Modèles multi-échelles et réseaux de neurones convolutifs

Stéphane Mallat
rapporté par Edwige Cyffers

21 janvier 2020

1 Organisation du cours

Ce cours se tient pour la troisième année au Collège de France. Son organisation diffère d'un cours de master classique, constitué d'un bloc de connaissance dont on valide via un examen final l'assimilation par les étudiants. On garde ici à l'esprit qu'on transmet des connaissances en construction, et ce cours est donc déjà orienté vers la recherche. Les élèves de Master souhaitant valider le cours doivent faire un challenge, qui consiste en l'exploitation de données réelles pour répondre à une problématique réelle proposée par une entreprise. Plus généralement, toute personne intéressée par le cours est invitée à faire de même.

Les challenges sont accessibles sur `challenge.data@ens.fr`.

Détail pratique : les étudiants souhaitant valider le cours doivent envoyer un mail à `dominique.bidois@college-de-france.fr` avec leur nom prénom et master et l'objet "Inscription cours S. Mallat".

2 Introduction : Comprendre un système d'apprentissage

2.1 Questions mathématiques sur la compréhension des systèmes d'apprentissage

Lorsque l'on parle de système d'apprentissage, on désigne différents types de procédés, qui vont être plus ou moins sophistiqués, et qui visent à extraire de l'information des entrées. Les comprendre peut se référer à plusieurs aspects, il faut donc déjà se demander ce qu'on entend par comprendre dans ce contexte.

D'un point de vue d'ingénieur, ces systèmes d'apprentissage sont parfaitement compris. Ils sont en effet entièrement spécifiés, ils sont implémentés d'une certaine manière et n'évoluent ou ne pensent pas par eux-mêmes. Cependant, d'un point de vue mathématique, il reste plusieurs problèmes, notamment sur l'origine de leur capacité de généralisation. A titre d'exemple, on verra dans ce cours des systèmes neuro-biologiques, comme le fonctionnement de l'appareil

auditif. Ce sont des systèmes si complexes qu'une solution qui s'impose rapidement est de les regarder avec un certains niveau d'abstraction, mais on va voir que cela ne rend pas forcément leur analyse plus aisée pour autant, et il en est de même pour les réseaux de neurones. On peut déjà lister trois grands groupes de problèmes mathématiques.

Robustesse On souhaite connaitre leur capacité à rester cohérent sur des entrées proches. C'est un problème très important pour certaines applications comme les voitures autonomes, et souvent les systèmes ne possèdent pas une telle robustesse

Efficacité Les systèmes d'apprentissage sont souvent coûteux, à la fois en termes de nombres de données nécessaires et en termes de temps pour les traiter. Actuellement, certains apprentissages demandent un tel coût de calcul qu'on peut se demander s'ils sont optimaux ou adaptés aux tâches apprises.

Contrôle On décide d'un système avec une architecture donnée pour un problème. En quelle mesure et de quelle façon s'effectue ce choix est un problème avec très peu de réponses pour l'instant. On aimerait comprendre comment ces architectures apprennent, c'est-à-dire quel est le rôle effectif de chaque partie dans le traitement de la donnée, sachant que le design de cette architecture représente l'écrasante majorité du temps de travail d'un ingénieur actuellement.

2.2 Quelle architecture pour quel apprentissage ?

A l'intérieur de ce troisième groupe, on peut voir émerger plusieurs problématiques. On aimerait savoir comment l'architecture joue un rôle dans la généralisation du problème appris. On peut aussi utiliser un système d'apprentissage pour tenter de quantifier la complexité d'un problème. On considère une tâche, et on entraine un système d'apprentissage – pas forcément un réseau de neurones d'ailleurs – et on voit si on arrive effectivement à obtenir des prévisions correctes. C'est fait pour certains systèmes chimiques : on ne sait pas comment le réseau réussit la tâche, car on ne voit pas apparaitre les équations de Schrödinger par exemple, mais ça peut nous donner de nouvelles idées sur la difficulté de la tâche. En chimie, on fait notamment des liens croissants avec la physique statistique. Enfin, l'architecture elle-même est un ensemble de paramètres, qu'on peut vouloir apprendre via un deuxième système d'apprentissage, par exemple via des approches d'algorithmes génétiques. Nous ne détaillerons pas ce point pendant ce cours, mais il s'agit d'un domaine de recherche actif.

Le panorama qui précède a pour but de vous donner un aperçu du fait qu'il y a de nombreuses questions possibles, même si on va se concentrer sur le lien entre architecture et capacité d'apprentissage, notamment avec le rôle de l'information a priori. On utilise les mathématiques dans la suite du cours comme langage pour décrire ces systèmes. Une question qui est souvent posée est "Comment se fait-il que les mathématiques soient si efficaces pour décrire le monde ?". Le

point de vue ici, celui de mathématiques appliquées, est que le langage évolue en fonction des questions que l'on pose. D'une certaine façon le langage est un système d'apprentissage, qui évolue avec les données qui sont les problèmes posées, et donc de ce point de vue il n'est pas surprenant que les mathématiques soient adaptées au problème de la physique car elles sont aussi le système d'apprentissage qui a été utilisé pour les résoudre. ici, les mathématiques de la très grande dimension, qui touchent à la fois les probabilités, les statistiques, d'analyse, sont donc aussi en évolution pour répondre aux problèmes posés.

Dans ces problèmes de grande dimension, on va suivre plusieurs principes d'organisation pour traiter ces données.

séparabilité Le cas le plus simple est celui de séparabilité des variables entre elles. Par exemple dans un système physique, les interactions sont essentiellement locales et donc on peut espérer avoir une séparabilité locale. Cependant très souvent on ne peut pas avoir quelques choses d'aussi fort, à part sur des cas très simples. On peut aussi avoir une séparabilité hiérarchique, et c'est ce qu'on aura généralement : . Pour étudier cette séparabilité, on utilisera des outils comme l'analyse harmonique, l'analyse des fréquences et des ondelettes.

Symétrie Elles sont un exemple d'informations a priori. On va prendre en compte les symétries telles les invariances par translation ou par rotation. Un objet reste le même lorsqu'on le déplace dans l'image. D'un point de vue mathématique, c'est la théorie des groupes qui nous vient en aide.

Parcimonie Il y a un certain nombre de structures élémentaires ou irréductibles qui composent le système étudié. Dans le cas des images, il y a un certain nombre de *patterns* intrinsèques comme la forme d'un visage, le présence des yeux... On va essayer de décomposer notre système selon ces structures.

Principe d'évolution Méta-principe. A partir d'un système, on apprend un minimisant un objectif que l'on a défini, et un peut voir le chemin de minimisation comme une évolution temporelle.

Ainsi, une des données que nous aurons à étudier est le temps. On peut le voir principalement de trois manières différentes. Il y a un premier cas où on peut finalement l'assimiler à de l'espace, il ne constitue qu'une dimension supplémentaire avec une continuité selon cette coordonnée. C'est le cas d'acquisition temporelle, où chaque image va avoir un sens séparément, être analysable. Une deuxième manifestation possible du temps est le temps tel qu'il apparait en physique. On va déterminer le comportement d'un objet en fonction d'équation faisant intervenir le temps. On peut citer l'exemple de la physique quantique, où on a la description du système en fonctions de l'hamiltonien:

$$\frac{dx}{dt} = H_t x$$

On va donc pouvoir décrire le phénomène via l'analyse d'un opérateur relativement simple puisque linéaire. Une dernière interprétation du temps, qui peut être aussi très intéressante dans le cas de l'apprentissage, est de pouvoir décrire des discontinuités. Il a été mis en évidence que les nouveaux-nés reconnaissent très vite les mains, en une dizaine de jours seulement, alors même qu'il s'agit d'un objet particulièrement complexe, qui change de forme, qui peut être différent selon les usages. Une explication possible de cette très bonne performance est que la main, qui sert à la préhension, est souvent à l'origine de discontinuités : on saisit l'objet avec la main, et donc il bouge. Le système d'apprentissage de l'enfant détecte cette corrélation et reconnait donc la main avec finalement très peu d'exemples.

2.3 L'information a priori

Nous finissons l'introduction avec ce concept qui sera très utilisé dans la suite du cours. On ne peut pas apprendre sans a priori, car il s'agit de définir la classe d'hypothèses que l'on va considérer. On aura des données x dont on cherche à extraire une information y, on pour cela on va chercher chercher des fonctions

$$f : x \to y = f(x), \mathcal{H} = \{f \mid \dots\}$$

Le choix de $\mathcal{H}$ est primordial: sans information a priori, on a l'ensemble de toutes les fonctions possibles, on va donc chercher à minimiser une fonction de risque $\mathcal{R}(f, \hat{f})$ dans un très grand espace, et on ne pourra pas dépasser la malédiction de la très grande dimension. On doit donc chercher un ensemble $\mathcal{H}$ suffisamment petit, ce qui, comme on le verra par la suite, consiste à expliciter les régularités de la fonction. Plus vous avez d'informations a priori, plus vous pouvez réduire la dimension. Mais l'information a priori est par essence inexacte, et peut donc vous mener à une solution trop inexacte. A un moment, il devient plus intéressant d'augmenter la taille de votre espace, pour avoir un ensemble de fonctions moins fausses, et donc trouver une meilleure approximation dans un espace plus grand.

Il faut donc choisir où on met le curseur. Jusqu'en 2010, le curseur était l'extrême, c'est-à-dire de définir tout le système et d'utiliser uniquement une régression linéaire à la fin. On verra que ce système a des limites : informations a priori trop fausses, insuffisantes.

Il faut que ce soit le thème des projets également: il faudra commencer par définir une représentation et obtenir le résultat le meilleur possible avec un simple classificateur linéaire. Vous serez éventuellement bloqué à un moment. Alors vous irez vers un système plus complexe, et vous verrez ce que le système a trouvé qui n'était pas compris dans l'information a priori.

3 Problèmes supervisés

Cette année on considère uniquement l'apprentissage supervisé. La grosse complexité vient de la taille des données $x = (x(1), \dots, x(d)) \in \mathbb{R}^d$, où d est très

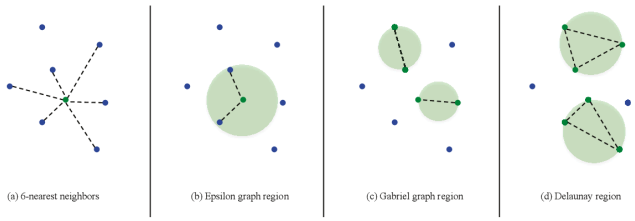

Figure 1: Une méthode naturelle pour estimer la fonction est de prendre les points d'un certain voisinage, et de réaliser une moyenne pondérée. Ici quelques exemples de voisinages possibles

grande, typiquement de l'ordre de plusieurs millions. A partie de ces données on peut effectuer des tâches de classification – par exemple déterminer quel animal est présent sur une image, on retourne une classe discrète – ou de régression – par exemple estimer l'énergie d'un système, on retourne donc une valeur continue.

Peut-on calculer l'énergie d'un système à partir de données ? En physique, on utilise les équations fondamentales de la physique. Ici on a de l'information a priori via les symétries, les rotations: l'énergie reste la même lors d'une rotation. On procède très différemment de la physique : on part d'exemples, et on interpole à partir de ceux-ci.

Cette interpolation est conceptuellement très simple. Imaginons qu'on nous donne des valeurs d'une fonction à certains points, et qu'on nous donne un nouveau point, on va prendre la moyenne des points qui sont dans l'entourage.3

Cette méthode tombe en défaut en grande dimension : pour avoir ne serait-ce qu'un point de connu dans une voisinage de taille d, il faut $n \geq e^{-d}$ points, ce qui est bien sûr impossible. C'est la malédiction de la très grande dimension. Faire une interpolation naïve comme la précédente demande donc de prendre une fonction très régulière et ce qui se révèle donc impossible en pratique. Il y a un cas où on peut encore espérer s'en sortir : c'est la cas où les données sont en fait accumulées dans un espace de dimension beaucoup petite3 Cela peut correspondre à certains cas très structurés : c'est vrai pour des images de chiffres. Mais ce n'est pas possible pour une photo quelconque, comme un photo prise dans la salle par exemple.

Prenons par exemple un nuage de points colorés avec deux couleurs définissant deux labels. On souhaite que le problème soit séparable, c'est-à-dire qu'il existe une fonction linéaire qui permet de prendre la décision d'appartenance à une des deux classes3. Cela peut être vu comme un vote : on prend un vote sur des informations faibles, les features x_i et on en déduit le label. Au niveau géométrique on peut regarder la frontière qui peut être interprétée comme une ligne de niveau, et on effectue un changement de variable pour changer cette frontière en un simple point, un seuil. On a donc le calcul suivant :

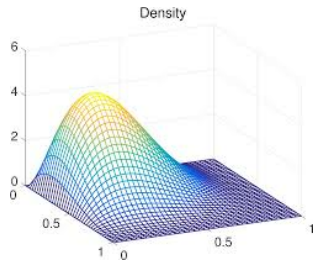

Figure 2: Exemple de nuages de points qui sont regroupés dans un espace de plus petite dimension

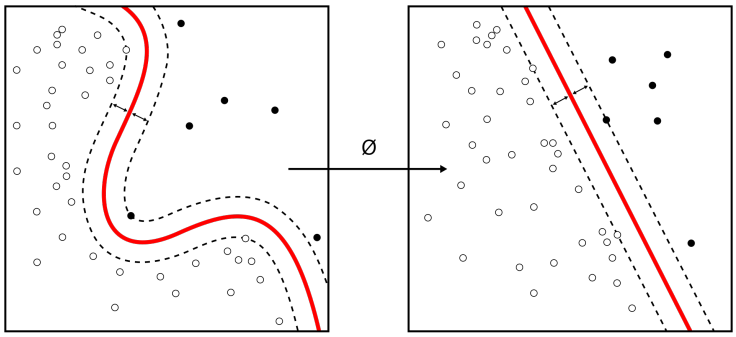

Figure 3: Transformation de données initiales afin de discriminer les deux classes de façon linéaire

$$\hat{f}(x) = sign(\langle w, \phi(x) \rangle + b) = sign\left(\sum_k w_k v_k + b\right)$$

La complexité de ce changement de variable ϕ nous donne donc d'une certaine façon la difficulté du problème. La difficulté est donc de trouver le meilleur changement de variable ϕ Le point de vue précédent était de prendre toutes les informations possibles a priori. L'autre extrême est de prendre un réseau de neurones et d'obtenir en dernière couche le résultat.

3.1 Sur les challenges

Il y aura un ensemble variés de données : images, séries temporelles, textes... Il s'agira toujours de faire des prédictions, pour lesquels on dispose d'un certain nombre de couples données/réponses. Ensuite on soumet la méthode qui est testée sur d'autres données.

Pour lutter contre l'apprentissage indirecte des données de test, au plus

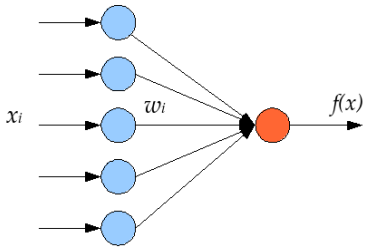

Figure 4: Schéma d'un perception

deux soumissions par jour sont possibles durant les deux mois : cela donne déjà beaucoup de soumissions potentiels, si bien qu'il est possible d'inférer les données du jeu de test. Le résultat final sera évalué sur d'autres données, complètement cachés jusqu'alors. L'évaluation sur des données cachées a lieu le 19 mars, et un classement final est établi le 15 décembre. Oraux entre le 23 et 27 mars pour les étudiants de master.

Pour ceux-ci, le challenge doit être l'occasion de confronter la théorie à la pratique. Une première étape est de développer une représentation fondée sur des a priori, et seulement ensuite, s'il y a assez de données, de les comparer avec une solution fondée sur des réseaux de neurones. Vous verrez qu'en pratique, sur des vraies données, les réseaux de neurones ne marchent pas très bien, notamment à cause de la présence de bruit.

3.2 Rappels sur les perceptrons multicouches

La motivation vient d'études sur la grenouille dans les années cinquante. L'idée est extrêmement simple, il s'agit de calculer un hyperplan séparateur. On va calculer le produit scalaire, c'est-à-dire la somme pondérée $\sum_k v_k w_k$. Si le résultat est inférieur au seuil b, on retourne 0, et on retourne la différence avec b dans le cas contraire, ce qui donne une sortie assez parcimonieuse.3.2

On passe ensuite à un réseau de neurones à plusieurs couches, dont on va donner les sorties d'un neurone comme entrée d'une autre couche de neurones 3.2. L'architecture correspond donc à comment on relie les neurones entre eux. L'apprentissage correspond au choix des paramètres de ces réseaux. Malgré l'obstacle a priori de multiples minimas locaux, le miracle est que l'on obtient une très bonne minimisation. La difficulté est donc de choisir l'architecture, et c'est là que l'on passe le plus de temps.

L'idée de Yan le Cun (prix Turing) a été de mettre en avant que le problème est très souvent invariant par convolution : si vous voulez reconnaitre un effaceur d'une bouteille d'eau, la réponse ne dépend pas de la position dans l'image. On va donc fixer les paramètres qui doivent correspondre aux mêmes tâches pour qu'ils aient les mêmes paramètres. La seconde idée est que la première couche doit s'intéresser à une information locale, c'est-à-dire qu'on va mettre en entrée une zone très étroite, typiquement 3 par 3 pixels. Mais dès que

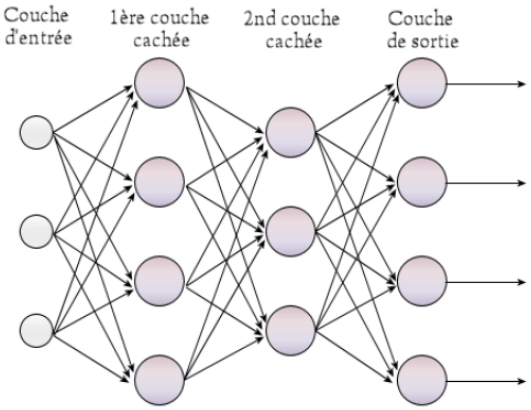

Figure 5: Schéma d'un réseau de neurones constitué de quatre couches

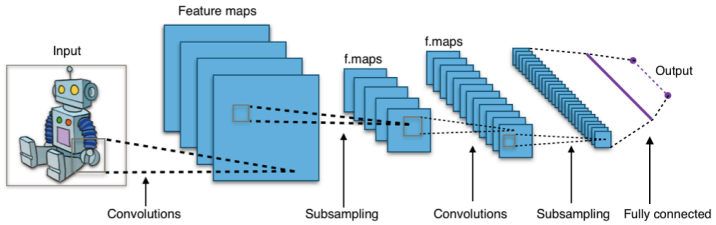

Figure 6: Schéma d'un réseau de neurones convolutif

l'on passe à la seconde couche, les neurones vont prendre en entrée des canaux correspondant à la couche précédente, et c'est là que l'analyse mathématique commence à être ardue. On parle de réseaux de neurones convolutifs pour désigner ces architectures.6

Ce qui est particulièrement formidable, c'est que cette architecture fonctionne très bien pour des problèmes extrêmement variés, comme on le verra via les séminaires. Le point important est que le $\phi(x)$ correspondant à l'image du réseaux a réussi à aplatir les structures, pour retourner le y attendu en sortie. On a différentes couches : on sous échantillonne et on filtre : un neurone voit au fur et à mesure des zones de plus en plus larges. A la fin les neurones voient donc toute l'image, on les appelle parfois neurones grand mère. Très souvent, leur réponse sera donc zéro, et ils vont flasher de temps en temps, et on essaie de comprendre pourquoi, ce qu'ils ont reconnu comme pattern.

ImageNet est un exemple très important car c'est le premier qui a montré la puissance des réseaux de neurones. Historiquement, la communauté étaie très réticente, il n'y avait pas de données qui étaient mieux traitées avec des réseaux de neurones plutôt qu'avec des systèmes sans apprentissage. C'est un

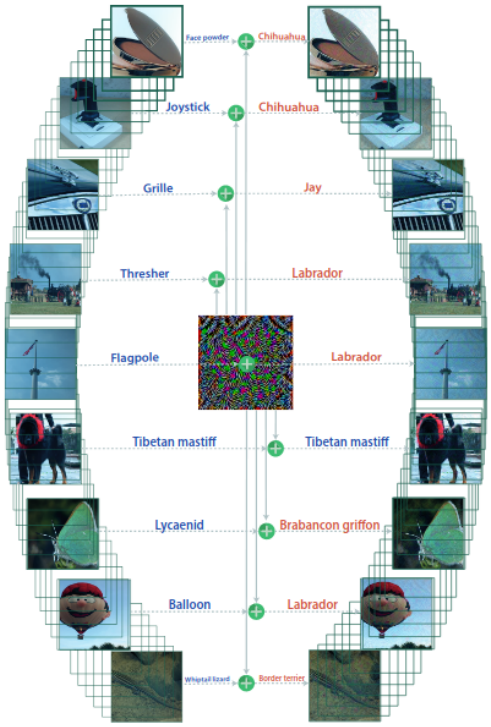

Figure 7: Exemple de manque de robustesse de réseaux de neurones : l'ajout d'un filtre invisible pour l'homme entraine des erreurs de classification

exemple de grande dimension, on ne peut pas comprendre les réseaux de neurones uniquement avec des données de petites dimensions comme MNIST, car seulement la grande dimension permet d'obtenir de meilleurs résultats que ce qu'on peut obtenir par des méthodes plus simples. Cependant, décrire le choix de l'architecture est complexe. Quand on regarde l'architecture d'AlexNet, on peut difficilement dire quel type de connaissances a été appris. Les poids ne fournissent pas la réponse, ils sont dépendants de l'initialisation. On n'est aujourd'hui en dessous des erreurs humaines, mais ce score est à mettre en perspective. Certaines classes sont mal connues des humains donc mal classifiées. Par contre l'humain est bien plus robuste. Il a été montré que tous les réseaux ont des instabilités, c'est-à-dire que des petites perturbations mènent à une mauvaise classification.7

3.3 Exemples d'application

La reconnaissance de la parole a obtenu de très bons résultats avec les systèmes d'apprentissage, et ces méthodes sont aujourd'hui utilisées par les téléphones portables pour les systèmes de reconnaissance vocale.

Prenons des données d'un spectrogramme. On a un axe de temps et les fréquences, qui correspondent aux harmoniques des différents phonèmes. On écoute deux élocutions d'une même phrase. Les spectrogrammes sont très différents, comment peut-on malgré tout faire le lien entre les deux ? La réponse classique utilisait des modèles de mixtures de gaussiennes sur une chaine de Markov, dont les performances plafonnaient.

De même, le problème de séparation de sources, qui est crucial pour les aides auditives. On ne veut pas amplifier le bruit ambiant, mais uniquement la source qui nous intéressent. En 2018, Luo et Mesgarani ont développé un réseau de neurones qui réussit cette tâche, alors même qu'elle semblait impossible, et on ne sait pas comment ça marche, excepté que la première couche ressemble à un spectrogramme.

Le cours et les séminaires présenteront aussi un ensemble de problèmes qui connaissent actuels des progrès via l'utilisation de réseaux de neurones. En physique, on est en train de s'emparer de ces méthodes, pour faire des simulations en 3D, pour trouver de nouveaux matériaux,... On peut aussi faire des liens avec la neurophysiologie. Si on prend le cortex, on distingue un certain nombre de zones. S'il existe réellement un principe mathématique sous-jacents de simplicité, même si le hardware est différent, il doit exister des correspondances avec nos réseaux de neurones sur ordinateur. On a vu qu'un être humain peut reconnaitre un animal en moins de 15 millisecondes. On sait que les neurones sont très lents, et donc il ne peut pas y avoir des boucles de rétro-action en un temps aussi cours. De façon caricaturales, on peut donc imaginer que c'est un simple feed forward qui nous permet de distinguer l'animal, même si cela ne signifie pas que l'on utilise uniquement un feed forward durant l'apprentissage.

Ainsi, le cours de cette année se concentre sur le lien entre l'architecture et l'apprentissage. On cherchera quelle information a priori est incluse dans une architecture, quels sont les filtres possibles. Enfin, on verra quel est le rôle des singularités.