

Apports des alternatives à la rétropropagation dans l'apprentissage des réseaux de neurones binaires

Ben Crulis*, Barthelemy Serres*, Cyril de Runz*, Gilles Venturini*

*LIFAT, Université de Tours, Tours
<https://lifat.univ-tours.fr/>, nom.prenom@univ-tours.fr
**CETU ILIAD3, Université de Tours, Tours,
<https://iliad3.univ-tours.fr>

Résumé. Les modèles de réseaux de neurones profonds utilisent actuellement des paramètres encodés avec des nombres flottants utilisant beaucoup d'espace mémoire au moment de l'inférence. Ces modèles devenant de plus en plus volumineux, il devient difficile d'envisager entraîner directement ces modèles sur des appareils portables tels que les smartphones. Les réseaux de neurones binaires promettent de réduire la taille des réseaux de neurones artificiels tout en diminuant le temps d'inférence et l'énergie consommée, permettant le déploiement de modèles plus puissants sur les appareils portables. Cependant, les réseaux de neurones binaires sont encore actuellement difficiles à entraîner en utilisant la rétropropagation classique. Nous fournissons des tests comparatifs pour 3 algorithmes dont la rétropropagation permettant d'entraîner des réseaux de neurones binaires sur MNIST ainsi que CIFAR-10. Les résultats montrent que les réseaux de neurones binaires peuvent être entraînés en utilisant des alternatives à la rétropropagation, et donner lieu à de meilleures performances.

1 Introduction

Les réseaux de neurones artificiels sont connus pour leur bonnes performances sur de nombreux types de tâches, mais au prix de coûteuses phases d'apprentissage préalables. De nos jours, la recherche se concentre sur le passage à l'échelle pour de plus gros modèles et jeux de données puisque l'ajout de paramètres supplémentaires a tendance à améliorer les performances. Cette tendance rend les modèles plus encombrant et lent à l'utilisation. Leur utilisation est, de ce fait, plus difficile sur des appareils contraints en mémoire et en capacité de calcul tels que les smartphones. Pourtant, déployer les modèles directement sur les smartphones promet plusieurs avantages tel que réduire les communications réseaux et tout en réduisant la dépendance à d'éventuels serveurs de calcul distant. La perspective d'avoir des algorithmes d'apprentissage moins coûteux et des modèles moins volumineux ouvre la possibilité d'effectuer l'entraînement ou le ré-entraînement des modèles directement sur les appareils des utilisateurs, ce qui amène de nouvelles possibilités de personnalisation et d'adaptation à l'utilisateur.

Les réseaux de neurones binaires ont été proposés pour rendre les modèles de réseaux de neurones plus économes en mémoire tout en améliorant leur vitesse d'inférence et leur coût