Projet d'appentissage statistique
Facteur de régulation des ResNets profonds

SHEN Pingya, VIN Charles

December 3, 2023

Chapter 1

Introduction

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 avait été développée pour pallier les problèmes d'apprentissage liés à l'augmentation de la profondeur de VGG, une autre architecture proéminente.

Un réseau plus profond peut offrir de meilleures performances dans certaines conditions, mais il est aussi sujet à des problèmes tels que l'explosion ou l'évanouissement du gradient de la loss. Durant la rétropropagation, les grandes ou petites valeurs de gradient peuvent s'amplifier à travers les couches du réseau, entraînant un gradient bien plus grand ou plus petit dans les dernières couches par rapport aux premières. Cet effet est multiplicatif et dépend donc de la profondeur du réseau.

Pour un réseau d'une profondeur L, on modélise ces états cachés de dimension d par une séquence $(h_k)_{1\leq k\leq L}$ avec $h_k \in \mathbb{R}^d, \forall 0 \leq k \leq L$. L'explosion du gradient peut être décrite mathématiquement par, 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.

GoogLeNet, bien qu'offrant des une légère amélioration des performance par rapport à VGG, était encore relativement complexe et sa profondeur comparable à celle de VGG, passant de 22 à 16 couches. En 2015, ResNet a introduit un modèle allant jusqu'à 152 couches, divisant par deux le nombre d'erreurs de GoogLeNet. Son innovation réside dans l'intégration de *skip connections* entre les couches successives, facilitant le passage du gradient au sein du réseau. Mathématiquement, cela donne la relation récurrente suivante pour la séquence $(h_k)_{1\leq k\leq L}$:

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

où $f(\cdot, \theta_{k+1})$ représente les transformations effectuées par la couche k et paramétrées par $\theta_{k+1} \in \mathbb{R}^p$.

Les ResNets sont devenus la base de nombreux modèles d'apprentissage profond de pointe, s'étendant au-delà du traitement d'images pour inclure des 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 courante dans la conception des réseaux neuronaux profonds.

Figure 1.1: Illustration du modèle ResNet avec la présence de *skip connections* dans chaque bloc.

Malgré ces avancées, ResNet rencontre toujours des problèmes de gradient durant l'apprentissage. La méthode traditionnelle pour contrer cela est la normalisation des états cachés après chaque couche (*batch normalization*). Cependant, cette approche a un coût computationnel et dépend fortement de la taille du *batch*. Une alternative est d'incorporer un facteur d'échelle α_L devant le terme résiduel, conduisant au modèle suivant :

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

Le choix de α_L est crucial et dépend naturellement de la profondeur L du réseau. Il assure que la variance du signal reste stable lors de sa propagation à travers les couches. Cependant, il n'existe actuellement ni preuve formelle ni justification mathématique solide pour le choix de ce facteur de régularisation.

Dans ce cours, nous examinerons les fondements mathématiques pour choisir la valeur de α_L en fonction de L et de la distribution initiale des poids, dans le but d'éviter les problèmes d'apprentissage. Deux axes principaux d'étude seront abordés :

1. Le facteur α_L à l'initialisation : L'initialisation des paramètres est cruciale pour la phase d'apprentissage d'un modèle et influe même sur ses capacités de généralisation. Une mauvaise initialisation peut entraîner une divergence ou une disparition rapide du gradient, voire un blocage dans l'apprentissage. L'étude du rôle de α_L lors de l'initialisation est donc pertinente. Nous considérerons que, à

l'initialisation, les poids de chaque couche $(\theta_k)_{1\leq k\leq L}$ sont choisis de manière indépendante et identique selon une loi, typiquement gaussienne ou uniforme sur $\mathbb{R}^p$.

2. L'approche continue :

Le facteur α_L à l'initialisation

Dans cette section, notre objectif est d'examiner comment le facteur de mise à l'échelle α_L affecte la stabilité des ResNets lors de leur initialisation, en supposant que les poids sont des variables aléatoires indépendantes et identiquement distribuées (i.i.d.). Nous analyserons la modélisation, l'initialisation des paramètres et les hypothèses nécessaires à cette démarche.

1.1 Modèle et hypothèses

1.1.1 Modèle

Le modèle est basé sur un ensemble de données composé de n paires $(x_i, y_i)_{1\leq i\leq n}$ avec $x_i \in \mathbb{R}^{n_{\text{in}}}$ comme vecteur d'entrée et $y_i \in \mathbb{R}^{n_{\text{out}}}$ comme vecteur de sortie à prédire (soit en valeurs continues soit en format *one-hot*). Soit $F_\pi(x) \in \mathbb{R}^{n_{\text{out}}}, x \in \mathbb{R}^{n_{\text{in}}}$ la sortie du ResNet définie par

$$\begin{aligned} h_0 &= Ax, \\ h_{k+1} &= h_k + \alpha_L V_{k+1} g(h_k, \theta_k), \quad 0 \leq k \leq L-1, \\ F_\pi(x) &= Bh_L, \end{aligned}$$

où $\pi = (A, B, (\theta_k)_{k\leq L}, (V_k)_{1\leq k\leq L})$ sont les paramètres du modèle avec $A \in \mathbb{R}^{d\times n_{\text{in}}}, B \in \mathbb{R}^{n_{\text{out}}\times d}, \theta_k \in \mathbb{R}^p$ et $V_k \in \mathbb{R}^{d\times d}$ pour $k = 1, \ldots, L$. La fonction $g : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^d$ représente le choix de l'architecture d'un bloc du ResNet. Nous nous intéressons principalement à la suite des états cachés $(h_k)_{0\leq k\leq L}$ et non aux changements de dimension permis par les matrices A et B. Finalement, on définit $l : \mathbb{R}^{n_{\text{out}}} \times \mathbb{R}^{n_{\text{out}}} \to \mathbb{R}_+$ comme la fonction de *loss*, différentiable par rapport à son premier paramètre. Cette *loss* peut être une perte quadratique ou une entropie croisée. L'objectif de l'apprentissage est de trouver le paramètre optimal π qui minimise le risque empirique $\mathcal{L}(\pi) = \sum_{i=1}^n l(F_\pi(x_i), y_i)$ à travers une descente de gradient stochastique ou l'une de ses variantes.

1.1.2 Initialisation des paramètres

Nous rappelons que $\theta_k \in \mathbb{R}^p$ et $V_k \in \mathbb{R}^{d\times d}$ sont les paramètres des couches cachées de notre modèle pour tout $k \in [\![1, L]\!]$. Ces paramètres sont choisis à l'initialisation comme la réalisation de variables aléatoires i.i.d., généralement suivant une distribution uniforme ou gaussienne. Cette initialisation est indépendante de L et donc du modèle représenté par g, permettant de considérer plusieurs architectures différentes dans notre étude. Nous examinerons également d'autres approches dépendantes du modèle pour étudier le choix de α_L (par exemple, Yang et Schoenholz, 2017 ou Wang et al., 2022).

1.1.3 Hypothèses

Pour notre première hypothèse, nous avons besoin de la définition suivante :

Définition 1 (Variable aléatoire s^2 sub-gaussienne) *En théorie des probabilités, une distribution s^2 sub-gaussienne est une distribution de probabilité caractérisée par une décroissance rapide des queues de distribution. Bien qu'il existe de nombreuses définitions et propriétés, nous retiendrons dans ce cours la suivante : soit X une variable aléatoire réelle,*

$$\forall \lambda \in \mathbb{R}, \mathbb{E}[\exp(\lambda X)] \leq \exp\left(\frac{\lambda^2 s^2}{2}\right).$$

De manière informelle, les queues d'une distribution sub-gaussienne sont dominées par celles d'une distribution gaussienne, c'est-à-dire qu'elles décroissent au moins aussi rapidement.

Pour tout $1 \leq k \leq L$

Hypothèse 1 *Pour un certain $s \geq 1$, les entrées de $\sqrt{d}V_k$ sont des variables aléatoires symétriques i.i.d., s^2 sub-gaussiennes, indépendantes de d et L et de variance unitaire.*

Hypothèse 2 *Pour un certain $C > 0$, indépendant de d et L, et pour tout $h \in \mathbb{R}^D$*

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

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

A TERMINER UN PEU MIEUX

1.2 Limite probabilistique de la norme des états cachés

Proposition 1 (Admise ?) *Considérons un ResNet (4) tel que les hypothèses 1 et 2 soient satisfaites. Si $L\alpha_L^2 \leq 1$, alors, pour tout $\delta \in (0, 1)$, avec une probabilité d'au moins $1 - \delta$,*

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

Proposition 2 (Admise) *Considérons un ResNet (4) tel que les hypothèses 1 et 2 soient satisfaites.*

(i) Supposons que $d \geq 64$ et $\alpha_L^2 \leq \frac{2}{(\sqrt{C}s^4+4\sqrt{C}+16s^4)d}$. Alors, pour tout $\delta \in (0, 1)$, avec une probabilité d'au moins $1 - \delta$,

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

à condition que

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

(ii) Supposons que $\alpha_L^2 \leq \frac{1}{\sqrt{C}(d+128s^4)}$. Alors, pour tout $\delta \in (0, 1)$, avec une probabilité d'au moins $1-\delta$,

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

Corollaire 1 *Considérons un ResNet (4) tel que les hypothèses 1 et 2 soient satisfaites, et soit $\alpha_L = 1/L^\beta$, avec $\beta > 0$.*

(i) Si $\beta > \frac{1}{2}$, alors

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

(ii) Si $\beta < \frac{1}{2}$ et $d \geq 9$, alors

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

(iii) Si $\beta = \frac{1}{2}$, $d \geq 64$, $L \geq \left(\frac{1}{2}\sqrt{C}s^4 + 2\sqrt{C} + 8s^4)d + 96\sqrt{C}s^4\right)$, alors, pour tout $\delta \in (0, 1)$, avec une probabilité d'au moins $1 - \delta$,

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

à condition que

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

Preuve 1 (Preuve :) *L'affirmation (i) est une conséquence de la Proposition 1. Nous avons $L\alpha_L^2 = \frac{L}{L^{2\beta}} = L^{1-2\beta}$, comme $\beta > 1/2 \Leftrightarrow 1 - 2\beta < 0$ nous avons $L^{1-2\beta} = \frac{1}{L^{2\beta-1}} \underset{L\to+\infty}{\longrightarrow} 0$. Ainsi*

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

L'affirmation (ii) est une conséquence de la Proposition 2.