Cours

Charles Vin

Date

1 Notation

- L number of layer

2 Neural ODE

Comme je connais vraiment rien en équa dif je vais sum up une vidéo.

- Neural ODE : mieux que les RNN pour prédire les time series
- Considère une NN non plus de manière discrète avec des layers/block de neurone mais d'une manière continue
- Proche des Equa dif, et on peut utiliser la théorie de ça pour train à la place de la descente de gradiant
- Pas appris beaucoup plus

3 Introduction

- Problème pour train les resnet en fonction de la profondeur: vanishing / exploding gradients.
- Solution classique : batch norm $\rightarrow$ fix la variance du signal $\rightarrow$ mais apporte d'autre problème
- Solution 2 : scaling factor dépendant de L mais comment ?? == objectif du papier
- 3 chapitres
  - Scaling at initialization :
    * Initialisation importante : avoid grad problems, larger learning rate (better generalization)
    * Exploding gradients **during backprop**== $\left\| \frac{\partial \mathcal{L}}{\partial h_0} \right\| \geq\geq \left\| \frac{\partial \mathcal{L}}{\partial h_L} \right\|$ with an high probability
    * $\rightarrow$ they study distibution and the choice of $\alpha_L = \frac{1}{\sqrt{L}}$
  - The continuous approach :
    * Si on pose $\alpha_L = 1/L \rightarrow$ ResNet= ODE
    * Mais contradictoire avec le résultat de la section précédente $\alpha_L = \frac{1}{\sqrt{L}}$
    * En faite $\alpha_L = \frac{1}{\sqrt{L}}$ correspond au bon choix pour neural stochastic pdifferential equation (SDE). Qui correspond à un ResNet avec une initialisation particulière
  - Section 4 : test de differente valeur de α_L dans le cadre SDE
- Related work :
  - Plein de papier sur α_L, plein de solution possible
  - Nous on analise α_L au moment de l'initialisation des paramètres
  - D'autre gens on trouvé $\alpha_L = \frac{1}{\sqrt{L}}$ mais sans donner trop de math et sans fouiller les autres cas $\alpha_L \ll \frac{1}{\sqrt{L}}$, $\alpha_L \approx \frac{1}{\sqrt{L}}$, $\alpha_L \gg \frac{1}{\sqrt{L}}$ et sans faire le lien avec les équa dif
  - Des gens on déjà fait le lien avec les equa diff mais dans des cas moins général je crois

4 Scaling at initialization

4.1 Model and assumptions

4.1.1 Probabilistic setting at initialization

- Les paramètre du modèle est une collection iid de variable aléatoire $\rightarrow$ donc les états cachés $h_0, \ldots, h_L$ aussi (mais il sont martingale eux mais osef pour l'instant)
- La distribution initial des paramèter n'est pas dépendante de L, donc indépendante de l'architecture du modèle considéré. Pratique !
-

4.1.2 Assumptions

- s^2 sub-Gaussian : $\forall \lambda \in \mathbb{R}, \mathbb{E}(\lambda X) \leq \exp(\frac{\lambda^2 s^2}{2})$ a sub-Gaussian distribution is a probability distribution with strong tail decay.

Proposition 4.1. *les resnet du tableau vérifie A_1 et A_2*

4.2 Probabilistic bounds on the norm of the hidden states

Part3 du corrolaire : Coeur du sujet, on vas regarder le comportement de $\|h_L - h_0\| \,/\, \|h_0\|$ en fonction de $L\alpha_L^2$. Seul $\beta = 1/2$ donne une distribution non dégénéré à l'initialisation

- $L\alpha_L^2 \ll 1$ identity function
- $L\alpha_L^2 \gg 1$ explosion du gradient avec forte proba
- $L\alpha_L^2 \approx 1$, h_L fluctue autour de h_0 avec borne

Illustration en figure 1.
Figure 2