Apport des réseaux de neurones convolutionnels basés graphes (graph-convolutional neural networks) pour l'analyse de données multi-omiques et multi-modales de population.

Mots clés

Réseaux de neurones convolutionnels, graphes, neuro-imagerie, multi-omique, études de population.

Profil et compétences recherchées

Expertise en machine learning et intelligence artificielle.

Présentation détaillée du projet doctoral

Résumé :

Ces dernières années, des efforts internationaux ont permis la constitution de grandes cohortes d'imagerie cérébrale de population générale, qui comportent de plus en plus souvent des données génomiques. A NeuroSpin, des travaux de recherche en neurosciences sont menés par exemple sur UK-Biobank (500.000 sujets) qui est à ce jour la cohorte internationale la plus fournie en population générale. Des travaux d'intérêt clinique sont aussi menés sur des cohortes plus thématiques avec entre autres IMAGEN (2,000 sujets, trois visites) pour les comportements à risques chez les adolescents, ou MEMENTO (2,500 sujets) pour la maladie d'Alzheimer.

L'exploration de ces données multi-omiques et d'imagerie multi-modales visent à trouver des marqueurs diagnostiques et pronostiques originaux. Ces études sont réalisées au moyen de diverses analyses multivariées utilisant les outils du machine learning. Ces études tendent de plus en plus à fusionner les multiples sources d'information dont certaines se représentent naturellement sous forme de graphes. Par ailleurs ces études nécessitent d'intégrer des informations de référence (connaissances académiques ou mesures sur biopsie de issus) qui à nouveau sont décrites sous forme de graphes.

Nous proposons un travail d'adaptation et de mise en œuvre des réseaux de neurones convolutionels basés sur des graphes (GraNN) pour réaliser des analyses dans des cohortes de population. Cette classe de réseaux neuronaux du domaine de l'apprentissage profond est à même de réaliser les objectifs de fusion et d'intégration de l'information lorsque celle-ci se présente sous forme de graphes. Le formalisme des GraNN est bien adapté à la représentation de très nombreuses mesures biologiques et médicales et ce travail devrait fournir des résultats novateurs en termes de méthodes ou de phénotypes. Enfin, comme les couches de sortie des GraNN ou les variables latentes qu'ils permettent d'extraire se présentent sous forme de graphes, ce travail offrira aussi des perspectives nouvelles d'interprétation des résultats.

Summary :

Recently, international initiatives have led to the building of large cohorts in general population of brain imaging studies. Most of them now include genomic data. At NeuroSpin, research in neurociences is being carried out on the UK-Biobank (500,000 subjects), which is to date the largest international cohort in the general population. Research of clinical interest is also being conducted

based on more thematic cohorts with, among others, IMAGEN (2,000 subjects, three visits) for risk behaviors in adolescents or MEMENTO (2,500 subjects) for Alzheimer's disease.

Data mining of these multi-omic and multi-modal imaging data aim to find original diagnostic and prognostic markers. These studies are carried out by various multivariate analyses using learning machine tools. The latter increasingly tend to merge multiple sources of information, some of which are naturally represented in the form of graphs. In addition, these studies require the integration of reference information (academic knowledge or measurements in biopsies) which again, are described in the form of graphs.

We propose a work to adapt and implement graph-based convolutional neural networks (GraNN) to carry out novel analyses in population cohorts. This class of neural networks of the deep learning domain is able to achieve the objectives of information fusion and information integration when this information is represented by a graph. The graphical formalism of GraNNs is well adapted to the representation of very numerous measurements of biological or medical phenomena, and this work should provide innovative results in terms of methods or phenotypes. Finally, since GraNNs provide graphical objects as output layers or latent variables, this work also should also offer new perspectives for the interpretation of the results.

Thématiques Domaine Contexte

Machine learning, Intelligence artificielle, Contraintes structurelles sous forme de graphes, Données multi-omiques et multi-modales en neuro-imagerie, Imagerie génétique, Application en oncologie et maladies psychiatriques.

Contexte et objectifs

Contexte : imagerie cérébrale et études de population.

L'imagerie cérébrale fournit des marqueurs in-vivo uniques pour diagnostiquer et étudier des pathologies neurologiques (comme les maladies neurodégénratives: sclérose en plaques, Parkinson, Alzheimer) ou des syndromes psychiatriques (comme l'autisme ou la schizophrénie). Plus précisément, l'imagerie cérébrale contribue à différents axes de recherche sur ces maladies, à savoir : 1) la détection précoce de la maladie et une prise en charge adaptée du patient ou de son entourage, 2) la mise au point de biomarqueurs prédictifs et/ou pronostics et 3) la compréhension des mécanismes biologiques de la maladie, qui permettra d'en traiter les causes. La génomique, qui produit des mesures à haut débit sur des entités moléculaires, contribue également à ces trois axes de recherche et apporte des informations complémentaires à l'imagerie.

Ces dernières années, des efforts internationaux ont permis la constitution de grandes cohortes d'imagerie cérébrale de population générale, qui comportent de plus en plus souvent des données génomiques. L'émergence de ces jeux de données, toujours plus vastes et plus représentatifs de nos populations, ouvre de nouvelles voies de recherche.

En comparaison à des études à effectif limités, ces cohortes de population générales ont des critères d'inclusion larges lors du recrutement des sujets. Cette spécificité des cohortes de population est à notre sens très largement compensée par les caractéristiques propres de ces nouveaux instruments de recherche. A savoir, ces cohortes inaugurent des apports spécifiques dont le plus important à ce jour est un complément aux tableaux cliniques classiques, sous la forme de phénotypes d'imagerie continus, contextuels et biologiquement proches des processus sous-jacents à une maladie ou des processus du développement d'un individu. Ces phénotypes, étalonnés à l'échelle d'une population,

pourront remplacer avantageusement des étiquettes diagnostiques dichotomiques (cas/contrôle), en les désambiguïsant, pour une prise en charge personnalisées des patients. Ces phénotypes étalonnés pourront être transposés à d'autres cohortes et confèreront une puissance statistique accrue aux analyses. Enfin, ces phénotypes, mesurés au plus près des organes et de de leur fonctionnent, permettront l'étude fondamentale de processus sous-jacents anormaux dans une maladie ou dans le développement de l'individu.

L'intérêt de ces cohortes est renforcé lorsque des données de génomique fonctionnelle sont présentes car elles enrichissent la palette des phénotypes à disposition lors de l'analyse. Enfin ces cohortes, lorsqu'elles comprennent des données génétiques, permettent de mener des études d'association entre des variations génétiques à l'échelle de la population et des traits phénotypiques complexes issus de l'imagerie en prenant en compte le style de vie des sujets de la cohorte pour étudier l'influence des facteurs environnementaux. Cela rend possible l'étude de la part relative de l'environnement et de la génétique dans les phénotypes complexes étudiés.

L'analyse de telles cohortes est en cours actuellement à NeuroSpin pour des fins de recherche fondamentale. Il en est ainsi pour des études concernant l'impulsivité sur les données **IMAGEN** (Luo et al., 2019), pour des études sur la sulcation du cerveau sur les données **HCP** (Le Guen et al., 2018) et **UK-BioBank** (Le Guen et al., 2019). De même, en recherche médicale, la cohorte **EU-AIMS** est utilisée pour élaborer des biomarqueurs prédictifs de l'autisme (Moessnang et al., 2020) et les données **ADNI** pour la caractérisation de la maladie d'Alzheimer (Guigui et al., 2019).

Objectifs : apport des modèles graphiques en intelligence artificielle.

Nous proposons d'utiliser des approches d'apprentissage profond utilisant les réseaux de neurones basés graphes (GraphNN) pour traiter des corpus de données issus de cohortes de population. Ces GraphNN auront trois objectifs. Premièrement, il s'agit d'intégrer les données multi-omiques et les données d'imagerie multi-modales disponibles sur un échantillonnage cartésien (espace image) ou représentées sous forme de graphes (c'est le cas des données de connectivité structurelle et fonctionnelle tirées des IRM de diffusion ou des IRM fonctionnelle de repos). Deuxièmement, le travail visera à intégrer les connaissances disponibles en neurosciences et en biologie, potentiellement décrites par des graphes. Enfin une tâche sera dévolue à la mise en avant des parties discriminantes de graphe lors de nos analyses de population.

En s'appuyant sur la récente formalisation par la communauté scientifique des réseaux de neurones basés graphes (Wu et al., 2020), nous proposons d'explorer la capacité des GraphNN à analyser des données biologiques et médicales présentées sous forme de graphes. Ce support est très adapté pour tenter de représenter et d'expliquer les processus du monde vivant. Des cadres d'apprentissage sont en cours d'élaboration pour ces GraphNN, qui reprennent les principes du transfer learning et des modèles génératifs. Ces cadres offrent des opportunités pour palier le faible nombre de mesures (compté en milliers cependant) et proposer des phénotype originaux et interprétables.

Atouts :

Le groupe d'imagerie génétique de NeuroSpin (BrainOmics) dispose de plusieurs atouts dans ces travaux. Tout d'abord, grâce à notre expertise en traitement d'images IRM du cerveau, nous extrayons des phénotypes riches et originaux en imagerie IRM multi-modales, allant des caractéristiques des sillons corticaux aux réseaux fonctionnels ou structurels. Nous avons déjà démontré notre capacité à passer nos traitements à l'échelle pour des cohortes d'intérêt clinique en oncologie comme BIOMEDE ou sur des pathologies neurodégénératives comme CATI/MEMENTO.

Actuellement NeuroSpin accueille plusieurs doctorants qui travaillent sur diverses applications des réseaux de neurones comme la classification, la segmentation, le recalage, le traitement du langage naturel ou la reconstruction d'images IRM.

Méthode et données

Données. Les chercheurs de NeuroSpin accèdent à de grandes cohortes de population internationales du fait de leur implication dans divers consortiums. Par ailleurs, nous accédons et nous utilisons les nombreuses ressources de référence disponibles dans la communauté. Ces données de haute qualité requièrent des infrastructures de très grandes importance tant en terme de calcul (utilisation des centres nationaux de calcul comme le CCRT) que de stockage (utilisation de la plateforme de données de santé de Bruyère Le Chatel, PFRDS). Ces cohortes rassemblent des mesures qui couvrent plusieurs niveaux d'organisation du vivant (séquence, expression des gènes, épigénétique, imagerie, dossier clinique et mesures comportementales).

Cohortes généralistes :

- IMAGEN : Cohorte européenne sur la santé mentale des adolescents (~2.000 sujets, trois visites). Etude longitudinale avec IRM anatomique et fonctionnelle, génotype et méthylation, comportement - https://imagen-europe.com/.
- AIMS : Cohorte européenne d'étude sur l'autisme (~1.000 sujets, deux visites). Etude longitudinale avec imagerie IRM anatomique et fonctionnelle, génotype, Eye tracking, EEG, comportement - https://www.eu-aims.eu/.
- HCP : Cohorte US sur sujets adultes (~1.200 sujets). Données imagerie IRM de hautes qualité anatomiques, fonctionnelle et de diffusion, génotype - https://www.humanconnectome.org/study/hcp-young-adult.
- UK-BioBank : cohorte UK en population générale en santé publique (~500.000 sujets). IRM et génétique (puce et séquençage) - https://imaging.ukbiobank.ac.uk/.

Cohortes sur pathologies :

- BIOMEDE: Cohorte sur les tumeurs pédiatrique de la ligne médiane - https://www.gustaveroussy.fr/fr/biomede.
- LPSNC : Cohorte sur les lymphome primaires du système nerveux central- https://www.reseauloc.org/lymphome-cerebral-primitif.
- SENIOR : Cohorte d'étude du vieillissement (NeuroSpin).

A des fins d'annotation des données, nous accédons aussi à des ressources de référence comme AllenBrain, TCGA, GTEx, EGA. Ces ressources contiennent des mesures dont l'acquisition n'est pas transposable pour des cohortes (par exemple les tissus biologiques).

Méthodes

Ces données relativement peu fournies en échantillons (n ~ 1K à 100K) au regard des tailles en big data actuel, sont échantillonnées dans des espaces de très grandes dimensions (plusieurs blocs de données avec p ~ 100K à 1.000K) très atypiques. La production et l'exploitation scientifique des phénotypes

riches sont aujourd'hui réalisées par des analyses multivariées que permettent les outils du machine learning classiques partant des caractéristiques extraites *a priori* des images ou de la génomique : entraînement de classifieur, méthodes de clustering avec des pénalisations permettant de régulariser les problèmes.

Structure de corrélation décrites par des graphes. Les travaux méthodologiques avancés en cours concernent la prise en compte de la structure de corrélation particulière à l'intérieur de chacun des blocs (modalités) de données qui constituent les cohortes. En plus de la parcimonie, les méthodes d'apprentissage sont dotées de pénalités structurées à même de rendre compte par exemple de l'organisation 3D des tissus corticaux ou du déséquilibre de liaison sur le génome. Des formes générales de structure de corrélation décrites sous forme de graphes sont aujourd'hui prises en compte.

Intégration de données décrites par des graphes. Un second front méthodologique concerne l'intégration des multiples blocs (modalités) de données. Des formes généralisées de l'analyse canonique des corrélations permettent de trouver des jeux de variables latentes (patterns d'images, pattern d'expression) résumant l'information des blocs et qui sont corrélés entre eux. Une forte attente existe concernant les méthodes modélisant des interactions riches entre les blocs de données. C'est le cas lorsqu'on désire intégrer un bloc d'IMRf de reposavec un bloc de connectivité structurelle qui se présente sous forme d'un graphe. C'est aussi le cas pour des blocs de données de co-expression des gènes.

Le vivant comme réseau d'entités biologiques. La biologie moléculaire et cellulaire, les sciences médicales établissent que le domaine du vivant est fondé sur des niveaux d'organisation eux-mêmes composés de réseaux d'interaction fortement résilients entre entités propres à un niveau. Sans vouloir modéliser l'ensemble de cette organisation, il est nécessaire cependant d'envisager de modéliser au moins partiellement les réseaux d'un niveau et leur variabilité : par exemple le réseau de co-expression des gènes propres à un tissu ou encore le réseau des aires du cerveau impliquées dans un traitement. Les cohortes généralistes pourraient permettre d'apprendre de tels réseaux (ou au moins certaines de leur caractéristiques). Ces réseaux pourraient ensuite être comparés à ceux construits (ou prolongés) à partir de cohortes de patients afin de déterminer les interactions impliquées dans une pathologie.

L'intelligence artificielle et plus particulièrement l'apprentissage profond a récemment donné lieu à des améliorations importantes dans de nombreuses applications. Cela s'explique en partie par l'arrivée de dispositifs de calculs dédiés (les GPUs), la disponibilité de grandes quantités de données pour l'apprentissage, et aussi l'efficacité de la phase d'apprentissage pour extraire les représentations latentes dans des données (Convolutional NN). Dans la résurgence des réseaux de neurones, un nombre croissant d'applications cherchent à intégrer des contraintes de graphes permettant à la fois de modéliser les données mais également certains processus biologiques. La description et les propriétés de ces réseaux particuliers (Graphical neural networks) ainsi que les cadres de leur entraînement commencent à être normalisées (Wu et al., 2020).

La capacité naturelle des réseaux de neurones basés graphes à apprendre des schémas biologiques dans les données nous engage à proposer ce travail de thèse. Malgré la difficulté que représente le nombre « restreint » d'échantillons dans une cohorte (y compris pour les plus ambitieuses disponibles à ce jour que nous accédons), nous pensons que ces outils présentent une opportunité pour apprendre des biomarqueurs plus performants, voire complètement novateurs.

Les travaux méthodologiques concerneront des outils d'analyse comme :

- la reconstruction de matrice de connectivité anatomique à partir de données de diffusion,

- l’intégration des données IRMf de repos avec des matrices de connectivité fonctionnelles,
- l’extraction de caractéristiques d’intérêt comme l’épaisseur cortical, l’indice de gyrification,
- ....

D’autre travaux méthodologiques pourront viser des applications cliniques comme la stratification des individus d'une population.

Résultats attendus

Adapter les outils des réseaux neurones basés graphes pour l’analyse des cohortes de population multi-modales et multi-omiques.

Produire des exemples de construction de phénotypes à support graphique pour la classification ou la stratification dans des cohortes médicales.

Précision sur l'encadrement

L’encadrement est assuré par V. Frouin (HdR) plus particulièrement pour la partie intégration de données imagerie-génétique et A. Grigis plus particulièrement pour la partie intelligence artificielle.

Le doctorant est accueilli au sein du laboratoire GAIA de Neurospin. Le travail du doctorant est suivi lors de réunions hebdomadaires avec le directeur de thèse et les encadrants. Le doctorant effectue des présentations lors des réunions de laboratoire. Un Comité de suivi de thèse sera établi.

Conditions scientifiques du projet de recherche

Le thésard accède les moyens de calcul et les cohortes décrites dans le projet et participe aux conférences du domaine scientifique.

Objectifs de valorisation des travaux de recherche

Diffusion par communications orales et écrites dans des congrès nationaux et internationaux dans le domaine de l’imagerie, de la bioinformatique (génétique et génomique fonctionnelle) et des neurosciences cliniques. Publications dans des revues internationales.

Collaborations envisagées

Collaboration avec :

- J.F. Mangin (NeuroSpin) pour les études sur maladies neuro-degénératives.
- S. Jamain (IMRB, INSERM U955 Hôp. Mondor)pour les études impliquant la génétique.
- Alentorn (MD Hôp Pitié-Salpêtrière) pour les études en neuro-oncologie.
- J. Grill (MD Hôp Gustave-Roussy) pour les études en neuro-oncologie.
- J. Houenou (NeruoSpin, Hôp. Mondor) pour les études en psychiatriques.

Ouverture Internationale

Participation aux projets internationaux du groupe, en particulier l'étude de la cohorte LEAP du projet AIMS2-TRIAL sur l'autisme.

Références bibliographiques

- **celles du groupe**

Guigui, N., Philippe, C., Gloaguen, A., Karkar, S., Guillemot, V., Löfstedt, T., & Frouin, V. (2019). Network regularization in imaging genetics improves prediction performances and model interpretability on Alzheimers's disease. In *ISBI 2019 - Proceedings of the IEEE International Symposium on Biomedical Imaging*. Venice, Italy. Retrieved from https://hal-cea.archives-ouvertes.fr/cea-02016625

Le Guen, Y., Leroy, F., Auzias, G., Riviere, D., Grigis, A., Mangin, J.-F. F., … Frouin, V. (2018). The chaotic morphology of the left superior temporal sulcus is genetically constrained. *NeuroImage*, *174*, 297–307. https://doi.org/10.1016/j.neuroimage.2018.03.046

Le Guen, Y., Leroy, F., Philippe, C., Consortium, I., Mangin, J.-F., Dehaene-Lambertz, G., & Frouin, V. (2019). A DACT1 enhancer modulates brain asymmetric temporal regions involved in language processing. *BioRxiv*, 539189. https://doi.org/10.1101/539189

Luo, Q., Chen, Q., Wang, W., Desrivières, S., Quinlan, E. B., Jia, T., … Feng, J. (2019). Association of a Schizophrenia-Risk Nonsynonymous Variant with Putamen Volume in Adolescents: A Voxelwise and Genome-Wide Association Study. *JAMA Psychiatry*, *76*(4), 435–445. https://doi.org/10.1001/jamapsychiatry.2018.4126

Moessnang, C., Baumeister, S., Tillmann, J., Goyard, D., Charman, T., Ambrosino, S., … Meyer-Lindenberg, A. (2020). Social brain activation during mentalizing in a large autism cohort: The Longitudinal European Autism Project. *Molecular Autism*, *11*(1), 17. https://doi.org/10.1186/s13229-020-0317-x

- **extérieures au groupe**

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, 1–21. https://doi.org/10.1109/tnnls.2020.2978386