Chapter 1

Introduction

1.1 L'architecture ResNet

Avant l'introduction de ResNet en 2015 par He et al, l'architecture GoogLeNet était le dernier gagnant des challenges de vision par ordinateur. Cette architecture venait contrer les problèmes d'apprentissage lié à une augmentation de la profondeur de VGG, une autre architecture proche de ResNet.

En effet, un réseau plus profond permet apporte dans certaine condition de meilleurs performance, mais aussi des problèmes comme l'explosion ou l'évanouissement du gradient de la loss. Lors de la backpropagation, les grandes ou petites valeurs peuvent s'amplifier à chaque couche du réseau donnant un gradient beaucoup plus grand/petit à la dernière couche du réseau en comparaison de la première. C'est un effet multiplicatif et donc en lien avec la profondeur du réseau.

Pour un réseau d'une profondeur L, on modélise ces états cachés de dimension d par une séquence $(h_k)_{0\leq k\leq L-1}$ avec $h_k \in \mathbb{R}^d, \forall 0 \leq k \leq L$. On peut décrire mathématiquement l'explosion du gradient tel que, avec une forte probabilité, $\left\| \frac{\partial \mathcal{L}}{\partial h_0} \right\| \gg \left\| \frac{\partial \mathcal{L}}{\partial h_L} \right\|$ où $\mathcal{L}$ représente la loss et $\|\cdot\|$ la norme euclidienne.

La solution apporté par GoogLeNet n'améliorait pas tant les performances comparé à VGG et restait assez complexe. Sa profondeur était comparable à celle de VGG, passant de 22 à 16 couches. En 2015, ResNet arriva avec un modèle allant jusqu'à 152 couches, divisant par deux le nombre d'erreur de GoogLeNet. Son inovation est la présence de *skip connections* entre les couches successives, permetant un meilleurs passage du gradient au sein du réseau. Mathématiquement, on peut alors écrire la relation récurente pour la suite $(h_k)_{0\leq k\leq L-1}$ tel que

$$h_{k+1} = h_k + f(h_k, \theta_{k+1}).$$

avec $f(\cdot, \theta_{k+1})$ représente plusieurs transformation faites par la couche k en question et paramétrisé par $\theta_{k+1} \in \mathbb{R}^p$.

ResNets est devenus la base de nombreux modèles d'apprentissage profond de pointe, dépassant le traitement d'images pour s'étendre à divers domaines tels que le traitement du langage naturel et l'apprentissage par renforcement. L'idée des *skip connections* a inspiré de nombreuses autres architectures et est devenue une pratique standard dans la conception des réseaux neuronaux profonds.

Figure 1.1: Illustration du modèle ResNet. Notons la présence de la *skip connection* au sein de chaque bloc.

Malgrès cette technique, ResNet soufre toujours de problème de gradient lors de l'apprentissage. L'approche historique pour éviter cela est de normaliser les état cachés à la sortie de chaque couche (*batch normalization*). Toutefois, cela à un cout computationnel et une forte dépendance par rapport à la taille du *batch*. Une alternative consiste à incorporer un facteur d'échelle α_L devant le terme résiduel, ce qui conduit au modèle suivant :

$$h_{k+1} = h_k + \alpha_L f(h_k, \theta_{k+1}).$$

Le choix de ce facteur α_L est cruciale et dépend naturellement de la profondeur du réseau. Il garantit que la variance du signal ne change pas radicalement lorsqu'il se propage à travers les couche. Mais il n'y a pour l'instant pas de preuve formel ni de preuve mathématique dans le choix de ce facteur de régularisation.

LIEN AVEC LES ODE A INTRODUIRE ICI ?

Dans ce cours, nous étudirons le rôle de valeur de α_L dans les problèmes de gradient à travers deux grand axes d'exploration.

1.

For any $1 \leq k \leq L$

Assumptions 1 For some $s \geq 1$, the entries of $\sqrt{d}V_k$ are symmetric i.i.d., s^2 sub-Gaussian random variable, independent of d and L, with unit variance.

Assumptions 2 For some $C > 0$, independent of d and L, and for any $h \in \mathbb{R}^D$

$$\frac{\|h\|^2}{2} \leq \mathbb{E}[\|g(h, \theta_k)\|^2] \leq \|h\|^2 .$$

$$\mathbb{E}[\|g(h, \theta_k)\|^8] \leq C \|h\|^8 .$$

Proposition 2 [Admited ?] Consider a ResNet (4) such that Assumptions (A1) and (A2) are satisfied. If $L\alpha_L^2 \leq 1$, then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\frac{\|h_L - h_0\|^2}{\|h_0\|^2} \leq \frac{2L\alpha_L^2}{\delta}.$$

Proposition 3 [Admited] Consider a ResNet (4) such that Assumptions (A1) and (A2) are satisfied.

(i) Assume that $d \geq 64$ and $\alpha_L^2 \leq \frac{2}{(\sqrt{C}s^4+4\sqrt{C}+16s^4)d}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\frac{\|h_L - h_0\|^2}{\|h_0\|^2} > \exp\left(\frac{3L\alpha_L^2}{8} - \sqrt{\frac{11L\alpha_L^2}{d\delta}}\right) - 1,$$

provided that

$$2L \exp\left(-\frac{d}{64\alpha_L^2 s^2}\right) \leq \frac{\delta}{11}.$$

(ii) Assume that $\alpha_L^2 \leq \frac{1}{\sqrt{C}(d+128s^4)}$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\frac{\|h_L - h_0\|^2}{\|h_0\|^2} < \exp\left(L\alpha_L^2 + \sqrt{\frac{5L\alpha_L^2}{d\delta}}\right) + 1.$$

Corollaire (4)**.** *Consider a ResNet (4) such that Assumptions (A1) and (A2) are satisfied, and let $\alpha_L = 1/L^\beta$, with $\beta > 0$.*

(i) If $\beta > \frac{1}{2}$, then

$$\frac{\|h_L - h_0\|}{\|h_0\|} \xrightarrow{\mathbb{P}} 0 \text{ as } L \to \infty.$$

(ii) If $\beta < \frac{1}{2}$ and $d \geq 9$, then

$$\frac{\|h_L - h_0\|}{\|h_0\|} \xrightarrow{\mathbb{P}} \infty \text{ as } L \to \infty.$$

(iii) If $\beta = \frac{1}{2}$, $d \geq 64$, $L \geq \left(\frac{1}{2}\sqrt{C}s^4 + 2\sqrt{C} + 8s^4)d + 96\sqrt{C}s^4\right)$, then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\exp\left(\frac{3}{8} - \sqrt{\frac{22}{d\delta}}\right) - 1 < \frac{\|h_L - h_0\|^2}{\|h_0\|^2} < \exp\left(1 + \sqrt{\frac{10}{d\delta}}\right) + 1,$$

provided that

$$2L \exp\left(-\frac{Ld}{64s^2}\right) \leq \frac{\delta}{11}.$$

Proof: Statement (i) is a consequence of Proposition 2. We have $L\alpha_L^2 = \frac{L}{L^{2\beta}} = L^{1-2\beta}$, as $\beta > 1/2 \Leftrightarrow 1 - 2\beta < 0$ we have $L^{1-2\beta} = \frac{1}{L^{2\beta-1}} \underset{L\to+\infty}{\longrightarrow} 0$. Thus

$$\frac{\|h_L - h_0\|^2}{\|h_0\|^2} \leq \frac{2L\alpha_L^2}{\delta}. \underset{L\to+\infty}{\xrightarrow{\mathbb{P}}} 0$$

Statement (ii) is a consequence of Proposition 3.

□