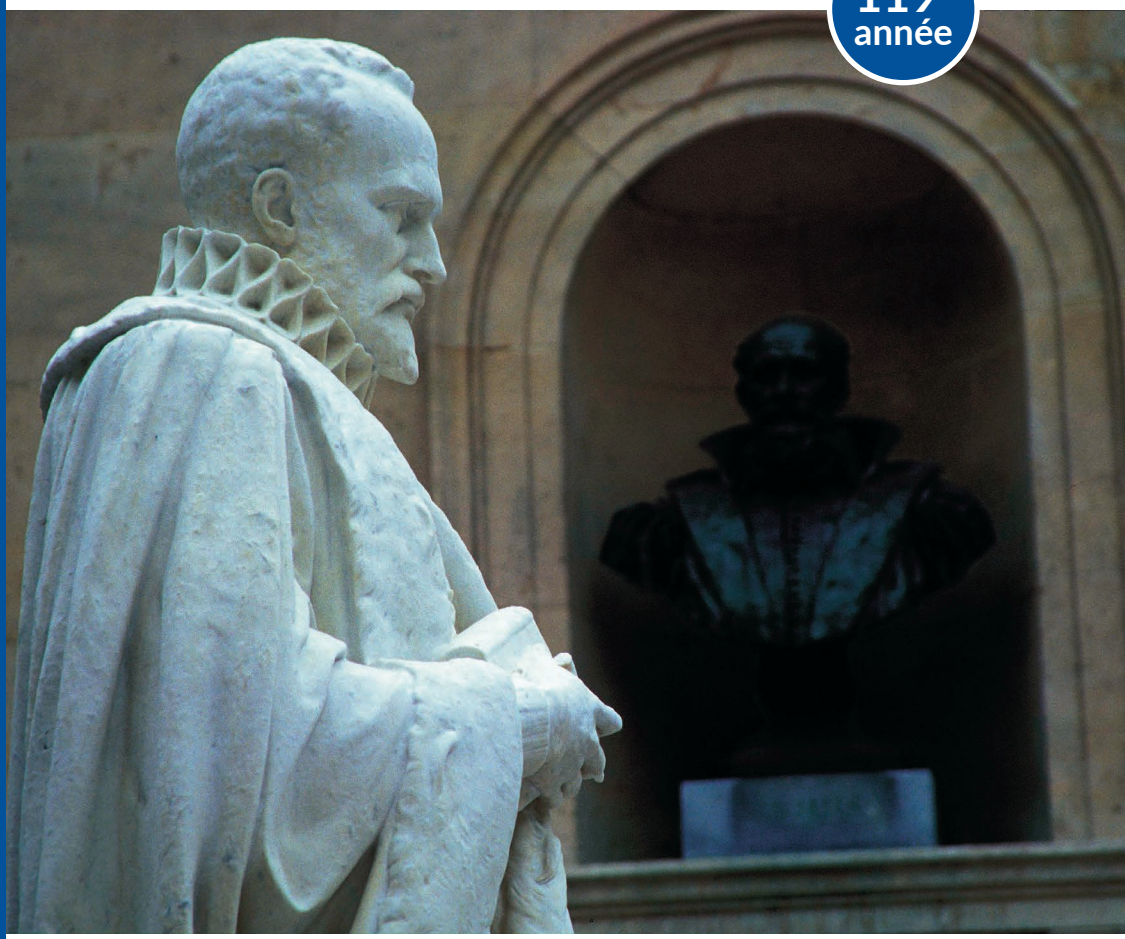
ANNUAIRE du **COLLÈGE DE FRANCE** 2018 - 2019

Résumé des cours et travaux

119e année



COLLÈGE DE FRANCE
1530

Sciences des données

Stéphane Mallat

Membre de l'Institut (Académie des sciences,
Académie des technologies),
professeur au Collège de France

Mots-clés : apprentissage, données, réseaux de neurones

La série de cours et séminaires « L'apprentissage par réseaux de neurones profonds » est disponible, en audio et/ou vidéo, sur le site internet du Collège de France (https://www.college-de-france.fr/site/stephane-mallat/course-2018-2019.htm).

Enseignement

Cours et séminaires – Apprentissage par réseaux de neurones profonds

Introduction

Les réseaux de neurones profonds ont des applications spectaculaires dans des domaines très divers dont la vision par ordinateur, la compréhension de la parole, l'analyse de langages naturels, mais aussi pour la robotique, la prédiction de phénomènes physiques divers, le diagnostic médical ou des jeux de stratégie comme le Go. Ce premier cours sur les réseaux de neurones présente leurs applications, les architectures de ces réseaux, les algorithmes permettant d'optimiser leurs paramètres, et, enfin, les questions mathématiques sur l'optimisation et la capacité de généralisation des réseaux de neurones. Nous verrons que les théorèmes connus ne répondent à ces questions que dans des cas simplifiés qui sont souvent loin des conditions d'applications de ces réseaux. La compréhension mathématique des réseaux de neurones profonds reste donc essentiellement un problème ouvert. Outre

S. Mallat, « Sciences des données », *Annuaire du Collège de France 2018-2019. Résumé des cours et travaux*, 119e année, Paris, Collège de France, 2022, p. 41-53, https://doi.org/10.4000/annuaire-cdf.16767.

les challenges de données, les séminaires sont dédiés à des applications spécifiques des réseaux de neurones profonds.

Cours 1 – Introduction aux réseaux de neurones profonds

23 janvier 2019

Le premier cours fait un rappel des principes mathématiques des algorithmes d'apprentissage supervisés et non supervisés, et présente les architectures des réseaux de neurones profonds, ainsi que leurs applications à la reconnaissance d'images. L'apprentissage supervisé consiste à estimer la réponse $y = f(x)$ à une question, à partir d'une donnée x de dimension d. On utilise une base d'exemples d'entraînements où pour des données xi, on connaît la valeur de $yi = f(xi)$. L'apprentissage non supervisé revient à estimer la distribution de probabilité $p(x)$ des données x, à partir d'une famille d'exemples xi qui sont considérés comme des réalisations indépendantes suivant cette distribution. La difficulté principale de ces problèmes vient de la grande dimension d des données x. Les réseaux de neurones sont des architectures de calcul qui incluent un très grand nombre de paramètres afin d'approximer $f(x)$ pour l'apprentissage supervisé ou $p(x)$ pour l'apprentissage non supervisé.

Les réseaux de neurones prennent en entrée la donnée x et calculent une approximation de $y = f(x)$ avec une cascade d'opérateurs linéaires suivis de non-linéarités ponctuelles comme des sigmoïdes ou des rectificateurs. Les réseaux de neurones ont été introduits dans les années 1950 avec une motivation biologique. Cependant, ce n'est qu'à partir des années 2010 que ces réseaux ont obtenu des résultats spectaculaires, grâce à l'augmentation massive des données d'entraînement et à l'augmentation de la vitesse des ordinateurs. Cela a permis d'entraîner des réseaux de grande taille. Des applications impressionnantes ont été faites dans de nombreux domaines dont la vision par ordinateur, la reconnaissance de la parole, l'analyse de sons, de langages naturels, le contrôle de robots, la prédiction de quantités physiques, le diagnostic médical ou pour des compétitions de jeu d'échecs ou de Go. Le fait qu'un même type d'architecture puisse approximer des problèmes aussi différents indique que ces problèmes partagent des formes de régularités que l'on ne comprend pas mathématiquement. Le cours présentera des architectures de réseaux et les algorithmes d'apprentissage, mais essayera aussi d'expliquer la performance de ces algorithmes, ou du moins les questions ouvertes sur ce sujet.

La vision par ordinateur est un domaine important d'applications des réseaux de neurones. Il s'agit ici de reconnaître une scène, ou un objet et sa localisation dans une image ou une vidéo, ou de segmenter l'image en un ensemble de structures identifiées. Jusqu'à récemment, les algorithmes de vision par ordinateurs étaient souvent fondés sur l'extraction de structures comme des contours, des coins ou des éléments de textures, qui étaient agrégés avec des règles. Ces approches ne fonctionnaient, cependant, que sur des images relativement simples. La performance des réseaux de neurones profonds à partir de 2012 fut une grande surprise, car ils ont obtenu des résultats remarquables sur des problèmes que l'on pensait inatteignables avant longtemps. Ces réseaux peuvent maintenant reconnaître des visages mieux qu'un humain, faire de la reconnaissance en temps réel pour guider des voitures, reconnaître des objets ou segmenter des images complexes. Il faut cependant les entraîner sur des très grandes bases de données et ils introduisent parfois des erreurs importantes. Les propriétés de ces algorithmes sont encore mal comprises.

Séminaire 1 – Challenges de données (partie I)

23 janvier 2019

Le site web Challengedata.ens.fr met à disposition des challenges de traitement de données par apprentissage supervisé. Ces challenges sont proposés par des entreprises ou des scientifiques, et sont issus de problématiques concrètes qu'ils rencontrent dans leur activité. Ils s'inscrivent dans un esprit d'échange scientifique, avec un partage de données et d'algorithmes.

Chaque challenge fournit des données labélisées, ainsi que des données de test. Les participants soumettent sur le site web leurs prédictions calculées sur les données de test. Le site calcule un score avec une métrique d'erreur qui est spécifiée. Il fournit un classement aux participants, ce qui permet d'évaluer leurs résultats dans une large communauté. Les challenges commencent le 1er janvier 2019. Une clôture intermédiaire a lieu en juin par une évaluation des prédictions sur de nouvelles données de test. La clôture finale est en décembre, avec une remise des prix en janvier 2020.

Cette année, les challenges ont été organisés et supervisés à l'ENS par Tomas Anglès, Louis Thiry, Roberto Leonarduzzi et John Zarka. L'organisation de ces challenges de données est soutenue par la chaire CFM de l'École normale supérieure, et par la fondation des Sciences mathématiques de Paris. Lors de cette première session, les sept challenges suivant sont présentés :

– « Prédiction des mouvements quotidiens d'actions américaines », présenté par Éric Lebigot de la société Capital Fund Management. L'objectif du challenge est de prédire le signe du rendement d'actions américaines au cours d'une période de 30 minutes en fin de journée, à partir de leur historique de rendements sur des fenêtres de 5 minutes en début de journée ;

– « Détection de métastases du cancer du sein », présenté par Charlie Saillard de la société Owkin. L'objectif du challenge est de déterminer la présence de métastases ganglionnaires lymphatiques au sein d'images de coupes histologiques de patients atteints d'un cancer du sein. Ce challenge est « faiblement supervisé », la présence ou non de métastases étant indiquée localement pour quelques images, mais uniquement globalement à l'échelle de l'image entière pour la plupart des coupes ;

– « Prédiction de profils dynamiques d'électricité », présenté par Pierre Cauchois de la société Enedis. L'objectif du challenge est d'estimer sept séries de profils dynamiques d'électricité représentant la forme de la consommation de catégories de clients différents du marché de masse, à partir de données énergétiques et météorologiques ;

– « Prédiction du ratio de Sharpe de mélanges de stratégies quantitatives », présenté par Stefan Duprey de la société Napoléon Crypto. L'objectif du challenge est de prédire le ratio de Sharpe, un indicateur de rendements ajusté par le risque, de combinaisons déterminées de 7 stratégies quantitatives au cours d'une période de 5 jours ouvrés, à partir de l'historique sur 21 jours ouvrés de rendements de ces 7 stratégies ainsi que de 3 indicateurs financiers pertinents ;

– « Prédiction de l'activité d'ondes lentes du cerveau au cours du sommeil profond », présenté par Valentin Thorey de la société Dreem. L'objectif du challenge est de prédire, à partir de 10 secondes de signaux EEG d'une onde lente du cerveau au cours d'un sommeil profond, ainsi que de divers indicateurs de la qualité du sommeil jusqu'alors, si une autre onde lente de faible ou grande amplitude suivra ;

– « Prédiction de la réponse attendue à des questions de pharmaceutique », présenté par Emmanuel Bilbault de la société Posos. L'objectif du challenge est de catégoriser des questions de pharmaceutique selon le type de réponse attendue ;

– « Résolution du Rubik's cube $2 \times 2 \times 2$ », présenté par Julien Peyras de la société LumenAI. L'objectif du challenge est de prédire, à partir d'une certaine configuration initiale d'un Rubik's cube $2 \times 2 \times 2$, le nombre minimum de mouvements à effectuer afin d'arriver à la solution.

Cours 2 – Applications des réseaux de neurones profonds

30 janvier 2019

Pour comprendre l'impact des réseaux de neurones et les questions que cela pose, ce cours présente des applications très diverses : reconnaissance de la parole, traitement du langage naturel, prédiction de phénomènes physiques, neurophysiologie de la perception, ainsi que la modélisation et la génération de données complexes comme des images ou des sons.

L'analyse de la parole est un des domaines les plus anciens d'analyse de signaux, qui a émergé dès les années 1960. Les algorithmes d'analyse de la parole étaient souvent fondés sur une représentation du signal par spectrogramme suivi d'un modèle de mixture de Gaussiennes optimisé avec une chaîne de Markov. Ces algorithmes ont été considérablement améliorés par les réseaux de neurones profonds, y compris pour séparer des signaux audio mélangés, ce que l'on appelle de la séparation de sources. Une des questions est de comprendre le lien entre ces réseaux et les algorithmes plus classiques de reconnaissance, et d'expliquer l'amélioration importante qu'ils apportent.

Les théories du traitement du langage naturel ont considérablement influencé notre compréhension de la notion même d'intelligence. Ces théories ont beaucoup évolué depuis le structuralisme qui essaye de caractériser l'agrégation de structures élémentaires du langage jusqu'aux théories de grammaires formelles introduites par Chomsky. Le traitement informatique du langage naturel a commencé avec des modèles statistiques fondés sur des chaînes de Markov et des modèles informatiques des grammaires formelles. Cependant, là encore les réseaux de neurones ont apporté des améliorations considérables pour des applications telles que la traduction automatique ou la génération de phrases pour répondre à des questions. Ce sont ces réseaux de neurones qui sont actuellement utilisés dans la plupart des applications commerciales.

L'application des réseaux de neurones à la physique est un sujet qui se développe rapidement. En physique, on essaye de prédire l'évolution d'un système en essayant de résoudre des équations fondamentales. Cela peut cependant devenir très compliqué lorsqu'il y a de nombreux corps en interaction. La prédiction de phénomènes physiques par réseaux de neurones peut s'assimiler à une forme de physique statistique. On prédit un phénomène à partir d'une base de données d'exemples, en essayant de capturer des formes de régularité de cette évolution. Des résultats de plus en plus précis sont obtenus avec des réseaux de neurones, en physique quantique, en dynamique des fluides ou en science des matériaux.

Le lien entre la neurophysiologie de la perception et l'architecture des réseaux de neurones profonds est aussi un sujet de plus en plus exploré. Les modèles de réseaux donnent actuellement de bons modèles de calculs pour expliquer certaines réponses de neurones dans les aires visuelles V1, V2, V4 ainsi que dans les aires corticales

auditives A1 et A2. De nombreuses recherches sont actuellement menées sur cette interface.

Pour des applications robotiques ou pour des jeux de stratégie comme les échecs ou le Go, les réseaux de neurones sont utilisés avec des algorithmes d'apprentissage par renforcement. L'optimisation du réseau est faite en utilisant l'information donnée par une récompense, qui évalue la performance d'une action. C'est ce principe qui a permis de mettre au point des réseaux de neurones capables de battre le champion du monde du jeu de Go, ou d'adapter les trajectoires de robots.

Les réseaux de neurones profonds ont aussi permis de créer de nouveaux modèles stochastiques afin de générer des données. Cela concerne des phénomènes physiques complexes comme de la turbulence en dynamique des fluides ou des textures d'images ou d'audio, aussi bien que des données très structurées comme des images de visages ou des musiques. Ces derniers sont synthétisés avec des paires de réseaux de neurones (GAN ou auto-encodeurs) qui, d'un côté, analysent les propriétés statistiques de ces données et, de l'autre, génèrent des données ayant les mêmes propriétés. Ces réseaux sont optimisés de façon conjointe à partir d'une base de données d'exemples. Des applications surprenantes ont récemment été montrées par exemple pour synthétiser des tableaux suivant des styles de peintres différents.

Séminaire 2 – Challenges de données (partie II)

30 janvier 2019

Lors de cette deuxième session, sept autres challenges du site web Challengedata.ens.fr sont présentés :

– « Prédiction de la concentration spatiotemporelle en particules fines PM10 », présenté par Grégoire Jauvion de la société Plume Labs. L'objectif du challenge est de prédire les taux de particules fines PM10 mesurés par certaines stations de surveillance de la qualité de l'air, à partir des mesures fournies par les stations de surveillance adjacentes ainsi que certaines caractéristiques urbaines ;

– « Prédiction de l'activité cérébrale d'un rat à partir de patterns temporels de potentiels d'action », présenté par Ilya Prokin du Group for Neural Theory (GNT) de l'ENS. L'objectif du challenge est de prédire l'état d'activité cérébrale d'un rat à partir des temps d'occurrence de potentiels d'action au sein d'un certain neurone de l'hippocampe ;

– « Prédiction de consommation électrique pour la tarification de l'électricité fournie », présenté par Alexis Lucido de la société BCM Energy. L'objectif du challenge est de prédire la consommation électrique sur une année de deux clients potentiels, à partir d'un ensemble d'historiques de consommation d'autres clients présentant des caractéristiques similaires mais distants géographiquement, de données géographiques et météorologiques ;

– « Dépistage et diagnostic du cancer de l'œsophage à partir d'images *in vivo* », présenté par Fanny Louvet-de Verchère de la société Mauna Kea Technologies. L'objectif du challenge est de classifier des images endoscopiques de l'œsophage parmi quatre classes représentant différents stades d'avancement d'un cancer de l'œsophage ;

– « Prédiction de la sinistralité d'un immeuble », présenté par Clémence Devries de la société Generali. L'objectif du challenge est de prédire la sinistralité relative aux dégâts des eaux sur une période d'un an d'un immeuble assuré, à partir de ses caractéristiques ;

– « Classification et optimisation de la qualité de vie au travail », présenté par Sylvain Le Corff de la société Oze-Energies. L'objectif du challenge est de prédire le confort ressenti par les occupants d'un bâtiment, à partir de données environnementales mesurées en temps réel par des capteurs au sein du bâtiment ;

– « *Pricing* d'options exotiques par interpolation non linéaire multidimensionnelle », présenté par Olivier Croissant de la société Natixis. L'objectif du challenge est de prédire la valeur d'options exotiques contenues dans des titres de créances, à partir de 23 paramètres les caractérisant.

Cours 3 – Approximations par réseaux de neurones et régularité

6 février 2019

Un réseau de neurones transforme les données x d'entrée par une cascade d'opérateurs linéaires représentés par des matrices de coefficients, suivis de non linéarités ponctuelles comme des sigmoïdes ou des rectificateurs. Cela implémente donc une classe de fonctions qui est paramétrée par les matrices utilisées pour calculer les couches successives. L'apprentissage optimise ces paramètres afin de minimiser l'erreur d'approximation d'une fonction $y = f(x)$. Cette erreur est évaluée sur les exemples d'entraînement. On fait face à deux types de problèmes. Le problème d'approximation consiste à montrer qu'il existe une fonction dans la classe des fonctions des réseaux de neurones, qui approxime précisément $f(x)$. Le second problème est d'optimiser les paramètres du réseau afin de calculer la meilleure approximation qui minimise l'erreur d'approximation. Cette optimisation se fait avec un algorithme de descente de gradient qui ajuste progressivement les paramètres afin de réduire l'erreur à chaque itération. Ce cours se concentre sur le problème d'approximation.

L'erreur d'approximation dépend typiquement de la régularité de la fonction $f(x)$ que l'on approxime. Si cette fonction est Lipchitz, on démontre que pour atteindre une erreur e, il faut un nombre d'exemples qui croit exponentiellement comme e^{-d}. C'est la malédiction de grande dimension d. Pour éviter cette malédiction, il faut que la fonction $f(x)$ soit beaucoup plus régulière, et que le réseau puisse utiliser cette régularité sous-jacente. Un enjeu mathématique est de comprendre la nature de la régularité qui est exploitée par les réseaux de neurones profonds.

En grande dimension, il est nécessaire d'utiliser des contraintes de régularité globale. Cette régularité peut être capturée par le groupe de symétrie de $f(x)$. Une symétrie est un opérateur g qui ne modifie pas la valeur de f : $f(g \cdot x) = f(x)$ pour tout x. L'ensemble des symétries a une structure de groupe. On a souvent des informations *a priori* sur ces symétries. Ainsi, de nombreux problèmes de reconnaissance d'images sont invariants par translation, par certaines rotations ou certaines déformations. Pour le son, ces symétries incluent des transpositions fréquentielles on des déformations dans le plan temps-fréquence.

L'architecture d'un réseau de neurones convolutif incorpore une information sur ces symétries en imposant que les poids du réseau sont invariants par translation. Des expériences numériques montrent que les réseaux de neurones reproduisent d'autres symétries en calculant des coefficients qui sont de plus en plus invariants lorsque la profondeur du réseau augmente. Une question importante est de comprendre le lien entre les coefficients appris par le réseau et les groupes de symétries.

L'existence de séparation d'échelles est une autre source importante de régularité. En physique, l'interaction de d particules peut souvent s'approximer par des

interactions de groupes de particules de tailles variables, ce qui permet de passer de d variables à $O(\log d)$ variables, qui représentent chacun des groupes. Ces propriétés de séparation d'échelles se retrouvent dans la plupart des applications dont la reconnaissance d'images, de sons, l'analyse du langage ou la physique. Dans les cas les plus simples, elles peuvent s'exprimer par des relations hiérarchiques pouvant se représenter par des arbres. Cependant, ces structures sont souvent trop rigides pour expliciter la complexité des interactions à travers les échelles. Les mathématiques permettant de représenter de telles interactions sont fondées sur la transformée en ondelettes. Dans les réseaux de neurones, la séparation d'échelles s'observe dans la structure hiérarchique des calculs à travers les couches.

Une troisième source de régularité est capturée par la notion de parcimonie. Il s'agit de comprendre s'il existe des « prototypes » de formes qui jouent un rôle important dans la valeur prise par $f(x)$. Cela peut se formaliser par une approximation parcimonieuse dans un dictionnaire de vecteurs. On retrouve cette parcimonie à la sortie des neurones d'un réseau, dont les réponses sont souvent nulles. Il s'agit de comprendre si un réseau de neurones a la capacité d'apprendre des dictionnaires, et s'il approxime la fonction $f(x)$ à partir d'une représentation parcimonieuse dans un dictionnaire.

L'enjeu mathématique est de relier ces notions de régularités aux capacités d'optimisation des réseaux de neurones et de comprendre les propriétés des classes de fonctions qui sont bien approximées par un tel réseau.

Séminaire 3 – Présentation des gagnants des Challenges 2018

6 février 2019

Au cours de la première partie, certains gagnants des Challenges 2018 présentent leurs algorithmes ainsi que les résultats obtenus :

– Cyrille Delabre, gagnant France du challenge « Prédiction de la volatilité sur des marchés financiers » proposé par la société Capital Fund Management ;

– Didier Gitton, gagnant du challenge « Classification de stade de sommeil » proposé par la société Dreem ;

– Benoît Schmauch, gagnant du challenge « Prédiction de maladie à partir du génome » proposé par la société Owkin et l'Inserm ;

– Dan Constantini et Tom Hayat, gagnants du challenge « Identification de célébrités » proposé par la société Reminiz.

Une remise des prix a été faite pour les gagnants des autres challenges :

– Étienne Seckinger, gagnant du challenge « Prédiction de la production d'énergie éolienne », proposé par la société Engie ;

– Élie Salem, gagnant des challenges (i) « Prédiction de réclamations lors de transactions *e*-commerce », proposé par la société PriceMinister-Rakuten France, et (ii) « Prédiction de la réponse attendue à des questions de pharmaceutique », proposé par la société Posos ;

– Omar Seck, gagnant du challenge « Prédiction de l'équipe vainqueur d'un match de NBA » proposé par la société LumenAI ;

– Adrien Le Franc, gagnant du challenge « Prédiction des performances énergétiques de bâtiments », proposé par la société Oze-Energies ;

– Jonathan Siaut, gagnant des challenges (i) « Prédiction d'approbation de publications », proposé par la société Dassault Systèmes, et (ii) « Prédiction de la

production électrique horaire par unité de production en France », proposé par la société Wattstrat ;

– Hervé Durand, gagnant du challenge « Prédiction de la saturation d'huile résiduelle », proposé par l'IFPEN.

Cours 4 – Les origines : la cybernétique et le perceptron

13 février 2019

Ce cours revient sur les idées à l'origine des réseaux de neurones, d'abord la théorie de la cybernétique initiée par Wiener, l'importance des structures hiérarchiques, et le perceptron de Rosenblatt. La cybernétique donne une perspective de systèmes dynamiques. L'intelligence est définie comme une capacité d'adaptation dans le temps. Cette adaptation optimise une trajectoire pour atteindre un but. En cybernétique, l'adaptation se fait par une boucle de rétroaction qui adapte les paramètres de contrôle afin de réduire une mesure d'erreur relativement au but à atteindre. Contrairement à un système en boucle ouverte, il n'est pas nécessaire de modéliser l'environnement mais juste de réagir aux perturbations qu'il introduit sur la trajectoire pour atteindre le but. Les algorithmes d'apprentissage par descente de gradient d'un réseau de neurones suivent ce principe. Ils optimisent progressivement les poids du réseau afin de réduire l'erreur de prédiction.

L'article « The architecture of complexity » de H. Simons en 1962 montre que l'existence de structures hiérarchiques est un autre élément qui permet de simplifier l'analyse et le contrôle des systèmes dynamiques. Ces hiérarchies se retrouvent dans la plupart des systèmes en sciences, en sciences humaines et dans les systèmes symboliques. On les retrouve dans l'architecture des réseaux de neurones profonds convolutifs.

Le perceptron de Rosemblatt introduit en 1957 définit un premier algorithme d'apprentissage sur un réseau de neurones. Il a une seule couche et une sortie binaire afin de classifier des données dans deux classes possibles. L'apprentissage se fait par une descente de gradient qui minimise une moyenne des écarts à la frontière de décision. On montre que cette descente de gradient suit la règle de Hebb, observée en biologie. Celle-ci observe que deux neurones qui sont excités simultanément vont renforcer le lien qui les unit. On démontre aussi que l'algorithme de Rosemblatt converge vers une solution qui dépend des conditions initiales si les données d'entraînement sont séparables linéairement, et ne converge pas si elles ne sont pas séparables.

Afin d'éviter ces problèmes de convergence, il faut régulariser la fonction de coût optimisée par le perceptron. Ainsi les « support vector machines » de Vapnik introduisent un critère de marge qui garantit que la frontière sépare au mieux les points de deux classes différentes, ce qui implique l'unicité du point de convergence et élimine la non convergence dans le cas de données non séparables.

Séminaire 4 – Apprentissage faiblement supervisé pour la reconnaissance visuelle

Josef Sivic (Inria), le 13 février 2019

Les succès actuels de la reconnaissance visuelle sont en grande partie dus à l'apprentissage de nouvelles représentations d'images, grâce aux techniques d'apprentissage supervisé et à l'existence de grandes bases de données d'images

annotées. Cette présentation explique que pour élaborer des algorithmes capables de comprendre les évolutions du monde visuel qui nous entoure, la difficulté principale est maintenant de développer des représentations visuelles capables de généraliser dans des environnements différents de ceux qui apparaissent dans la base de données d'entraînement. Il faut aussi qu'ils puissent apprendre avec une supervision faible, avec des données bruitées et annotées partiellement. Plusieurs éléments permettent d'avancer dans cette direction. L'existence de données multimodales qui permettent de recouper des informations visuelles, auditives ou textuelles sans annotation, et l'utilisation de modèles physiques appris sur des données. Cet exposé présente des directions de recherche qui permettent d'aborder ces problèmes avec des applications pour la compréhension du contenu de vidéos ou pour trouver des correspondances visuelles.

Cours 5 – Approximation universelle par un réseau à une couche cachée

20 février 2019

Le théorème d'approximation universelle des réseaux de neurones ayant une seule couche cachée est un premier résultat théorique important. La première partie du cours présente ce théorème dans le cas simplifié où les données x comportent d valeurs binaires et où la valeur de $f(x)$ est aussi binaire. Dans ce cas binaire, on démontre que la fonction $f(x)$ peut être exactement représentée par un réseau de neurones ayant une seule couchée cachée de taille , en utilisant la fonction $signe(x)$ comme non-linéarité.

La seconde partie du cours aborde le théorème général d'approximation universelle qui démontre que toute fonction $f(x)$ continue s'approxime par un réseau de neurones ayant une seule couche cachée. L'erreur uniforme (maximum sur x) converge vers 0 lorsque la taille de la couche cachée tend vers l'infini. Ce théorème est valable pour des réseaux de neurones implémentés avec des non-linéarités ponctuelles $s(t)$ qui sont continues mais qui ne sont pas des polynômes.

Dans le cadre du cours, le théorème est démontré pour un rectificateur $s(t) = max(t, 0)$. La démonstration se fait en montrant d'abord que toute fonction $f(x)$ continue s'approxime comme combinaison linéaire de sinus et de cosinus avec une erreur uniforme qui tend vers 0 lorsque le nombre de termes tend vers l'infini. On démontre ensuite qu'un sinus et un cosinus peuvent s'approximer par des fonctions linéaires par morceaux obtenues avec des combinaisons linéaires de rectificateurs dilatés et translatés, avec une erreur qui tend vers 0 lorsque le nombre de termes tend vers l'infini.

Séminaire 5 – Le langage naturel

Piotr Bojanowski (Facebook), le 20 février 2019

Ce séminaire présente des applications des réseaux de neurones profonds pour le traitement du langage naturel. Ces architectures profondes ont permis des avancées spectaculaires pour de nombreuses tâches, et un changement complet de paradigme dans le domaine. Le séminaire commence en discutant des modèles de langages et des architectures de réseaux adaptées (réseau de neurones récurrent, LSTM, GRU et transformer). Piotr Bojanowski présente l'optimisation des paramètres de ces modèles, ainsi que les approximations nécessaires afin d'appliquer ces méthodes à grande échelle. Dans une seconde partie il introduit des systèmes de traduction

automatique et des améliorations algorithmiques qui ont permis de construire les modèles performants d'aujourd'hui.

Cours 6 – Erreur d'approximation avec une couche cachée et régularité

27 février 2019

Le théorème d'approximation universelle d'un réseau de neurone à une couche cachée garantit que l'erreur d'approximation d'une fonction $f(x)$ continue va décroître vers 0, mais il ne spécifie pas la vitesse de décroissance de cette erreur. Cette vitesse de décroissance est liée à la régularité de $f(x)$. On verra que si $f(x)$ est seulement localement régulière alors l'erreur décroît très lentement et souffre de la malédiction de la dimensionnalité.

Le cours considère d'abord le cas de fonctions localement régulières, qui sont m fois différentiables au sens de Sobolev. On démontre des bornes supérieures sur l'erreur d'approximation en fonction du nombre M de neurones utilisés dans la couche cachée. On montre qu'une erreur e s'obtient avec $M = O(e^{-d/m})$ neurones. Cette décroissance est très lente si m est petit devant d ce qui est toujours le cas en grande dimension.

On considère aussi le cas où la transformée de Fourier de f est parcimonieuse, ce qui s'impose avec un critère L^1 proposé par Barron. Dans ce cas, on démontre que la décroissance de l'erreur est beaucoup plus rapide et qu'il suffit de $M = O(e^{-1/2})$ neurones pour atteindre une erreur e. Cependant, cette propriété de parcimonie en Fourier est rarement satisfaite dans les applications.

En dehors d'exemples particuliers, aucun théorème général ne permet d'expliquer l'augmentation des performances d'approximations obtenues avec des réseaux de neurones ayant plus de couches cachées, pour les fonctions que l'on rencontre dans les applications. Ce problème reste donc ouvert.

Séminaire 6 – Analyse automatique de vidéos

Cordelia Schmid (Inria), le 27 février 2019

Les progrès récents des réseaux de neurones profonds ont permis d'avancer de façon significative la compréhension automatique d'actions dans des vidéos. Le séminaire commence par donner une vision globale des algorithmes utilisés pour la classification de vidéos, puis il présente plusieurs algorithmes pour localiser dans le temps et dans l'espace les actions d'une vidéo. Il montre comment les « tublets » d'actions permettent d'obtenir l'état de l'art pour la localisation spatio-temporelle d'actions, et pourquoi la modélisation des relations entre les objets et les humains peut améliorer ces performances. Une grande base de données de vidéos d'actions est présentée. On décrit un algorithme faiblement supervisé afin d'apprendre les actions humaines dans des vidéos. Cet algorithme réduit de façon très significative le coût des annotations nécessaires pour entrainer des algorithmes de classification de vidéos.

Cours 7 – Optimisation d'un réseau par maximum de vraisemblance

13 mars 2019

L'optimisation d'un réseau de neurones consiste à estimer un vecteur de paramètres thêta qui minimise un risque calculé sur les exemples d'entraînement. Cela se fait par une descente de gradient si bien que le risque doit être différentiable.

Pour la classification, le principe du maximum de vraisemblance permet de définir un risque différentiable. Le maximum de vraisemblance cherche les paramètres θ qui maximisent le log de la probabilité conditionnelle de $y' = f(x')$ pour les exemples x', y' de la base de données d'entraînement. On démontre que ce maximum de vraisemblance maximise la distance de Kullback-Liebler entre la distribution conditionnelle des données et la distribution paramétrée par θ. Le risque est donc défini par cette distance. Dans le cas où le modèle de probabilité conditionnelle est Gaussien, on obtient un risque de régression quadratique.

La classification d'un réseau de neurones se fait le plus souvent en choisissant un modèle de probabilité conditionnelle défini par un *softmax*. Il attribue une distribution de probabilité à un ensemble de valeurs *zk* calculées pour chaque classe *k*, où la probabilité de *zk* est proche de 1 lorsque que *zk* a la plus grande valeur parmi tous les autres *zk'*. On peut alors calculer analytiquement le maximum de vraisemblance en fonction des *zk*, et c'est une fonction différentiable.

La régression logistique est un classificateur multi-classe pour lequel les sorties *zk* sont des fonctions affines de la donnée d'entrée *x*. La maximisation de la vraisemblance calculée avec un *softmax* est une fonction convexe des paramètres et admet donc une solution unique. On montre que l'unicité de la solution vient de l'introduction d'un critère de marge qui optimise la position des frontières.

Séminaire 7 – Apprentissage profond par renforcement

Yann Olivier (CNRS), le 13 mars 2019

Ce séminaire présente les principales approches de l'apprentissage par renforcement, en insistant sur les algorithmes effectifs, et en comparant les avantages de différentes approches (avec modèle du monde, sans modèle, etc.). Yann Olivier discute l'état de la compréhension mathématique de ces algorithmes au-delà de problèmes jouets, ainsi que les limites à la fois théoriques et pratiques des approches actuelles.

Cours 8 – Descente de gradient et rétro-propagation du gradient

20 mars 2019

Ce cours étudie les algorithmes de descentes de gradient par *batch* et de gradient stochastique, ainsi que leur implémentation dans un réseau de neurones avec l'algorithme de rétro-propagation du gradient.

L'algorithme de descente de gradient ajuste des paramètres pour minimiser une fonction de coût, qui quantifie l'erreur entre la réponse estimée et la bonne réponse sur des données d'entraînement. Ces paramètres sont modifiés itérativement en soustrayant le gradient de la fonction de coût. Cet algorithme nécessite donc de calculer la variation du coût à la sortie du réseau relativement à la variation des paramètres du réseau. Dans un réseau de neurones, le coût se calcule par composition d'opérateurs linéaires ou de non-linéarités ponctuelles différentiables. La différentiation d'une composition d'opérateurs s'obtient en appliquant la règle de Leibniz. La propagation des dérivées se fait donc dans le réseau par multiplications successives de matrices jacobiennes.

Le gradient est normalement calculé en moyenne sur un *batch* qui contient toutes les données d'entraînement, ce qui nécessite beaucoup de calculs. L'algorithme de descente de gradient stochastique réduit ce *batch* à un seul exemple pris au hasard

dans la base de données. Dans le cas où la fonction de coût est strictement convexe, on démontre que l'algorithme de descente de gradient converge vers un minimum unique. La vitesse de convergence est exponentielle, avec un exposant qui dépend du conditionnement du Hessien.

L'algorithme de descente de gradient stochastique a une convergence beaucoup plus lente que l'algorithme par *batch*, car chaque itération est bruitée par la variabilité du gradient qui dépend de l'exemple particulier qui a été choisi. Cependant, lorsque la base de données est suffisamment grande, cet algorithme peut nécessiter moins d'opérations que la descente de gradient par *batch* pour approcher le minimum de la fonction de coût.

Cours 9 – Convergence de la descente de gradient stochastique

20 mars 2019

La convergence de la descente de gradient par *batch* et de gradient stochastique est démontrée dans le cas où la fonction de coût est Lipchitz et fortement convexe relativement à ses paramètres.

Pour une descente de gradient, on montre une décroissance exponentielle en t de la distance euclidienne entre le vecteur de paramètres calculé à une itération t et le vecteur optimal qui minimise la fonction de coût. L'exposant de cette décroissance dépend de la borne de Lipchitz et de la constante de convexité forte.

L'algorithme du gradient stochastique peut être interprété comme une version bruitée de l'algorithme de descente de gradient. Ce bruit dépend de la variance de la fonction de coût pour des exemples pris aléatoirement. On démontre un premier théorème qui calcule une borne supérieure de la distance euclidienne entre le vecteur de paramètres obtenu à l'itération t et le vecteur optimal qui minimise la fonction de coût. Un choix optimal du pas de gradient permet d'obtenir une borne supérieure qui décroît comme $1/t$. Ces théorèmes s'étendent au cas de fonctions de coût non différentiables et convexes, en remplaçant le gradient par la notion de sous-gradient, et on obtient alors le même type de résultat.

Les algorithmes de descente de gradient ont une précision limitée par la variabilité aléatoire de la fonction de coût. Pour converger vers la solution optimale, il faut utiliser des algorithmes hybrides qui se comportent comme une descente de gradient stochastique tant qu'on est loin de la solution optimale, puis qui incorpore les exemples vus dans le passé pour se comporter comme une descente de gradient par *batch* lorsque le nombre d'itérations devient grand. Cela permet de converger vers la solution optimale.

Les propriétés de convergence des paramètres de réseaux de neurones profonds restent mal comprises car les fonctions de coût ne sont pas convexes. L'optimisation de ces réseaux garde donc une forte composante expérimentale. Pour que l'optimisation donne de bons résultats, il faut qu'il y ait une faible densité de minima locaux qui ont un coût bien plus important que les minima globaux. Cela dépend de l'architecture du réseau. De nombreuses recherches sont dédiées à la compréhension de ce problème et à l'étude de la géométrie du paysage d'optimisation.

Recherche

Stéphane Mallat dirige l'équipe de recherche « Data » à l'École normale supérieure, qui étudie des problèmes de mathématiques appliquées aux sciences des données. Cela couvre l'apprentissage supervisé, l'apprentissage non supervisé ainsi que des problèmes inverses de traitement du signal.

L'équipe travaille sur des modèles mathématiques permettant d'expliquer la performance des réseaux de neurones profonds. Ils se fondent notamment sur la transformée de *scattering* qui est un modèle non linéaire de réseaux de neurones, fondé sur la transformée en ondelettes. En 2018-2019, une partie importante de la recherche était dédiée à la construction de modèles stochastiques pour la synthèse d'images, de musiques, de finance, mais aussi pour modéliser des phénomènes de turbulence en physique et des propriétés de réseaux de télécommunication.

Publications

ANDÉN J., LOSTANLEN V. et MALLAT S., « Joint time-frequency scattering », *IEEE Transactions on Signal Processing*, vol. 67, nº 14, 2019, p. 3704-3718, https://doi.org/10.1109/TSP.2019.2918992.

ANDREUX M. et MALLAT S., « Music generation and transformation with moment matching scattering inverse networks », *in* : *Proceedings of the 19th ISMIR (International Society for Music Information Retrieval) Conference* (Paris, 23-27 septembre 2018), 2018, p. 327-333.

BRUNA J. et MALLAT S., « Multiscale sparse microcanonical models », *Journal of Mathematical Statistics and Learning*, vol. 1, nº 3, 2018, p. 257-315, https://www.doi.org/10.4171/MSL/7.

BROCHARD A., BŁASZCZYSZYN B., MALLAT S. et ZHANG S., « Statistical learning of geometric characteristics of wireless networks », *in* : *IEEE INFOCOM 2019: IEEE Conference on Computer Communications*, 2019, p. 2224-2232.

LEONARDUZZI R.F., MALLAT S., BOUCHAUD J.-P., ROCHETTE G., « Maximum entropy scattering models for financial time-series », *in* : *ICASSP 2019: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, p. 5496-5500, http://dx.doi.org/10.1109/ICASSP.2019.8683734.

MALLAT S., « Quelles limites pour l'intelligence artificielle au travail ? », *in* A. SUPIOT (dir.), *Le Travail au XXIe siècle*, Ivry-sur-Seine, Les Éditions de l'Atelier, 2019.