

Rapport projet long
Équipe HAL9000

Hamelain Christian, Hasan Pierre-Yves, Gaspart Quentin, Trestour Fabien

9 juin 2016

Table des matières

I	Introduction	2
0.1	Contexte	3
0.2	Présentation du sujet	3
0.3	Déroulement	3
II	Le Perceptron	4
1	Architecture d'un perceptron	5
1.1	Introduction	5
1.2	La brique de base : le neurone	5
1.3	Organisation des couches	7
2	Algorithmes d'apprentissage	10
2.1	L'objectif des algorithmes d'apprentissage	10
2.2	La méthode de rétropropagation du gradient	11
3	Influence des paramètres sur l'apprentissage.	13
III	Machines de Boltzmann	14
4	Introduction	15
5	Les machines de Boltzmann restreintes	16
5.1	Principe	16
5.2	Les lois d'apprentissage	17
6	Les Deep Belief Networks	18
7	Aller plus loin	19
7.1	La visualisation des spectres du réseau	19
A	La base MNIST	21
A.1	Présentation de la base de données	21
A.2	Limites de cette base	21
B	Présentation du code	24

Première partie

Introduction

0.1 Contexte du projet

CentraleSupélec propose à ses étudiants en deuxième année de participer à des projets longs qui se réalisent en équipe, tout au long de l'année sur un thème mêlant plusieurs compétences que l'ingénieur se doit de posséder.

Dans notre cas, c'est l'élaboration de réseaux neuronaux et les méthodes de Deep Learning que 12 élèves ont étudié, répartis en 3 groupes, dont HAL9000.

Ce rapport présente les avancements de ce dernier groupe sur le sujet tout au long de l'année.

0.2 Présentation du sujet

Ce projet aborde le sujet du Deep Learning. Cette méthode spécifique de machine learning est dérivée du concept de réseau de neurones.

Les réseaux de neurone sont une modélisation simple du fonctionnement cérébral à l'échelle cellulaire. Cette approche a vu le jour avec les études de McCulloch et Pitts dès la fin des années 50. Malgré certains écueils dans les années 70, l'approche connexionniste des réseaux de neurones a su se développer et devenir un sujet de recherche populaire dans les dernières années, notamment grâce aux capacités d'adaptabilité et de généralisation de ces réseaux qui en font d'excellents candidats pour des applications telles que la reconnaissance d'image ou la classification.

0.3 Déroulement du projet

Tout d'abord, il s'agissait pour nous de comprendre le fonctionnement de ces structures et commencer à coder des structures élémentaires afin de comprendre les enjeux du deep learning. Nous avons pour cela travaillé avec la base de données MNIST de Yann LeCun et en JAVA. Afin de travailler en groupe de manière efficace, nous avons aussi utilisé l'outil GIT.

Par la suite nous nous sommes intéressés aux architectures profondes, chaque groupe se penchant sur une architecture spécifique. Notre équipe s'est en particulier intéressée aux machines de Boltzmann, en commençant par les machines de Boltzmann restreintes (RBM), puis la structure de Deep Learning associée, les Deep Belief Networks.

Cette étude a été encadrée par Joanna Tomasik et Arpad Rimmel, enseignants à CentraleSupélec, campus de Gif-sur-Yvette.

Deuxième partie

Première approche du problème : le Perceptron

Chapitre 1

Architecture d'un perceptron

1.1 Introduction

Ici, l'objectif était la reconnaissance des caractères de la base de données MNIST grâce à un perceptron.

Le perceptron fait partie des architectures de réseaux neuronaux les plus simples. Son étude nous a donc permis de s'introduire à la problématique du machine learning avant d'approfondir en étudiant des structures plus complexes.

1.2 La brique de base : le neurone

Le neurone est le composant élémentaire des réseaux neuronaux. Il est une modélisation du fonctionnement des neurones du système nerveux humains.

Chaque neurone reçoit un signal via une entrée, qui correspond aux dendrites des systèmes biologiques. Le neurone prend en compte la valeur de toutes ses entrées et en déduit la valeur de sortie. Cette sortie est ensuite propagée par le biais d'un axone vers un autre neurone.

La sortie du $j^{\text{ième}}$ neurone est donnée par la formule :

$$s_j(x) = f\left(\sum_{k=1}^N w_{j,k} * x_k + w_0\right) \quad (1.1)$$

où :

- s est la valeur de la sortie
- f est la fonction d'activation.
- N est la dimension du vecteur d'entrée.
- $w_{j,k}$ est la $k^{\text{ième}}$ composante du vecteur de poids w_j du $j^{\text{ième}}$ neurone. w_0 est le biais du neurone. Cette valeur correspond au poids d'une entrée fictive valant toujours 1.
- x_k est la $k^{\text{ième}}$ composante du vecteur d'entrée x .

L'entrée x est un vecteur défini par un ensemble de caractéristiques que l'on choisit pour représenter les données d'entrées. Dans le cas d'une image par exemple, ce vecteur d'entrée peut être composé de la valeur de tous les pixels

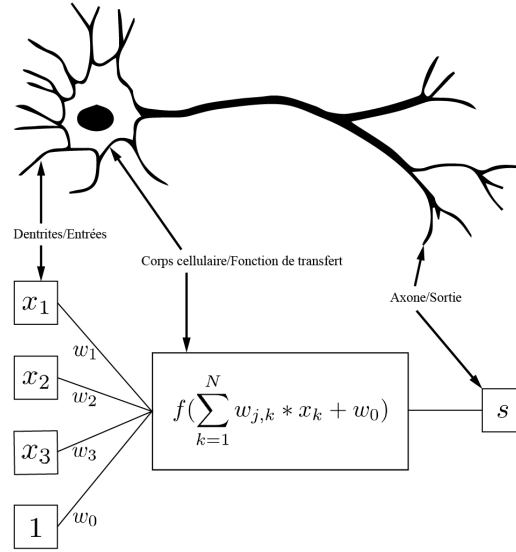


FIGURE 1.1 – Analogie entre neurone biologique et neurone formel.

de l'image. D'autres représentations moins triviales peuvent être choisies afin de réduire la quantité d'information à traiter et de s'approcher au mieux des données significatives de la problématique traitée.

La fonction d'activation permet de définir le comportement du neurone. Selon la définition de cette fonction on aura une sortie à valeurs discrètes ou continues, centrées en 0 ou en 0,5. Le choix de la fonction est donc étroitement lié avec le problème à traiter. Il existe différents types de fonctions d'activation. Parmi les plus utilisées figurent :

— La fonction échelon

$$f(x) = \mathbf{1}_{\mathbb{R}_+^*} \quad (1.2)$$

— Les fonctions linéaires

$$f(x) = \alpha * x + \beta \quad (1.3)$$

— La fonction sigmoïde

$$f(x) = \frac{1}{1 + e^{-\lambda * x}} \quad (1.4)$$

— La fonction tangente hyperbolique

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (1.5)$$

Le vecteur de poids du $j^{\text{ième}}$ neurone représente la pondération de chacune des entrées de ce neurone. C'est ce paramètre qui permet de modifier le réseau de neurone sans avoir à modifier son architecture. Toutes les propriétés d'un réseau neuronal sont donc issues de ce vecteur de pondération et de ses variations.

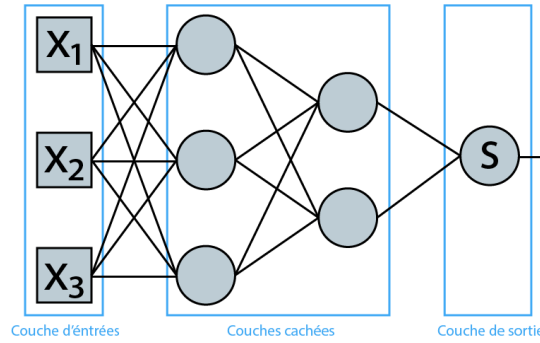


FIGURE 1.2 – Architecture d'un perceptron multi-couches.

1.3 Organisation des couches

Il existe de nombreuses architectures de réseau de neurone. Elles peuvent être récurrentes ou non, entièrement connectées ou seulement partiellement, organisées en couches, ... De nombreuses propriétés permettent de caractériser une architecture de réseau de neurone.

Le Perceptron est un modèle assez élémentaire de réseau. Il est constitué en couches totalement connectées.

On peut distinguer trois types de couches : la couche d'entrée, les couches cachées et la couche de sortie.

La couche d'entrée est assez élémentaire. C'est tout simplement une couche dont la valeur sera le vecteur d'entrée x . Cette couche doit donc être constituée d'autant de neurones que le vecteur d'entrée a de dimensions. Les neurones de cette couche ont pour fonction d'activation l'identité.

Les couches cachées sont celles qui effectuent les calculs. Les principales propriétés du réseau sont héritées de cet empilement de couches. On comprend alors mieux l'intérêt actuel pour le Deep Learning, c'est-à-dire l'apprentissage des réseaux avec un grand nombre de couches cachées. Le nombre de neurones dans chaque couche et les fonctions d'activation utilisées par les neurones sont choisies par le concepteur du réseau selon le problème traité par le réseau.

Afin d'explicitier l'importance de l'architecture du réseau, abordons l'exemple usuel du ou exclusif.

Essayons de réaliser avec un unique neurone la fonction logique XOR. Ce neurone aura deux entrées x_1 et x_2 à valeurs dans $\{0;1\}$ et une sortie s , elle aussi à valeurs dans $\{0;1\}$. On prendra comme fonction d'activation la fonction de Heaviside, c'est à dire $\mathbf{1}_{\mathbb{R}_+^*}$. Ce neurone aura par conséquent un fonctionnement totalement binaire. Il s'agit donc ici de déterminer les pondérations w_1 et w_2 , respectivement associées aux entrées x_1 et x_2 , qui conviennent pour obtenir en sortie la valeur $x_1 \oplus x_2$.

Pour un tel problème, le neurone agit comme un séparateur linéaire de l'espace des entrées. L'équation de la droite séparant le demi-espace définit par

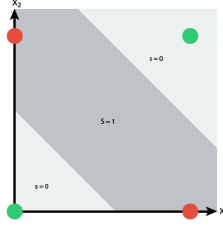


FIGURE 1.3 – Fonction XOR à réaliser par le perceptron.

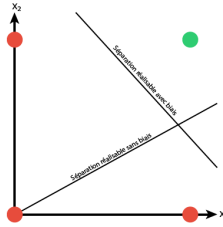


FIGURE 1.4 – Intérêt de l'introduction d'un biais.

$s = 0$ de celui défini par $s = 1$, découle alors directement de l'équation 1.1 :

$$w_2 * x_2 + w_1 * x_1 + w_0 = 0 \quad (1.6)$$

On constate ici l'intérêt du biais, qui permet de réaliser des séparations affines de l'espace des entrées et pas seulement des fonctions linéaires.

Les limites d'un neurone seul sont alors évidentes : toutes les séparations de l'espace des entrées ne sont pas affine, et le cas du XOR en est déjà une illustration.

Il est alors nécessaire d'introduire la structure de réseau. En prenant une couche d'entrée, une couche cachée et une couche de sortie, respectivement composées de deux, deux et un neurone, on peut contourner le problème sus-cité. Cette structure permet en fait de générer deux séparations de l'espace des entrées.

La séparation que l'on cherche à effectuer est en effet de la forme :

$$f(x_1, x_2) = 0 \Leftrightarrow \begin{cases} x_1 + x_2 - 0,5 < 0 \\ x_1 + x_2 - 1,5 > 0 \end{cases} \quad (1.7)$$

Ce premier exemple simpliste manifeste donc la nécessité d'adapter la structure du réseau au problème traité.

De manière plus générale, le nombre de neurones du réseau fait varier le nombre de poids du réseau, et donc la capacité du réseau à séparer des ensembles

multiples et complexes. Ainsi, un réseau sous-dimensionné mène à des résultats pas assez précis et un réseau sur-dimensionné mène à du sur-apprentissage. On a donc environ la loi suivante :

$$N < \frac{1}{10} * T * \dim(s) \quad (1.8)$$

- N est le nombre de neurones dans le réseau.
- T est le nombre de vecteurs de la base d'apprentissage.
- s est le vecteur de sortie.

De plus, le nombre de neurones dans une couche doit être de moins de trois fois celui de la couche précédente.

Ces deux approximations donnent donc une première idée de la structure à adopter pour un réseau de neurones.

Chapitre 2

Algorithmes d'apprentissage

2.1 L'objectif des algorithmes d'apprentissage

L'objectif des algorithmes d'apprentissage est simple : optimiser les poids du réseau de neurones afin qu'il "apprenne".

Les poids sont en effet la seule variable disponible après avoir défini l'architecture du réseau (il existe des algorithmes permettant de modifier l'architecture durant l'apprentissage mais on ne les étudiera pas ici). On va donc les modifier progressivement, après avoir présenté au réseau un nombre fixé d'exemples, afin que le réseau réagisse au mieux aux stimulations qui lui sont présentées.

Il existe en réalité deux types d'algorithmes d'apprentissage.

Le premier cas est celui de l'apprentissage supervisé. Il nécessite de connaître "l'étiquette" des exemples entrés dans le réseau pour effectuer l'apprentissage.

On observe ainsi quelle est la réponse proposée par le réseau après propagation de l'entrée et on la compare à l'étiquette, correspondant à la réponse attendue. On peut alors "récompenser" les comportements positifs en renforçant les connections mises en jeu lors de cette propagation, ou "punir" les mauvais comportements en affaiblissant ces dernières.

On s'appuie alors sur une classification a priori des données : les classes sont créées par le concepteur du problème et du réseau. Cela est à double tranchant. En effet, les classes choisies seront a priori plus pertinentes vis-à-vis du résultat attendu mais pas nécessairement vis-à-vis des données et de la structure du réseau.

Le second cas est celui de l'apprentissage non supervisé. Comme l'indique son nom, ce genre d'apprentissage n'a pas besoin d'une intervention extérieure pour apprendre. En pratique cela se manifeste par l'absence d'étiquettes dans les bases d'apprentissage.

Ce genre d'apprentissage revient plus ou moins à mettre en concurrence les différentes classes de données à chaque présentation d'exemple. Ainsi, la classe qui correspond la mieux à l'entrée devient la sortie et la classe est modifiée pour ressembler encore plus à l'entrée qui a permis l'activation.

Au final, dans ce cas, la définition de chaque classe n'est pas explicite : c'est

le réseau qui détecte lui même une structure de classe et il l'amplifie durant l'apprentissage afin d'améliorer la classification qu'il a détecté. Les classes ne sont alors pas forcément celles que l'on attend mais plutôt les classes les plus "naturelles" étant données la base d'apprentissage et l'architecture du réseau. En ce sens les algorithmes d'apprentissage s'approchent de la notion de "clustering".

2.2 La méthode de rétropropagation du gradient

La rétropropagation du gradient correspond à une descente de gradient dans l'espace des poids.

Le principe de l'algorithme est en fait d'assigner à chaque neurone sa responsabilité dans l'erreur commise dans le calcul de la sortie. Les poids individuels des neurones sont ainsi corrigés en fonction de cette responsabilité.

2.2.1 Initialisation des poids

Les valeurs sont initialisées au départ de manière aléatoire, avec des valeurs centrées en 0.

Le choix de l'intervalle possible est très important : un intervalle trop petit va beaucoup ralentir l'apprentissage tandis qu'un intervalle trop grand donne des données trop éparpillées qui rendent plus difficile l'apprentissage.

Ainsi, l'intervalle choisi est le suivant :

$$\forall i, w_i \in \left[\frac{-2.4}{N_i}, \frac{2.4}{N_i} \right] \quad (2.1)$$

où N_i est le nombre d'entrées du neurone.

2.2.2 Calcul des variations de poids

Les exemples sont présentés consécutivement au perceptron, et à chaque nouvel exemple on calcule les nouvelles variations de poids.

Une fois l'entrée propagée dans le réseau, on obtient une sortie $y^{obtenue}$. Notons par ailleurs la sortie théorique $y^{theorique}$ et h le produit scalaire entre les poids w et les entrées x . Nous pouvons alors calculer l'erreur commise par le réseau, pour chaque neurone de sortie :

$$\forall i, e_i^{sortie} = f'(h_i^{sortie}) * (y_i^{theorique} - y_i^{obtenue}) \quad (2.2)$$

On propage ensuite l'erreur de la couche n vers la couche $n - 1$ en assignant à chaque neurone sa responsabilité dans l'erreur commise e^{sortie} :

$$\forall j, e_j^{(n-1)} = f'(h_j^{(n-1)}) * \sum_k w_{i,j} * e_i^{(n)} \quad (2.3)$$

Une fois le calcul de l'erreur rétropropagé pour toutes les couches du réseau, il ne reste plus qu'à assigner les différentes variations de poids selon un taux d'apprentissage λ :

$$\Delta w_{i,j}^{(n)} = \lambda e_i^{(n)} x_j^{(n-1)} \quad (2.4)$$

2.2.3 Potentielle modification de l'algorithme

L'une des modifications possibles de cet algorithme est l'ajout d'une inertie à la modification des poids.

La variation de poids devient alors :

$$\Delta w_{i,j}^{(n)}(t) = \lambda e_i^{(n)} x_j^{(n-1)} + \alpha \Delta w_{i,j}^{(n)}(t-1) \quad (2.5)$$

Le coefficient α est le coefficient d'inertie. Il est dans l'intervalle $[0;1]$ et ajoute ainsi une influence des variations de poids précédentes sur les variations de poids actuelles.

Cette technique permet d'éviter certains minima locaux et de se stabiliser dans des configurations plus optimales. Cela peut aussi accélérer la convergence de l'apprentissage si le coefficient α est bien choisit.

Chapitre 3

Influence des paramètres sur l'apprentissage.

Troisième partie

Approfondir le sujet : les
machines de Boltzmann.

Chapitre 4

Introduction

Les réseaux de Boltzmann sont des réseaux qui ont un fonctionnement particulier que nous allons présenter dans ce chapitre. Les machines de Boltzmann constituent des réseaux neuronaux. La principale différence entre les réseaux neuronaux comme les réseaux neuronaux de convolution (CNN) et les réseaux de Boltzmann va être le fait qu'un réseau de Boltzmann n'est pas *feed-forward* ; comme l'était le perceptron, précédemment présenté dans ce rapport.

Ce rapport traitera d'abord des machines de Boltzmann restreintes en les présentant et en expliquant les lois qui régissent leur fonctionnement pour ensuite s'intéresser aux Réseaux de Boltzmann Profond.

Chapitre 5

Les machines de Boltzmann restreintes

Les machines de Boltzmann restreintes, ou RBM, ont d'abord été pensées par P. Smolensky (1986) mais réellement mise en oeuvres et étudiées par G. Hinton(2006). Nous allons présenter dans cette partie leur constitution et leurs fonctionnement.

5.1 Principe

Les RBM sont des réseaux neuronaux qui donnent une probabilité d'activation de chaque neurone.

Une RBM possède deux couches de neurones : l'une sera qualifiée de couche visible et l'autre de couche cachée. La couche visible correspond à l'entrée de notre réseau, ce sont les neurones de la couche d'entrée. On va leur assigner les valeurs des exemples de notre jeu de données pour entraîner le réseau.

Au sein d'une même couche, aucun neurone n'est relié à un autre, c'est ce caractère restreint qui donne sa particularité et son nom à ce type de réseau de Boltzmann. Cependant, chaque neurone de la couche visible (respectivement cachée) est connecté à tous les neurones de la couche cachée (respectivement visible).

On forme alors un graphe, non orienté, qui est un modèle de la distribution de probabilité. Chaque neurone est une variable aléatoire qui prend sa valeur dans 0; 1. Le fait qu'aucun neurone v ne soit connecté aux autres de la couche traduit le fait que les variables aléatoires associées sont indépendantes.

La probabilité qu'un neurone v_i de la couche visible soit activé est On introduit maintenant l'échantillonnage de Gibbs, qui est un algorithme de la classe des Metropolis-Hasting. Cet algorithme est un Monte Carlo d'une chaîne de Markov dont l'idée principale est d'échantillonner les variables en se basant sur la valeur des autres variables. C'est une marche aléatoire sur les chaînes de Markov.

A expliquer : MRF et les propriété \rightarrow chaîne de Markov Particularité et donc convergence \rightarrow contrastive divergence.

La particularité de ce réseau est donc le fait qu'il n'est pas *feed-forward*, même si l'on considère la couche cachée comme entrée du réseau.

5.2 Les lois d'apprentissage

Un point critique de l'implémentation d'une RBM est comme bien souvent la loi d'apprentissage. Si l'on trouve dans la littérature plusieurs solutions, la forme générale du calcul est la suivante :

$$\Delta w_{i,j} = \alpha (< v_i h_j >_{data} - < v_i h_j >_{model}) \quad (5.1)$$

où α est le *learning rate*

Dans ce calcul, on distingue l'échantillonnage de la machine à partir des données (*data*), de l'échantillonnage de la machine après la reconstruction par pas de Gibbs, $< v_i h_j >_{model}$.

Pour extraire des informations à partir de l'*input*, il est important de ne pas reconstruire les deux couches depuis les *inputs*. Ainsi le calcul du produit $< v_i h_j >_{data}$ s'obtient par l'état de l'entité visible v_i et l'évaluation de l'entité cachée sous forme réelle $:p(h_j|v^{(0)})$.

Chapitre 6

Les Deep Belief Networks

Chapitre 7

Aller plus loin

7.1 La visualisation des spectres du réseau

Lors du développement de la RBM, nous avons été confrontés au problème de la quantification de l'apprentissage. Comment savoir si oui ou non le réseau s'adapte aux exemples qu'on lui a présenté ?

Suite à l'absence de réelle mesure de cette notion, l'idée nous est venue d'implémenter une mesure bien plus qualitative du phénomène : l'objectif était de visualiser l'importance que le réseau accorde à chaque pixel de l'image.

Dans le cas de la RBM, cela correspond tout simplement à l'ensemble des pondérations des connections entre la couche visible et la couche cachée du réseau. Ainsi, à chaque neurone de la couche cachée, on peut associer une image correspondant aux dites pondérations. Ce sont ces images que l'on qualifie de "spectres".

Ces spectres représentent les caractéristiques que le réseau essaie de détecter dans une image pour y reconnaître un caractère. Selon le nombre de neurones dans la couche cachée, ces caractéristiques sont assez différentes. En effet, plus il y a de neurones cachés, plus le réseau pourra détecter de nombreuses caractéristiques précises, qui une fois assemblées formeront la représentation du



FIGURE 7.1 – Spectre d'une RBM entraînée à vingt neurones cachés.

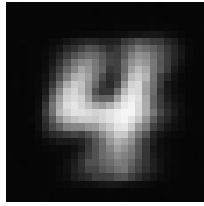


FIGURE 7.2 – Spectre d'une RBM ayant apprise la "forme" 4.

caractère complet. Ces caractéristiques peuvent être, par exemple, des angles ou des boucles. A contrario, quand le réseau a peu de neurones cachés, ces spectres sont beaucoup plus proches de la forme "usuelle" du caractère associé. Ainsi, avec un seul neurone caché, on obtient très exactement la représentation d'un chiffre idéal pour ce réseau.

L'intérêt d'une telle initiative est simple : utiliser l'intuition visuelle qu'on a des chiffres pour juger de la qualité de l'apprentissage.

Au début de l'apprentissage, l'image sera exclusivement composée de bruit blanc (c'est une conséquence de l'algorithme d'apprentissage), tandis qu'à la fin elle ressemblera au négatif d'un des exemples de la base de donnée (toujours dans le cas d'un réseau à un seul neurone caché).

Annexe A

La base MNIST

A.1 Présentation de la base de données

La base de données MNIST, Mixed Mixed National Institute of Standards and Technology, est un ensemble de chiffres manuscrits extraits d'une base antérieure, NIST. La base est organisée de la manière suivante :

- La base d'apprentissage est constituée de 60000 images.
- La base de test est constituée de 10000 images.
- Chaque image est constituée de 28*28 pixels en nuances de gris.
- Chaque image est normalisée.

Cette base est aujourd'hui devenue une référence permettant de tester les algorithmes d'apprentissages ou de reconnaissance de forme divers et variés. Cet aspect facilite ainsi la comparaison de nos résultats avec ceux d'autres personnes ayant fait la même démarche que nous.

On peut trouver cette base à l'adresse : <http://yann.lecun.com/exdb/mnist/>

A.2 Limites de cette base

Cette base, malgré tous ses avantages, n'est cependant pas parfaite.

En effet, une première limite apparaît assez rapidement quand on travaille avec MNIST : cette base n'est pas "universelle". Les caractères utilisés dans la base de donnée sont fortement connotés anglosaxons. On peut constater cet aspect dans les spectres réalisés par nos RBM.

L'image idéale de chaque chiffre a toujours une prédominance américaine, et quand il s'agit ensuite d'entrer des données européenne, la limitation de la base se révèle.

Une autre limite apparaît aussi : c'est celle de la simplicité de la base de donnée.

La reconnaissance de caractères est en effet une tâche simpliste en comparaison avec certains des problèmes auxquels peuvent être confrontés les réseaux



FIGURE A.1 – Exemples de chiffres de la base de donnée MNIST.

1

(a) "1" en français

/

(b) "1" anglosaxon

FIGURE A.2 – Exemple d'une situation de litige entre notation anglosaxone et française.

issus du deep learning. Cela se constate par les résultats très bons obtenus pour des réseaux peu complexes, tel que notre perceptron qui atteignait les 3% d'erreur sur MNIST.

Il devient alors nécessaire de trouver un test plus discriminatoire de la qualité des réseaux.

Faute de temps nous n'avons pas eu la possibilité d'approfondir ce point de vue mais semble nécessaire pour tester les réseaux dans un domaine plus complexe.

Annexe B

Présentation du code