Big data cherche mathématiques

Stéphane Mallat, école Normale Supérieure, SMAI, Matapli n101

1 Introduction

L'importance des problèmes de big data est maintenant un sujet tarte à la crème, et pourtant, au-delà des statisticiens, peu de mathématiciens y ont goûté. Les enjeux sociétaux, industriels et scientifiques sont considérables, avec un besoin criant de mathématiques. On a beaucoup dit que le XXIème siècle serait celui du cerveau, mais le traitement des masses de données jouera aussi un rôle important, et ces deux domaines se rejoindront peut être.

Rappelons que la production de données numériques double tous les 3 ans depuis 1980, et a atteint le chiffre astronomique de $2\,10^{21}$ octets par jour. Ces données se trouvent au coeur de la plupart des industries et services. Cela ne concerne pas seulement Google, Amazon ou Facebook, mais aussi les industries pharmaceutiques, pétrolières, les assurances, banques, les télécommunications, le monde médical, le marketing... La recherche est aussi affectée, en bioinformatique, astronomie, géophysique, météorologie, physique des hautes énergies, neuro-sciences, vision par ordinateur, reconnaissance de signaux, sciences sociales... Ces enjeux ont motivé le lancement d'une "Big Data Initiative", qui finance de nombreux projets de recherche aux États-Unis, et l'Europe en fera probablement une priorité à l'horizon 2020.

Beaucoup de problèmes de big data sont technologiques et informatiques : organiser des bases de données inhomogènes, faire du calcul distribué, visualiser et sécuriser les données. Cependant, un des goulots d'étranglement est la capacité d'exploitation et d'analyse automatique de ces masses de données. Cela pose des problèmes mathématiques profonds, souvent liés à la malédiction de la très grande dimension.

Ces dernières années, l'algorithmique a fait des progrès importants, grâce aux grandes bases de données d'apprentissage et à de nouvelles approches non-linéaires parfois déroutantes. Cependant, la multiplication des algorithmes avec leurs variantes exotiques, a transformé ce domaine en une jungle épaisse. L'éclairage des mathématiques est devenu d'autant plus nécessaire. Pour l'instant, les statistiques et l'optimisation sont les seuls domaines mathématiques qui se sont adaptés avec succès. Pourtant ce ne sont pas les seuls concernés, loin de là. L'analyse, les probabilités, la géométrie, la topologie, la théorie des groupes ont beaucoup à apporter. Je n'essaierai pas de donner une couverture exhaustive des problèmes, mais plutôt un éclairage biaisé par mon expérience.

2 La malédiction de la dimensionalité

Un ensemble de données numériques, que ce soit une image, des mesures biochimiques ou des données marketing, peut être représenté par un vecteur $x \in \mathbb{R}^d$, où d est souvent supérieur à 10^6.

La malédiction de la dimensionalité rend très difficile la comparaison, la détection ou la classification de ces données. Elle vient du fait que le volume est une fonction exponentielle de la dimension. En dimension 1, avec 100 points, on échantillonne un segment de taille 1 avec des intervalles de 10^{-2}. Dans un espace de dimension 10, il faudrait 10^{20} points pour échantillonner un cube de largeur 1 avec des points dont la distance est 10^{-2}. En dimension d il en faut 10^{2d}, autrement dit un nombre inimaginable pour $d = 10^6$.

La taille monstrueuse d'un espace de grande dimension se voit en comparant le volume d'un cube de largeur $2r$, soit $(2r)^d$, et le volume de l'hypersphère de rayon r inscrite dans ce cube, soit $2r^d\pi^{d/2}d^{-1}\Gamma(d/2)^{-1}$. Le rapport des deux volumes est $\pi^{d/2}d^{-1}2^{1-d}\Gamma(d/2)^{-1}$. Il tend exponentiellement vers 0 quand d aug-

mente. Cela veut dire que l'essentiel du volume du cube est loin de son centre, et se concentre dans les coins, ce qui est totalement contre-intuitif. À cause de cette malédiction, la distance euclidienne perd sa capacité de discrimination. Étant donnée une distribution de p points dans l'espace, le rapport entre la distance maximum et la distance minimum de ces points, $\max_{i\neq j} ||x_i - x_j|| / \min_{i\neq j} \|x_i - x_j\|$, tend vers 1 quand la dimension d augmente [6].

3 Le far-west

Les algorithmiciens et traiteurs de signaux ont relevé les défis posés par le fléau de la dimensionalité, avec de nombreux succès récents. Ce qui a changé, ce sont les capacités de mémoire et de calcul des ordinateurs, et les ordres de grandeurs des bases de données. Le recherche autour des big data est devenue un far-west expérimental, extraordinairement riche et créatif. Un monde où s'armer d'un gros ordinateur pour implémenter une belle idée est souvent plus efficace que de manipuler des concepts mathématiques élaborés. Passer par cette étape était sans doute nécessaire pour débroussailler ce nouveau domaine. Des solutions suprenantes sont apparues, prenant à contre-pied beaucoup d'intuitions mathématiques "raisonnables".

Les réseaux de neurones sont un exemple d'architecture algorithmique, longtemps décriée par la communauté scientifique bien pensante, dont je faisais partie. Leur efficacité est maintenant indiscutable, pour résoudre des problèmes en très grande dimension. Un neurone computationnel prend en entrée le long de ses "synapses" des variables $(a_i)_{1\leq i\leq j}$ qui sont les sorties d'autres neurones. Il calcule en sortie une valeur $\rho(\sum_i w_i a_i)$ où les w_i sont des poids qui sont appris à partir des données d'entrainement. La fonction $\rho(u)$ peut être une sigmoïde, une valeur absolue, un seuillage ou d'autres fonctions non-linéaires. L'art un peu mystérieux de ces réseaux de neurones consiste à organiser l'architecture d'un tel réseau, choisir les non-linéarités, et définir un algorithme d'apprentissage qui optimise les paramètres w_i de chaque neurone. Il existe de nombreuses approches et architectures différentes [6] Récemment, les réseaux de neurones profonds ont obtenu des succès remarquables, pour la reconnaissance d'images sur des bases de données de plus de 10^7 images avec 10^4 classes différentes, mais aussi pour la reconnaissance de la parole et le traitement de données biomédicales [2, 8]. Derrière la grande diversité des algorithmes développés, avec ou sans réseaux de neurones, quelques principes généraux ont émergé.

Mémoriser beaucoup d'informations et de paramètres apparaît comme nécessaire pour résoudre des problèmes de classification en grande dimension. Certain réseaux de neurones ont plus de 10 millions de paramètres [7], ce qui reste tout petit à côté du nombre de neurones et de synapses du cerveau, qui est 10^6 fois plus grand. En permanence, on mémorise inconsciemment une proportion importante des sons et des images auxquelles on prête attention. Les avancées récentes des traducteurs automatiques montrent qu'il est souvent plus efficace de stocker des exemples, avec des algorithmes d'indexation basés sur des modèles statistiques simples, plutôt que d'effectuer des analyses syntaxiques fines avec des représentations sémantiques complexes, comme cela s'est fait pendant longtemps.

La parcimonie est un autre principe important pour l'apprentissage de structures. Tout comme le rasoir d'Occam utilisé en recherche scientifique, parmi l'ensemble des possibles, il s'agit de trouver un nombre minimum de paramètres pour expliquer la variabilité des données, en vue d'une tâche particulière. Cela permet d'apprendre avec relativement peu de données d'entraînement. Dans un réseau de neurones, cela veut dire qu'une faible proportion des neurones vont être activés par une entrée particulière. Cette parcimonie est aussi observée dans le cerveau

La malédiction de la dimensionalité vient de la variabilité considérable des données. Réduire cette variabilité passe par la construction d'invariants. Cependant, ces sources de variabilité étant multiples et

complexes, l'élaboration de ces invariants et leur organisation deviennent compliquées. Une approche est de cascader des opérateurs contractants, afin d'obtenir des paramètres qui représentent des structures de plus en plus spécialisées, et de plus en plus invariantes [2, 4, 10]. L'analogie est parfois faite avec les "neurones grand-mères" en neurophysiologie de la perception [7]. De tels groupes de neurones semblent très spécialisés et invariants. Ils peuvent par exemple répondre à l'image d'une personne en particulier, quels que soit l'environement, la position ou l'expression du visage.

4 Représenter pour classifier

Les problèmes d'analyse de données sont du ressort des statistiques, mais pas seulement. La nécessité de construire des représentations adaptées aux grandes dimensions pose des problèmes mathématiques allant bien au-delà des statistiques.

Considérons un problème de classification. On veut classifier des données $x \in \mathbb{R}^d$ parmi K classes, qui sont des sous-ensembles $\{C_k\}_{1 \leq k \leq K}$ disjoints de $\mathbb{R}^d$. Si x est une image, ces classes peuvent correspondre à des chaises, des animaux, des visages, des fleurs, des paysages...

En termes probabilistes, on peut définir la distribution de probabilité $p_k(x)$ de chaque classe dans $\mathbb{R}^d$, dont le support est dans C_k. Le classificateur Bayésien optimal associe à chaque observation x la classe C_k qui maximise $p_k(x)$. Le problème est que l'on ne connaît pas les distributions $p_k(x)$. Par contre, on a des exemples labélisés $\{(x_i, y_i)\}_{i \leq n}$, où $y_i = k$ si $x_i \in C_k$. L'algorithme des plus proches voisins est une implémentation empirique du classificateur de Bayes, à partir de ces exemples. Il associe à x la classe C_k pour laquelle il y a le plus grand nombre de voisins x_i qui appartiennent à C_k parmi les K plus proches. Cela revient à faire une estimation de chaque $p_k(x)$ par un moyennage empirique au voisinage de x, et de choisir le plus grand.

Cependant, l'estimation empirique des distributions $p_k(x)$ se heurte au fléau de la dimensionnalité. Le plus souvent, on a moins de 10^3 exemples labélisés par classe. Or, pour contrôler l'erreur, il faudrait un nombre d'exemples exponentiel en la dimension d [6], à moins que les données n'appartiennent à un sous-espace de basse dimension. D'une facon générale toutes les méthodes de calculs locaux dans $\mathbb{R}^d$ sont inopérantes, sauf si les données sont dans un ensemble régulier de basse dimension, ce qui est hélas rarement le cas. Pourtant, si nos ancêtres avaient eu besoin de renconter 10^{10} mammouths pour en reconnaître un, ils auraient eu peu de chance de survivre. De fait, les expériences psychophysiques montrent que quelques images sont suffisantes pour apprendre à reconnaitre un animal en quasiment toutes circonstances.

L'apprentissage non-supervisé permet de résoudre cette apparente contradiction. On utilise des exemples non-labélisés $\{x_i\}_i$, n'incluant par forcément des exemples de la classe que l'on veut apprendre, mais d'autres qui ont des caractéristiques similaires [8, 3]. Des millions d'images ou de sons non-labélisés sont ainsi diponibles sur Internet, il existe aussi beaucoup d'archives d'examens médicaux sans diagnostic, d'enregistrements sismiques sans analyse géophysique, et ainsi de suite. Le problème est de "synthétiser" cette information afin de se préparer à un nouveau problème de reconnaissance avec peu d'exemples labélisés.

Pour apprendre avec peu d'exemples, il faut contourner la malédiction de la dimensionalité, ce qui est possible avec un classificateur global linéaire. Cela consiste à séparer deux classes C_k et $C_{k'}$ avec un hyperplan. Cependant, deux classes compliquées sont rarement séparées par un hyperplan. L'idée est donc d'optimiser une représentation $\Phi x \in \mathbb{R}^{d'}$, pour qu'une telle séparation linéaire génère une faible erreur. La difficulté de la classification est donc reportée sur le choix de Φ. Il s'agit de comprendre comment construire un opérateur Φ, qui agrège des informations "générales" sur les structures du monde, à partir

d'exemples non labélisés, et permet de quasiment séparer linéairement des classes que l'on ne connaît pas à l'avance.

Avant de considérer cette question très difficile, revenons au classificateur linéaire. Si w est un vecteur orthogonal à l'hyperplan séparateur, alors le classificateur définit un seuil $b \in \mathbb{R}$ et associé à x la classe

$$\hat{y}(x) = \begin{cases} k & \text{si } \langle \Phi x, w \rangle \geq b \\ k' & \text{si } \langle \Phi x, w \rangle < b \end{cases} . \tag{1}$$

On peut voir Φx comme un vecteur de d' caractéristiques de x. Le choix de $w \in \mathbb{R}^{d'}$ optimise un "vote" qui agrège linéairement ces indicateurs pour prendre une décision :

$$\langle \Phi x, w \rangle = \sum_{j=1}^{d'} \Phi x(j)\, w(j) .$$

Différentes approches comme les Support Vector Machines optimisent w et b en minimisant les taux d'erreurs sur les exemples d'entraînement, tout en régularisant l'estimation avec une pénalisation convexe [6].

On pourrait imaginer que ce problème est simple, car il existe toujours un hyperplan qui sépare les n exemples d'entraînement Φx_i s'ils sont linéairement indépendants en dimension $d' \geq n$. Il suffit donc que la dimension d' soit suffisamment grande. Cependant, cette erreur nulle à l'entrainement ne se généralise pas, au sens où l'erreur de classification sera importante sur d'autres exemples. Cela vient du fait que la complexité de la classification, mesurée par la dimension de Vapnick-Chernovenkis [6], augmente avec d'. L'optimisation de la représentation $\Phi x \in \mathbb{R}^{d'}$ est donc un problème complexe, au coeur de la classification.

5 Réduction de dimensionalité

Certes, x appartient à un espace de grande dimension $\mathbb{R}^d$, mais on peut espérer pouvoir résoudre un problème de classification en approximant x sur un ensemble de basse dimension d', ce qui permetrait de contourner la malédiction. Il existe diverses approches pour effectuer cette réduction de dimensionalité, qui utilisent de beaux résultats à l'interface de l'analyse harmonique, de la géométrie et des probabilités. Hélas, ces hypothèses de faible dimension sont rarement satisfaites. Il faut bien le comprendre pour ne pas sous-estimer les problèmes mathématiques de la grande dimension.

Supposons que x appartienne à une variété régulière de basse dimension d'. L'estimation d'une variété est basée sur des calculs de distances locales [1]. Ces distance ne sont significatives qu'en dimension d' typiquement plus petite que 10, à cause du fléau de la dimensionalité. On peut alors calculer les coordonnées intrinsèques de x sur cette variété en décomposant x sur les d' vecteurs propres du Laplacien sur la variété. On estime ce Laplacien à partir d'un graphe de voisinage calculé sur les exemples [1]. Ces techniques ont trouvé de belles applications pour caractériser des systèmes dynamiques de basse dimension.

Le compressed sensing est une autre approche de réduction de dimensionalité. Si x a une représentation parcimonieuse comme combinaison linéaire de p vecteurs quelconques d'une base, alors il peut être caractérisé par $d' = O(p \log p)$ produits scalaires avec des vecteurs aléatoires [5]. On suppose donc ici que x appartient à une union d'espaces linéaires de basse dimension p. L'utilisation de mesures aléatoire permet ainsi d'obtenir d' descipteurs linéaires qui characterisent une variété complexe de dimension p.

Cependant, la réduction brutale de dimensionalité a été la source d'un grand nombre d'impasses en recherche algorithmique. Jusque dans les années 2000, on a le plus souvent essayé de résoudre les

problèmes de classification en calculant un petit nombre de caractéristiques discriminantes. Pour détecter la présence d'un visage dans une image, il est ainsi naturel d'essayer de localiser la tête, les yeux, le nez, la bouche, par exemple avec des ellipses de tailles et de positions variables. Ces ellipses ont des paramètres dont les distributions de probabilité dépendent de la variabilité géométrique des visages. Si on n'est pas trop naïf sur la façon de détecter et d'ajuster la position de ces ellipses, on peut espérer obtenir de bons résultats. Je l'ai fait et je me suis trompé comme beaucoup.

Pour la détection de visages, il est plus efficace d'agréger progressivement beaucoup plus d'informations sur l'image. Une technique développée en 2001 [12] consiste à définir un très grand nombre de tests "faibles", qui sont des seuillages de produits scalaires de x avec des vecteurs de Haar de supports variables. Ces tests sont agrégés par combinaisons linéaires, pour définir des tests plus "forts" qui sont utilisés dans un arbre de décision. L'apprentissage à partir d'exemples labélisés choisit les meilleures combinaisons linéaires, à chaque noeud de l'arbre de décisions, avec un algorithme de boosting [12]. Cette cascade de tests permet de détecter précisément des visages malgré la grande variabilité des images. Cependant, il faut disposer d'un grand nombre d'exemples labélisés pour apprendre les paramètres de la cascade. Ceci est possible pour la détection de visages et cet algorithme est implémenté dans de nombreux appareils photos.

Cette architecture utilise des agrégations linéaires successives de classificateurs, qui prennent des "décisions" partielles et mélangent de plus en plus de variables. Ces résultats montrent que l'agrégation d'un grand nombre indicateurs de faible expertise, mais très divers, permet de répondre à des questions complexes. C'est une idée clef des algorithmes de classification en grande dimension. Malgré certaines avancées en statistique de la décision [11], on comprend encore mal les propriétés mathématiques de ces agrégations hiérarchiques.

6 Les bénéfices de la grande dimension

Il est étrange de se plaindre de la malédiction de la dimensionalité alors que l'on cherche à augmenter la résolution et donc la dimension des données, que ce soit des images, des données bio-médicales, géophysiques, économiques... Plus d'information devrait permettre de prendre des décisions plus fiables. Pour éviter la malédiction, cela demande cependant d'éliminer la variabilité inutile et de construire des invariants stables et discriminant.

Dans un cadre probabiliste, une classe C_k est modélisée par processus X_k dont chaque réalisation x est un élément de C_k. Pour différentier des classes C_k et $C_{k'}$ on définit une représentation Φx de façon à pouvoir trouver une combinaison linéaires $\langle \Phi x, w \rangle = \sum_{j=1}^{d'} \Phi x(j)\, w(j)$, qui soit quasiment invariante mais différente sur C_k et sur $C_{k'}$. Les invariants sur C_k peuvent s'écrire comme des espérances $E(H(X_k))$ où H est une fonctionelle typiquement non-linéaire. Le coefficient $\langle \Phi x, w \rangle$ peut donc s'interpréter comme l'estimateur d'une telle espérance. Cela indique l'existence d'une forme d'ergodicité qui permet de remplacer l'espérance par une combinaison linéaire des coefficients de Φx. Pour résoudre de multiples problèmes de classification, il faut que Φx ait le potentiel de créer de nombreux invariants différents par combinaisons linéaires. Cela doit donc être un vecteur de grande taille. La réduction de dimensionalité se fait lors de la projection sur le vecteur w. L'optimisation de w revient à choisir un invariant adapté à chaque problème de classification spécifique.

Prenons un example simple où les données $x \in \mathbb{R}^d$ sont les réalisations d'un vecteur aléatoire X_k dont les coordonnées sont des variables aléatoires indépendantes de même densité de probabilité $p_k(u)$. Ces données vivent dans la quasi-totalité de $\mathbb{R}^d$ mais l'information encapsulée par ces données se résume à la densité p_k. Celle-ci peut s'estimer avec un histogramme Φx. Plus d est grand plus cet histogramme sera

précis, d'où l'intérêt de la grande dimension. On peut savoir si cet histogramme correspond à p_k ou à $p_{k'}$ par example en testant le signe du produit scalaire avec $w = p_k - p_{k'}$. Evidemment cet exemple est un cas d'école trop simple car n'inclut pas de structure de correlation entre les données.

Les processus stationaires fournissent un deuxième exemple déjà plus intéressant. Les textures, visuelles ou auditives, peuvent être modélisées comme des réalisations de processus stationaires, qui vivent dans un espace de grande dimension. La stationnarité exprime l'invariance de la distribution du processus sur le groupe des translations. Les moments du second ordre $E(X_k(u_1)\, X_k(u_2))$ ne dépendent que de $u_2 - u_1$ et la transformée de Fourier définie la puissance spectrale. Avec une hypothèses faible de décorrélation, les moments du second ordre peuvent être estimés par moyennage temporel, à partir d'une seule réalisation. Cette information n'est cependant souvent pas suffisante car des textures totalement différentes ont les mêmes moments d'ordre deux. On peut différencier des processus stationnaires avec des moments d'ordre supérieur, qui offrent des invariants plus riches. Hélas, les espérance de puissances supérieures sont difficilement estimables par moyennage temporel sur une seule réalisation, car les estimateurs résultants ont une trop grande variance et sont donc imprécis. Dans ce cadre de processus stationnaires, on voit la difficulté de calculer des invariants suffisamment discriminant.

Il y a évidemment d'autres sources de variabilités au-delà du groupe des translations. Par exemple, les textures sont déformées par les effets de perspectives sur les surfaces tri-dimensionelles. Il faut alors construire des invariants relativement à l'action de difféomorphismes, autrement dit de groupes de Lie non-commutatifs de grande dimension. Les outils rigides comme la transformée de Fourier ne sont plus adaptés, ce qui ouvre des nouveaux problèmes d'analyse harmonique. De tels invariants peuvent être construits en cascadant des transformées en ondelettes définies sur le groupe de Lie, avec des non-linéarités contractantes [9]. Des similarités frappantes apparaissent avec les architectures des réseaux de neurones profonds [2].

Il n'y a pas de bon modèle probabiliste pour des classes complexes d'objets à la fois structurés et très variables comme des chaises, les prononciations différentes d'un mot ou les images de poumons. Les distributions sont typiquement non-Gaussiennes, non-Markoviennes, pas multifractales, enfin rien de ce que l'on sait manipuler mathématiquement. L'enjeu est de construire des représentations Φx permettant de calculer des invariants multiples, malgrés l'absence de modèle probabiliste. Les algorithmes récents d'apprentissage non-supervisé sont partiellement capables d'apprendre de telles représentations [3]. Parmi eux, les réseaux de neurones profonds cascadent des opérateurs contractants, dont les paramêtres sont estimés avec des grandes bases de données [2]. La parcimonie joue un rôle important mais mystérieux dans tous ces algoritmes d'apprentissage, avec des résultats encourageant sur divers problèmes de big data.

L'état de l'art algorithmique souffre cependant de beaucoup de limitations. La quasi absence de feedback dans les architectures d'apprentissages, pourtant très présents dans le cerveau [10], limite l'adaptivité des représentations. Malgrés des résultats expérimentaux prometteurs, l'apprentissage en grande dimension reste un champs mathématique totalement ouvert, avec de nombreuses ramifications notamment en probabilité, analyse harmonique, géométrie, théorie des groupes et systèmes dynamiques.

7 Rêvons un peu

Au-delà des applications, les problématiques de big-data ouvrent des questions profondes sur la grande dimension. On peut maintenant expérimenter numériquement, et donc tester de nombreuses approches. L'élaboration d'outils mathématiques adaptés aura probalement un impact bien au-delà du traitement des données. Il est notamment possible que cela aide à mieux comprendre certains principes du traitement

de l'information sensorielle par le cerveau. En effet, le cerveau est une machine extraordinairement efficace pour le traitement de données gigantesques.

L'analyse en grande dimension est aussi un défi en physique. En particulier, l'analyse de la turbulence tridimensionelle reste un sujet ouvert, malgrés les nombreux travaux de recherche qui ont suivi les résultats de Kolmogorov sur le décroissance du spectre de Fourier. Pour les grands nombres de Reynolds, l'équation de Navier-Stokes définit un système dynamique ayant un grand nombre de degrés de liberté, qui génère des structures complexes. Comprendre les propriétés de ces structures et leurs dynamiques est un problème d'analyse en grande dimension, pour lequel on manque toujours d'outils mathématiques.

On peut espérer que tous ces problèmes de grande dimension se rapprochent à terme, au travers d'une meilleure compréhension des mathématiques sous-jacentes. C'est en tout cas une raison de plus pour encourager des jeunes mathématiciens à travailler dans ce domaine.

Références

[1] M. Belkin, and P. Niyogi, "Laplacian Eigenmaps for dimensionality reduction and data representation". Neural Computation, 15 (6) : 1373-1396, 2003.

[2] Y. Bengio, A. Courville, P. Vincent, "Representation Learning : A Review and New Perspectives", IEEE Trans. on PAMI, 2013.

[3] Y-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. "Learning Mid-Level Features For Recognition". In IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[4] J. Bouvrie, L. Rosasco, T. Poggio : "On Invariance in Hierarchical Models", NIPS 2009.

[5] E. J. Candès, J. Romberg and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," Comm. Pure Appl. Math., 59 1207-1223, 2006.

[6] T. Hastie, R. Tibshirani, J. Friedman, "The elements of statistical learning", Springer, 2001.

[7] Q. Le, M.A. Ranzato, R. Monga, M. Devin, G. Corrado, K. Chen, J. Dean, A. Ng, "Building high-level features using large scale unsupervised learning,"Proc. ICML 2012.

[8] Y. LeCun, K. Kavukvuoglu and C. Farabet : "Convolutional Networks and Applications in Vision", Proc. of ISCAS 2010.

[9] S. Mallat "Group Invariant Scattering", Communications in Pure and Applied Mathematics, vol. 65, no. 10. pp. 1331-1398, October 2012.

[10] Y. Liu, J.J. Slotine, and A. Barabasi "Controllability of Complex Networks," Nature, 473(7346), 2011.

[11] A. Tsybakov, "Optimal aggregation of classifiers in statistical learning", The Annals of Statistics, vol. 32, no. 1, 135-166, 204.

[12] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", Proc. IEEE CVPR, 2001