Intellectica, 1994/2, 19, pp. 217-252

Gilles LEONE, Jocelyn PATINEL, Maurice MILGRAM

Détection des symétries par réseaux de neurones formels : utilisation de représentations internes dans l'apprentissage

1. Modèles de détection de la symétrie

1.1. Approche neuro-psychologique

Qu'ils soient manufacturés ou non, la plupart des objets qui nous entourent possèdent un centre ou un axe de symétrie. Cette constatation, à la base des travaux de Mach (1897) et de l'école de la Gestalt, devient moins anecdotique quand on sait que la symétrie axiale est une des propriétés structurelles des objets qui est la plus aisément détectée par le système visuel humain, comparée par exemple à la répétition de motifs (Julesz, 1971 ; Corballis et Roldan, 1975). La symétrie axiale pour les formes bidimension-nelles ou tridimensionnelles a d'ailleurs fait l'objet de nombreuses études de psychophysique, différentes les unes des autres par les conditions expérimentales et/ou le type de stimuli employés (Corballis et Roldan, 1975 ; Goldmeier, 1972 ; Julesz, 1971 ; Palmer et Hemenway, 1978 ; Pellizer et col., 1992 ; Rock et Leaman, 1963 ; Royer, 1981 ; Shiffrar et Shepard, 1991).

En particulier, le temps de présentation du stimulus nécessaire à la détection de la symétrie peut être très bref, de l'ordre de 25 millisecondes (Carmody et col., 1977). Locher et Nodine (1973) ont montré que les mouvements oculaires d'exploration du contour d'une forme étaient fortement influencés, en ce qui concerne les points de fixation, par la présence d'une symétrie bilatérale dans le

stimuli. Ce résultat semble confirmer que la détection d'une symétrie préprogramme les mouvements oculaires. L'intérêt de détecter précocement un axe de symétrie résiderait dans une réduction considérable de l'information à traiter et à mémoriser (Barlow et Reeves, 1979). En outre, quand deux stimuli ne diffèrent que par la présence d'un axe de symétrie, le stimuli symétrique est généralement considéré comme moins complexe et ayant une "bonne" forme (Chipman, 1977). Cette simplification de l'information est mise en évidence dans des tâches de rappel et de discrimination (Freyd et Tverski, 1984) qui concluent que les représentations mentales issues des formes à peu près symétriques sont toujours plus symétriques que les formes elles-mêmes.

On peut donc penser que la détection de symétrie est un processus holistique intervenant très tôt dans les principes organisationnels des formes, et en conséquence serait influencé par les conditions de présentation (orientation de l'axe de symétrie, présence de faces cachées...). Ainsi, Mach (1897) remarquait déjà que lorsque l'axe de symétrie d'une figure est vertical, il est détecté plus rapidement qu'un axe incliné. Cette constatation se trouva confirmée par toutes les études suivantes aussi bien pour les temps de détection que pour les taux d'erreurs. Même si le problème de la définition de la "verticale" intervenant dans la détection n'est pas encore résolu (certaines études considérant que seule la verticale rétinienne intervient (Corballis et Roldan, 1975) alors que d'autres considèrent l'intervention d'une verticale dite "phénoménale", élaborée par différentes informations sensorielles et cognitives (Rock, 1974)), les résultats différent surtout par les courbes des temps de réponses et d'erreurs pour différentes inclinaisons de l'axe de symétrie. Ces courbes de temps de réponse sont en fait la seule voie, à l'heure actuelle, pour inférer les mécanismes sous-jacents à la perception de la symétrie. Ainsi, elles sont généralement soit strictement croissantes quand la différence d'orientation (DO) entre la verticale et l'axe de symétrie croit de 0° à 90° (Corballis et Roldan, 1975 ; Pellizer et col., 1992), soit en cloche, possédant un minimum absolu quand DO vaut 0°, un minimum relatif quand DO vaut 90° et un maximum quand DO vaut environ 45° (Barlow et Reeves, 1979). Ce dernier type de relation entre temps de réponse et orientation de l'axe de symétrie est classiquement dénommé "effet

d'oblique". Un tel effet d'oblique se rencontre dans beaucoup de taches visuo-spatiales (jugement de parallélisme...) et chez beaucoup d'espèces animales (voir Appelle 1972 pour une revue assez complète).

Cette différence entre les courbes des temps de réaction s'explique selon certains auteurs (Palmer et Hemenway, 1978 ; Royer, 1981) par le fait que les stimuli employés par Corballis et Roldan comportent déjà l'axe supposé de symétrie. Dans le cas ou l'axe de symétrie n'est pas explicite, Palmer et Hemenway (1978) supposent l'existence d'un autre mécanisme. Le modèle de détection qu'ils proposent est en deux étapes, comme le confirment leurs résultats expérimentaux sur des polygones plans fermés. La première étape consiste en une sélection grossière des axes de symétrie les plus probables. Chaque axe ainsi sélectionné est affecté d'une certaine probabilité d'être traité par la seconde étape. Les courbes de réponses des sujets semblent indiquer que l'axe vertical est affecté d'une forte probabilité alors que les axes diagonaux sont affectés de la plus faible probabilité de choix. La seconde étape du modèle consiste dans le choix d'un axe pour lequel la symétrie est précisément testée. Si le stimulus n'est pas symétrique suivant cette ligne, un autre axe est choisi à nouveau parmi l'ensemble des droites présélectionnées soustrait de celle précédemment testée. Les auteurs de ce modèle séquentiel n'expliquent cependant pas en quoi consiste la première étape de détection grossière. Le modèle de Royer (1981), qu'il teste en utilisant comme stimulus des matrices de points noirs et blancs (identiques à nos stimuli) suppose que le codage ou le décodage d'un stimulus symétrique résulte du parcours mental d'un arbre hiérarchique de décision dont les noeuds représentent les sous groupes des transformations isométriques. Un modèle parallèle a aussi été proposé où le stimulus visuel serait traité par plusieurs canaux, sélectifs à une orientation, mais avec des temps de traitement de l'information sensiblement différents. En outre, pour rendre compte de certains résultats, ce modèle doit intégrer un mécanisme réduisant le taux de traitement dans chaque canal en fonction du nombre de canaux actifs en parallèle. En conséquence, le mécanisme doit avoir une capacité de traitement limitée.

Ces modèles de psychophysique sont basés sur l'hypothèse, *a priori* subjective, que le traitement d'une orientation de l'axe est

indépendant des autres (que le processus soit séquentiel ou parallèle). Certaines considérations neurophysiologiques, quant au traitement cortical de l'information visuelle, supportent cependant assez bien une telle hypothèse. Depuis les travaux de Hubel et Wiesel (1962, 1974), il est connu qu'il existe dans le cortex visuel primaire des neurones qui répondent à une orientation préférée du stimulus dans leur champ récepteur. De surcroit, Orban et col. (1984) ont montré que la population de neurones répondant à une orientation verticale du stimulus était plus importante que celle répondant à une orientation horizontale, cette dernière population étant elle même plus importante que la population de neurones répondant à une orientation oblique du stimulus. Certains auteurs (Barlow et Reeves, 1979) suggèrent que l'effet d'oblique serait une conséquence de ce biais dans la distribution des directions préférées des cellules sélectives à l'orientation. Néanmoins, cette explication nous semble difficile à concilier avec l'effet d'oblique constaté lors d'une tâche de détection de symétrie. En effet, une orientation verticale de l'axe de symétrie d'un polygone plan, par exemple, n'implique en rien que les contours de ce même polygone aient principalement des orientations verticales ou horizontales.

Il semble intéressant de savoir si en faisant apprendre à un réseau de neurones formels à détecter des symétries axiales dans des figures planes, la structure interne de ce réseau se stabilise en différents neurones sélectifs à une orientation préférée de l'axe de symétrie, d'une façon assez similaire au cortex visuel, et sinon quelle(s) stratégie(s) utiliserait un tel réseau. Cette question est le point de départ de notre étude.

1.2. Approche booléenne

Si on considère une image I de taille n∞n et un ensemble de symétries possibles $S_1,...,S_k$, nous dirons que I est symétrique si elle est globalement invariante sous *au moins* une symétrie S_i.

On peut définir sur l'ensemble E des images (qui est une partie de IR^p avec p=n∞n) une fonction booléenne F telle que F(I)="vrai" si I est symétrique et F(I)="faux" sinon. Cette fonction F peut être développée en une disjonction de k termes F_i, chaque terme i

indiquant si la symétrie S_i est vérifiée. Chacun de ces termes F_i étant une conjonction de termes indiquant l'égalité de deux points de I (symétriques sous S_i). Cf. § 4.1.

A priori, la fonction booléenne F se présente donc comme une disjonction de k termes, chaque terme étant une conjonction d'environ $n^2/2$ clauses. Nous verrons plus loin qu'il existe des réseaux de neurones formels ayant la même table de vérité que la fonction F et dont la *taille de la couche cachée* est *indépendante de n*.

Notons aussi que le problème posé n'est pas *linéairement séparable* , c'est-à-dire qu'il n'existe pas d'hyperplan H de IR^p tel que toutes les figures symétriques (resp. non symétriques) appartiennent au même demi-espace (le complémentaire de H dans IR^p est en effet constitué de deux demi-espaces). Rappelons que le XOR (le "ou" exclusif) est la fonction booléenne la plus simple qui ne soit pas linéairement séparable (plus exactement, les vecteurs dont l'image est +1 ne peuvent pas être séparés de ceux dont l'image est -1 par un hyperplan). Une généralisation à N dimensions du XOR consisterait à attribuer la valeur +1 aux vecteurs dont le nombre de +1 est impair et -1 aux autres. Ce XOR généralisé (en fait c'est un test de parité effectivement utilisé en transmission) n'est pas non plus "linéairement séparable" mais pourra être réalisé par un réseau possédant des couches cachées.

Plus généralement, le problème de la réalisation de fonctions booléennes par des réseaux de neurones formels a été posé depuis longtemps, en premier lieu par MacCullogh et Pitts (1943). Plus récemment, (Bochereau et Bourgine, 1990) certains auteurs ont abordé cette question sous l'angle de l'apprentissage de règles logiques à partir d'exemples. Le problème plus spécifique de la détection de symétrie par des méthodes connexionnistes a déjà fait l'objet de quelques études. Ainsi, Sejnowsky et coll. (1986) ont utilisé des machines de Boltzmann pour catégoriser des images symétriques carrées en fonction de l'orientation de leur axe de symétrie (horizontal, vertical ou diagonal). Leurs réseaux n'étaient pas entraînés à séparer les stimuli symétriques et non symétriques. En effet, la motivation principale des auteurs était d'examiner si "the boltzmann learning algorithm is capable of findings sets of weights

which turn hidden units into useful higher order feature detectors capable of solving symmetry problem". Contrairement au perceptron, cette étude montrait qu'une machine de Boltzmann est capable de résoudre des "prédicats du second ordre", tels que les ont définis Minsky et Papert (1969). Un autre résultat intéressant de l'étude de Sejnowsky et coll fut l'apparition à la fin de la phase d'apprentissage de patrons de poids de connexion entre la couche d'entrée et les couches cachées. Ces patrons révélaient des propriétés de symétrie et/ou d'antisymétrie. Cependant, certains poids de connexion étaient extrêmement faibles en valeur absolue dans la couche cachée. Les auteurs suggérèrent que les poids ont tendance à décroître dés que l'apprentissage a atteint un état stable. Néanmoins, ils ne confirmèrent pas cette suggestion par des études supplémentaires.

Cohen et coll. (1986) utilisèrent des techniques similaires afin de comparer les performances humaines et celles de machines de Boltzmann dans une tâche de détection de symétrie. En résumé, les auteurs trouvèrent des effets d'amorçage rétroactif (*priming repetition effects*) et de généralisation similaires.

Cependant, il semble que personne n'ait abordé le problème de la détection des symétries axiales dans des images carrées, c'est à dire comment séparer des images symétriques et non symétriques avec un réseau de neurones artificiels en utilisant l'algorithme de rétropropagation du gradient. McClelland et Rumelhart (1986) ont étudié la classification d' "input strings as to whether or not they are symmetrical about their center" (p. 341). Avec ce problème à une dimension, ces auteurs ont montré que les réseaux de neurones utilisent uniquement deux neurones dans la couche cachée qui ont des patrons de connexion antissymétrique par rapport à leur centre. De telles propriétés sont-elles généralisables pour des stimuli plans ? Nous nous proposons de répondre à cette question dans cet article.

2. Un réseau de neurones pour détecter la symétrie, premiers résultats

2.1. Les exemples

Un exemple est une image I carrée formée de n∞n pixels de valeur +1 ou -1 (blanc ou noir). On dit que I est symétrique si elle est globalement invariante lorsqu'on lui applique une des quatre symétries suivantes :

H : symétrie / axe horizontal, V : symétrie / axe vertical, D1 : symétrie / première diagonale, D2 : symétrie / seconde diagonale,

Figure 1 : Exemples d'images carrées 6x6 binaires constituant notre base d'apprentissage ou de généralisation. Un carré noir symbolise un bit négatif et un carré blanc un bit positif. L'exemple A est symétrique par rapport à l'axe

vertical V, B par rapport à l'axe horizontal H, C par rapport à la diagonale D2, D par rapport à la diagonale D1 et enfin E n'est symétrique par rapport à aucun de ces axes.

chaque axe passant par le point central si n est impair, par le centre du carré si n est pair (figure 1).

Chaque image est associée à la sortie désirée du réseau (sortie à +1 si symétrique ou à -1 si non symétrique).

2.2. Réseaux de neurones et apprentissage

2.2.1. Structure des réseaux

Nous avons utilisé dans nos simulations des réseaux à couches, entièrement connectés et sans boucles. Les neurones ont une fonction de transfert de type sigmoïde dont les sorties appartiennent à l'intervalle [-1, 1]. La sortie du neurone s'écrit :

$$S_i = f(?_j W_{ij} S_j + t_i) \qquad (1)$$

où S_i est la sortie du neurone i ; S_j la sortie du neurone j, entrée du neurone i ; W_{ij} le poids de la connexion du neurone j vers le neurone i ; t_i le seuil du neurone i ;

et f la fonction sigmoïde de transfert : $f(x) = (e^{kx} -1)/(e^{kx} +1)$ où k représente la raideur et est égal à 1 dans la suite (quand $k \varnothing 8$, on obtient des réseaux à seuil).

Le nombre de couches cachées varie de 1 à 3, le nombre de neurones de ces couches est également variable d'un réseau à un autre. Le nombre de cellules en entrée est fixe et dépend de la taille des images que l'on présente (Cf. § 2.1). Le réseau ne comporte qu'une cellule de sortie, qui doit prendre la valeur +1 ou -1 selon que l'exemple présenté est symétrique ou non-symétrique. Or les valeurs de cette cellule appartiennent à l'intervalle [-1 ; 1], nous devons donc choisir une règle de décision externe pour déterminer la classe de l'exemple. Ainsi :

$-1 < S_i < -S$ signifiera "exemple non symétrique"

$-S < S_i < +S$ signifiera "exemple non déterminé"

$+S < S_i < 1$ signifiera "exemple symétrique".

S représente le seuil de détection, il sera pris égal à 0,75 dans toutes nos simulations. La performance d'un réseau est le pourcentage d'exemples correctement classés, relativement à ce seuil S. Il est en effet illusoire d'utiliser l'écart entre la sortie désirée et la sortie

obtenue pour mesurer la performance, cet écart dépendant beaucoup de l'ordre de grandeur des poids multipliés par la raideur.

2.2.2. Algorithme d'apprentissage

Nos simulations sont basées sur un apprentissage supervisé : une série d'exemples de patrons symétriques et non-symétriques est présentée au réseau, en lui fournissant à chaque fois la sortie désirée S_d (resp. 1 et -1).

L'initialisation des poids se fait de manière aléatoire selon une distribution uniforme entre -0,5 et 0,5 , les seuils sont initialisés à 0.

L'algorithme utilisé est la rétropropagation du gradient de l'erreur (Le Cun, 1987 ; Hinton, 1986) avec un terme de moment destiné à accélérer la convergence (Vogl et col., 1988).

La modification du poids W_{ij} (allant du neurone j vers le neurone i) s'écrit :

$$? W_{ij}^{(t)} = \lambda . ?E/?W_{ij} + \mu . ? W_{ij}^{(t-1)} \quad (2)$$

où $\lambda = 0,7$ et $\mu = 0,2$

avec $E = (S - S_d)^2$, erreur commise par le réseau.

La modification des poids est appliquée après un passage complet de la base d'exemples, ce qui rend l'apprentissage insensible à l'ordre de présentation ; l'apprentissage est conduit jusqu'à quasi stabilisation de l'erreur.

2.3. Simulations avec des images de 6x6 pixels

2.3.1. Ensemble d'exemples

Dans une première série de simulations nous nous sommes restreint à des images de 6x6 pixels, ce qui donne 2^{36} (~6,87 E+10) images possibles. Parmi celles-ci on compte 2^{18} (~2,62 E+5) images invariantes par la symétrie H, autant par V, 2^{21} (~2,10 E+6) par D1 et autant par D2, le reste (99,3 %) étant non symétrique.

Nous avons cependant choisi de constituer notre base d'apprentissage de 50 % d'exemples non symétriques et de 50 % d'exemples symétriques parmi lesquels nous conservons les proportions "statistiques" de 1 symétrie H ou V pour 8 symétries D1 ou D2. Les exemples sont générés de manière aléatoire, le nombre total d'exemples est de 2592.

Les capacités de généralisation des réseaux seront testées sur une base de généralisation de 5184 exemples, générée avec les mêmes proportions que la base d'apprentissage.

2.3.2. Résultats

Un total de 10 simulations ont été effectuées (Cf. tableau 1). Les résultats sont exprimés en pourcentage du nombre d'exemples bien classifiés, compte tenu de la règle de décision précédemment décrite (§ 2.2.1).

Pour un nombre croissant de neurones (de 3 à 25) dans une unique couche cachée (et à durée d'apprentissage équivalente) la performance augmente légèrement, ceci est surtout sensible en généralisation. D'autre part à nombre de connexions égal (comparer SYM3_1, SYM4_1 et SYM5_1) l'adjonction d'une seconde couche cachée améliore les résultats tandis qu'une troisième n'apporte rien de plus. Cependant nous ne disposons pas d'assez de simulations pour pouvoir généraliser ces résultats, de plus ils s'avèrent non valables pour les simulations réalisées avec des images de 10x10 pixels (Cf. tableau 2).

Réseaux de neurones			% correct en apprentissage			% correct généralisation		
Nom	Architecture	ta	sym	ns	Total	sym	ns	Total
SYM3_2	36/25/1	1	100	99	99,5	97	98	97,5
SYM3_3	36/19/1	1	97	96,5	97	92	96	94
SYM3_1	36/12/1	1	94,5	97,5	96	88	95	91
SYM3_S1	36/12/1	2	100	100	100	97	97	97
SYM3_4	36/6/1	2	100	98,5	99	96	97	96,5
SYM3_S4	36/6/1	2	99	98,5	99	92	96	94
SYM3_51	36/3/1	1	?	?	?	86	92	89
SYM3_52	36/3/1	2	97,5	97	97	94	98	96
SYM4_1	36/10/7/1	1	100	99	99,5	98	97	97,5
SYM5_1	36/10/5/2/1	1	99,5	99,5	99,5	96	98	97

Tableau 1 : Résultats des simulations avec des images de 6x6 pixels

Une ligne représente une simulation.

L'architecture de chaque réseau est donnée par le nombre de neurones de chaque couche. La colonne "ta" donne la durée de l'apprentissage en valeur relative. Les résultats sont exprimés en % du nombre d'exemples bien classés par rapport au nombre total d'exemples (ici 2592 pour l'apprentissage et 5184 pour la généralisation) ; la colonne "sym" donne les résultats sur les seuls exemples

symétriques, la colonne "ns" sur les exemples non symétriques, la colonne "Total" sur l'ensemble des exemples.
Les réseaux SYM3_S1 et S4 ont appris sans utiliser de seuil à l'entrée des neurones.
Le réseau SYM3_52 est le réseau SYM3_51 après une nouvelle série de passages de la base d'exemples.

Réseaux de neurones			% correct en apprentissage			% correct en généralisation				
Nom	Architecture	nb iter.	sym	ns	Total	H& V	D1 &2	sym	ns	Total
SYM3_11	100/10/1	9114	100	99,5	99,5	70	89	86,2	80,2	83,2
SYM3_20	100/19/1	7009	100	99,5	100	49	72	69,3	69,1	69,2
SYM4_10	100/9/3/1	19693	100	99,5	100	79	93	90,9	80,6	85,7
SYM4_20	100/19/5/1	6695	100	100	100	61	85	78,9	82,0	80,5
SYM5_15	100/14/4/2/1	8582	100	100	100	84	97	94,8	89,1	92,0

Réseaux de neurones			% correct en apprentissage			% correct en généralisation				
Nom	Architecture	nb iter.	sym	ns	Total	H&V	D1&2	sym	ns	Total
SYM3_11	100/10/1	9114	100	99,5	99,5	70	89	86,2	80,2	83,2
SYM3_20	100/19/1	7009	100	99,5	100	49	72	69,3	69,1	69,2
SYM4_10	100/9/3/1	19693	100	99,5	100	79	93	90,9	80,6	85,7
SYM4_20	100/19/5/1	6695	100	100	100	61	85	78,9	82,0	80,5
SYM5_15	100/14/4/2/1	8582	100	100	100	84	97	94,8	89,1	92,0

Tableau 2 : Résultats des simulations avec des images de 10x10 pixels
Cf. la légende du Tableau 1.
La colonne "nb iter" donne le nombre d'itérations pendant lequel le réseau a appris. En généralisation nous présentons (colonnes "H&V" et "D1&2") des résultats détaillés pour les deux ensembles de symétrie : Horizontale et Verticale d'une part et par rapport aux diagonales d'autre part. Ici les résultats sont donnés pour 4320 exemples en apprentissage et 2160 en généralisation.

Nous nous sommes également intéressés à la *structure interne du réseau après apprentissage*, en visualisant les connexions afférentes à la première couche cachée (diagramme de Hinton), suivant en cela l'exemple de Lehky et Sejnovski (1989). Certaines distributions de poids approximativement antisymétriques (au sens des matrices) par

rapport aux 4 axes, se retrouvent dans tous les réseaux à la fin de l'apprentissage (Cf. figure 2) ; les valeurs de ces connexions sont distribuées de manière très spécifique : elles sont soit inférieures à -2, soit supérieures à 2, soit très proches de 0. Ces poids peuvent être considérés comme formant un masque avec lequel sont convoluées les entrées, réalisant ainsi un codage de celles-ci dans certaines cellules de la première couche cachée. Le nombre de telles cellules est seulement de 2 ou 3 par réseau : ce point sera discuté par la suite.

Enfin nous définissons la "contribution" de chaque neurone des couches cachées comme étant la différence de performance du réseau en généralisation avant et après destruction du neurone considéré (tous les poids sortant du neurone sont fixés à 0) ; les cellules ayant la plus forte contribution sont celles que nous avions jugées "structurées" ; par contre la destruction d'autres neurones peut parfois augmenter la performance du réseau (voir la figure 3).

Partant de ces constatations nous avons cherché d'une part à définir la notion de cellule structurée de façon objective et quantifiable et d'autre part à valider nos résultats sur un ensemble comportant plus d'exemples et permettant un taux de généralisation plus important. Nous avons donc effectué une série de simulations avec des images de 10x10 pixels, dont les résultats seront donnés dans la partie suivante.

Figure 2 : Exemples de patterns de poids afférents à quelques neurones structurés de la première couche cachée. Ces patterns sont relativement antisymétriques et se retrouvent dans la plupart des réseaux ayant à déterminer la symétrie dans une image 6x6. Ces figures rendent compte de la structure spatiale des patterns, tout en montrant la valeur relative des connexions.

Figure 3 : Contribution des neurones de la première couche cachée du réseau SYM3_20 pour une base d'exemples de généralisation, symétriques ou non symétriques (Cf. texte). Les neurones ayant une contribution inférieure à 1 en valeur absolue pour chaque classe d'exemples n'ont pas été représentés. Les neurones 3, 9 et 17 montrent la plus forte contribution et sont aussi les seuls à être structurés. L'élimination des autres neurones non structurés peut parfois entraîner un gain dans les performances du réseau. Cela est vrai essentiellement pour les exemples symétriques et cela se traduit alors par une contribution négative.

3. Étude de la structuration des neurones de la première couche cachée

3.1. Le critère de structuration

Les résultats, précédemment décrits, indiquent l'existence de cellules de la première couche cachée qui semblent être structurées. Néanmoins, il restait à définir un indicateur numérique rendant compte de cette structuration, afin de pouvoir quantifier l'influence

d'une représentation interne au sein de la couche cachée sur les performances du réseau.

Une unité i de la première couche cachée est reliée avec les n^2 cellules (j) de la rétine par des poids W_{ij}. Pour chaque unité i de la couche cachée, on peut écrire la matrice de poids de connexion à la rétine W(i) dont les éléments sont les W_{ij} pour j=1 à n^2, où n^2 représente le nombre de pixels de la rétine. Ainsi définie, W(i) est une matrice de $M_n(IR)$. Il est donc possible d'utiliser les notions usuelles de distance et de norme dont est muni cet espace métrique, et en particulier la norme $L^2(IR^n)$ définie par la relation :

$$N(W(i)) = \sqrt{(\Sigma_j((W(i)_j)^2))} \qquad j= 1..n^2. \qquad (3)$$

Comme il a été décrit précédemment, les matrices W(i) semblent antisymétriques par rapport à chacun des axes de symétrie (H, V, D1, D2). La notion d'antisymétrie pour une matrice A de $M_n(IR)$ se définit généralement en rapport à sa première diagonale et se traduit par la relation suivante entre ses éléments :

$a_{i,j} = - a_{j,i}$ pour i= 1 à n, j = 1 à n, où i représente la ligne et j la colonne.

Désormais, nous appellerons "transposée de A suivant D1", notée ${}^tA_{D1}$, la matrice ayant pour élément $a_{j,i}$, l'indice D1 fait référence à la première diagonale. De même, par extension la transposée de A suivant la seconde diagonale, notée ${}^tA_{D2}$, est la matrice ayant pour éléments $a_{n+1-j,n+1-i}$; la transposée de A suivant l'axe vertical, notée ${}^tA_V$, est la matrice ayant pour éléments $a_{n+1-i,j}$; enfin la transposée de A suivant l'axe horizontal, notée ${}^tA_H$, est la matrice ayant pour éléments $a_{i,n+1-j}$. En conséquence, la distance de W(i) à $-{}^tW(i)_{D1}$, notée $d(W(i), -{}^tW(i)_{D1})$ et qui équivaut à $N(W(i) + {}^tW(i)_{D1})^2$, sera nulle si et seulement si $W(i) = -{}^tW(i)_{D1}$. En outre, l'inégalité du triangle donne la relation (4a) :

$$0 = d(W(i),-{}^tW(i)_{D1}) = d(W(i),M0) + d(M0,-{}^tW(i)_{D1}) \qquad (4a)$$

où M0 est la matrice de $M_n(IR)$ d'éléments 0. Mais comme $d(W(i),M0) = N^2(W(i)) = d(M0,-{}^tW(i)_{D1})$, on obtient la relation (4b) :

$$0 = d(W(i),-{}^tW(i)_{D1}) = 2xN^2(W(i)) \qquad (4b)$$

Cette relation (4b) est vérifiée pour chacun des quatre axes de symétrie. Nous obtenons en faisant la somme des inéquations (4b) pour chacun des axes, l'inéquation (4c) :

$$0 = 1/4.(d(W(i),-{}^tW(i)_{D1}) + d(W(i),-{}^tW(i)_{D2}) + d(W(i),-{}^tW(i)_{H}) + d(W(i),-{}^tW(i)_{V})) = 2xN^2(W(i)) \qquad (4c)$$

En supposant W(i) différente de la matrice nulle, ce qui est toujours vérifié dans nos simulations, on peut diviser la relation (4c) par $N^2(W(i))$, ce qui nous fournit la relation (4d) ainsi que la définition de notre critère CRIT(i) :

$$CRIT(i) = (d(W(i),-{}^tW(i)_{D1}) + d(W(i),-{}^tW(i)_{D2}) + d(W(i),-{}^tW(i)_{H}) + d(W(i),-{}^tW(i)_{V}))/ 4xN^2(W(i))$$

d'après les inégalités précédentes : $0 = CRIT(i) = 2$ (4d)

L'égalité CRIT(i) = 0 ne peut avoir lieu que dans le cas où W(i) est antisymétrique par rapport à chacun des 4 axes de symétrie. Néanmoins, plus ce critère est proche de 0, plus W(i) tend, au sens mathématique, vers une matrice totalement antisymétrique et, en conséquence plus le patron des poids de connexion semble structuré. C'est la valeur de ce critère qui rend désormais compte de ce que nous appelons "structuration des cellules de la couche

cachée". Dans la suite de cette étude, nous considérons qu'une cellule de la première couche cachée est structurée si son critère CRIT est inférieur à 0,7.

Ce critère peut permettre de corréler les performances d'un réseau à la structuration de certaines unités cachées, c'est à dire à la construction d'une représentation interne adéquate au problème à résoudre au niveau de la couche cachée. Enfin, l'évolution de ce critère au cours de l'apprentissage d'un réseau peut nous permettre de comprendre comment le réseau construit cette représentation interne.

3.2. Apprentissages

3.2.1. Les exemples

En utilisant des images 10x10, nous avons cette fois 2^{100} (~1,27 E+30) configurations possibles dont 2x(2^{50} (~1,13 E+15) + 2^{60} (~1,15 E+18)) ~2,31 E+18 symétriques. Nous avons conservé les proportions des différents types de symétries déjà utilisées précédemment. Pour ces simulations le nombre d'exemples a été porté à 4320 pour l'apprentissage et 2160 pour la généralisation.

3.2.2. Résultats des simulations

Nous remarquons ici (Cf. tableau 2) que les résultats en généralisation sont inférieurs à ceux obtenus avec des images plus petites. D'autre part une analyse des résultats en généralisation par type de symétrie montre une meilleure performance des réseaux sur les symétries diagonales (+15 à 25% selon les réseaux) par rapport aux symétries horizontale ou verticale. Bien que notre étude n'ait pas pour objectif de tester cette hypothèse, il nous semble que ceci découle de la proportion plus importante d'exemples de ce type dans les fichiers d'apprentissage.

3.3. Étude de la structuration au œurs du temps, en relation avec d'autres caractéristiques des réseaux

3.3.1. Structuration et contribution

Nous pouvons maintenant au vu des valeurs de notre critère, pour les neurones de la première couche cachée de l'ensemble des réseaux, séparer ces neurones en deux classes bien distinctes : neurones structurés (pour lesquels le critère varie de 0,1 à 0,7, la majorité étant inférieurs à 0,3) et neurones non structurés (pour lesquels le critère varie de 0,9 à 1,6, la majorité étant supérieurs à 1,4). Nous trouvons *2 ou 3 neurones structurés par réseau en fin d'apprentissage* : il est remarquable de noter que *le passage d'images 6x6 à 10x10 pixels ne modifie pas ce nombre* ; ce point sera analysé au § 4.1.

De même que précédemment nous avons observé la dégradation des résultats en généralisation lorsque l'on supprime des neurones.

Nous avons en particulier cherché à voir si les neurones structurés seuls suffisaient pour assurer de bonnes performances au réseau. Les résultats (Cf. tableau 3) montrent que pour 3 des 5 réseaux on peut ne garder dans la première couche cachée que 2 neurones structurés et que les performances sont même améliorées par rapport au réseau pris dans sa totalité.

Réseau	Neurones restant	sym	non sym	Total
SYM3_11	Tous	82,5	80,5	81,5
	3, 5, 7	100	66,1	83
	3, 5	97,8	74,2	86
	3, 7	84,7	61,1	73
	5, 7	66,1	58,1	62
	3	0	44,4	22
SYM3_20	Tous	63,9	69,2	66,5
	5, 15	89,4	45,8	68
SYM4_10	Tous	86,7	81,7	84,2
	5, 6, 10	94,7	75,6	85
	5, 6	54,2	71,9	63
	5, 10	97,5	75,3	86
	6, 10	0	73,6	37
	5	0	46,9	24
SYM4_20	Tous	76,7	80,0	78,3
	12, 18	64,4	69,7	67
SYM5_15	Tous	94,4	89,4	91,9
	2, 10	75,3	86,1	81
	2	0	99,2	50
	10	71,1	63,6	67

Tableau 3 : Effet de la destruction de neurones, sans réapprentissage, sur des réseaux ayant appris des images de 10x10 pixels

Tous les résultats sont calculés sur des exemples de généralisation (360 symétriques et 360 non symétriques). On ne garde pour chaque réseau que 1 à 3 neurones parmi ceux qui sont structurés, les résultats sont à comparer avec ceux obtenus sur le réseau entier (neurones restants : tous).

Le détail des résultats pour les exemples symétriques d'une part et non symétriques d'autre part montre que, avec les quelques

neurones structurés restant, les réseaux reconnaissent nettement plus d'exemples symétriques qu'avec la totalité des neurones. Par contre ils reconnaissent moins d'exemples non symétriques. On peut en conclure que les neurones structurés codent la symétrie, tandis que les autres ont pour effet de globalement déplacer la valeur de la sortie vers les valeurs négatives.

3.3.2. Étude de la variation simultanée des caractéristiques des réseaux

Nous avons étudié les variations simultanées de différents paramètres, dans le but de déterminer les origines de la structuration :

Critère de structuration (CRIT) ;

Norme moyenne des poids entrant dans une cellule de la première couche cachée (NPE) ;

Norme moyenne des poids sortant d'une cellule de la première couche cachée (NPS) ;

Taux de généralisation du réseau (TG) ;

Taux de généralisation du réseau après destruction du neurone étudié (TGD) ;

Contribution du neurone étudié : PTG = TG - TGD ; sur les exemples symétriques seulement : PTS, ou sur les non symétriques : PTNS.

Figure 4 : A) Évolution au cours de l'apprentissage des paramètres d'un neurone structuré, *appartenant au réseau SYM3_11. L'échelle à gauche se rapporte à l'évolution du critère de structuration (Crit), de la norme moyenne des poids en entrée (NPE) et de la contribution du neurone en généralisation (PTG) en pourcentage. L'échelle de droite se rapporte à l'évolution de la norme des poids en sortie (NPS). On remarque qu'il faut attendre la 350ème*

itération de la base d'apprentissage avant de constater une augmentation de la contribution du neurone alors que la structuration est engagée dès la 250ème itération.

État des réseaux en fin d'apprentissage. Il apparaît nettement en fin d'apprentissage deux populations de neurones : les neurones structurés qui correspondent aux maxima de NPE et NPS ainsi que des contributions, et les neurones non structurés. L'étude des corrélations entre les différents paramètres indiquent, qu'à l'exception de PTNS, tous ceux-ci sont étroitement corrélés (taux supérieurs à 0,92). Ceci confirme que les neurones structurés codent la symétrie.

Figure 4 : B) Évolution au cours de l'apprentissage des paramètres d'un neurone non structuré*, appartenant au réseau SYM3_11. On remarque cette fois que le critère de structuration, la norme moyenne des poids en entrée et la contribution en généralisation ne varient quasiment pas pendant l'apprentissage. La contribution reste en moyenne nulle. La norme des poids de sortie augmente mais dans des proportions 2,5 fois moindre que pour un neurone structuré.*

Évolution simultanée des paramètres. L'étude de cette évolution montre que la divergence des neurones entre les deux classes a lieu relativement tôt au cours de l'apprentissage : en général après

environ 200 passages de la base d'exemples (Cf. figures 4A et 4B). Cette divergence s'accompagne d'une augmentation brutale de l'ensemble des paramètres (NPE, NPS, PTG), quoique nettement moins prononcée dans le cas d'un neurone restant non structuré ; dans ce dernier cas la contribution du neurone n'augmente quasiment pas. L'étude des coefficients de corrélation entre paramètres sur la durée de l'apprentissage montre, dans le cas structuré, une étroite dépendance entre tous les paramètres, et dans le cas non structuré que CRIT et PTG sont tous deux indépendants de l'ensemble des autres.

Nous pouvons déterminer précisément le moment où apparaît définitivement pour chaque paramètre la distinction entre les deux classes de neurones. Elle apparaît d'abord pour CRIT (200ème passage) puis pour PTS, NPS et NPE (vers les 350 ou 400ème passages). On peut donc supposer que CRIT est à l'origine de l'évolution des autres paramètres au moment crucial de l'apprentissage. L'origine de l'évolution d'un neurone vers une classe ou l'autre reste indéterminée car les neurones ayant les valeurs initiales de CRIT (ou d'autre paramètres) les plus faibles (ou les plus fortes) ne sont pas forcément les mêmes que ceux qui ont les valeurs finales les plus faibles (ou les plus fortes).

3.4. Apprentissage "épigénétique"

Partant de la constatation qu'il suffit de très peu de neurones dans la première couche cachée pour assurer une bonne performance, nous avons cherché à obtenir de tels réseaux directement par l'apprentissage. Quelques tentatives de simulations nous ayant montré qu'un réseau ayant trop peu de neurones dans la couche cachée au départ ne peut apprendre, nous avons élaboré une nouvelle procédure que nous nommons "apprentissage épigénétique" (s'inspirant de l'"optimal brain damage" de Le Cun, 1989). Celle-ci consiste, à partir d'un réseau ayant un nombre de neurones en première couche cachée habituel (10 à 25), à lancer l'apprentissage, puis, avec une certaine fréquence (tous les 60 passages de la base d'apprentissage dans nos simulations), à

détruire le neurone le moins structuré si les deux conditions suivantes sont vérifiées :

C1 : il y a au moins un neurone dont le critère de structuration est inférieur à 0,7 ; ceci afin de laisser s'engager le phénomène de structuration,

C2 : il y a au moins un neurone dont le critère de structuration est supérieur à 0,7 ; ceci afin de ne pas détruire tous les neurones.

Ainsi dès lors que le processus de structuration est engagé, l'apprentissage se continue avec de moins en moins de neurones dans la première couche cachée. A la fin de celui-ci, il ne reste que des neurones fortement structurés. Il faut remarquer que la décrémentation du nombre de neurones de la couche cachée est basée sur la structuration de ceux-ci et non sur leur performance. Ces deux types de décrémentation auraient conduit à des réseaux différents, car en début d'apprentissage le neurone le moins structuré n'est pas forcément celui ayant la plus faible contribution en généralisation.

L'intérêt d'un algorithme "décrémental" d'apprentissage peut se situer à deux niveaux. Tout d'abord, il y a un gain appréciable du temps d'apprentissage sachant que le temps d'une itération de l'algorithme de rétropropagation du gradient est en o(n), où n représente le nombre de neurones de la couche cachée.

Réseau	Architecture	H & V	D1 & D2	sym	non sym	Total
SYM3_11 E	100/10->3/1	92	99,5	99	95	97
SYM4_13 E	100/12->4/4/1	95	99,5	100	95,5	98
SYM4_13 C	100/12/4/1	56	87	82	86,5	84
SYM4_13 D	100/4/4/1	30	35	35,5	64,5	50

Tableau 4 : Résultats des simulations épigénétiques

Les colonnes ont la même signification que dans les précédents tableaux. Les résultats concernent des exemples de généralisation (720 pour les deux premières colonnes, 1440 pour les deux suivantes et 2880 pour le total).

Deux réseaux ont subi un apprentissage épigénétique : leur nom est suivi de la lettre E. Le réseau SYM4_13 C partant du même état initial que SYM4_13 E, a subi un apprentissage classique, avec le même nombre d'itérations (voir texte). Le réseau

SYM4_13 D dérive de SYM4_13 C par conservation des seuls neurones cachés existant dans SYM4_13 E (voir texte).

Le gain d'un tel apprentissage peut aussi résider dans la sélection de quelques neurones particulièrement adaptés à la résolution du problème et dans l'amélioration des performances du réseau. Afin de tester ces hypothèses, nous avons comparés deux réseaux (SYM4_13E et SYM4_13C), tout à fait identiques quant à leur état initial (nombres de neurones et valeurs initiales des connexions) et quant aux paramètres d'apprentissages (nombre d'itérations de la base d'apprentissage, terme du gradient...). Les réseaux soumis à un apprentissage épigénétique, désormais appelés réseaux épigénétiques, ont des résultats nettement meilleurs que ceux des réseaux obtenus auparavant. Ainsi, comme l'indique le tableau 4, un réseau épigénétique montre des performances plus grandes (d'au moins 9%) en généralisation.

Le réseau épigénétique SYM4_13E, qui ne possède que 4 neurones dans la couche cachée (en comparaison de 12 au départ), montre aussi des performances améliorées par rapport au réseau normal n'ayant que ces 4 neurones actifs : SYM4_13D (et cela de plus de 30%). La figure 5 montre l'évolution simultanée des critères de structuration, en fonction du nombre d'itérations de la base d'apprentissage, pour les 4 neurones présents dans le réseau épigénétique. Ce graphique montre clairement que l'apprentissage épigénétique entraîne une accélération et une amélioration de la structuration des neurones cachés et ceci dans un rapport d'environ 2. En outre, il confirme bien que la structuration commence vers le 200ème passage de la base d'apprentissage. Il faut constater que les neurones conservés par l'algorithme épigénétique sont ceux qui ont les meilleurs critères de structuration dans le réseaux SYM4_13C, à l'exception du neurone caché numéro 10 qui ne se structure pas au cours de l'apprentissage normal. On peut aussi remarquer que les neurones qui vont se structurer ne sont pas nécessairement ceux ayant initialement une valeur de critère faible (Cf. § 3.3.2.2). D'autre part, le nombre de 4 neurones obtenus dans la première couche cachée est supérieur à celui trouvé précédemment (2 ou 3) ; cependant une modification des paramètres de l'apprentissage

épigénétique (C1=0,7 et C2=0,3), nous permet d'obtenir un réseau de 3 neurones dans la couche cachée (SYM3_11E) et ayant de très bonnes performances (Tableau 4).

Ces résultats indiquent qu'un nombre restreint (3) de neurones dans la couche cachée permet d'obtenir une excellente performance pour ce problème de détection des symétries et que cette performance est corrélée à la valeur des critères de structuration des neurones. Enfin, l'algorithme épigénétique décrit ci-dessus fait preuve d'une efficacité supérieure à celle de l'algorithme de rétropropagation du gradient classique.

Figure 5 : Différence d'évolution du critère de structuration pendant un apprentissage classique (noté C) et un apprentissage épigénétique (noté E) (cf. texte). Les deux réseaux avaient le même état initial et les mêmes paramètres d'apprentissage. On compare les critères des 4 neurones cachés qui sont conservés par l'apprentissage épigénétique. Notez qu'après la 1500ème itération de la base d'apprentissage, il n'y a plus d'évolution notable de la valeur des critères.

Ces résultats suggèrent que l'association de quelques patrons de poids structurés différents, dans lesquels *les poids faibles sont non*

nuls (figure 2) est certainement suffisante pour classifier la quasi-totalités des images possibles.

4. Le problème des configurations symétriques d'une orbite

4.1. Analyse théorique du problème des symétries

4.1.1. Présentation du problème

Nos simulations ont indiqué qu'un réseau présentant de bonnes performances dans la détection des symétries possède un nombre restreint de neurones structurés (2 ou 3) dans la première couche cachée. En outre, nos résultats ont montré que la majeure partie des performances du réseau reposait sur ces seules cellules structurées. Même si nous n'avons pas pu fixer de bornes précises à la quantité de neurones nécessaire et suffisante dans la première couche cachée pour résoudre notre problème, il semble que ce nombre soit indépendant de la taille de la rétine d'entrée. En effet, des simulations avec une rétine 6x6 pixels ou 10x10 pixels ont conduit à une quantité similaire de cellules structurées. En conclusion, il semble que la taille de l'entrée n'influe pas sur le nombre de neurones nécessaires à la discrimination de la symétrie dans le cas de quatre axes de symétrie. La démonstration de ce fait est donnée ci-après :

Soit une image I carrée formée de MxM points de valeur +1 ou -1 (blanc ou noir). Nous rappelons que I sera dite symétrique si elle est globalement invariante lorsqu'on lui applique une des quatre symétries suivantes :

H: symétrie / axe horizontal
V: symétrie / axe vertical
D1: symétrie / première diagonale
D2: symétrie / seconde diagonale

chaque axe passant par le point central si M est impair, par le centre du carré si M est pair.

Considérons la partition de I en orbites O_j ; une orbite est obtenue en appliquant les 4 symétries à un point quelconque de I. Il est facile de voir qu'en général, une orbite possède 8 points.

Si le point choisi est sur un axe de symétrie, l'orbite n'aura plus que 4 points. Si le point choisi est sur tous les axes (cas M impair), l'orbite est réduite à un point.
Le problème de décision concernant I se ramène à N problèmes (N étant le nombre d'orbites) concernant chaque orbite séparément, *la seule contrainte étant que la même symétrie doit opérer pour toutes les orbites*. Pour une orbite typique de 8 points, il existe 256 configurations dont 54 symétriques. Si on note $x_0, x_1,..., x_7$ les 8 valeurs des points d'une orbite donnée, numérotés dans le sens direct, la condition de symétrie s'écrira :

$$x_i=x_{1-i} \text{ ou } x_i=x_{3-i} \text{ ou } x_i=x_{5-i} \text{ ou } x_i=x_{7-i} \quad \text{pour } i=0,...,7$$

les indices étant tous pris modulo 8. Si on considère l'image comme un vecteur à M^2 composantes, la condition de symétrie H s'écrira comme une conjonction de conditions du type : $x_i=x_{K-i}$ pour un K fixé dans 1,3,5,7 et pour i=0,...,7 avec 8 ou 4 conditions par orbites.

4.1.2. Résolution dans le cas d'un axe de symétrie

La première question posée est la suivante : combien de cellules cachées (en supposant une seule couche) faut-il à un réseau pour, ayant l'image I à M^2 valeurs en entrée, donner en sortie +1 si I admet la symétrie H et -1 sinon.
Nous allons montrer que la réponse est "2 cellules", et cela indépendamment de la valeur de M. Pour établir ce résultat, commençons par transformer le problème en un problème géométrique.

Considérons l'ensemble $E=\{-1 ; +1\}^{MxM}$ de toutes les images possibles et H_M le sous ensemble des images symétriques pour H. Si nous trouvons un hyperplan de IR^{MxM} qui rencontre E selon H_M, nous aurons une forme linéaire u (celle qui est nulle sur l'hyperplan) telle que :

$$X \in E \text{ et } \&u(X)\&< \varepsilon \quad X \in H_M$$

car E n'a qu'un nombre fini de points, donc la valeur minimale non nulle de &u(X)& quand X est dans E-H_M est bien strictement positive.
Il suffira alors de prendre comme poids les coefficients de la forme linéaire u pour la première cellule cachée C^+ et leurs opposés pour la deuxième cellule cachée C^-. On choisit de plus les seuils θ de ces deux cellules de manière à obtenir deux sorties proches de +1 quand l'entrée vérifie u(X)=0. Pour obtenir une sortie voisine de +1 si et seulement si X H_M , on réalise ensuite un ET logique entre les deux cellules cachées et la cellule de sortie. Si X n'appartient pas à H_M , on a u(X)>ε ou u(X)<-ε et l'une des sorties d'une cellule cachée sera assez différente de +1 pour donner une sortie globale du réseau proche de -1.

Il nous reste à montrer qu'un tel hyperplan (ou la forme linéaire associée) existe bien, cela va découler du lemme d'évitement.

Lemme d'évitement

Soit p points de IR^n tous différents du vecteur nul et n=2 , il existe alors une forme linéaire u et un réel ε>0 tels que :

pour i=1,...,p ∧u(x_i)∧ > ε

Preuve : ar récurrence sur p : c'est évident pour p=1 et nous le démontrons pour p+1 en le supposant vérifié pour p.

Soit u la forme linéaire apportée par l'hypothèse sur les p premiers points. Si x_{p+1} n'annule pas u, c'est fini. Si u(x_{p+1})=0, nous allons construire une nouvelle forme linéaire v en faisant tourner l'hyperplan pour qu'il évite x_{p+1} sans pour autant passer par les p premiers points. Posons :

v(x) = u(x)+ α.<x , x_{p+1}>

D'après la deuxième inégalité du triangle, on a :

&v(x)& > & &u(x)& - α.&<x , x_{p+1}>& & i=1,...,p

et en prenant $\alpha < Min \{ \frac{\&u(x_i)\&}{\&<x_i, x_{p+1}>\&} \}$ étendu aux indices i où $<x_i, x_{p+1}> ? 0$, on aura :

$$\&v(x_i)\& > 0 \qquad i=1,...,p$$

et donc il existe $\varepsilon > 0$ tel que : $v(x_i)\& > \varepsilon' \qquad i=1,...,p$

puisque $u(x_i)$ est toujours non nul. Pour x_{p+1}, on aura :

$$\&v(x_{p+1})\& = \alpha . | | x_{p+1} | |^2 > \varepsilon''$$

Il suffit de prendre $\varepsilon = Min(\varepsilon',\varepsilon'')$. On a ainsi construit une forme linéaire v qui évite les p+1 points. La figure 6 est une description géométrique de ce lemme d'évitement.

Figure 6 : Le nombre de points ¥i étant fini, il est toujours possible de trouver un hyperplan H évitant complètement les ¥i ; en "épaississant" cet hyperplan d'un épaisseur d, on obtient une tranche d'espace vide de ¥i. L'épaississement

est toujours possible considérant que le ¥i le plus proche de H est à une distance d'>d de H.

Pour utiliser ce lemme, introduisons l'application linéaire D de IR^{MxM} dans IR^n qui associe à une image I le vecteur D(I) :

$D(I)=(x_{i1}-x_{i2},...,x_{in}-x_{in+1})$

où les paires d'indices $(i_1,i_2),...,(i_n,i_{n+1})$ correspondent aux conditions de la symétrie par rapport à H (par exemple). Ces paires d'indices peuvent être regroupées par orbite, chaque orbite donnant autant de paires qu'elle comporte de points (8 ou 4).
Dire que I est H-symétrique est équivalent à : D(I) = (0,0,...,0).

L'ensemble E de toutes les images est fini et son image par D l'est aussi, D(E) a par exemple p éléments de IR^n et comme H_M est le noyau de D, l'image de E-H_M ne contient pas 0. On peut appliquer le lemme d'évitement à D(E-H_M) ce qui nous fournit une forme linéaire u. La forme linéaire $U=u_oD$ de IR^{MxM} satisfait ainsi à la condition : X E et &U(X)&< ε X H_M et ceci prouve que la propriété de symétrie par rapport à H peut être trouvée par un réseau à deux cellules cachées seulement.

4.1.3. Généralisation à 4 axes de symétrie

La deuxième question posée est la suivante : combien de cellules cachées (en supposant une seule couche) faut-il à un réseau pour, ayant l'image I à M^2 valeurs en entrée, donner en sortie +1 (à ε près) si I admet une des 4 symétries H, V, A, D et -1 (à ε près) sinon, et ceci quelquesoit ε >0.
Pour l'instant, nous savons seulement que ce nombre est inférieur ou égal à 8 (en utilisant le résultat précédent) alors que les expériences décrites plus haut donnent un nombre inférieur.
Remarquons que nous n'avons pas utilisé la régularité de l'ensemble D(E-H_M) = $\{-2,0,+2\}^n$ mais seulement le fait que c'est un ensemble fini Il faut peut être y voir l'explication de la différence entre la prédiction théorique (8 cellules cachées en tout) et expérimentale (2 à 3 cellules, mais avec une faible erreur résiduelle).

Nous nous proposons de montrer que, en ce qui concerne une seule orbite, le nombre d'hyperplans suffisants n'est pas de 8, mais est égal à 4.

4.2. Détermination expérimentale du nombre de cellules pour résoudre le problème de l'orbite

Afin de fixer une borne supérieure au nombre d'hyperplans (ou de cellules cachées) nécessaires à la séparation des configurations symétriques et non symétriques pour un vecteur de IR^8, représentant les valeurs des pixels d'une orbite, nous avons essayé de résoudre ce problème avec des réseaux neuronaux. Nous rappelons qu'il existe 256 configurations distinctes parmi lesquelles seulement 54 configurations symétriques pour ce problème.

Nous avons fait l'apprentissage de notre réseau avec les mêmes techniques que précédemment. La performance du réseau est évaluée en terme de tout ou rien, c'est à dire qu'un réseau n'est considéré comme ayant résolu le problème qu'à l'instant où il classifie correctement les 256 exemples au seuil 0,75. Cela signifie que la valeur en sortie du réseau à la présentation d'une configuration symétrique appartient à [0,75 ,1] et que celle correspondant à une configuration non symétrique appartient à [-1, -0,75]. Toutes nos simulations ont porté sur des réseaux à trois couches, dont la première se compose de 8 neurones, la dernière d'un seul neurone et la couche cachée d'un nombre variant de 3 à 10 neurones.

Quant un réseau possédant N neurones dans la couche cachée est capable de résoudre ce problème, nous savons alors que le nombre minimum de neurones nécessaires et suffisants est inférieur ou égal à N. Les résultats théoriques présentés ci-dessus nous donnent une borne supérieure de 8. Les simulations effectuées indiquent qu'avec 5 neurones dans la couche cachée, un réseau résout ce problème. Toutes nos simulations avec des réseaux ayant un nombre inférieur à 5 neurones dans la couche cachée, et dont les poids initiaux sont générés aléatoirement, ont échoué : les réseaux ne pouvant généralement pas classer correctement 2 configurations symétriques.

Pour les réseaux ayant plus de 5 neurones, les patrons des poids de connexions en fin d'apprentissage sont à nouveau antisymétriques.

4.3. Une solution particulière

Les huit points de l'orbite peuvent se décomposer en sommets de 2 rectangles R1 et R2 (figure 7A). Si l'on suppose que le réseau doit détecter les configurations symétriques par rapport à H ou V des sommets de R1, il est facile de trouver un patron de poids afférents d'un neurone caché tel que son activité soit nulle si et seulement si les sommets de R1 représentent une configuration symétrique par rapport à H ou V.

Ce patron peut par exemple être celui représenté dans la figure 7B, c'est à dire des poids identiques en valeur absolue (1 par exemple) mais dont les signes sont alternés. Dans le cas d'une configuration non symétrique, l'activité du neurone caché ne peut valoir en valeur absolue que 1 ou 2. Si on combine un tel patron avec un patron identique pour le rectangle R2 (avec un rapport adéquat : 1/4 dans notre exemple), l'activité du neurone caché résultant est nulle si l'orbite est symétrique par rapport à H ou V, et non nulle si l'un des 2 rectangles au moins ne présente aucune des 2 symétries. Enfin, quand ce n'est pas la même symétrie qui opère sur R1 et R2 simultanément, la sortie est encore nulle, cependant l'orbite est encore bien symétrique par rapport à une des diagonales (figure 7C).

Figure 7 : A) Pour 4 axes de symétrie (H, V, D1, D2) une orbite se réduit à huit points (xi). Ces points peuvent être regroupés en sommets de deux triangles (Ri).
B) Pattern de poids d'un neurone caché dont la sortie sera nulle si l'image est symétrique par rapport à H ou V (cf. texte).
C) Exemple d'une configuration H-symétrique pour R2, V-symétrique pour R1 et D1 symétrique pour l'orbite.

Une couche cachée composée d'un neurone ayant un tel patron de poids couplé à un neurone ayant un patron de poids opposés peut résoudre le problème de détection d'une symétrie H ou V. Si on y ajoute 2 autres neurones dont les poids sont obtenus par une permutation circulaire, on peut alors détecter, en théorie, les 4 types de symétries. Pour ce faire, nous avons initialisé un réseau à 4 neurones dans la couche cachée avec de tels poids, les autres poids étant initialisés à 0. Ce réseau a résolu aisément le problème, en conservant la même organisation des poids. En fin d'apprentissage,

les poids initialement à 1 sont à 5,03 (±0,1), les poids initialement à 0,25 sont à 3,24 (±0,1), les seuils de tous les neurones de la couche cachée valent -1,75 , celui du neurone de sortie 8,34, enfin les connexions entre la couche cachée et la couche de sortie valent 8,24.

Il est donc possible théoriquement et expérimentalement de trouver un réseau avec 4 neurones dans la couche cachée, pour résoudre le problème de l'orbite, dans le cas de 4 axes de symétrie. La difficulté de trouver un tel réseau avec une initialisation aléatoire peut s'expliquer par le nombre "réduit" de solutions ou par des particularités géométriques du graphe de la fonction coût dans l'espace des poids.

5. Discussion

Nous avons un réseau possédant quatre neurones dans une seule couche cachée, résolvant sans erreur le problème de la symétrie pour une orbite (c'est à dire un ensemble de huit points fermés pour les quatre opérations de symétrie axiale). Néanmoins, nous n'avons pas pu trouver le nombre de neurones *minimum* nécessaire pour résoudre ce problème de l'orbite.

De plus, le problème global de la détection d'une symétrie axiale dans une image carrée n'est pas la simple conjonction de ce même problème posé indépendamment pour chaque orbite, car il faut que la même symétrie opère sur chacune des orbites. Ceci rend l'analyse mathématique du problème plus difficile ; nous avons pu démontrer l'existence théorique d'une borne supérieure égale à huit neurones cachés. Cependant, nos résultats expérimentaux indiquent qu'avec 2 ou 3 neurones cachés on résout le problème avec une erreur résiduelle très faible pour l'ensemble de généralisation (dont la représentativité n'est pas assurée).

Relation avec l'approche neuro-psychologique

Il faut considérer que nous n'avons pas essayé, lors de cette étude, de faire une simulation "plausible" de la détection de symétrie

par le cortex visuel. En effet, nos neurones formels n'ont aucune des propriétés connues des neurones réels (dans la rétine ou dans le cortex visuel primaire). Néanmoins, nous pouvons confronter nos résultats à ceux obtenus en psychologie humaine et tenter d'en comprendre les convergences ou les divergences.

Le résultat principal de notre article est le petit nombre de neurones formels nécessaires pour détecter des stimuli symétriques, et cela avec une assez grande précision (environ 97 % de réussite). Il faut se rappeler que la symétrie est très facilement et rapidement détectée par le système visuel humain. On peut donc supposer qu'à l'instar de nos réseaux de neurones formels, un nombre assez restreint de neurones soient dédiés à la détection des symétries axiales. Cependant, nous avons simplifié le problème en ne présentant que quatre axes possibles de symétrie (vertical, horizontal et diagonaux). Il est probable qu'un plus grand nombre d'orientations possibles de l'axe de symétrie aurait accru le nombre de neurones nécessaires à la détection des symétries, mais dans des proportions raisonnables au vu de nos résultats sur 4 orientations et à la partie théorique.

Un autre résultat important que nous avons obtenu est que chaque neurone de la couche cachée est sélectif à toutes les orientations de l'axe de symétrie, et ne possède pas une orientation préférée, à la différence de certains neurones du cortex visuel primaire. Ce point est donc une différence majeure entre nos simulations et tous les modèles de la psychologie expérimentale de détection des symétries qui reposent sur un traitement différentiel de chaque orientation de l'axe de symétrie.

Il faut rappeler que ces modèles proviennent de résultats qui indiquent tous une anisotropie directionnelle dans la détection des symétries ("effet d'oblique"). Nos réseaux de neurones formels ne simulent pas du tout un tel comportement, au contraire, les résultats indiquent que le taux d'erreurs est plus important quand l'orientation de l'axe de symétrie des stimuli est vertical ou horizontal, comparé à une orientation oblique. Nous suggérons que cette divergence provient des proportions différentes des stimuli possédant une symétrie horizontale/verticale versus diagonale dans notre base d'apprentissage. En effet, nous avons essayé de respecter la proportion "mathématique" des différents types de symétrie (c'est-à-

dire il y a beaucoup plus de stimuli symétriques par rapport aux diagonales que par rapport à la verticale ou à l'horizontale). Cependant, cette proportionnalité "mathématique" ne correspond absolument pas à la proportion "naturelle" des différentes orientations des axes de symétrie. Ainsi, comme le soulignait Barlow et Reeves (1979), il y a plus d'objets ayant un axe de symétrie vertical ou horizontal que d'objets ayant un axe oblique dans la vie courante. Quoique nous n'ayons pas mené une telle étude, il nous semble qu'en modifiant la proportion des différentes catégories de stimuli dans la base d'apprentissage, nous puissions simuler un "effet d'oblique" avec des réseaux de neurones formels.

Conclusion

En conclusion, nous avons présenté, dans cet article, un ensemble de résultats théoriques et expérimentaux. Sur le plan théorique, nous n'avons pu établir qu'une borne supérieure assez pessimiste pour le nombre de neurones cachés (8 neurones) nécessaire pour résoudre le problème des 4 axes de symétrie. Néanmoins, le fait que ce nombre soit indépendant de la taille de la rétine nous paraît déjà un résultat très intéressant.

Sur le plan expérimental, il apparaît clairement qu'un réseau possédant 2 ou 3 cellules cachées est capable de classifier correctement quasiment toutes les figures symétriques ou non symétriques.

Il est à noter que *l'apprentissage par rétropropagation du gradient a conduit les réseaux à utiliser un nombre faible de neurones cachés sans qu'aucun mécanisme de compétition ou de sélection n'intervienne.* Ces quelques neurones sont spécifiques de par la structure spatiale de leurs connexions à la couche d'entrée. Par la définition d'un indicateur numérique rendant compte de cette structuration, nous avons pu corréler les performances d'un réseau à l'émergence de ces structures. Nous avons aussi défini un algorithme sélectif d'apprentissage qui permet d'entraîner des réseaux plus rapidement et avec une amélioration des performances.

Il nous semble que la structuration des poids et la possibilité d'améliorer l'apprentissage par une destruction sélective des cellules sont d'une portée plus large que celle du problème posé.

En l'occurrence, nous avons réussi dans un cas particulier à synthétiser une fonction booléenne complexe (si on prend comme mesure le nombre de termes de sa forme normale disjonctive par exemple) à l'aide d'un réseau beaucoup plus simple.

Il reste donc à affiner les résultats théoriques et à examiner les possibilités d'applications de ce type de synthèse à d'autres fonctions booléennes.

Gilles LEONE
Laboratoire de Physiologie de la Perception et de l'Action
Collège de France, CNRS-Paris
Jocelyn PATUREL
Laboratoire d'Intelligence Artificielle
CEMAGREF,
92185 Antony Cedex
Maurice MILGRAM
PARC, Université Paris VI
75252 Paris Cedex 05

Bibliographie

APPELLE S. (1972) Perception and discrimination as a function of stimulus orientation : the oblique effect in man and animals, *Psychological Bulletin,* **78**, pp. 266-278.

BARLOW H.B., REEVES B.C. (1979) The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research,* **19**, pp. 783-793.

BOCHEREAU L., BOURGINE P. (1990) Extraction of semantic features from a multilayer neural network. *Proc of IJCNN90, Washington DC*.

CARMODY D.P., NODINE C.F., LOCHER P.J. (1977) Global detection of symmetry. *Perceptual and Motor Skills*, **45**, pp. 1267-1273.

CHIPMAN S.F. (1977) Complexity and structure in visual patterns. *Jounal of Experimental Psychology General*, **106**, pp. 269-301.

COHEN N.J., ABRAMS I.T., HARLEY W.S., TABOR L., SEJNOWSKY T.J. (1986) Skill learning and repetition priming in symmetry detection: parallel studies

of human subjects and connectionist models in *Proc. of cognitive science society*, **8**, pp. 23-44.

CORBALLIS M.C., ROLDAN C.E. (1974) On the perception of symmetrical and repeated patterns. *Perception & Psychophysics*, **16**, pp. 136-142.

CORBALLIS M.C., ROLDAN C.E. (1975) Detection of symmetry as a function of angular orientation. *Journal of Experimental Psychology: Human perception and Performance*, **1(3)**, pp. 221-230.

FREYD J., TVERSKY B. (1984) Force of symmetry in form perception. *American Journal of Psychology*, **97(1)**, pp. 109-126.

GOLDMEIER E. (1936) Uber ahnlichkeit bei gesehenen figuren. *Psychologische Forschung*, **21**, 146-208.

HINTON G.E. (1986) Learning distributed representations of Concepts. *in Proc of the 8th ann conf of the Cognitive Science Society*. Hillsdale, Erbaum.

HUBEL D.H., WIESEL T.N. (1962) Receptive fields, binocular interaction and functional architecture of monkey striate cortex. *J. Physiol. (London)*, **160**, pp. 106-154.

HUBEL D.H., WIESEL T.N. (1974) Uniformity of monkey striate cortex: a parallel relationship between field size, scatter and magnification factor. *J. Comp. Neurol*, **158**, pp. 295-306.

JULESZ B. (1971) *Foundations of cyclopean perception*. Chicago: University of Chicago Press.

LE CUN Y. (1987) Modèles connexionnistes de l'apprentissage. *Thèse de Doctorat de l'Université Pierre et Marie Curie*, France.

LE CUN Y. and alii (1989) Optimal brain damage. *in Advances in Neural Information Processing Systems , Denver, Colorado ,* San Mateo, C.A. Kaufmann.

LEHKY S.R., SEJNOWSKI T.J. (1989) Network model for computing surface curvature from shaded images in *Sensory processing in the mammalian brain : neural substrates and experimental strategies.* J.S. Lund (éd.). New York: Oxford University Press, pp. 331-344.

LOCHER P.J., NODINE C.F. (1973) Influence of stimulus symmetry on visual scanning patterns. *Perception & Psychophysics*, **13**, pp. 408-412.

MACH E. (1897) *The analysis of sensation*. Chicago: Open Court Publishing House.

McCLELLAND J.L., RUMELHART D.E. (éds) (1986) *Parallel distributed processing*. Vol 1, Chap 8. MIT Press.

McCULLOGH W., PITTS W. (1943) A logical calculus for the ideas immanent in nervous activity. *Bull. Math. Biophysics*, **5** , pp. 115-133.

MINSKY M., PAPERT S. (1969) *Perceptrons*. Cambridge: MIT Press.

ORBAN G., VANDENBUSSCHE E., VOGELS R. (1984) Meridional variations and other properties suggesting that accuity and orientation discrimination rely on different neural mechanisms. *Ophthalmic Physiol Opt*, **4**, pp. 89-93.

PALMER S.E., HEMENWAY K. (1978) Orientation and symmetry : effect of multiple, rotational, and near symmetries. *Journal of Experimental Psychology: Human perception and Performance*, **4, n° 4**, pp. 691-702.

PELLIZZER G., LEONE G., GEORGOPOULOS A.P. (1992) Moving to the symmetrical direction from a visual stimulus. *Proceedings of the 18thAmerican Neurosciences Congress*, part 2, pp. 1550.

ROCK I. (1974) *Orientation and form.* New York: Academic Press.

ROCK I, LEAMAN R. (1963) An experimental analysis of visual symmetry. *Acta Psychologica*, **21**, pp. 171-183.

ROYER F.L. (1981) Detection of symmetry. *Journal of Experimental Psychology: Human perception and Performance*, **7, n° 6**, pp. 1186-1210.

SEJNOWSKY T.J., KIENKER P.K., HINTON G.E. (1986) Learning symmetry groups with hidden units: Beyond the perceptron. *Physica* , **22 D**, pp. 260-275.

SHIFFRAR M.M., SHEPARD R.N. (1991) Comparaison of cube rotations around axes inclined relative to the environment or to the cube. *Journal of Experimental Psychology: Human perception and Performance*, **17, n° 1**, pp. 44-54.

VOGL T.P., MANGIS J.K., RIGLER A.K., ZINK W.T., ALKON D.L. (1988) Accelerating the convergence of the back-propagation method. *Biological Cybernetics*, **59**, pp. 257-264.