" near start] & \mathbb{R} \arrow[loop above, "f"] \arrow[r, "\alpha"] \arrow[dr, "\beta" near start] & \mathbb{R} \\
\mathbb{R} \arrow[r, "d"'] \arrow[ur, "c" near start] & \mathbb{R} \arrow[loop below, "f"'] \arrow[r, "\delta"'] \arrow[ur, "\gamma" near start] & \mathbb{R}
\end{tikzcd}
$$

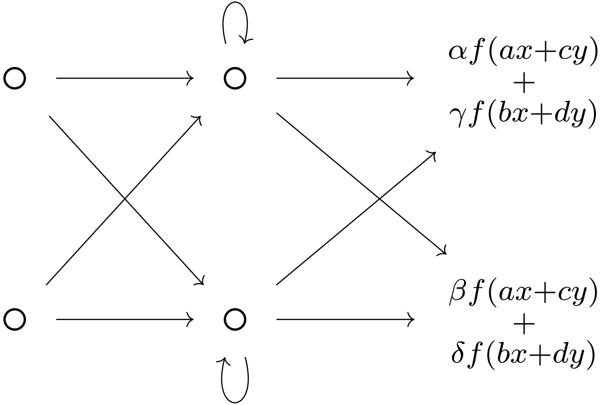
$$
\begin{tikzcd}
x \arrow[r, "a"] \arrow[dr, "b" near start] & \mathbb{R} \arrow[loop above, "f"] \arrow[r, "\alpha"] \arrow[dr, "\beta" near start] & \mathbb{R} \\
y \arrow[r, "d"'] \arrow[ur, "c" near start] & \mathbb{R} \arrow[loop below, "f"] \arrow[r, "\delta"'] \arrow[ur, "\gamma" near start] & \mathbb{R}
\end{tikzcd}
$$

$$
\begin{tikzcd}
\bigcirc \arrow[r, "ax"] \arrow[dr, "bx" near start] & \mathbb{R} \arrow[loop above, "f"] \arrow[r, "\alpha"] \arrow[dr, "\beta" near start] & \mathbb{R} \\
\bigcirc \arrow[r, "dy"'] \arrow[ur, "cy" near start] & \mathbb{R} \arrow[loop below, "f"] \arrow[r, "\delta"'] \arrow[ur, "\gamma" near start] & \mathbb{R}
\end{tikzcd}
$$

$$
\begin{tikzcd}
\circ \arrow[r] \arrow[dr] & ax+cy \arrow[loop above, "f"] \arrow[r, "\alpha"] \arrow[dr, "\beta"] & \mathbb{R} \\
\circ \arrow[r] \arrow[ur] & bx+dy \arrow[loop below, "f"] \arrow[r, "\delta"'] \arrow[ur, "\gamma"] & \mathbb{R}
\end{tikzcd}
$$

$$
\begin{tikzcd}
\circ \arrow[r] \arrow[dr] & f(ax+cy) \arrow[loop above] \arrow[r, "\alpha"] \arrow[dr, "\beta"] & \mathbb{R} \\
\circ \arrow[r] \arrow[ur] & f(bx+dy) \arrow[loop below] \arrow[r, "\delta"'] \arrow[ur, "\gamma"] & \mathbb{R}
\end{tikzcd}
$$


$\alpha f(ax+cy)$
$\mathbb{R}$
$\beta f(ax+cy)$
$\gamma f(bx+dy)$
$\mathbb{R}$
$\delta f(bx+dy)$


$\alpha f(ax+cy)$
$+$
$\gamma f(bx+dy)$
$\beta f(ax+cy)$
$+$
$\delta f(bx+dy)$

$$\mathbb{R}^2 \xrightarrow{\Psi(W,f)} \mathbb{R}^2$$

$$\begin{tikzcd} \mathbb{R}^2 \arrow[r, "\Psi(W,f)"] \arrow[d, "\varphi(W,f)"'] & \mathbb{R}^2 \\ \mathrm{Rep} Q/_{\dot{\cong}} \arrow[ur, "\Psi(\cdot,1)(1)"'] & \end{tikzcd}$$

$$
\begin{tikzcd}
\mathbb{R}^2 \arrow[r, "\Psi(W,f)"] \arrow[d, "\varphi(W,f)"'] & \mathbb{R}^2 \\
\mathrm{Rep}Q/_{\dot{\cong}} \arrow[r, "\pi"'] & \mathrm{Mat}_{2\times 2}(\mathbb{R}) \arrow[u, "\mathrm{ev}_1"']
\end{tikzcd}
$$

$$
\begin{tikzcd}
\mathbb{R}^2 \arrow[r, "\Psi(W,f)"] \arrow[d, "\varphi(W,f)"'] \arrow[dr, dashed, "\mathbb{M}(W,f)"] & \mathbb{R}^2 \\
\mathrm{Rep}Q/_{\dot{\cong}} \arrow[r, "\pi"'] & \mathrm{Mat}_{2\times 2}(\mathbb{R}) \arrow[u, "\mathrm{ev}_1"']
\end{tikzcd}
$$

Attaques adversariales

Figure 1: Visual illustration of adversarial examples crafted by EAD (Algorithm 1). The original example is an ostrich image selected from the ImageNet dataset (Figure 1 (a)). The adversarial examples in Figure 1 (b) are classified as the target class labels by the Inception-v3 model.

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi et Cho-Jui Hsieh. “EAD : Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples”. In : AAAI Press, 2018. URL : https://arxiv.org/abs/1709.04114

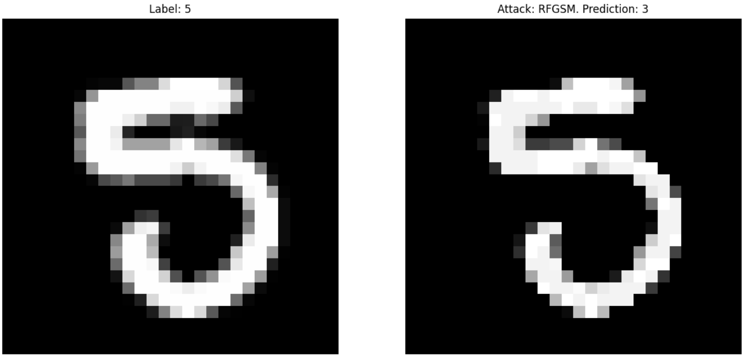
Attaques adversariales

Figure 8. Real-life example of a backdoored stop sign near the authors' office. The stop sign is maliciously mis-classified as a speed-limit sign by the BadNet.

Tianyu Gu, Brendan Dolan-Gavitt et Siddharth Garg. *BadNets : Identifying Vulnerabilities in the Machine Learning Model Supply Chain*. 2019. arXiv : 1708.06733 [cs.CR]. URL : https://arxiv.org/abs/1708.06733

MNIST — Exemple d'une détection d'attaque

Le réseau (W, f) a été entrainé jusqu'à $\approx 98\%$ de précision.



Si on note par $\mathbb{M}_3$ la matrice moyenne de la classe 3, Ψ_3 le vecteur de sortie moyen de la classe 3 et x l'image attaquée, alors

$$\|\mathbb{M}(W, f)(x) - \mathbb{M}_3\|_\infty \approx 84 \text{ et } \|\Psi(W, f)(x) - \Psi_3\|_\infty \approx 4.$$

CODE : *simple_adversarial_detection*. https://github.com/samueleblanc/simple_adversarial_detection. 2024