Notes sur l'abstraction des réseaux de neurones

Paul LANDRIER

1 Généralité

1.1 Idée Générale

On souhaite abstraire le fonctionnemet d'un réseau de neurones.

Si on appelle I le vecteur en entrée du réseau, l'exécution s'écrit naïvement :

$$\phi_1(A_1\phi_2(A_2\phi_3(\dots A_{n-1}\phi_n(A_nI)\dots))) \tag{1}$$

où les (A_i) sont des matrices et les (ϕ_i) sont des fonctions d'activations, typiquement ReLU ou sigmoid coordonnée par coordonnée.

On lui préferera en général l'écriture :

$$\mathbb{R}^{d_1} \xrightarrow{\varphi_1} \mathbb{R}^{d_2} \xrightarrow{\varphi_2} \dots \xrightarrow{\varphi_n} \mathbb{R}^{d_{n+1}} \tag{2}$$

qui met plus en avant les transformtions successives, qui est le processus que l'on cherche à abstraire.

Quelques remarques :

- L'équation 2 s'instancie en l'équation 1 en remplaçant φ_1 par $X \mapsto \phi_n(A_nX)$, φ_2 par $X \mapsto \phi_{n-1}(A_{n-1}X)$ et plus généralement φ_i par $X \mapsto \phi_{n-i+1}(A_{n-i+1}X)$.
- Les transformations pourront être considérées dans le sens direct (input $\rightarrow$ output), comme dans l'exemple précédent, ou indirect (output $\rightarrow$ input), comme pour la *layer-wise relevant propagation*.

1.2 Morphisme de variété

En s'inspirant de l'exemple de la provenance par semi-anneaux, nous aimerions trouver quelle structure est transportée par les applications φ_i. L'ensemble $\mathbb{R}^d$ est naturellement muni d'une structure d'espace vectoriel, conservée dans l'équation 1 par la multiplication par la matrice A_i mais cette structure est complètement ignorée par les activations non-linéaires, par définition. Cela nous conduit à chercher, dans un premier temps, l'abstraction de ces application d'un point de vu topologique.

Dans *Geometric Understanding of Deep Learning*, (http://arxiv.org/abs/1805.10451), Na Lei, Zhongxuan Luo, Shing-Tung Yau et David Xianfeng Gu proposent une interprétation des réseaux de neurones fondée sur les variétés.

On aimerait ainsi de généraliser l'équation 2 à travers une structure topologique sous la forme :

$$V_1 \stackrel{\varphi_1}{\to} V_2 \stackrel{\varphi_2}{\to} \dots \stackrel{\varphi_n}{\to} V_{n+1} \tag{3}$$

où les V_i sont des variétés et les φ_i des morphismes de variété.

Cependant, bien que les activations linéaires conservent pour la plupart la structure de variété (à l'exception notable de ReLU, elles induisent des homéomorphismes) les applications linéaires, elles, ne conservent pas cette structure. On remarque ainsi que la courbe $\{(x, x^2), x \in \mathbb{R}^2\}$, qui est une sous variété de dimension 1 de $\mathbb{R}^2$, est envoyée sur $\mathbb{R}^+$, qui n'est pas une sous variété de $\mathbb{R}$, par l'application $x \to < x, e_2 >$ où $e_2 = (0, 1)$. Dans ce premier cas, l'ensemble d'arrivée est toujours munissable d'une structure de variété à bord. Un exemple plus problématique est par exemple le cas de la variété[1]

$$\begin{aligned} V = &\{(\theta, r\sin(\arctan(\theta))^2, r\cos(\arctan(\theta))^2), \theta \in \mathbb{R}^-, r \in \mathbb{R}\} \\ &\cup \{(\theta, 0, r), \theta \in [0, 1], r \in \mathbb{R}\} \\ &\cup \{(\theta, r\sin(\arctan(\theta - 1))^2, r\cos(\arctan(\theta - 1))^2), \theta \in \mathbb{R}^+, r \in \mathbb{R}\}. \end{aligned} \tag{4}$$

Cet ensemble est une variété différentielle et il est representé en figure 1.2.

En revanhce, sa projection orthogonale sur le plan (x, y) est $\mathbb{R}_{<0} \times \mathbb{R} \cup [0, 1] \times \{0\} \cup \mathbb{R}_{>1} \times \mathbb{R}$ représenté en figure 1.2.

On a ainsi exhibé un exemple de variété dont la structure ne serait pas préservée par les transformations faites par un réseau de neurones.

[1] L'exemple a été trouvé à partir de l'échange https://math.stackexchange.com/questions/1932215/image-of-a-submanifold-under-a-linear-map et d'une discussion avec Antoine Groudiev.

Figure 1: La variété V.

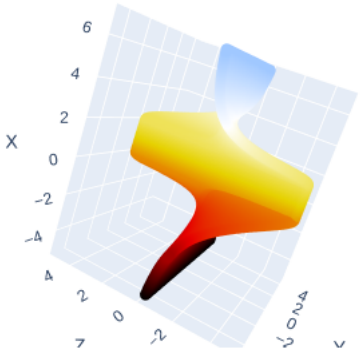

Figure 2: Projection orhtogonale de V sur le plan (x, y)