Mise en Cohérence des Objectifs du TIPE

Maximilien de Dinechin

Décembre 2017

1. Positionnements thématiques et mots-clés

Thèmes

- Informatique pratique
- Automatique

Mots clés

- Apprentissage automatique (*Machine learning*)
- Réseaux de neurones (*Neural networks*)
- Vision par ordinateur (*Computer vision*)
- Attaques adversaires (*Adversarial examples*)

2. Bibliographie commentée (650 mots)

Introduction

Introduits dès les années 50, les réseaux de neurones ont connu un déclin des années 80 aux années 2010, causé par leur énorme coût calculatoire, pour connaître ensuite un regain de popularité spectaculaire à partir de 2012, où ils se sont imposés comme leaders en classification d'images puis reconaissance de parole. Le livre *Deep Learning* [1] se veut une synthèse du domaine.

I. L'entraînement d'un réseau de neurones

À l'image (simplifiée) du cerveau humain, un réseau de neurones est un assemblage de neurones formels reliées entre eux par des connexions pondérées. Chaque neurone réalise une opération simple : il calcule la somme de ses entrées, pondérées par les poids des connexion, lui applique ensuite une fonction de transfert, et transmet l'information aux neurones suivants auxquels il est rélié.

Un réseau de neurones s'entraîne à partir d'une base de données étiquetée, c'est à dire pour laquelle on connaît déjà le résultat attendu. L'entraînement d'un réseau de neurones consiste alors à trouver les poids optimaux qui minimisent l'erreur commise par le réseau : On la diminue par des descente de gradient successive sur des petits lots d'images (*mini-batches*) choisis au hasard. Cette méthode s'appelle la descente stochastique de gradient (*SGD*), démontrée et explicité en détail dans le livre *Deep Learning and Neural Networks* [2].

Plusieurs amélioration de cette méthode existent, la plus utilisée aujourd'hui étant *Adam* introduite en 2014 [3]. Enfin, la normalisation des sorties des neurones sur chaque mini-batch (*Batch Normalization*) [4] s'est révélée une technique efficace pour entraîner bien plus vite un réseau (c'est à dire en moins d'étapes).

Tout l'enjeu des réseaux de neurones consiste à généraliser les résultats appris à de nouvelles entrées. On cherche donc cherche à éviter la sur-adaptation (*overfitting*), phénomène où un réseau apprend trop les

spécifités des images de sa base de données d'entraînement, au détriment de la généralisation de ses résultats à de nouvelles entrées. Pour pallier celà, le *Dropout* [5] est largement utilisé.

II. La classification d'images

Chaque année depuis 2010 est organisé le concours *ILSVRC*, qui consiste à concevoir un algorithme qui classifie correctement une image parmi 1000 catégories possibles. La performance est mesurée en *erreur Top 5*, qui correspond au pourcentage d'échec de l'algorithme à proposer la bonne étiquette parmi ses 5 prédictions possibles.

Les réseaux de neurones n'y avaient jamais été efficaces, jusqu'en 2012, où le réseau *AlexNet* [6], premier réseau à utiliser le Dropout, participe au concours et pulvérise la concurrence, avec une erreur Top 5 de 15.3% (par comparaison le deuxième meilleur était à 26%).

Ce coup de tonerre provoque un bouleversement du domaine de la vision par ordinateur, qui adopte ces techniques très vite. Quatre ans plus tard, le réseau *ResNet* de Microsoft atteint 3% d'erreur sur ce concours, et ce résultat sera encore probablement amélioré dans le futur, par exemple grâce récente innovation majeure, les *Capsule Networks* [7].

III. Les attaques adversaires

Cependant, en 2014, une équipe de chercheurs remarque une "propriété intriguante des réseaux de neurones" [8] (puis [9] l'année suivante) : les associations entrées-sorties apprises par les réseaux sont fortement discontinues au niveau de l'espace des données, permettant de trouver des entrées très proches mais retournant deux résultats différents.

Pour mettre en évidence ce phénomène sur un réseau classificateur, ils modifient de manière imperceptible une image, et obtiennent une classification erronée avec une assurance élevée : une telle image appelée un *exemple adversaire*.

La faiblesse des réseaux de neurones face aux attaques par exemples adversaires devient alors un domaine important de la recherche : dans une utilisation concrète, une telle faiblesse est potentiellement dangereuse, par exemple dans le cas de la conduite autonome.

Il n'y a toujours pas aujourd'hui de solution satisfaisante à ce problème : régulièrement paraissent des publications qui proposent une solution, contredites peu de temps après.

3. Problématique retenue (50 mots)

L'efficacité exceptionnelle des réseaux de neurones étend leur utilisation à de nombreux domaines, en particulier celui de la conduite autonome. Pourtant, certains travaux remettent en cause leur fiabilité : des attaques ciblées malveillantes les trompent avec succès. En quoi (*technique à insérer*) permet-elle de se prémunir contre ces attaques ?

4. Objectifs du travail

5. Liste des références bibliographiques

[N°1] I. Goodfellow, Y. Bengio & A. Courville. **Deep Learning** Chapitres 6 à 9, The MIT Press, 2016

[N°2] Michael Nielsen. **Neural Networks and Deep Learning.** Determination Press, 2015

[N°3] D. Kingma & J. Ba. **Adam: A method for stochastic optimization.** ICLR (2014)

[N°4] S. Ioffe & C. Szegedy. **Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.** (2015)

[N°5] G. Hinton, N. Srivastava, A. Krizhevsky & al. **Improving neural networks by preventing co-adaptation of feature detectors.** (2012)

[N°6] A. Krizhevsky, I. Sutskever & G. Hinton. NIPS'12 Proceedings, **ImageNet Classification with Deep Convolutional Neural Networks.** Volume 1 (2012), Pages 1097-1105

[N°7] S. Sabour, N. Frosst & G. Hinton NIPS'17 Proceedings, **Dynamic Routing Between Capsules.** Volume ? (2017), Pages ?-?

[N°8] C. Szegedy, I. Goodfellow & al. ICLR 2014, **Intriguing Properties of Neural Networks.**

[N°9] I. Goodfellow & al. **Explaining and Harnessing Adversarial Examples.** (2015)