Understanding Deep Learning Requires Rethinking Generalization
Présentation et Analyse

Mariem BOUHAHA Eden BELOUADAH

18 janvier 2018

1 Présentation de l'article

Il s'agit d'un article publié en Novembre 2016 dans arXiv, ayant pour auteurs Chiyuan Zhang (MIT), Samy Bengio (Google Brain), Moritz Hardt(Google Brain), Benjamin Recht (UC) et Oriol Vinyals (DeepMind). Dans leur article, les auteurs répondent à la question suivante :

Par quoi exliquer la bonne généralisation des Réseaux de Neurones en pratique ?

Selon les auteurs, répondre à cette question va permettre de rendre les réseaux de neurones plus interprétables et aboutir à un design plus fiable de l'architecture du modèle.

2 Principales Contributions

A travers plusieurs expérimentations, les auteurs montraient comment les approches traditionnelles échouent dans l'explication du phénomène que les réseaux de neurones généralisent bien en pratique. Par "généraliser bien" les auteurs voulaient simplement dire "Qu'est-ce qui cause un réseau de neurones qui performe bien sur les données d'apprentissage à performer aussi bien sur les données de test ? » En expérimentant sur les data sets CIFAR 10 (50.000 images d'apprentissage, 10.000 images de validation, 10 classes) et ImageNet (1.281.167 000 images d'apprentissage, 50,000 images de validation, 1000 classes) et en se basant sur l'architecture Inception V3 et une de ses versions (AlexNet) ainsi qu'en utilisant des MLPs, les auteurs trouvaient que :

- les réseaux de neurones apprennent facilement des labels aléatoires.
- La régularisation "explicite" peut améliorer la généralisation du réseau de neurones, mais elle n'est ni nécessaire ni suffisante pour contrôler l'erreur de généralisation.
- Ils complétaient leur approche expérimentale par une construction théorique montrant qu'il existe un réseau de neurones à 2 couches avec des activations ReLU et 2n+d poids qui peut représenter n'importe quelle fonction sur un échantillon de données de taille n et de dimension d.

Et en faisant appel aux modèles linéaires, les auteurs analysaient comment la Descente

de Gradient Stochastique (SGD) agit comme une régularisation « implicite » et suggéraient qu'une investigation plus approfondie devrait être faite pour comprendre les propriétés des modèles qui utilisent la SGD dans leur apprentissage.

3 Analyse

Selon la Théorie d'Apprentissage Statistique, un modèle généralise bien s'il n'a pas la capacité d'overfitter des données aléatoires de même taille que les données d'apprentissage réelles. Dans cet article, les auteurs discutaient de la capacité des réseaux de neurones et des méthodes de régularisation les plus utilisées en Machine Learning et trouvaient que la théorie classique d'apprentissage statistique ainsi que les stratégies de régularisation ne peuvent pas expliquer la remarquable capacité des réseaux de neurones à bien se généraliser.

Cet article présente un ensemble d'expérimentations, dont **la randomisation des labels** et **l'ajout de bruit**, qui ont mis en relief le pouvoir énorme des réseaux de neurones larges. Encore, ces mêmes modèles arrivent à généraliser même quand on supprime toute forme de **régularisation explicite** ou **implicite**. Ces observations sont utilisées pour montrer l'incapacité de la théorie classique (dimension de VC, complexité de Rademacher, stabilité uniforme) à expliquer le pouvoir de généralisation.

Comme l'ont mentionné les auteurs, dès qu'une famille de modèles arrive à mémoriser toutes les données d'apprentissage, la théorie classique ne permet pas de percevoir le comportement de ces modèles en termes de généralisation, ce qui ne laisse que le choix d'expérimentation et d'étude empirique.

Bien que ce travail ne propose pas assez d'explications pour ces capacités de généralisation, il oblige le lecteur à penser au problème de généralisation d'un nouvel angle, différemment de la façon traditionnelle dont on le comprend.

4 Critique

D'abord, l'article montre qu'une bonne généralisation n'est pas souhaitée pour un problème à labels aléatoires. Ceci est certainement correct, mais nous pensons que c'est inintéressant en pratique, du moment où on n'a pas intérêt à travailler avec des labels aléatoires, pour une finalité de classification par exemple. Ensuite, bien que l'article incite à la réflexion, à la lecture et à la discussion, on n'a cependant pas très bien clarifié de quelles "approches traditionnelles" parlait-on et pour quoi est-il si surprenant qu'elles échouent à expliquer la performance des réseaux de neurones. Les auteurs expriment leur surprise à propos de certaines capacités des réseaux de neurones, mais ils ne donnaient pas de références ni d'arguments qui expliquent pourquoi devrait-on être surpris. On s'est surpris aussi du fait que la SGD arrive à optimiser les poids pour fitter les labels aléatoires sans aucun changement d'hyperparamètres. On sait tous que les réseaux de neurones larges ont suffisamment de paramètres pour mémoriser les labels aléatoires, donc ce n'est pas si évident que ce soit surprenant. Enfin, il aurait été mieux de présenter des explications alternatives au phénomène de généralisation, au lieu de seulement remettre en question les approches

traditionnelles.

5 Conclusion

Les expérimentations présentées dans cet article mettent l'accent sur l'importance de comprendre la capacité effective des modèles de réseaux de neurones, qui sont capable de mémoriser facilement tout l'ensemble d'apprentissage, même lorsqu'on casse le lien entre les données et les labels. Une optimisation empirique facile n'est à priori pas la vraie raison de généralisation. C'est pour cela qu'il est nécessaire de comprendre les vraies notions qui se cachent derrière la généralisation pour pouvoir bien comprendre l'apprentissage profond.

Références Zhang, Bengio, Hardt, Recht, Vinyals, "Understanding Deep Learning Requires Rethinking Generalization", arXiv, 2016.