# Présentation de l’article :

Il s’agit d’un article uploadé en 11/2016 dans arXiv, ayant pour auteurs Zhang, Chiyuan; Bengio, Samy; Hardt, Moritz; Recht, Benjamin et Vinyals, Oriol. Dans leur article, les auteurs répondent à la question suivante:

*« Qu’est-ce qui distingue les Réseaux de Neurones qui généralisent bien de ceux qui ne le font pas ?*»

Selon eux, 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.

# 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 images, 1000 classes) et en se basant sur l’architecture *Inception* et certaines de ses versions, 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 que :

* **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.

# 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 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 modèle 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.

# Critique :

* 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.
* 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. Quelques exemples sont :

« *To our surprise, several properties of the training process for multiple standard architectures is largely unaffected by this transformation of the labels. This poses a conceptual challenge. Whatever justification we had for expecting a small generalization error to begin with must no longer apply to the case of random labels.”* De quel challenge conceptuel parlait-on?

« *Surprisingly, stochastic gradient descent with unchanged hyperparameter settings can optimize the weights to fit to random labels perfectly, even though the random labels completely destroy the relationship between images and labels.*" Pourquoi est-ce surprenant? 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.