UNIVERSITÉ PARIS-SUD

ÉCOLE DOCTORALE 427 :
INFORMATIQUE PARIS SUD

Laboratoire : *Equipe Parietal, Inria Saclay, CEA Saclay*

Synthèse en français

THÈSE DE DOCTORAT

INFORMATIQUE

par

Michael EICKENBERG

Evaluation de modèles computationnels de la vision humaine en imagérie par résonance magnétique fonctionnelle

Date de soutenance : 21/09/2015

Composition du jury :

Directeur de thèse :	**Bertrand** THIRION	DR, Inria Saclay, Palaiseau
Rapporteurs :	Marcel VAN GERVEN	DR, Donders Institute, Nijmegen, NL
	Nikolaus KRIEGESKORTE	DR, Cambridge University, UK
Examinateurs :	Stéphane MALLAT	Professeur, ENS, Paris
	Yves FREGNAC	DR, CRNS Gif-sur-Yvette, Gif-sur-Yvette
	Balazs KEGL	DR, Université de Paris Sud, Orsay
	Alexandre GRAMFORT	DR, CRNS, Télécom ParisTech, Paris

Evaluation de modèles computationnels de la vision en imagerie par résonance magnétique fonctionnelle

Introduction

L'imagerie par résonance magnétique fonctionnelle (IRMf) est un outil puissant pour étudier le cerveau humain en fonction, de façon globale et non-invasive. Ici nous nous focalisons sur l'étude du fonctionnement du système visuel humain. Les dernières deux décennies ont vu une série d'expériences confirmant des résultats obtenus au préalable chez des animaux ou des humains en condition pathologique, comme la rétinotopie ou la sélectivité à l'orientation dans le cortex visuel primaire. De plus, sa capacité à fournir des images globales nous a permis d'identifier des régions dédiées à une tâche spécifique de la vision. Les modèles les plus fins sont des modèles dits “d'encodage”, basés sur des représentations riches d'un stimulus visuel, qui emploient la méthodologie d'apprentissage statistique pour prédire l’activité cérébrale.

La contribution primaire de cette thèse est de pousser cette méthodologie plus loin en exploitant les représentations hiérarchiques d'informations visuelles créées par les réseaux de neurones convolutionnels de la dernière génération. En effet, ces réseaux de neurones artificiels, entrainés à la reconnaissance d'objet sur plus d'un million d'exemples, sont biologiquement inspirés dans leur architecture de chaines de traitements simples. Ils développent sans information plus précise que la catégorie d’objet de chaque exemple d'entrainement des représentations de plus en plus haute complexité - du détecteur de contours sur le niveau le plus proche à l'image vers une représentation distribuée de la catégorie d'objet dans le dernier niveau, qui se généralise facilement à des catégories non encore vues. Nous évaluons ces représentations quant à leur capacité de prédire l'activité cérébrale et nous découvrons une correspondance entre couches du réseau de neurones artificiel et des aires cérébrales du traitement visuel qui sont, elles aussi, hiérarchiquement organisées et représentent l'information visuelle à plusieurs niveaux d'abstraction.

Pour aboutir à ce résultat, nous présentons d'autres analyses au préalable. Dans un premier temps, nous examinons la structure temporelle de la réponse hémodynamique cérébrale à la stimulation. Dans le cadre d'un système linéaire invariant par le temps, nous proposons d'estimer la réponse impulsionnelle au lieu d'en utiliser une réponse canonique prédéfinie pour les analyses et nous montrons que ceci améliore les capacités prédictives des modèles d'encodage et de décodage. Ensuite, nous présentons une nouvelle pénalité convexe pour le décodage dans le contexte d'une minimisation du risque empirique. Cette pénalité favorise la parcimonie, la structure spatiale locale et fournit des solutions lisses tout en étant homogène en

ses hyperparamètres. Nous montrons une amélioration de la capacité prédictive des modèles de décodage.

Keywords: fMRI, vision, machine learning, deep learning, convolutional networks, object recognition, convex optimization, optimization, regularized risk minimization, hemodynamic response function

Publications:

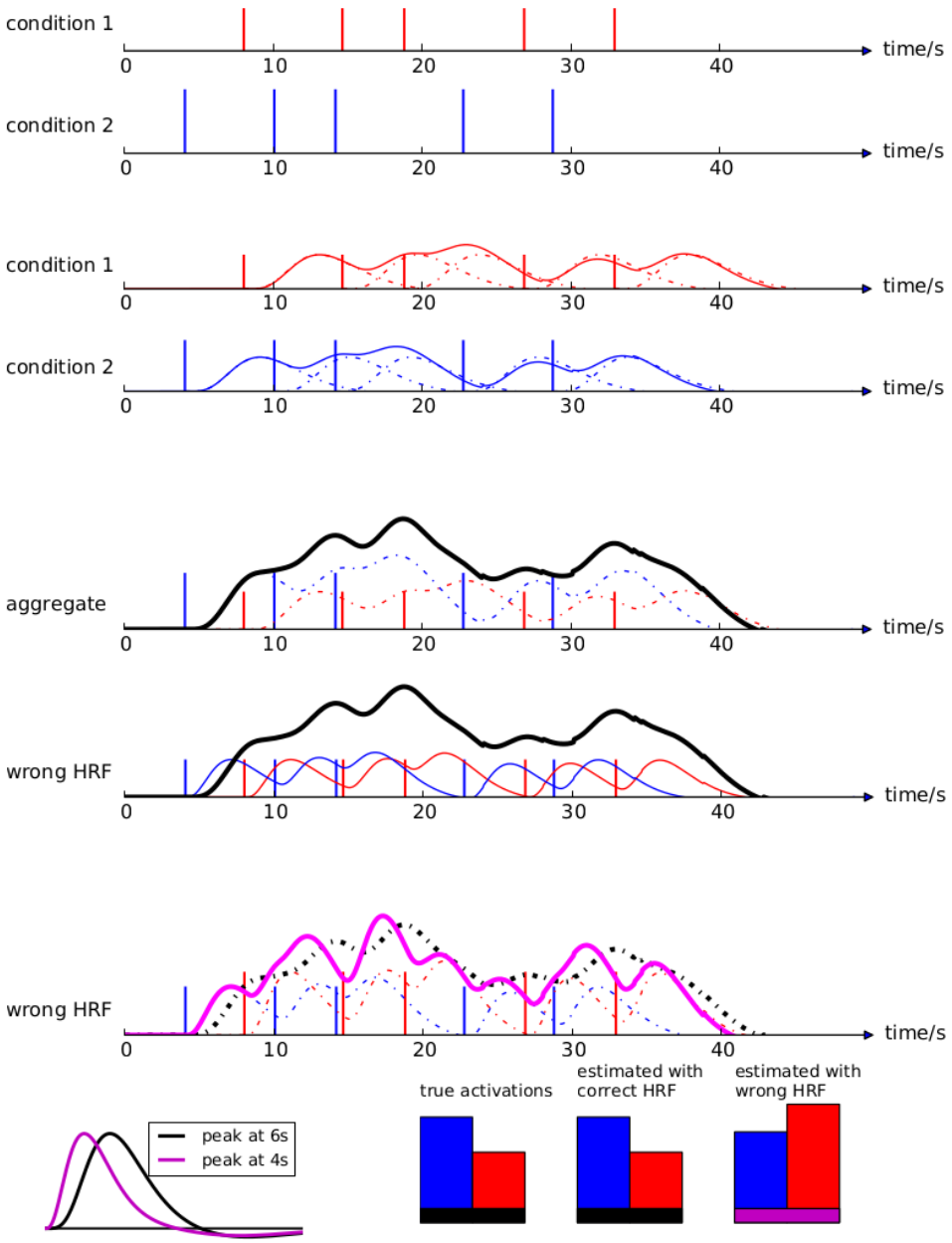
1. “Data-driven HRF estimation for encoding and decoding models”, NeuroImage 2015
2. “Total Variation meets sparsity: Statistical learning with segmenting penalties”, MICCAI 2015
3. “Seeing it all: Convolutional Networks Map the Architecture of the Human Visual System”, NeuroImage (submitted)

D’autres contributions à des conférences

Estimation de la réponse hémodynamique pour l'amélioration de modèles d'encodage et décodage

Une grande partie des expériences IRMf liées à des tâches cognitives est conçue de façon à pouvoir identifier les instants temporels de l'activité neuronale. En effet, une activité visuelle ou auditive est évoquée peu après la présentation d'un stimulus visuel / auditif. Une activité dans le cortex moteur précède l'appui d’un bouton par exemple. Nous savons que la réponse hémodynamique du cerveau liée à une activité neuronale est étalée et retardée dans le temps. Dans une analyse de données BOLD il faut tenir compte de ce délai, d'autant plus si les événements d'activité neuronale d'interêt sont peu espacés temporellement. La réponse hémodynamique à une activité neuronale ponctuelle est maximale entre 4 et 6 secondes après l'événement, elle devient légèrement négative à environ 10 secondes et se dissipe complètement au bout de 20s. Comme l'activité hémodynamique est, sauf cas extrême, un système linéaire invariant dans le temps avec comme entrée l'activité neuronale, elle est entièrement caractérisée par sa réponse impulsionnelle.

Dans les analyses classiques, une fonction de réponse hémodynamique (HRF) canonique fixe est utilisée pour estimer le niveau d'activité neuronale sachant les moments dans le temps lorsqu’elle a dû se produire: Le système linéaire à résoudre contient des régresseurs indicateurs d'un certain type d'événement convolués avec la HRF. Cette approche fonctionne toujours si les événements sont suffisamment espacés. Si les événements sont très proches dans le temps, la forme précise de la HRF devient importante.



Dans la figure ci-dessus, une série temporelle BOLD suite à une expérience IRMf avec deux types d'événements neuronaux a été simulée en supposant une réponse hémodynamique avec intensité maximale à un délai de 6s. L'analyse avec un modèle linéaire employant une HRF avec un délai de 4s entraine une mauvaise estimation des coefficients des deux conditions, à tel point que l'ordre des intensités des deux activations est inversé.

Ceci présente une motivation d'évaluer l'impact d'une estimation de la réponse hémodynamique dans l'analyse des données sur les analyses secondaires menées par la suite. Nous nous concentrons ici sur des analyses modernes d'encodage (prédicition de l'activité cérébrale à partir du stimulus) et de décodage (prédiction d'une variable externe, par exemple le stimulus, à

partir des données cérébrales). De plus, nous supposons que les événements neuronaux et les mesures IRMf se trouvent tous sur une grille temporelle d'espacement régulier.

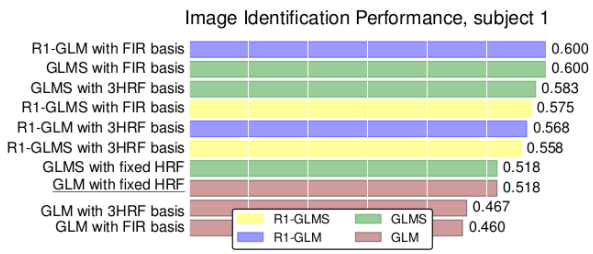
Nous estimons la HRF soit point par point en temps, soit en tant que combinaison linéaire d'un ensemble de fonctions - une base de HRF. X représente une matrice de dessin. Dans l'analyse classique, ses colonnes représentent les événements neuronaux convolués avec la HRF. Ici nous avons plusieurs colonnes par événement où l'indicatrice des événements est convoluée avec les fonctions de base. Nous supposons que chaque voxel a sa propre HRF qui sera la même à travers les types d'événements.

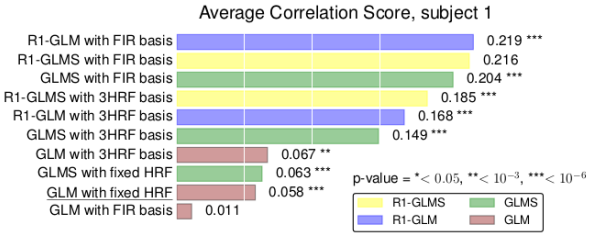
Ceci nous amène à un problème bilinéaire de moindres carrés que l'on peut formuler de façon succincte avec des produits de Kronecker:

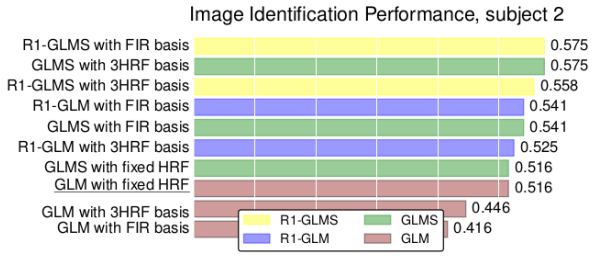
$$\frac{1}{2}\|\mathbf{y} - \mathbf{X_B}(\boldsymbol{\beta} \otimes \mathbf{h}) - \mathbf{Z}\boldsymbol{\omega}\|^2 =$$
$$\frac{1}{2}\|\mathbf{y} - \mathbf{X_B}(\mathbf{I} \otimes \mathbf{h})\boldsymbol{\beta} - \mathbf{Z}\boldsymbol{\omega}\|^2 =$$
$$\frac{1}{2}\|\mathbf{y} - \mathbf{X_B}(\boldsymbol{\beta} \otimes \mathbf{I})\mathbf{h} - \mathbf{Z}\boldsymbol{\omega}\|^2 \ .$$

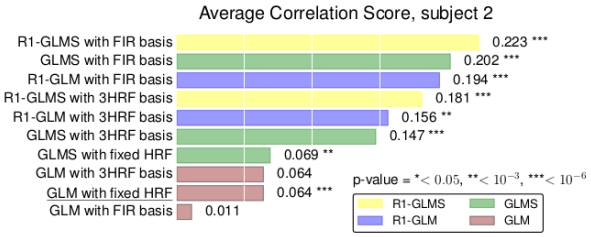
y représente la mesure BOLD, beta les activations et h la réponse hémodynamique. Z contient des régresseurs de nuisance en basse fréquence. Pour sa solution unique, nous imposons une contrainte de signe et de norme sur la réponse hémodynamique h. Nous proposons un algorithme d'optimisation jointe des valeurs recherchées. Nous étendons également le GLM séparé (Turner et al., 2012) à être conforme à ce modèle.

Sur un jeu de données d'encodage d'information visuelle (1750 images naturelles présentées à deux sujets toutes les 4 secondes dans un scanner IRMf) nous évaluons la performance des analyses secondaires menées après avoir appliqué ou non notre approche pour l'analyse primaire.



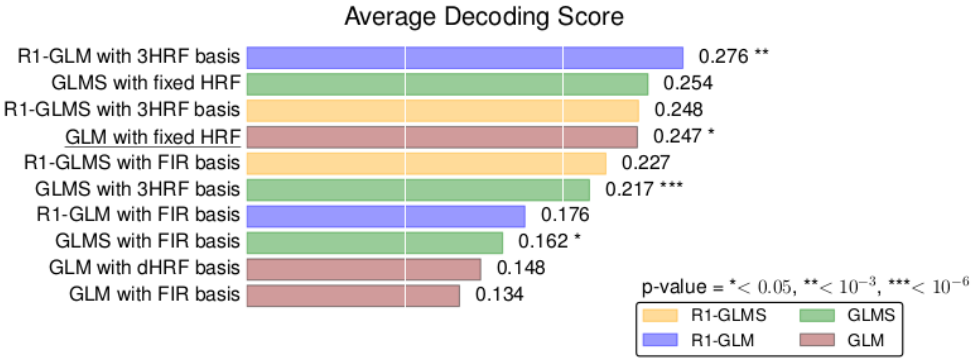






Nous observons que nos approches proposées (R1-GLM et R1-GLMS) produisent parmi les meilleurs résultats, mais que le choix de la base (FIR: point par point; 3HRF: base de HRF canonique et ses dérivées) joue un rôle. Ceci vaut pour la mesure du taux d'identification d'image présentée parmi 120 images possibles et pour la prédiction du signal BOLD brut sur des cas non observés.

Sur un jeu de données de décodage (prédiction du gain potentiel dans un jeu de pari sur gain et perte) nous avons évalué l'analyse primaire avec notre méthode contre les méthodes classiques en calculant les taux de prédiction du gain à partir des cartes cérébrales obtenues dans une analyse secondaire:



Similairement aux modèles d'encodage, l'estimation de la HRF apporte des gains significatifs.

Regroupement de variation totale et parcimonie: Une pénalité convexe promouvant la segmentation

Le décodage, dans le contexte de l'analyse de données cérébrales, signifie la prédiction d'une variable externe/phénotypique à partir d'une mesure du cerveau. Celle-ci pourrait être une propriété, la classe, ou l'identité exacte d'un stimulus présenté lors d'une acquisition fonctionnelle, le diagnostic d'une certaine condition clinique ou d'autres paramètres comme l'age ou le sexe.

Très souvent, le décodage se fait via un modèle statistique linéaire. Ceci est principalement dû à la très haute dimensionalité du problème (un voxel correspond à une variable et une image cérébrale en contient typiquement des dizaines de milliers de voxels) ce qui fait échouer la plupart des méthodes non-linéaires qui se caractérisent par une grande variabilité: Il n'y a simplement pas assez de données pour estimer ces méthodes de façon fiable.

Un avantage des modèles linéaires est que le méchanisme prédictif se fond sur un produit scalaire avec un vecteur de poids, qui peut être visualisé en tant qu'image cérébrale - puisqu'il a exactement la même dimensionalité - et amené à une interprétation. L'interprétation d'une carte de poids doit être faite de façon globale et non pas voxel par voxel, en tenant en compte le fait que les poids ne représentent pas uniquement des zones bien corrélées avec la cible de la prédiction.

Plusieurs problèmes compliquent l'estimation et l'interprétabilité. Le fait qu'il y ait beaucoup moins d'exemples que de variables dans la carte des poids rend le problème sous-déterminé et donc mal-posé. Les exemples disponibles sont de plus très correlés entre eux, ce qui rend le problème d'estimation mal-conditionné. Pour remédier à ces difficultés, le problème initial peut être régularisé. L'approche la plus simple est de contraindre la norme euclidienne de la solution. En formulation lagrangienne ceci s'écrit sous la forme d'une pénalité L2 qui est additionnée au problème d'optimisation.

Avec un peu de prudence, une régularisation additive peut souvent être interpreté comme une information a priori dans un contexte probabiliste, plus spécifiquement dans le cadre d'une estimation du maximum a posteriori. Cette information nous motive à faire un choix de penalité le plus renseigné possible sur la structure que nous attendons pour la solution. Notre information a priori se résume à deux aspects: La parcimonie du vecteur de poids et sa contiguité spatiale. La parcimonie rend compte du fait que ce sont souvent très peu de modules cérébraux qui portent de l'information concernant la tâche de prédiction. La composante spatiale traduit la contigüité spatiale des zones actives qui le plus souvent définissent des régions entières actives.

Il est important de noter que la seule parcimonie, induite par exemple par une pénalité L1, n'est pas satisfaisante dans ce contexte, puisque les poids actifs sont dispersés et laissent difficilement appréhender la structure sous-jacente. Comme les variables dans un même voisinage ont tendance à être très corrélées entre elles, seule une variable parmi elles sera sélectionnée par cette méthode, et ce sera une autre si on rééchantillonne les exemples. La pénalité "ElasticNet", conçue pour le cas des variables corrélées ne conduit pas à des résultats sensiblement supérieurs. Bien qu'elle rende possible une sélection de variables corrélées, la

sélection peut facilement être bruitée, car la nature de cette régularisation permet des corrélations entre toute paire de variables, peu importe leur écart spatial.
Pour cette raison, la régularisation spatiale a été introduite pour les modèles de décodage en IRMf. En 2009 et 2011, Grosenick introduit GraphNet, un régulariseur qui consiste en une somme pondérée de norme L1 et norme L2 au carré du gradient spatial de l'image. Ceci fournit des cartes parcimonieuses avec régions actives lisses. Encore en 2011, V. Michel introduit la régularisation par variation totale (norme euclidienne du gradient spatial en chaque point sommée sur toute l'image) du signal, privilégiant ainsi une estimation de cartes constantes par morceaux. En 2012, Baldassare rajoute à la variation totale une norme L1. La somme pondérée des deux s'appelle TV+L1. Cette pénalité enforce des régions parcimonieuses et constantes par morceaux donnant lieu à des cartes avec une majorité de poids à zéro et les zones actives étant constantes. La régularisation GraphNet et la régularisation TV+L1 représentent en réalité une classe de pénalités paramétrées par deux hyperparamètres: L'un règle l'intensité de la régularisation et l'autre détermine la pondération entre parcimonie et contiguité spatiale. Pour sélectionner un couple de paramètres, bien qu'il existe d'autres méthodes, la validation croisée sur des données manquantes en prédiction de la cible semble indiquée pour ce choix. En faisant ce type de sélection on observe que GraphNet, par l'inhomogénéité de la somme pondérée de ses deux composantes, rend la bonne sélection des paramètres plus difficile que celle de TV + L1. Les résultats sont souvent des estimations bruitées des cartes de poids, contrairement à la pénalité TV + L1. En revanche, la forme du signal enforcé par TV + L1 ne correspond pas tout à fait à ce que nous cherchons: Des activations lisses et non constantes par morceaux.

$$\Omega_{\mathrm{GN}}(w) = \lambda((1-\rho)\|\nabla w\|_2^2 + \rho\|w\|_1)$$

$$\Omega_{\mathrm{TV}-\ell_1}(w) = \lambda((1-\rho)\|\nabla w\|_{2,1} + \rho\|w\|_1)$$

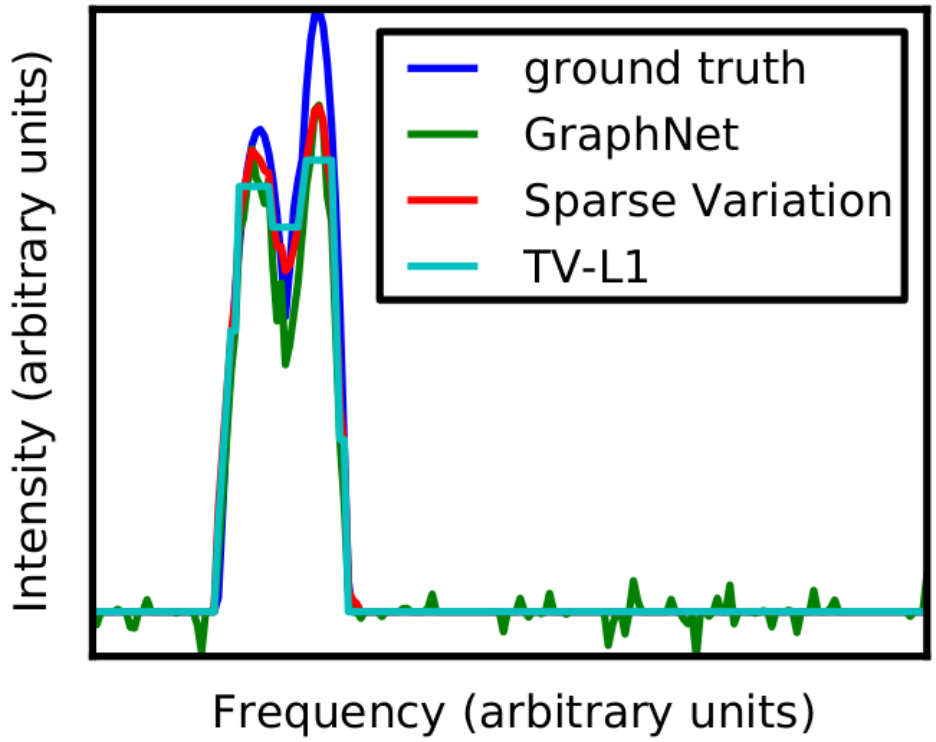
En résumé, les formules de Graphnet et TV + L1 sont
Nous sommes en mesure de remédier aux problèmes des deux pénalités en introduisant une nouvelle pénalité convexe, que nous appelerons “Sparse Variation”:

$$\Omega_{\mathrm{SV}}(w) = \lambda\|Kw\|_{2,1},$$

$$K = \begin{pmatrix} (1-\rho)\nabla \\ \rho\,\mathrm{Id}_p \end{pmatrix}$$

Cette nouvelle pénalité regroupe en une norme euclidienne les trois composantes de la dérivée spatiale ainsi que la coordonnée dans laquelle cette dérivée est prise. Ensuite, ces normes sont sommées spatialement. La parcimonie enforcée par cette structure est jointe en coordonnée et en dérivée: Soit la coordonnée ainsi que les dérivées sont nulles, soit la coordonnée et la dérivée peuvent être actives librement, évitant ainsi des situations dans lesquelles la coordonnée est active mais non la dérivée (effet escalier). Le résultat est un ensemble de zones lisses, spatialement contigües et parcimonieuses.

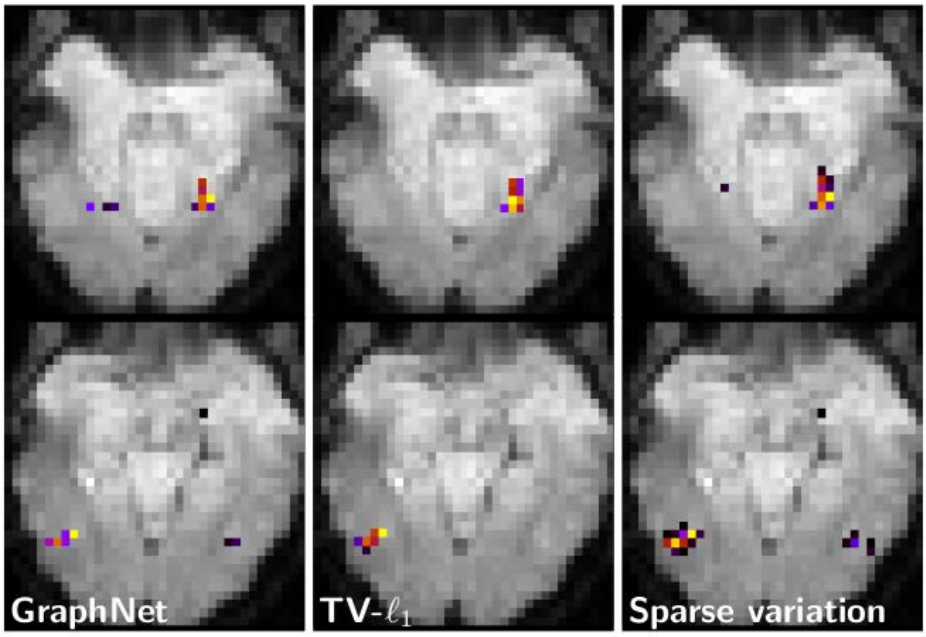
Pour visualiser les effets décrits nous montrons un exemple en une dimension:



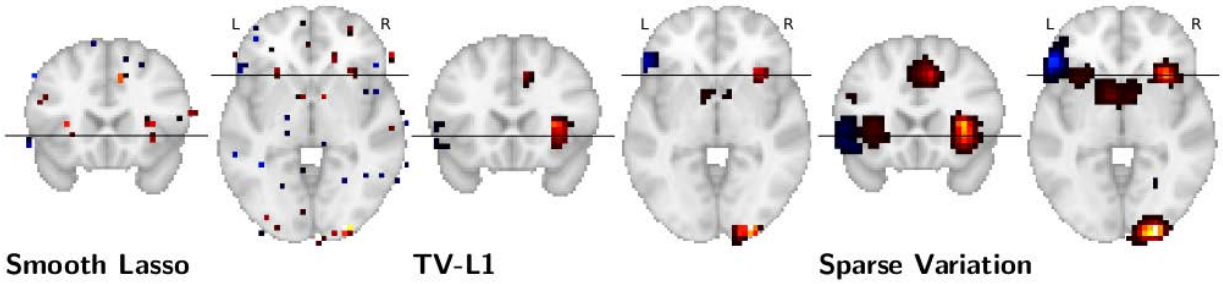
Dans le diagramme ci-dessus nous évaluons la reconstruction d'un signal parcimonieux, lisse, mesuré par 30% de ses coefficients DCT sous bruit Gaussien additif. La vérité-terrain est marquée en bleu et les paramètres sélectionnés pour chaque méthode sont ceux qui donnent la meilleure prédiction en validation croisée. Dans cette situation, nous observons que la solution GraphNet est la plus bruitée, bien qu'elle suive assez bien la forme de la vérité-terrain. En revanche, la solution TV+L1 est parfaitement débruitée, mais il devient évident que la forme imposée par cette pénalité ne correspond pas à ce que nous cherchons: Elle enforce des régions plates au lieu de lisses. C'est la solution pénalisée en Sparse Variation qui remédie aux problèmes des deux autres en présentant une solution non bruitée qui est lisse et qui suit bien la forme de la vérité-terrain. Une mesure de corrélation le confirme également.

Nous procédons à une évaluation sur des données IRMf en présentant trois exemples d'utilisation de Sparse Variation: 1) Une étude fonctionnelle visuelle [Haxby2001], 2) une étude fonctionnelle de comportement [Jimura2011], 3) une étude anatomique OASIS.

Sur le premier jeu de données, [Haxby2001], nous faisons de la classification entre catégories d'objets présentés à un sujet. Nous évaluons les contrastes classiques “faces vs places” et “objects vs scramble”.

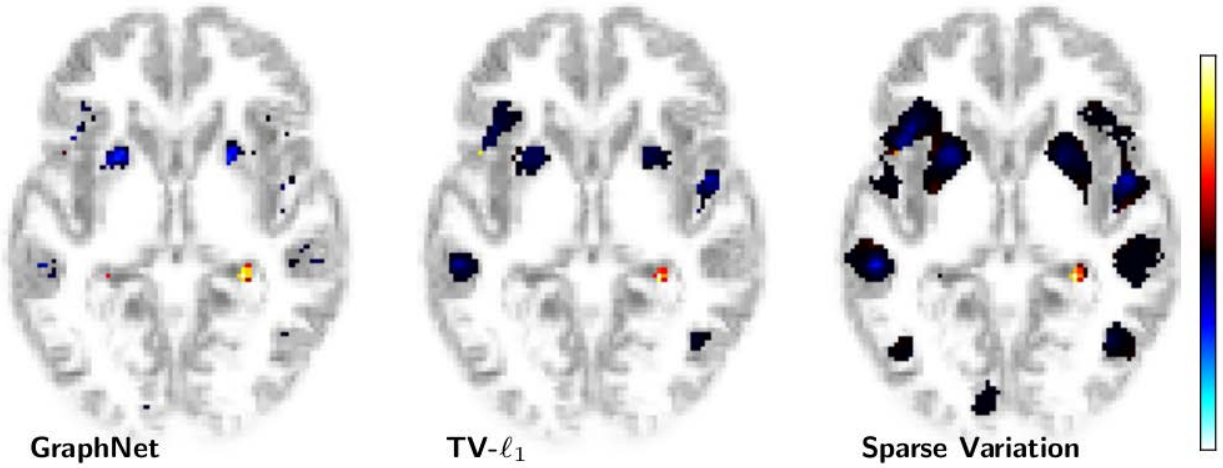


Sur la figure ci-dessus, dans la première ligne, nous montrons la carte spatiale obtenue avec des hyperparamètres optimaux pour la prédiction “faces vs places”, pour GraphNet, TV-L1 et Sparse Variation. Dans la deuxième ligne nous étudions le contraste “objects vs scramble”. Nous observons que les cartes sont assez similaires entre elles. Cependant, Sparse Variation sélectionne des régions légèrement plus grandes que les autres. Il s'agit d'une étude sur un seul sujet à la fois où les effets sont assez serrés spatialement, d'où la similarité entre cartes.



La deuxième étude est une expérience comportementale dans laquelle les sujets avaient à décider s'ils voulaient faire un pari étant donnés les potentiels de gains et de pertes qui pouvaient varier. Nous proposons de prédire les gains potentiels de chaque pari en un problème de régression linéaire. Avec des hyperparamètres optimaux pour la prédiction, nous observons que GraphNet sélectionne une parcimonie très penché vers une pénalisation L1 pure. Ici la sélection du modèle semble avoir échouée par rapport aux deux autres méthodes qui trouvent bien des zones prédictives et spatialement contigües. Il s'agit d'une étude multi-

sujet: Dû à la variabilité inter-sujet, on peut s'attendre à des zones prédictives étendues. Nous observons que Sparse Variation favorise des régions plus lisses.



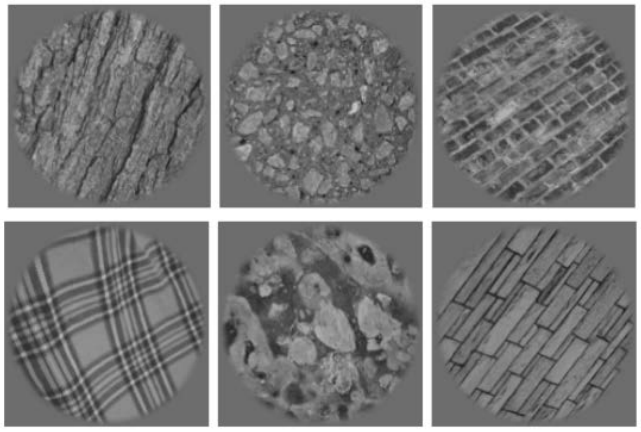
La figure ci-dessus montre les résultats d'une régression linéaire sur des données d'IRM structurelle (anatomique) avec comme cible prédictive l'age du sujet. L'étude des cartes obtenues avec des hyperparamètres optimaux montre une sélection parcimonieuse pour GraphNet, une sélection plus contigüe pour TV+L1, et une sélection lisse et symmétrique pour Sparse Variation.

En somme, nous avons introduit une nouvelle pénalité convexe qui favorise la parcimonie, la contiguité spatiale et des solutions lisses. Nous trouvons qu'elle permet une meilleure sélection du modèle que GraphNet et encode un a priori plus adéquat que TV-L1 pour l'estimation de cartes cérébrales.

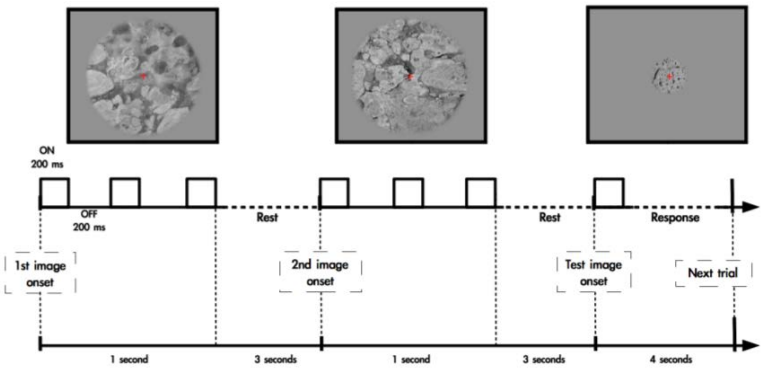
Etude IRMf de la réponse cérébrale aux textures visuelles

La reconnaissance et la compréhension des scènes visuelles qui entourent un individu est un méchanisme d'importance primordiale pour un grand ensemble d'espèces, y compris les humains. L'image retinale deux-dimensionnelle doit être analysée et interpretée pour en extraire de l'information. Ayant l'organisation spatiale de la lumière incidente dans les yeux, un certain nombre de calculs doit être effectué pour en tirer des renseignements. La première opération doit être effectuée sur les valeurs d'intensité elles mêmes. Un traitement bas-niveau, au niveau de la première aire visuelle V1, est le filtrage avec des filtres passe-bande orientés donnant ainsi la capacité d'extraire des contours visuels. Il se trouve que très souvent, la frontière entre un objet et le fond est caractérisée par un tel changement d'intensité. Cependant, il existe des cas plus compliqués, notamment quand un objet et son fond sont texturés. Dans cette situation,

il peut y avoir des contours d'intensité partout et ce n'est que dans une analyse de la variation de ces contours que l'on pourrait identifier une frontière entre textures. Ceci nécessite donc un modèle à deux niveaux. Le cerveau humain sépare bien différentes textures naturelles. De plus, il est aussi capable de les identifier (sans contraste avec une autre texture). Nous étudions ici la réponse du cerveau humain à la présentation de textures visuelles pour une analyse en encoding avec des modèles à deux couches. Nous étudions aussi les effets dûs aux différentes classes de texture en fonction des aires cérébrales considérées.

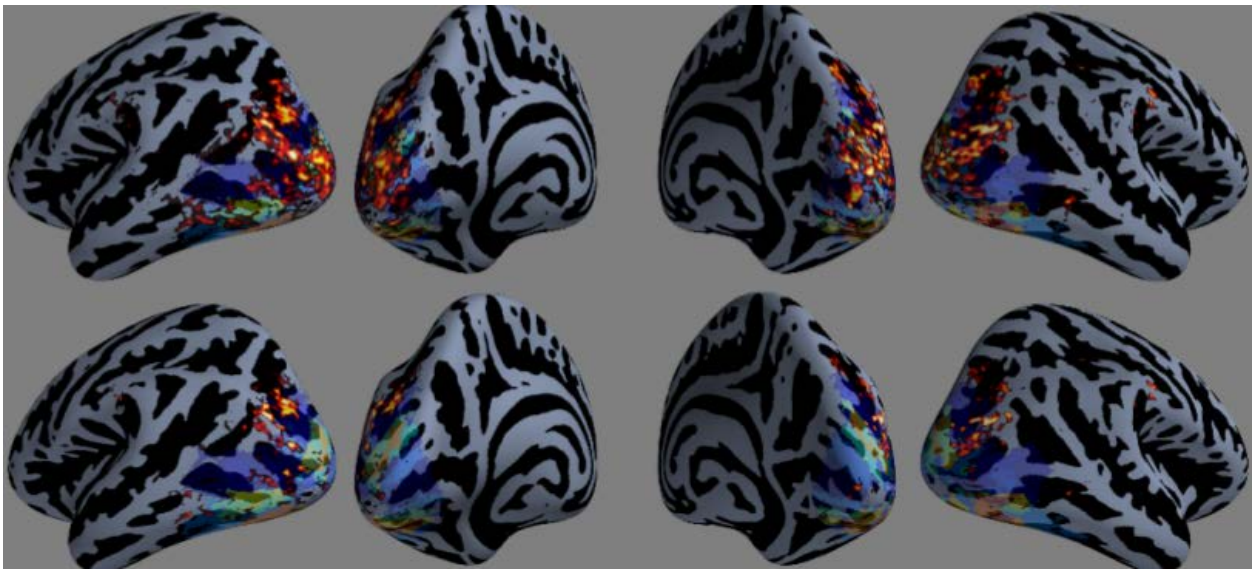


Ci-dessus les différentes classes de textures utilisées.



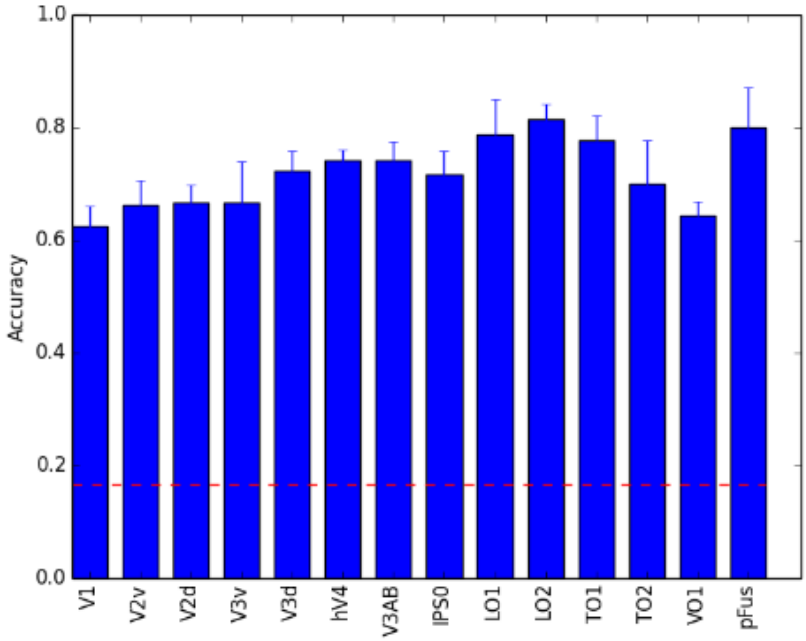
La figure ci-dessus décrit l'expérience comportementale effectuée dans l'IRM: Pour chaque essai, deux images différentes de la même classe (parmi 6) de texture sont présentées à un écart de 4s. Encore 4s après la deuxième image présentée, un extrait de soit la première, soit de la deuxième, soit d'une autre image est présenté et le sujet doit indiquer quelle correspondance a été perçue en appuyant sur un bouton.

Après une analyse avec un GLM, nous pouvons montrer les régions du cerveau qui répondent généralement à la présentation d'une texture. Ensuite, nous nous intéressons aux régions qui répondent differemment aux classes de texture.



Ci-dessus, dans la première ligne, nous observons l'effet “texture”. Les voxels qui montrent une réponse significative aux textures, seuillée à t=3.0 sont montrés sur un atlas cérébral. L'activité est ventrale et dorsale et les aires primaires de la vision sont actives. Dans la deuxième ligne nous montrons l'effet “différence entre textures”. On voit immédiatement que les zones répondant différemment aux textures sont plus restreintes, avec un accent sur la voie dorsale vers le sillon intrapariétal. Les aires primaires V1 et V2 ne montrent pas de distinction significative.

Pourtant, l’information permettant de décoder la classe de texture est aussi présente, au moins globalement, dans les aires primaires de la vision. Ceci devient visible en évaluant la capacité d'un décodeur de classe de texture depuis des aires visuelles différentes indiquées par un atlas.



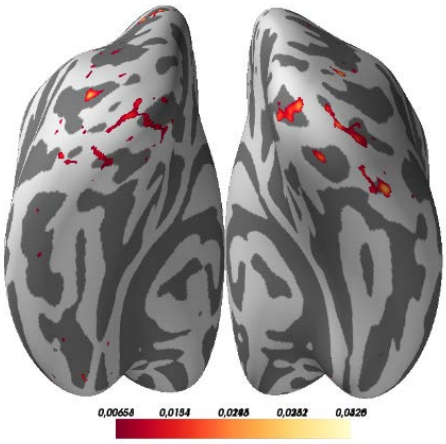
Ci-dessus sont affichés les taux de prédiction de la bonne classe de texture à partir des images cérébrales. Des aires bas-niveau V1 vers les aires plus haut-niveau LO et TO, nous voyons un accroissement des taux de prédiction: L'information sur la classe de textures est plus facilement accessible linéarement lorsque la région est plus haut-niveau.

De plus, nous effectuons une analyse d'encodage avec des coefficients de scattering extraits des images de texture. La transformée de scattering est une cascade de transformées en ondelettes complexes suivies par la prise du module complexe. En formules, pour une cascade de deux couches, cela se traduit en

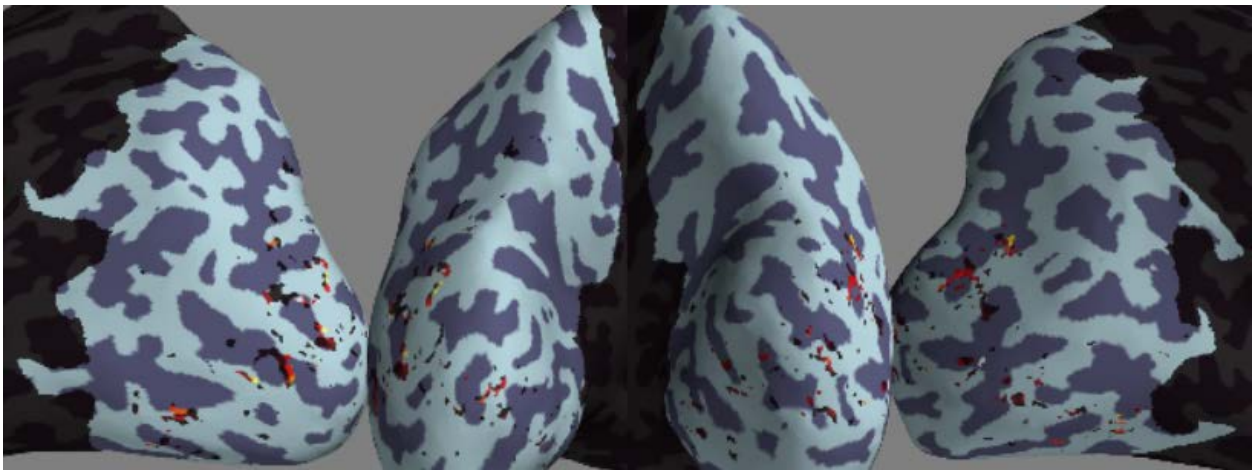
$$(W_{\gamma,j}u)(x) = |\psi_{\gamma,j} * u|(x)$$
$$(W_{\gamma_2,j_2}W_{\gamma_1,j_1}u_0)(x) = |\psi_{\gamma_2,j_2} * |\psi_{\gamma_1,j_1} * u_0||(x)$$

pour une image u(x).

En évaluant à travers les 6 séances acquises sur 3 sujets les voxels répondant le plus fort à la combinaison des couches 1 et 2 qu'à la seule couche 1, nous obtenons une carte indiquant des traitements du deuxième ordre:



Le diagramme ci-dessus indique les endroits où la deuxième couche de scattering apporte un gain significatif en prédiction (FDR < 1%). Mis à part V1, on identifie des régions dorsales similaires à celles du contraste des différences entre textures. En effectuant la même analyse sur le jeu de données d'images naturelles quelconques de [Kay2008], nous observons un effet similaire:



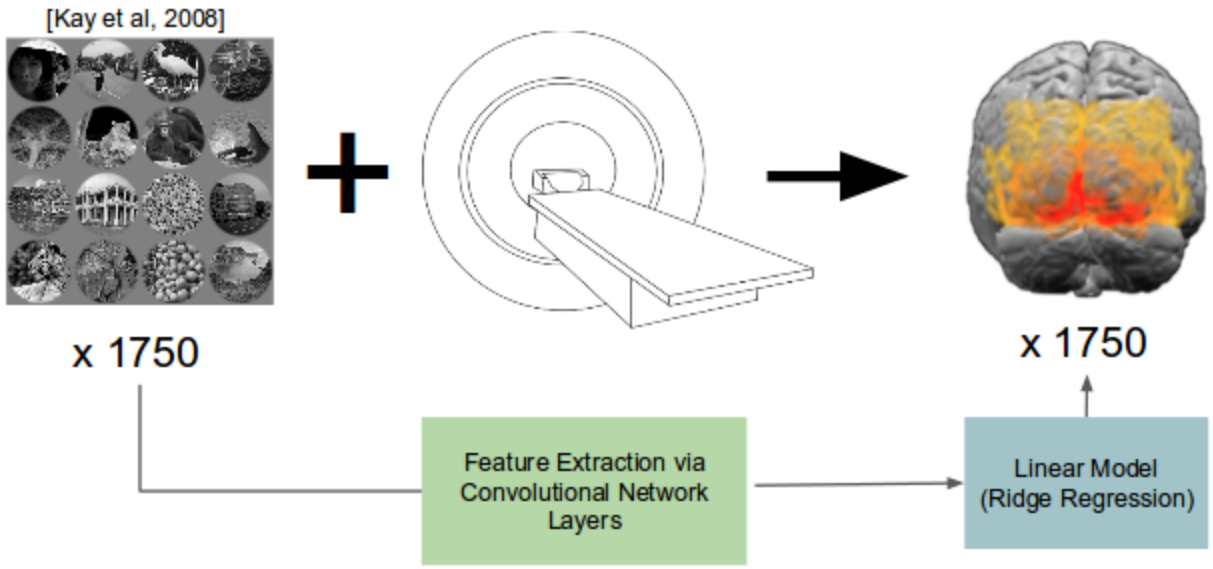
En conclusion, nous avons analysé la réponse cérébrale aux textures.

Liens entre l'architecture d'un réseau de neurones convolutionnel pour la détection d'objet et le système visuel humain étudiés en IRMf

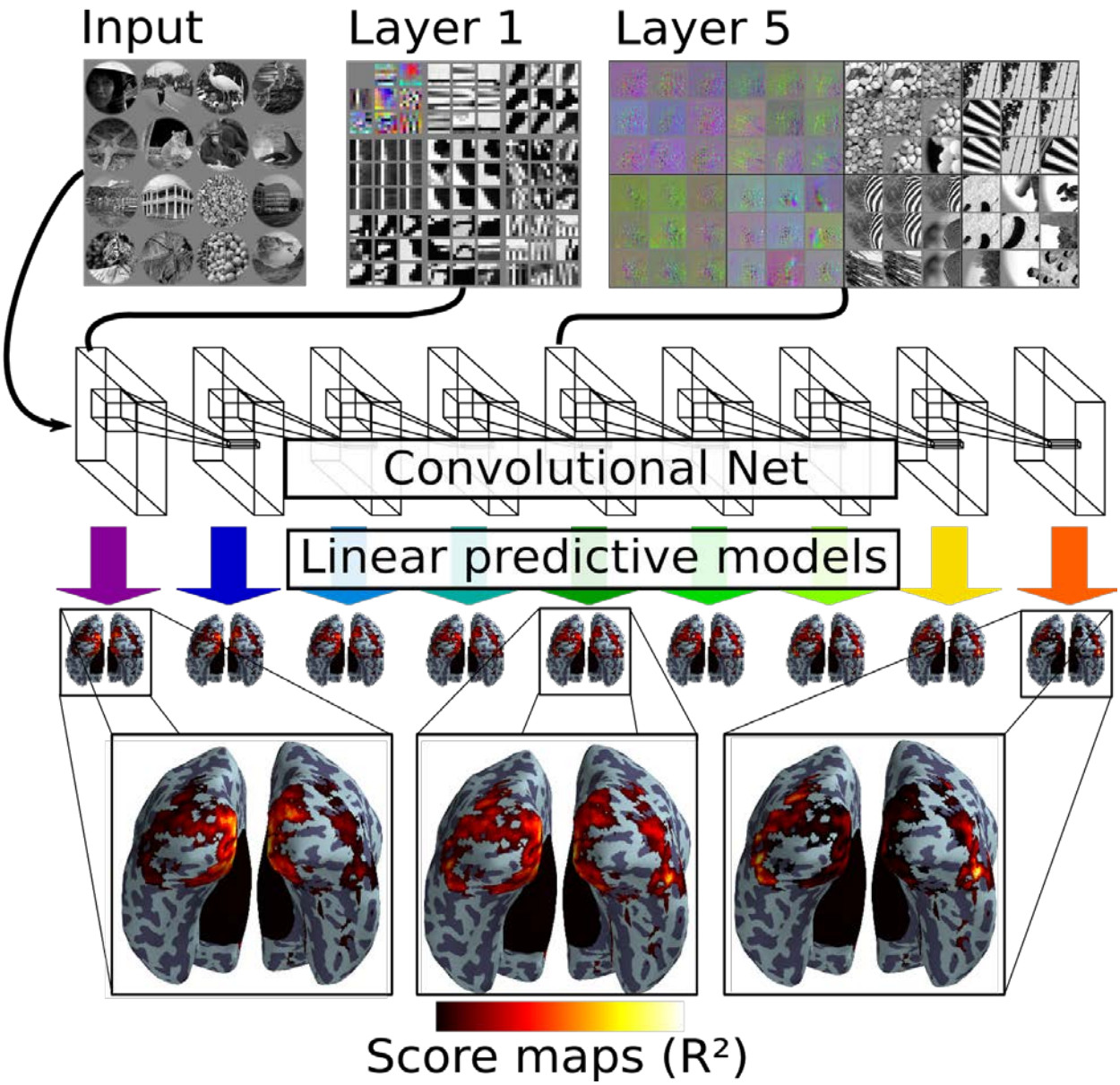
Récemment, la vision par ordinateur et notamment la reconnaissance d'objet a été révolutionnée par les réseaux convolutionnels profonds. En effet, dans des domaines auparavant associés à la supériorité des animaux et humains en particulier, les réseaux de neurones convolutionnels ont gagné beaucoup de terrain et dépassent sur certaines tâches spécifiques la capacité d'un cerveau humain.

Les réseaux de neurones convolutionnels sont inspirés par la biologie le long de leur développement. Bien que leur formalisme soit abstrait, nous y retrouvons une architecture de traitement multi-couche avec des opérations linéaires suivies par des non-linéaritées point par point ou très localisées. La réplication des filtres convolutionnels sur toute l'image rend compte de l'invariance par translation de l'ensemble des images naturelles qui est aussi exploitée en partie par le cerveau humain.

Les données de [Kay2008] fournissent 1750 images cérébrales IRMf en réponse à une stimulation avec des images naturelles. Nous allons explorer la capacité prédictive de chaque couche d'un réseau de neurones convolutionnel confronté à l'activation cérébrale. En inspectant la première couche d'un réseau convolutionnel, on observe des détecteurs de contour avec une fonction similaire à ce que fait la première aire visuelle V1. De plus, l'avant-dernière couche d'un réseau de neurones de reconnaissance d’objet représente l'image d'entrée de façon similaire à un ensemble de neurones IT dans le macaque. Ayant ainsi retrouvé des correspondances des deux côtés d'une architecture de réseau convolutionnel, nous nous intéressons à la capacité prédictive des couches intermédiaires dans des aires intermédiaires de la vision.

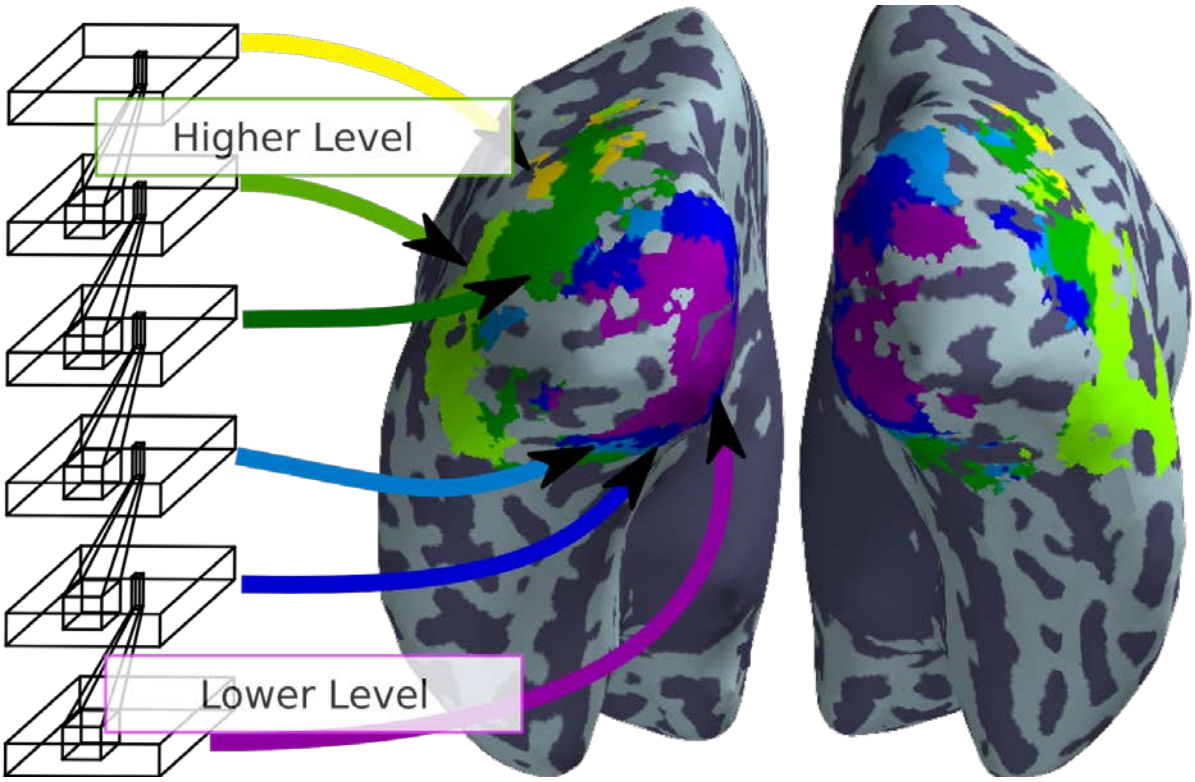


L'expérience IRMf conduisant au jeu de données utilisé consiste en la présentation d'un stimulus naturel toutes les 4 secondes. Chaque stimulus d'entrainement est présenté deux fois. Les stimuli de validation sont présentés 13 fois.



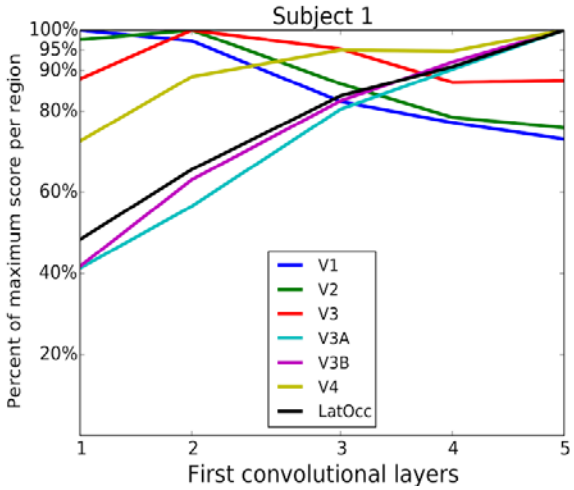
La figure ci-dessus décrit notre approche expérimentale. Pour chaque stimulus nous sauvegardons l'activation dans chaque couche du réseau de neurones OverFeat, un modèle pré-entrainé disponible sur Internet. Depuis la sortie de chaque couche, nous sommes en mesure de créer un modèle prédictif de l'activité cérébrale mesurée. Nous pouvons évaluer la capacité prédictive de chaque modèle dans tous les voxels du cerveau en affichant par exemple le taux R^2. En bas de la figure, trois exemples sont agrandis: A gauche, la première couche prédit bien dans une zone médiale occipitale qui contient V1 et V2 et à droite nous observons une capacité prédictive relativement réduite en V1/V2 mais un taux de prédiction plus élevé dans des aires latérales occipitales et dorsales.

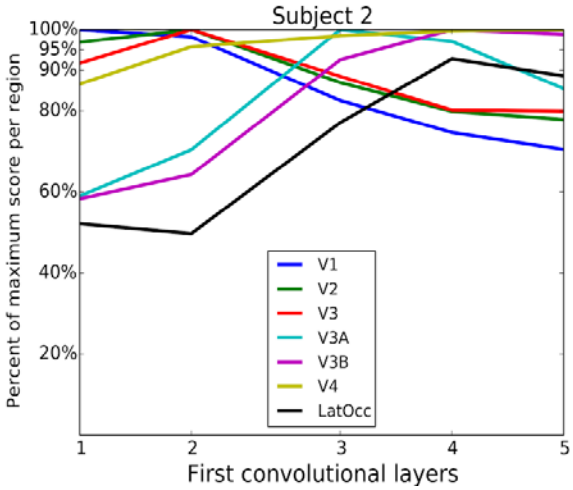
En suivant le code couleur des couches dans ce diagramme, nous montrons sur un seul volume cérébral la couche qui prédit le mieux chaque voxel:



La première couche présente une meilleure prédiction dans les zones médiales occipitales observées auparavant. La zone la mieux prédite par la deuxième couche est adjacente à celle la mieux prédite par la couche 1, et ce de façon bilatérale du côté ventral et dorsal. De même pour la troisième couche. Ceci nous amène à dire que la hiérarchie de la complexité des opérations visuelles présente dans le cerveau est bien approximée par les premières couches d'un réseau de neurones convolutionnel.

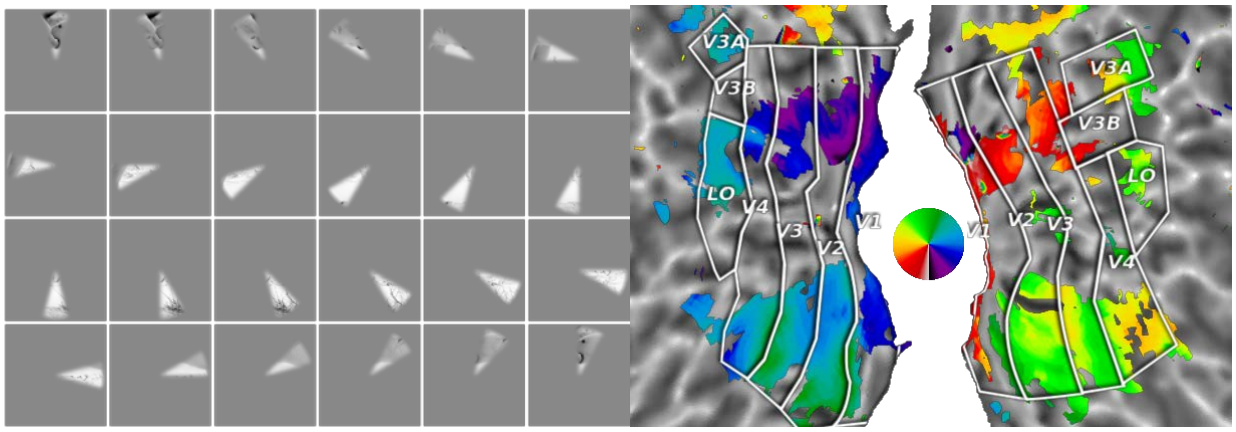
En utilisant les régions d'interêt fournies par le jeu de données, nous pouvons également évaluer les taux moyens de prédiction dans chaque aire visuelle en fonction de la couche employée pour la prédiction.



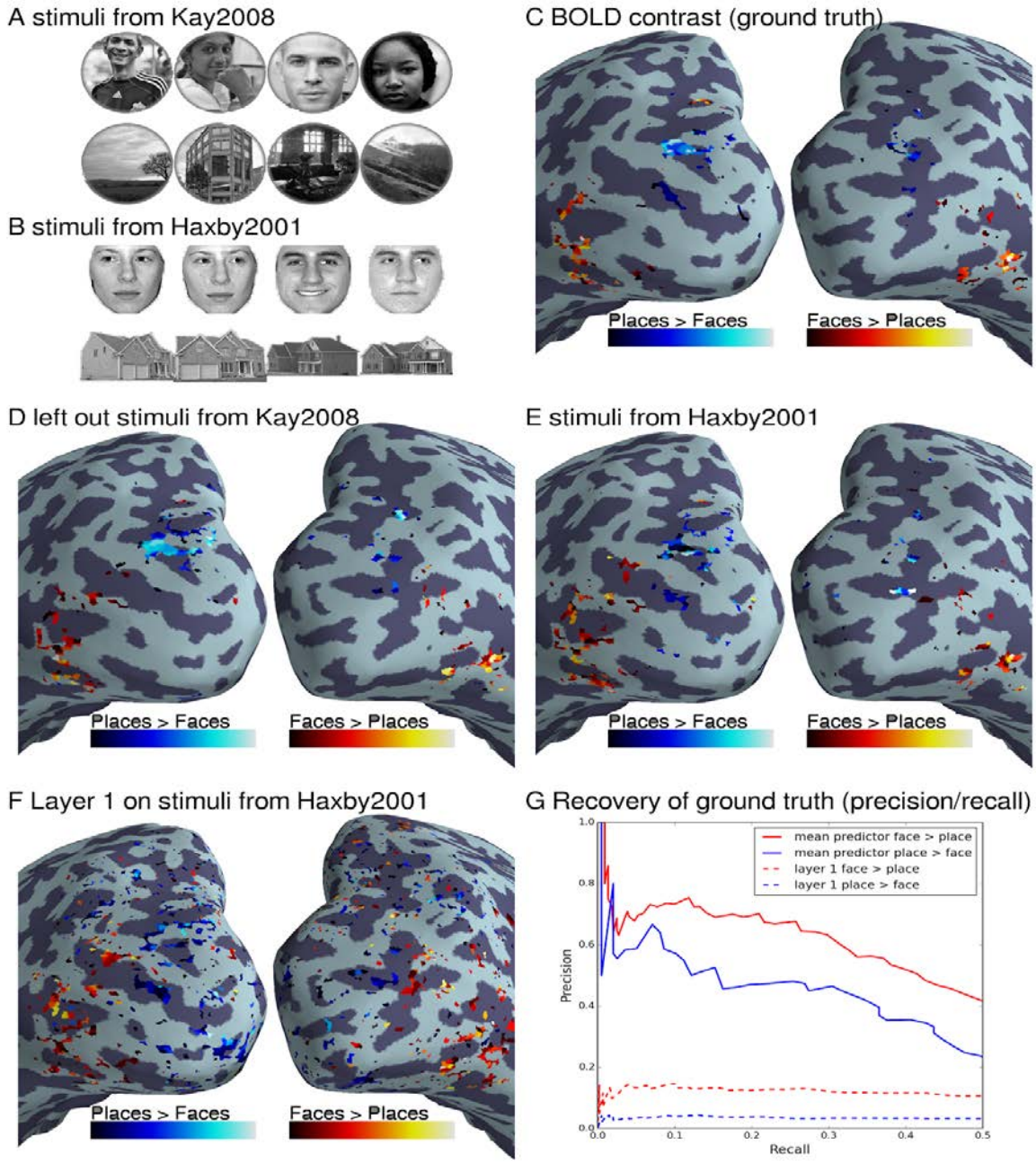


Pour les deux sujets disponibles, nous observons les mêmes dynamiques: L'aire visuelle V1 est la mieux prédite par la première couche et presque aussi bien par la deuxième couche, avec un déclin monotone pour les couches suivantes. L'aire V2 est la mieux prédite par la deuxième couche, mais presque aussi bien avec la première, avec un déclin vers les couches de haut niveau. L'aire V3 suit la même dynamique que V2 mais avec un déclin plus lent, suggérant ainsi une représentation plus complexe globalement. Pour V4 les taux sont ascendants et pour les aires de niveau plus haut, V3A/B et LatOcc, nous observons que les couches haut-niveau sont bénéfiques et que la croissance est quasi-linéaire. Nous en déduisons une deuxième fois une correspondance entre les niveaux d'abstraction créés par le réseau de neurones et ceux créés par le cerveau humain.

Si notre modèle prédictif est suffisamment puissant, nous devrions pouvoir synthétiser de nouvelles données. Ceci est effectivement possible et nous le démontrons à deux niveaux: La rétinotopie en sera une confirmation bas-niveau et un contraste sémantique entre visages et lieux sera une confirmation haut-niveau.
Pour la rétinotopie nous créons des stimuli appropriés tels que visualisés dans la partie gauche du diagramme suivant:



A droite nous obtenons, après un GLM de codage de phase, une carte de la correspondance des angles visuels sur le cerveau. Les changements de signe du gradient de la phase intersecte exactement les bords des aires visuelles qui sont par ailleurs définies ainsi par une méthode indépendante.
Pour le contraste visages/lieux nous écartons 45 photos de portrait de personnes du jeu de données ainsi que 45 photos de lieux vides. Nous les utilisons pour créer un contraste BOLD de "vérité-terrain" qui nous servira comme point de comparaison aux contrastes générés. Ensuite, l'activité cérébrale est prédite sur les photos écartées en utilisant un modèle entrainé sans ces images. De plus, nous faisons passer des stimuli d'une expérience totalement indépendante [Haxby2001].



Panel A représente les stimuli écartés du jeu de données, Panel B montre des stimuli tirés d'une autre expérience. Panel C montre le contraste BOLD entre les stimuli représentés par Panel A qui servira comme point de comparaison. Panel D montre le contraste synthétisé sur les images du Panel A, Panel E montre le contraste synthétisé sur les images du Panel B. Les résultats confirment leur similarités à la vérité-terrain. Pour montrer que ceci n'est pas atteignable avec les descripteurs bas-niveau seuls, Panel F montre que l'image est bruitée et ne permet pas d'isoler les zones actives. Ceci vaut pour tous les seuils, ce qui est indiqué par la ligne en pointillées dans le Panel G.

En somme, nous avons étudié la vision humaine en IRMf. En passant par des aspects de traitement de données et d’estimation du filtre impulsionnel hémodynamique des données BOLD de façon jointe avec le GLM, et l'introduction d'une nouvelle pénalité convexe, parcimonieuse et spatialement informée. Nous avons comparé des modèles computationnels de la vision, notamment des couches d'un réseau de neurones convolutionnel à l'architecture du système visuel humain et avons trouvé des correspondances. Notre modèle est suffisamment puissant pour générer de l'activité cérébrale plausible à partir de nouveaux stimuli. Deux tests de bas et haut niveau confirment que les cartes générées correspondent aux observations. Ce modèle pourra donc être utilisé pour formuler des hypothèses sur les activations du cerveau.